Hammill Institute on Disabilities

Journal of Learning Disabilities
2021, Vol. 54(1) 20–35

DOI: 10.1177/0022219420911375
journaloflearningdisabilities.sagepub.com
SAGE

Using Direct Observation to Document "Practice-Based Evidence" of Evidence-Based Mathematics Instruction

Christian T. Doabler, PhD[1], Ben Clarke, PhD[2], Derek Kosty, PhD[3], Jessica E. Turtura, PhD[2], Marah Sutherland, MS[2], Steven A. Maddox, MEd[1], and Keith Smolkowski, PhD[3]

Abstract

Implementation of evidence-based practices (EBPs) is paramount to students' development of mathematics proficiency. This study investigated "practice-based evidence" of interventionists' actual use of explicit mathematics instruction, a well-established EBP. Specifically, this study analyzed direct observation data collected in a federally funded efficacy trial involving a Tier 2 first-grade mathematics intervention to examine whether the quantity and quality of explicit mathematics instruction was associated with the mathematics outcomes of 470 first-grade students with or at risk for mathematics learning disabilities. Associations between group-level pretreatment skill levels and the quality and quantity of explicit mathematics instructional practices used in the intervention were also explored. Findings suggested significant associations between positive gains in student mathematics outcomes and (a) lower rates of incorrectly answered mathematics-focused questions, and (b) the rate in which interventionists delivered group-level practice opportunities and offered academic feedback. Significant associations were also found between initial student mathematics performance and rates of student errors and the quality of explicit instruction. Implications for using direct observation to document enacted EBPs are discussed.

Keywords

mathematics, multitiered systems of support, response to intervention (Tier 2/Tier 3), design of, instruction

The medical field is a forerunner in establishing and implementing evidence-based practices (EBPs). Areas across the field of medicine benefit from the use of EBPs. Nursing practice, for example, has deep roots in EBPs and myriad validated practices are available for nurses working in the field (Nettina, 2019). Practices such as proper hand hygiene and wound care are essential for nurses to provide optimal patient care and offer infection prevention.

The field of education science has a growing history of establishing EBPs (Cook et al., 2013, 2015). And like nursing, the use of EBPs in education is essential for producing beneficial outcomes for targeted populations (Smolkowski et al., 2019). Over the past few decades, the education field, particularly special education, has been instrumental in establishing EBPs through randomized controlled trials (RCTs) involving academic interventions (Cook et al., 2015). Although the number of EBPs made available from this intervention research is encouraging, little is known about how frequently and at what quality educators use EBPs within today's classrooms. There is also limited research on direct observation as a relevant measurement tactic for gathering "practice-based evidence" (Green, 2008) around the quantity and quality of teachers' use of EBPs. In this study, practice-based evidence refers to the real-time implementation of EBPs by teachers in authentic educational settings. The field of education has long been plagued with breaks in the implementation pipeline of EBPs (Carnine, 1997; Cook & Cook, 2011). Therefore, using direct observation to document

[1]The University of Texas at Austin, Austin, TX
[2]University of Oregon, Eugene, OR
[3]Oregon Research Institute, Eugene, OR

Corresponding Author:
Christian T. Doabler, Meadows Center for Preventing, Educational Risk, The University of Texas at Austin, 1 University Station, D5300 SZB 408B, Austin, TX 78712, USA.
Email: cdoabler@austin.utexas.edu

such practice-based evidence may serve as a valuable method for bridging the research-to-practice gap (Cook et al., 2013; Gersten & Dimino, 2001).

The purpose of this study was to analyze direct observation data on teachers' use of explicit mathematics instruction, a well-recognized EBP in the field of education (Hughes et al., 2017). Data analyzed were collected during a federally funded efficacy trial involving the Fusion mathematics program (Clarke, Doabler. & Fien, 2016), a Tier 2 first-grade intervention designed for students who are at risk for mathematics learning disabilities (MLD). Prior research reported the Fusion program to be effective for first-grade students who struggle early with mathematics (Clarke et al., 2014). In the current study, we examined specifically for whether the quantity and quality of explicit mathematics instruction delivered during the Fusion intervention was associated with increased outcomes for first-grade students who demonstrate academic risk in mathematics. In addition, we explored whether students' initial mathematics skills influenced the quality and quantity of explicit mathematics instructional practices used by teachers during Fusion instruction.

Explicit Instruction and the Facilitation of Instructional Interactions

Explicit instruction is an instructional approach known for directly teaching fundamental concepts and skills to students through empirically validated instructional design and delivery principles (Deshler, 2015; Hughes et al., 2017; Simmons, 2015). A growing body of evidence, primarily generated by RCTs involving Tier 2 mathematics interventions, suggests that explicit instruction is an effective means for increasing the mathematics achievement of students at risk for MLD (Dennis et al., 2016; Gersten et al., 2009). Although findings from these RCTs are important for buttressing the use of explicit mathematics instruction when teaching struggling learners, more information is needed to help pinpoint as to why this EBP produces positive effects on student mathematics achievement. One plausible mechanism behind the beneficial outcomes of explicit mathematics instruction is its systematic facilitation of instructional interactions. In the current study, instructional interactions are operationalized as explicit teaching events that occur between teachers and students, and among students, in authentic educational settings. Such interactions center on foundational mathematics content and are purposefully designed to positively influence student learning and mathematical development.

Instructional interactions are considered integral to students' mathematics development (National Research Council, 2001; Pianta & Hamre, 2009). In early mathematics (kindergarten to second grade), instructional interactions are essential for engaging struggling learners in critical mathematics concepts and skills, such as whole number and operations. Theoretically, explicit instructional interactions represent a dynamic interplay of three EBPs: (a) overt teacher modeling (Gersten et al., 2009), (b) independent and guided student practice opportunities (Hughes & Riccomini, 2019), and (c) timely, academic feedback (Halpern et al., 2007). We operationally define each practice below and provide evidence for their empirical backing.

Overt teacher modeling represents teachers using clear demonstrations and explanations to present new mathematical concepts, procedures, and skills to students. For example, direct teacher modeling allows teachers to unambiguously demonstrate and explain what students are expected to do in a mathematical activity on place value. Research suggests that direct modeling is an effective way to build initial understanding and promote higher order thinking of mathematical concepts and skills (Alfieri et al., 2011).

Student practice opportunities, which typically follow the teacher demonstrating a mathematical concept or skill, comprise individuals and groups of students completing written exercises, manipulating visual representations of mathematics, and verbalizing their mathematical understanding. Initial practice opportunities often involve guided support from the teacher (Riccomini & Morano, 2019). Then, as students develop an understanding of the targeted content, the teacher's support is systematically withdrawn to increase students' opportunities to independently practice (Hughes & Riccomini, 2019). A growing body of evidence suggests that student practice opportunities are essential for building conceptual understanding and promoting procedural fluency (Clements et al., 2013; Gersten et al., 2009).

Academic feedback consists of teachers providing timely, informational feedback to students on their performance with solving mathematical problems (Halpern et al., 2007). Such feedback is often delivered through teacher verbalizations, where teachers draw students' attention to previous incorrect responses or misconceptions (Gersten et al., 2009). For example, when a student incorrectly verbalizes the solution of a multidigit addition problem, the teacher will immediately correct the error via explanation and demonstration, and then provide additional opportunities for the student to practice with a similar problem type.

Role of Initial Mathematics Achievement in Explicit Instructional Interactions

Explicit mathematics interventions that contain scripted lessons offer an instructional platform for teachers to provide overt demonstrations, facilitate frequent student practice opportunities, and deliver timely academic feedback. Thus, as a function of these explicit lesson guidelines, implementing a scripted mathematics intervention with fidelity will naturally engage at-risk learners in important instructional interactions around targeted mathematics content. In addition to the instructional design of interventions (Simmons, 2015), another variable that may come into play with teachers' facilitation of instructional interactions is the initial mathematics skill levels that students bring to the intervention table.

Initial mathematics achievement is an important factor in setting students' trajectory of mathematical learning (Morgan et al., 2009). As such, student skill levels may play a key role in the quantity and quality of instructional interactions delivered during explicit mathematics interventions. Teachers may need to adjust how often and at what quality they facilitate explicit instructional interactions based on students' mathematical performance. For example, in some instances, a teacher may have to overtly demonstrate how to break down a word problem-solving strategy into more manageable chunks to reduce the cognitive load for students. In others, a teacher may have to offer a richer explanation of a mathematical concept than that scripted in a mathematics intervention to better meet the instructional needs of at-risk learners. Moreover, if a group of students struggles to gain an early understanding of a targeted mathematics concept, a teacher may have to provide additional opportunities for students to verbalize their mathematical thinking and distribute such opportunities to students with the most significant learning difficulties. Because these instructional adjustments may serve as "instances of positive infidelity" (Munter et al., 2014, p. 95), research is needed to explore whether group-level, pretreatment mathematical performance levels are associated with teachers' facilitation of explicit instructional interactions (Doabler et al., 2020).

Documenting Explicit Instructional Interactions Through Direct Observation

The public nature of instructional interactions during classroom instruction lends documentation and interpretability through direct observation (Shavelson et al., 1986; Snyder et al., 2006). As such, direct observation is a viable approach for measuring instructional interactions and thus unpacking the "black box" for how EPBs work and function in authentic educational settings. Over the past four decades, researchers have developed and validated a variety of direct observation systems designed to document instructional interactions in real time (Brophy & Good, 1986; Pianta & Hamre, 2009; Smolkowski & Gunn, 2012; Vaughn & Briggs, 2003). This established line of research has focused on the beneficial role that both the quality and quantity of instructional interactions have on student mathematics outcomes.

Quality of Instructional Interactions

Building off of previous models of instructional quality (e.g., Pianta & Hamre, 2009), our conceptualization of the quality of explicit instructional interactions represents the manner and richness in which teachers provide overt demonstrations, facilitate frequent student practice opportunities, and deliver timely academic feedback. The current study operationalizes instructional richness as opportunities where teachers (a) overtly link and make connections, both between and within, concrete and symbolic ideas of mathematics; (b) provide mathematical meaning to targeted concepts and skills; and (c) make explicit how to use models and tools appropriately and strategically (Hill et al., 2008).

In the existing literature, instructional quality is typically documented through moderate to high inference instruments that include Likert-type rating scales. Such observation tools typically rely on observers' impressions to rate the quality of instructional interactions. Because of this coding structure, moderate to high inference instruments permit greater flexibility than low-inference instruments for observers to make informed evaluations and decisions about the quality of instruction (Gersten et al., 2005).

For example, Pianta and Hamre (2009) validated the Classroom Assessment Scoring System (CLASS; Howes et al., 2008; La Paro et al., 2009; Mashburn et al., 2008; Pianta & Hamre, 2009). The CLASS is moderate-level observation tool that requires coders to use a 7-point rating scale to rate 10 dimensions centered on three domains: emotional supports, classroom organization, and instructional supports. Observers' ratings from two studies involving the CLASS indicate that students are more likely to have stronger mathematics outcomes if they are placed in classrooms that offer higher quality instructional interactions than lower ones (Howes et al., 2008; Mashburn et al., 2008).

In our own research, we have administered a moderate-level instrument to examine the quality of explicit instruction facilitated during a Tier 2, kindergarten mathematics intervention (Doabler et al., 2020). The Quality of Explicit Mathematics Instruction (QEMI; Doabler & Clarke, 2012) employs a 4-point rating scale to gauge the quality of seven features of explicit instructional interactions. These features include group and individual practice opportunities, student participation, teacher modeling, academic feedback, efficiency of instructional delivery, and instructional scaffolding. In a recent study involving the QEMI, a major finding was that ratings of higher quality explicit instruction predicted increased student mathematics achievement (Doabler et al., 2020). In other words, kindergarten students in intervention groups that offered higher instructional interactions made greater gains across the school year than their peers in intervention groups with lower quality instructional interactions.

Quantity of Instructional Interactions

Prior research has also employed relatively low inference measurement approaches to document the instructional interactions that occur between teachers and students. These measures employ measurement tactics that record instructional interactions at a more molecular level than moderate inference instruments. Rather than relying on an observer's subjective ratings, low-inference instruments tend to employ strict coding procedures to document predetermined instructional events, activities, and behaviors (Baker et al., 2006; Gersten et al., 2005). These types of measures require observers to record the occurrences of observable instructional interactions. Researchers typically employ low-inference instruments to quantify the rates or frequencies of instructional interactions. And relative to global rating systems such as the CLASS and QEMI, low-inference instruments are often better able to minimize observer inference and control for variance due to observer characteristics (e.g., biases to participants) because they focus on clearly defined target behaviors that are less subjective to interpretation (Snyder et al., 2006). Finally, information generated by low-inference instruments is often reported in metrics that are highly interpretable and actionable for teachers to implement in authentic educational settings (Doabler et al., 2020).

Clements et al. (2013) administered a low-inference observation measure to document "teacher-directed" instructional practices during first- and second-grade mathematics instruction. The observation data were collected in over 600 first- and second-grade classrooms during a large-scale, RCT focused on the efficacy of four different mathematics curricula. Findings suggested that the frequency of individual student mathematics verbalizations were related to increased mathematics achievement in second grade. Nonsignificant findings were reported in first grade, however.

In a similar line of research, Doabler et al. (2015) explored whether the rate in which students engaged in explicit instructional interactions predicted gains in student mathematics achievement. Data analyzed were captured by a low-inference observation tool during an RCT focused on testing the efficacy of a Tier 1 mathematics program. The RCT included 400 observations conducted in 129 kindergarten classrooms from two different geographical regions of the United States. Results indicated that students in classrooms with more frequent individual practice opportunities, including working with visual representations and engaging in mathematical discourse, had stronger overall performance on two standardized mathematics outcome measures than students in classrooms with fewer practice opportunities.

Purpose of the Study

This study investigated direct observation data to examine "practice-based evidence" (Green, 2008) of interventionists' actual use of evidence-based mathematics instruction. Specifically, we investigated whether and to what extent the quantity and quality of explicit instructional interactions facilitated during the Fusion intervention, a Tier 2 mathematics intervention, predicted increased student mathematics achievement. Small-group mathematics instruction that provides frequent, high-quality instructional interactions may promote beneficial outcomes for students with or at risk for MLD. This study also explored whether the skill composition of Fusion groups, as established by students' mathematical performance at the onset of the Fusion intervention, influenced the frequency and quality in which interventionists facilitate explicit instructional interactions. Interventionists who deliver Tier 2 mathematics interventions, such as Fusion, may offer more overt teacher demonstrations, structured practice opportunities, and academic feedback in intervention groups that demonstrate lower preintervention skill levels. Two research questions were addressed:

1. Does group-level initial mathematics achievement, as established by pretest performances on standardized measures of whole numbers and operations, predict the quantity and quality of explicit instructional interactions during Fusion instruction?
2. Does the quantity and quality of explicit instructional interactions during Fusion instruction predict gains in student mathematics achievement from pretest to posttest?

Method

Research Design and Context

This study analyzed data collected during the first 2 years of a 4-year, federally funded efficacy project involving the Fusion intervention, a Tier 2 first-grade mathematics intervention. Three nonoverlapping cohorts of first-grade students participated in the first and second year of the project. The Fusion Efficacy Trial (Clarke et al., 2016) employed a partially nested RCT (Baldwin et al., 2011). Blocking on classrooms, 680 first-grade students were randomly assigned within first-grade classrooms to one of three conditions. In one condition, students were randomly assigned to receive the Fusion intervention in groups with 2:1 student–teacher ratios (2:1 Fusion groups), whereas a second condition provided Fusion in groups with 5:1 student–teacher ratios (5:1 Fusion groups). The third condition represented a no-treatment control condition (i.e., business-as-usual). In all, 131, 339, and 210 students were assigned to the 2:1 Fusion, 5:1 Fusion, and no-treatment control conditions, respectively. Students randomly assigned to the two treatment groups received the Fusion intervention in addition to district-approved core mathematics instruction. In total, 135 Fusion intervention groups were conducted in Years 1 and 2 of the Fusion Efficacy

Trial (67 = 2:1 Fusion, 68 = 5:1 Fusion). The current study focused specifically on mathematics outcomes data from the 470 Fusion students and direct observation data collected in the 135 Fusion groups.

Participants

Schools. A total of 18 elementary schools from five Oregon and Massachusetts school districts participated in the current study. One school district was located in the metropolitan area of Portland, two districts were located in suburban areas of western Oregon, and two districts were located in the metropolitan area of Boston. Across the five districts, student enrollment ranged from 5,492 to 40,495 students. Within the 18 participating schools, between 12% and 19% of students had disabilities, 4% and 38% were English learners, and 29% and 65% were eligible for free or reduced lunch. Between 1% and less than 1% were American Indian or Native Alaskan, 1% and 16% Asian, 1% and 7% were Black, 9% and 87% were Hispanic, 0% and 1% were Native Hawaiian or Pacific Islander, 7% and 73% were White, and 1% and 8% were more than one race.

Classrooms. The study took place in 77 first-grade classrooms taught by 60 certified teachers across three cohorts of students. All classrooms provided mathematics instruction in English and operated 5 days per week. Classrooms had an average of 24.1 students (SD = 5.2).

Students and inclusion criteria. In each participating classroom, all students with parental consent were screened in late fall of their first-grade year. The screening process, which included 1,600 first-grade students, comprised the three measures of the first-grade Assessing Student Proficiency in Early Number Sense (ASPENS) battery (Clarke et al., 2011). These measures included magnitude comparison, missing number, basic arithmetic facts, and base-10. Students were considered eligible for the Fusion intervention and thus considered at risk for MLD if they had an ASPENS composite score in the strategic or intensive categories based on winter benchmarks. Composite scores at or below the strategic category suggest that students have less than a 50% chance of meeting end-of-year grade level expectations in mathematics (Clarke et al., 2011).

Students with ASPENS composite scores in the strategic or intensive categories were rank ordered in each participating classroom by an independent evaluator. Within each classroom, the independent evaluator then randomly assigned the 10 students with the lowest ASPENS composite scores to one of three conditions: (a) a Fusion intervention group with a 2:1 student–teacher ratio, (b) a Fusion intervention group with a 5:1 student–teacher ratio, or (c) a control (i.e., business-as-usual) condition. A total of 1,600 first-grade students were screened for Fusion eligibility. Of these students, 680 met eligibility criteria and were randomly assigned within each of the 60 classrooms to 2:1 Fusion (n = 131), 5:1 Fusion (n = 339), or the control condition (n = 210). This study focused on the 470 Fusion students only. Demographic data for these students indicated that 17% received special education services, 17% were English learners, and 51% were females. Although the majority racial group of Fusion students was White (56%), 28% were Hispanic, 4% were Black, 3% Asian, 1% were American Indian, and 7% were multiple races.

Interventionists. Fusion intervention groups were taught by district-employed instructional assistants and by interventionists hired specifically for this study. The majority of interventionists identified as female (91%) and White (79%), with 6% identifying as Hispanic. The remaining 15% identified as another race or ethnicity or declined to respond. Most interventionists (88%) had previous experience teaching small groups, and 80% held a bachelor's degree or higher. Over half (62%) had taken a college level algebra course. On average, interventionists had 6.6 years of teaching experience (SD = 9.4) and 25% held a teaching certification.

Procedures

Fusion intervention. Fusion is a 60-lesson, Tier 2 first-grade mathematics intervention aimed at building students' proficiency with critical concepts and skills of whole number. Each 30-min lesson addresses mathematical content from two strands focused on whole number understanding of the first-grade Common Core State Standards for Mathematics: (a) operations and algebraic thinking and (b) number and operations in base-10. Fusion's scope and sequence introduces new concepts and skills in "tracks," with students practicing a variety of different skills in each lesson. Activities within lessons build over time as increasingly advanced content is introduced. This sequencing allows for frequent cumulative review and promotes students' maintenance of mathematical skills across time.

In the first half of Fusion (Lessons 1–30), students develop proficiency with numbers up to 100 through identifying, modeling, writing, and sequencing activities. Students are also explicitly taught strategies to fluently recall addition and subtraction number combinations within 20. As lessons progress, students encounter increasingly complex content to expand on skills taught earlier in the program. After building fluency with two-digit numbers, students are taught two-digit addition and subtraction, and to mentally add or subtract 10 from a two-digit number. To add to their repertoire of number combinations, students learn doubles facts and number families (e.g., 3, 4, and 7). Students also develop a deep understanding of mathematical problem-solving, learning the underlying structures of the various word problem types identified in the first-grade Common Core State Standards for Mathematics, including “add to, take from, put together, and take apart problems.”

Fusion incorporates mathematical models through a concrete-representational-abstract (CRA) sequence to build students’ conceptual understanding. For example, when teaching place value of two-digit numbers, students first use base-10 blocks and unit cubes to model the tens and ones in a given number (concrete model), then use place value charts with manipulatives or written numerals (representational model), and last verbalize or write the number of tens and ones in a given number (abstract model). In Fusion, the CRA sequence is typically introduced across multiple lessons, providing students with a high degree exposure to visual representations before fading to the exclusive use of mathematical symbols. Other mathematical models used in the program include number lines, number family cards, linking cubes, layered place value cards, and a hundreds chart to support students’ conceptual understanding.

To promote high-quality mathematics instruction, the Fusion intervention offers scripted lessons to support teachers in (a) delivering demonstrations and explanations of targeted mathematics content, (b) facilitating individual and group practice opportunities, and (c) offering timely academic feedback. The lesson scripting also enables teachers’ use of precise and consistent mathematical language and to promote high-quality instructional interactions centered on whole number concepts and skills. These interactions are intended to facilitate deep mathematical thinking and reasoning, through individual and group-level mathematics verbalizations (Gersten et al., 2009). For example, when teaching the commutative law of addition, the interventionist writes two problems on the board (e.g., 3 + 1 =, 1 + 3 =) and asks students to discuss with their partner how the two problems are alike. In the latter part of the lesson, students practice explaining the commutative law to their partner using their own words.

In this study, the Fusion intervention was delivered in 30-min, small group formats (i.e., two or five students per interventionist), 5 days per week for approximately 12 weeks. Because Fusion is designed as a supplemental intervention, instruction occurred at times that did not conflict with core Tier 1 mathematics instruction. For all students, instruction began in early winter and ended in the spring. The early winter start date was selected to provide students with opportunities to respond to core mathematics instruction and therefore minimize the false identification of typically achieving students during the screening process.

Professional development. All interventionists participated in two 4-hr professional development workshops delivered by project staff. The first workshop was held prior to the start of implementation of the Fusion intervention and focused on content from Lessons 1 to 30, whereas the second focused on Lessons 31 to 60. Both workshops centered on EBPs in early mathematics instruction, and strategies for managing small-group instruction and potential behavioral issues. Staff leading the workshops explicitly modeled instructional practices, such as group response signals, immediate correction of student errors, and pacing of activities within lessons. Interventionists were provided opportunities to practice and receive feedback on lesson delivery from project staff. To promote implementation fidelity and enhance the quality of instruction, all interventionists received, on average, two coaching visits during Fusion implementation. Coaching visits consisted of direct observations of lesson delivery, followed by feedback on instructional quality and fidelity of Fusion implementation.

Fidelity of implementation. Fidelity of Fusion implementation was measured on 393 occasions via direct observations. Each Fusion group was observed approximately three times during the course of the intervention. Trained research staff conducted all fidelity checks and observed instruction for the duration of each lesson targeted to assess fidelity (~30 min). On a 4-point scale (4 = *all*, 3 = *most*, 2 = *some*, 1 = *none*), observers rated the extent to which the interventionist (a) met the lesson’s instructional objectives, (b) followed the lesson’s teacher scripting, and (c) used the lesson’s mathematics models. Observers also recorded whether the interventionist taught the number of activities prescribed in the lesson. Interventionists were found to meet instructional objectives (M = 3.3, SD = 0.5), follow scripting (M = 3.2, SD = 0.5), and use prescribed models (M = 3.5, SD = 0.5). The majority of prescribed activities were also taught (M = 5.1 of 8 activities per lesson, SD = 0.5).

Student Mathematics Outcome Measures

Students were administered three mathematics outcome measures at pretest (T_1) and posttest (T_2). These measures focused on critical whole number concepts and skills. Trained research staff administered all student measures, and interscorer reliability criteria ≥.85 were met for all assessments.

ProFusion is a researcher-developed assessment designed to assess students’ conceptual and procedural knowledge of number and numeration, place-value concepts, basic number combinations, and problems involving multidigit addition and subtraction. In an untimed, small-group setting, students are asked write numbers from dictation and numbers missing from a sequence, write numbers matching base-10 block models, and decompose double-digit numbers. Moreover, students complete addition and subtraction problems, and two word problems. Students also complete

1-min, timed addition and subtraction fluency measures and work with proctors individually to complete a set of number-identification items. Criterion validity of ProFusion with other mathematics outcome measures, including the SAT-10, ranges (r) from .56 to .68 (Clarke et al., 2014).

Test of Early Mathematics Ability–Third Edition *(TEMA-3).* The TEMA-3 is a standardized, norm-referenced, individually administered measure of beginning mathematical ability (Ginsburg & Baroody, 2003). 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 math, 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 mathematics outcome measures, the TEMA-3 manual reports coefficients ranging from .54 to .91.

ASPENS. ASPENS is a set of curriculum based measures (CBMs) validated for screening and progress monitoring in first-grade mathematics (Clarke et al., 2011). Each 1-min fluency-based measure assesses an important aspect of early numeracy proficiency, including magnitude comparison, missing number identification, and arithmetic facts and base-10. Test authors report test–retest reliability ranges from .70 to .90. Criterion concurrent validity with the TerraNova 3 is reported as ranging from .51 to .63.

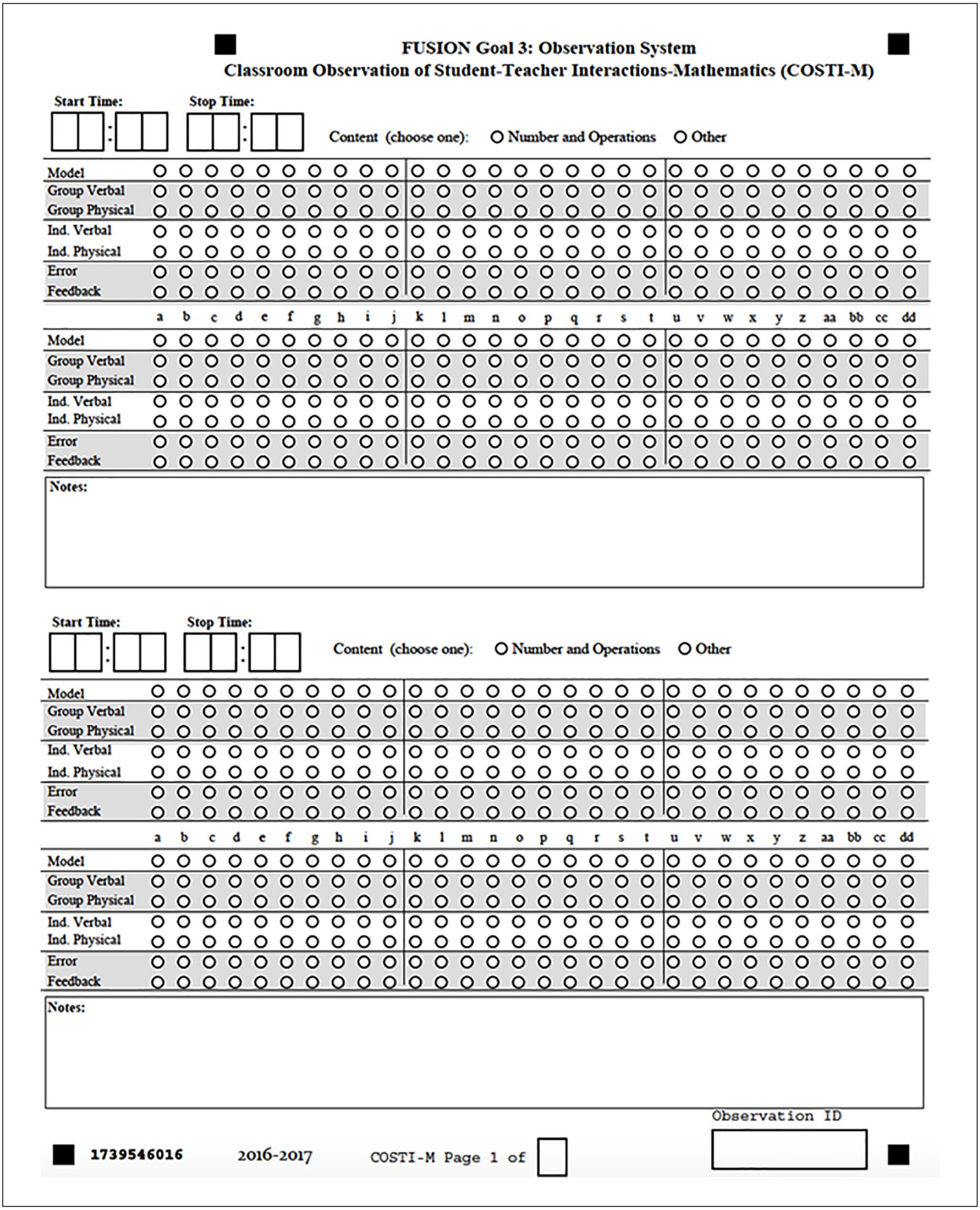
Observations of Fusion Instruction

Each Fusion group was observed approximately three times over the course of the intervention, with approximately 3 weeks separating each observation occasion. A total of 393 observations were conducted, of which 94 (24%) included two observers who simultaneously evaluated interobserver agreement. Observations were scheduled in advance and observers remained for the duration of Fusion instruction, with an average observation lasting 25.2 min (SD = 2.7). Trained observers, who were blind to our research hypotheses, used two observation measures to conduct all observations.

Classroom Observations of Student–Teacher Interactions–Mathematics *(COSTI-M)*. The COSTI-M is a low-inference observation instrument that has been empirically validated to document the frequency of teacher demonstrations, individual and group student practice opportunities, teacher-provided academic feedback, and student mistakes (Doabler et al., 2015; Smolkowski & Gunn, 2012). Figure 1 illustrates the COSTI-M, the EBPs that compose the instructional interactions targeted by the COSTI-M, the series of bubble coding columns to document the frequency of explicit instructional interactions, and the serial method of coding. As documented by the COSTI-M, teacher models represent a teacher's verbalizations of thought processes and physical demonstrations of mathematical content. For example, observers coded a teacher model if the teacher explicitly described the structural features of an "add to" word problem. Academic feedback was operationalized as a teacher's verbal reply or physical demonstration to affirm or correct a student response. For example, observers recorded an academic feedback code if the teacher verbally corrected a student mistake. Group practice opportunities were defined as a mathematics-related verbalization produced by two or more students in unison. Individual practice opportunities were coded whenever a single student had the opportunity to verbalize or physically demonstrate her mathematical thinking, such as when a teacher asked a specific student to answer a mathematical question (e.g., "Rafael, use the place value bocks to show 82?"). Rates per minute for each targeted behavior were computed as the frequency of the behavior divided by the duration of the observation in minutes. Doabler et al. (2015) reported predictive validity of the COSTI-M with the TEMA-3 (p = .004, Pseudo-R^2 = .08) and the EN-CBM (p = .017, Pseudo-R^2 = .05).

QEMI. The QEMI, a moderate-inference observation instrument, comprises seven items that target the quality of explicit instructional interactions, including group and individual practice opportunities, student participation, teacher modeling, academic feedback, efficiency of instructional delivery, and instructional scaffolding (Doabler & Clarke, 2012). Internal consistency of the measure is high, .93 (coefficient alpha). To rate the quality of each item, observers used a 4-point rating scale, with scores of 1 to 2 representing the *lower quality range* and 3 to 4 representing the *upper quality range*. Total QEMI scores were computed as the mean across all items. The mean across the observation occasions was used in subsequent analyses.

Observation training. Trained observers from Oregon and Massachusetts conducted all direct observations. The observers included former educators, doctoral students, faculty members, and experienced data collectors. Observers received approximately 10 hr of training, with an initial training lasting 6 hr and a 4-hr follow-up training prior to the third round of observations to recalibrate observers, help minimize observer drift, and increase interobserver reliability. Training focused on direct observation procedures, first-grade mathematics, and use of the COSTI-M and QEMI observation instruments. The training also addressed aspects of implementation fidelity specific to the Fusion program, focusing on the intervention's core components (i.e., explicit instructional design and delivery principles), targeted mathematical models (e.g.,



FUSION Goal 3: Observation System
Classroom Observation of Student-Teacher Interactions-Mathematics (COSTI-M)

Start Time: : Stop Time: :
Content (choose one): O Number and Operations O Other

Model
Group Verbal
Group Physical
Ind. Verbal
Ind. Physical
Error
Feedback

a b c d e f g h i j k l m n o p q r s t u v w x y z aa bb cc dd

Model
Group Verbal
Group Physical
Ind. Verbal
Ind. Physical
Error
Feedback

Notes:

Start Time: : Stop Time: :
Content (choose one): O Number and Operations O Other

Model
Group Verbal
Group Physical
Ind. Verbal
Ind. Physical
Error
Feedback

a b c d e f g h i j k l m n o p q r s t u v w x y z aa bb cc dd

Model
Group Verbal
Group Physical
Ind. Verbal
Ind. Physical
Error
Feedback

Notes:

Observation ID

1739546016 2016-2017 COSTI-M Page 1 of

Figure 1. COSTI-M, showing the EBPs, the series of bubble coding columns, and the serial method of coding.

place value blocks), and lesson structure. 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-min video of first-grade mathematics 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 agreement and stability intraclass correlations coefficients (ICCs). To estimate interobserver agreement in observation measures, we calculated ICCs to describe the proportion of variance in each observation measure occurring between versus within paired observation occasions. A total of 94 paired reliability observations were conducted. Interobserver agreement ICCs for COSTI-M and QEMI scores ranged from .63 to .99, which based on guidelines proposed by Landis and Koch (1977), represented substantial to nearly perfect agreement. To estimate stability across time, we calculated ICCs to describe the proportion of variance in each observation measure occurring between versus within Fusion groups. Stability ICCs for the COSTI-M were .11 for teacher demonstrations, .09 for individual practice, .37 for group practice, .33 for student mistakes, .45 for academic feedback, and .54 for the QEMI scale. Reliability of mean scores across the three observation occasions were fair and ranged from .23 (for individual practice) to .78 (for QEMI scores). The fair reliability of the mean score for individual practice opportunities suggests that additional observations beyond the three scheduled for each Fusion group may be necessary in future research.

Statistical Analysis

Following the examination of descriptive statistics for the study variables, we performed a series of random coefficients analyses (RCAs; Snijders & Bosker, 2012) designed to address our research questions. The statistical models accounted for pretest and posttest measures of mathematics achievement nested within students and Fusion groups. Specifically, we regressed mathematics achievement at pretest and posttest on time (coded 0 for pretest and 1 for posttest), a group-level quantity or quality of explicit instruction predictor variable (mean centered), and the cross-level Time × Predictor interaction. The effect of time represents the average change in outcome from pretest to posttest among groups given the average value of the predictor variable. The effect of the quantity or quality predictor variable addresses Research Question 1 and represents the association between group-level mathematics achievement at pretest and the specific measure of the quantity or quality of explicit instructional interactions. The Time × Predictor interaction addresses Research Question 2 and represents the difference in change in mathematics outcome from pretest to posttest due to a unit increase in the quantity or quality of explicit instructional interactions. To support interpretation of results, we reported r^2 equivalent (Rosnow & Rosenthal, 2003) for the fixed effects of the quantity or quality predictor variable (Research Question 1) and the Time × Predictor interaction (Research Question 2). Alpha was set to .05.

We performed analyses using SAS PROC MIXED version 9.4 (SAS Institute, 2016) and restricted maximum likelihood estimation. Maximum likelihood estimation uses all available data and produces potentially unbiased results even in the face of substantial missing data, provided the missing data were missing at random (Schafer & Graham, 2002). We considered this assumption tenable because missing data (<6% missing for each outcome measure) mostly involved students who were absent on the day of assessment (e.g., due to illness) or transferred to a new school (e.g., family mobility). The statistical model also assumes independent and normally distributed observations. We addressed the first of these assumptions by modeling the multilevel nature of the data. The outcome measures in the present study also did not markedly deviate from normality; skewness and kurtosis fell within ±.89.

Results

Table 1 provides mean values, standard deviations, and sample sizes for each student outcome and observed measure of the quantity and quality of instructional interactions. Measures of the quantity and quality of explicit instructional interactions were correlated between $r = \pm.01$ to .37 (see Table 2). Tables 3 to 5 summarize results of the RCAs for our two research questions.

Research Question 1 focused on the associations between group-level pretest mathematics achievement and measures of the quantity and quality of explicit instructional interactions. These associations were evaluated by the fixed effects of each quantity or quality predictor presented in the second row of data in Tables 3 to 5. Results demonstrated statistically significant associations between pretest mathematics performance and rates of student mistakes and QEMI scores. Specifically, the second row of Table 3 shows that lower pretest TEMA scores were associated with higher rates of student mistakes ($p < .0001$, $r^2_{\text{equivalent}} = .121$) and lower QEMI scores ($p = .0117$, $r^2_{\text{equivalent}} = .048$). Table 4 indicates that lower pretest ASPENS scores were associated with higher rates of student mistakes ($p = .0069$, $r^2_{\text{equivalent}} = .055$) and lower QEMI scores ($p = .0250$, $r^2_{\text{equivalent}} = .038$). Table 5 shows that lower pretest ProFusion scores were associated with higher rates of student mistakes ($p < .0001$, $r^2_{\text{equivalent}} = .163$) and lower QEMI scores ($p = .0318$, $r^2_{\text{equivalent}} = .035$). No

Table 1. Descriptive Statistics for Student- and Group-Level Variables.

Variable	*M* (*SD*)	*n*
Student outcomes		
TEMA-3 pretest	34.7 (7.2)	462
TEMA-3 posttest	42.3 (8.2)	427
ASPENS pretest	21.0 (11.5)	470
ASPENS posttest	43.4 (16.8)	437
ProFusion pretest	25.1 (9.2)	463
ProFusion posttest	45.3 (11.5)	436
Group-level quantity and quality variables		
Teacher demonstrations per minute	0.1 (0.1)	134
Individual practice opportunities per minute	2.5 (0.6)	134
Group practice opportunities per minute	1.2 (0.7)	134
Student mistakes per minute	0.3 (0.2)	134
Teacher-provided academic feedback per minute	1.7 (0.7)	134
Quality of explicit mathematics instruction	3.2 (0.5)	134

Note. Observation measures were aggregated across approximately three observation occasions per Fusion group. TEMA-3 = *Test of Early Math Achievement–Third Edition*; ASPENS = Assessing Student Proficiency in Early Number Sense.

Table 2. Pearson Correlation Coefficients Among Measures of the Quantity and Quality of Explicit Instructional Interactions.

Variable	1	2	3	4	5
1. Teacher demonstrations per minute					
2. Individual practice opportunities per minute	–.09				
3. Group practice opportunities per minute	.10	–.21*			
4. Student mistakes per minute	.10	.23**	–.10		
5. Teacher-provided academic feedback per minute	.08	.21*	.15	–.01	
6. Quality of explicit mathematics instruction	.15	–.04	.44***	.10	.37***

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

Table 3. Random Coefficients Analysis of the Quantity and Quality of Explicit Instructional Interactions Predicting TEMA-3 Scores.

		Predictor					
Model parameters		Teacher demonstrations	Individual practice	Group practice	Student mistakes	Academic feedback	QEMI
Fixed Effects	Intercept	34.8 (0.4)***	34.7 (0.4)***	34.7 (0.4)***	34.8 (0.4)***	34.7 (0.4)***	34.7 (0.4)***
	Predictor	–4.6 (5.3)	–0.2 (0.7)	1.2 (0.6)	–10.5 (2.5)***	–0.3 (0.6)	2.1 (0.8)*
	Time	7.4 (0.3)***	7.4 (0.3)***	7.4 (0.3)***	7.5 (0.3)***	7.5 (0.3)***	7.5 (0.3)***
	Predictor × Time	6.0 (3.9)	–0.4 (0.5)	1.2 (0.5)*	–4.4 (2.0)*	0.1 (0.5)	0.4 (0.6)
Variances	Residual	15.2 (1.2)***	15.3 (1.2)***	15.2 (1.2)***	15.2 (1.2)***	15.3 (1.2)***	15.3 (1.2)***
	Student gains	35.9 (3.4)***	35.9 (3.4)***	35.6 (3.4)***	35.1 (3.3)***	35.8 (3.4)***	36.0 (3.4)***
	Group intercept	7.0 (2.7)**	6.9 (2.7)**	5.9 (2.5)*	3.6 (2.2)	7.0 (2.7)**	5.3 (2.5)*
	Group gains	1.7 (0.9)	1.7 (0.9)*	1.6 (0.8)	1.8 (0.9)*	1.8 (0.9)*	1.7 (0.9)
p values	Predictor	.3890	.8094	.0651	<.0001	.6809	.0117
	Predictor × Time	.1262	.4614	.0101	.0299	.8984	.4744
$r^2_{\text{equivalent}}$	Predictor	.006	.000	.026	.121	.001	.048
	Predictor × Time	.018	.004	.050	.036	.000	.004

Note. Table entries show parameter estimates with standard errors in parentheses. Tests of fixed effects used 130 degrees of freedom. TEMA-3 = *Test of Early Mathematics Achievement–Third Edition*; QEMI = Quality of Explicit Mathematics Instruction.
*$p < .05$. **$p < .01$. ***$p < .001$.

Table 4. Random Coefficients Analysis of the Quantity and Quality of Explicit Instructional Interactions Predicting ASPENS Scores.

Model parameters		Predictor					
		Teacher demonstrations	Individual practice	Group practice	Student mistakes	Academic feedback	QEMI
Fixed Effects	Intercept	20.9 (0.8)***	21.0 (0.8)***	20.9 (0.7)***	21.0 (0.7)***	20.9 (0.8)***	21.0 (0.7)***
	Predictor	3.1 (9.3)	0.1 (1.2)	2.0 (1.1)	−12.5 (4.6)**	−1.9 (1.1)	3.3 (1.5)*
	Time	22.3 (0.7)***	22.3 (0.7)***	22.4 (0.7)***	22.5 (0.7)***	22.4 (0.7)***	22.4 (0.7)***
	Predictor × Time	13.9 (8.7)	–0.8 (1.1)	1.6 (1.1)	–9.6 (4.4)*	1.4 (1.1)	0.5 (1.4)
Variances	Residual	85.8 (6.7)***	85.7 (6.7)***	85.5 (6.7)***	85.5 (6.7)***	85.8 (6.7)***	85.8 (6.7)***
	Student gains	96.8 (11.3)***	97.0 (11.4)***	96.2 (11.3)***	96.1 (11.2)***	97.0 (11.4)***	97.2 (11.4)***
	Group intercept	15.4 (8.1)	15.8 (8.2)	13.3 (7.8)	9.1 (7.2)	15.3 (8.1)	12.0 (7.8)
	Group gains	6.3 (4.2)	6.7 (4.3)	6.8 (4.3)	6.4 (4.2)	6.2 (4.3)	6.8 (4.3)
p values	Predictor	.7373	.9247	.0724	.0069	.1008	.0250
	Predictor × Time	.1122	.4839	.1419	.0319	.1733	.6958
$r^2_{\text{equivalent}}$	Predictor	.001	.000	.025	.055	.021	.038
	Predictor × Time	.019	.004	.017	.035	.014	.001

Note. Table entries show parameter estimates with standard errors in parentheses. Tests of fixed effects used 130 degrees of freedom.
ASPENS = Assessing Student Proficiency in Early Number Sense; QEMI = Quality of Explicit Mathematics Instruction.
$*p < .05$. $**p < .01$. $***p < .001$.

Table 5. Random Coefficients Analysis of the Quantity and Quality of Explicit Instructional Interactions Predicting ProFusion Scores.

Model parameters		Predictor					
		Teacher demonstrations	Individual practice	Group practice	Student mistakes	Academic feedback	QEMI
Fixed Effects	Intercept	25.2 (0.6)***	25.2 (0.6)***	25.2 (0.6)***	25.3 (0.5)***	25.2 (0.6)***	25.3 (0.6)***
	Predictor	3.0 (7.1)	–0.8 (0.9)	1.5 (0.8)	−16.4 (3.3)***	–0.6 (0.9)	2.4 (1.1)*
	Time	20.1 (0.5)***	20.1 (0.5)***	20.0 (0.5)***	20.1 (0.5)***	20.1 (0.5)***	20.0 (0.5)***
	Predictor × Time	–5.4 (5.7)	0.4 (0.7)	1.6 (0.7)*	–4.1 (3.0)	1.5 (0.7)*	1.0 (0.9)
Variances	Residual	37.9 (3.0)***	37.8 (3.0)***	37.9 (3.0)***	37.9 (3.0)***	37.8 (3.0)***	37.8 (3.0)***
	Student gains	55.0 (5.9)***	55.0 (5.9)***	54.7 (5.9)***	53.9 (5.8)***	55.0 (5.9)***	55.3 (6.0)***
	Group intercept	13.1 (4.8)**	12.8 (4.8)**	11.1 (4.5)*	5.4 (3.7)	13.2 (4.8)**	10.4 (4.5)*
	Group gains	2.5 (1.9)	2.6 (1.9)	2.1 (1.8)	2.6 (1.9)	2.1 (1.8)	2.6 (1.9)
p-values	Predictor	.6684	.3927	.0688	<.0001	.4948	.0318
	Predictor × Time	.3506	.6269	.0245	.1697	.0337	.2902
$r^2_{\text{equivalent}}$	Predictor	.001	.006	.025	.163	.004	.035
	Predictor × Time	.007	.002	.038	.014	.034	.009

Note. Table entries show parameter estimates with standard errors in parentheses. Tests of fixed effects used 130 degrees of freedom. QEMI = Quality of Explicit Mathematics Instruction.
$*p < .05$. $**p < .01$. $***p < .001$.

significant associations emerged between pretest mathematics performance and rates of teacher demonstrations ($p \geq .3890$), individual practice opportunities ($p \geq .3927$), group practice opportunities ($p \geq .0651$), or academic feedback ($p \geq .1008$).

For Research Question 2, we examined whether the quantity or quality of explicit instructional interactions predicted gains in student mathematics achievement from pretest to posttest. We evaluated this question using the Time × Predictor interactions presented in the fourth row of data in Tables 3 to 5. Results indicated that gains in mathematics achievement were significantly associated with rates of group practice opportunities, student mistakes, and academic feedback. Specifically, the fourth row of Table 3 shows that greater gains in TEMA-3 scores were associated with higher rates of group practice ($p = .0101$, $r^2_{\text{equivalent}} = .050$) and lower rates of student mistakes ($p = .0299$, $r^2_{\text{equivalent}} = .036$). Table 4 shows that greater gains in ASPENS scores were associated with lower rates of student mistakes ($p = .0319$, $r^2_{\text{equivalent}} = .035$). Table 5 shows that greater gains in ProFusion scores were associated with higher rates of group practice ($p = .0245$, $r^2_{\text{equivalent}} = .038$) and academic feedback

($p = .0337$, $r^2_{\text{equivalent}} = .034$). No significant associations emerged between gains in mathematics achievement and rates of teacher models ($p \geq .1122$), individual practice ($p \geq .4614$), or QEMI scores ($p \geq .2902$).

Discussion

Teachers' use of EBPs in today's classrooms is integral to getting at-risk students on track for developing mathematical proficiency. This study examined "practice-based evidence" (Green, 2008) of teachers' actual use of explicit mathematics instruction, a well-established EBP. Specifically, we analyzed direct observation data from a recent federally funded efficacy trial to explore the quality and frequency in which teachers facilitated explicit instructional interactions around foundational mathematics concepts and skills. Two research questions were addressed. Findings from our research questions and implications for the field are discussed below.

Results Summary

Research Question 1. Our first research question addressed whether and to what extent group-level initial mathematics skill predicted the quantity and quality of explicit instructional interactions facilitated during Fusion instruction. Analyses focused on students' initial academic skill levels have garnered recent attention in the field (Fuchs & Fuchs, 2019). In the current study, initial mathematics achievement was based on group-level pretest performances on the three mathematics outcome measures focused on whole numbers and operations. Findings indicated significant associations between initial group-level mathematics performance and rates of student errors and quality ratings of explicit instruction. Specifically, students in Fusion groups with lower performance at the start of the intervention made more errors than students in groups that started with higher initial skill levels. Interestingly, we found that higher performing groups received higher quality explicit mathematics instruction, suggesting students in these groups received richer instructional interactions during Fusion instruction. For example, higher performing groups received higher quality teacher models, practice opportunities, and feedback. While not privy to students' pretest scores, it may be that interventionists of the higher performing groups recognized students in these groups were prepared to handle more in-depth instructional interactions, such as receiving more complex explanations of targeted whole number concepts and skills. This finding is also encouraging because it lends further credibility to the notion of instructional assistants or paraprofessionals serving as effective change agents in today's schools, particularly when they have access to well-designed, evidence-based mathematics interventions (Pellegrini et al., 2018).

Nonsignificant associations were found between students' group-level mathematics performances at pretest and rates of teacher models, individual practice, and group practice. These findings were similar to that reported in one of our previous studies, where we explored "practice-based evidence" of explicit mathematics instruction facilitated during small-group, kindergarten mathematics instruction (Doabler et al., 2020). In that earlier study, at-risk kindergartners received a validated kindergarten mathematics intervention focused on early number sense concepts. Parallel to this previous research, the current study found that intervention groups received similar rates of teacher models and group and individual practice opportunities, regardless of their initial mathematics skill levels at intervention onset.

Rates of academic feedback in the current study also did not produce significant results. Interestingly, this finding runs contrary to results from some of our earlier intervention work, which suggests that interventionists tend to go above and beyond the academic feedback opportunities scripted in explicit mathematics interventions when teaching lower performing groups (Doabler et al., 2017). In the current study, interventionists may have felt that the amount of scripting offered in Fusion for academic feedback was sufficient enough to meet the instructional needs of groups composed of at-risk students with more intensive learning needs. Future research, however, is warranted to unpack as to why some interventionists tend to provide more academic feedback than that prescribed in explicit mathematics intervention programs.

Research Question 2. Our second research question focused on associations between the quantity and quality of explicit instructional interactions and gains in student mathematics achievement across the Fusion intervention time period. Results from this question suggested that quality ratings of explicit instruction were not significantly associated with gains in student mathematics outcomes. Our analyses also indicated nonsignificant associations between student mathematics achievement and rates of individual practice. Interestingly, this finding is orthogonal to research involving the full range of learners and whole-class mathematics instruction (Doabler et al., 2015; Clements et al., 2013). Yet, it aligns with recent studies involving at-risk learners and small-group mathematics instruction (Doabler et al., 2017; Doabler et al., 2020). One conjecture as to why individual practice does not demonstrate high value for at-risk learners in small-group settings is based on confidence. It may be that students with MLD lack self-assurance in demonstrating their individual understanding of mathematics in front of their peers even when in small-group settings. Another possibility is that the Fusion intervention contains too few individual practice opportunities to predict increased student mathematics outcomes. Regardless of the reason, be it student confidence or insufficient practice opportunities, we encourage curriculum developers and interventionists to continue offering at-risk learners individual practice opportunities in small-group settings until future research can better unveil the value of individual practice for students with or at risk for MLD.

Our findings also indicated significant associations between positive gains in student mathematics outcomes and higher rates of group practice opportunities. Choral responses are a backbone of explicit mathematics instruction and when facilitated well they can effectively engage all students in practice opportunities (Hughes et al., 2017). As such, we encourage the continued use of this type of practice opportunity when working with groups of at-risk learners in small-group settings. Findings also indicated that Fusion groups with the lowest rates of errors during Fusion instruction made the greatest gains from pretest to posttest. While errors are a part of the learning process, lower rates of incorrect responses may be indicative of effective mathematics instruction. Finally, results indicated significant associations between positive gains in student mathematics outcomes and higher rates of academic feedback. Similar to group practice opportunities, academic feedback is an integral component of explicit mathematics instruction, with teachers using this EBP to rectify student errors and extend learning opportunities. Given the body of evidence behind academic feedback (Halpern et al., 2007), a recommendation is that teachers deliver timely, specific feedback when students with or at risk for MLD engage in group-level and individual practice opportunities.

Limitations

Several limitations should be considered when interpreting our results. One limitation of the current study was that it included only one mathematics intervention. Future research should expand this work to other programs and other grade levels. A second limitation was the number of observations conducted. Our research included three observations per Fusion intervention group. Direct observations can be cost intensive, particularly when conducted in the context of large-scale efficacy trials. Such studies often include multiple school districts located in different geographical regions. These aspects can quickly drive up the costs for conducting real-time observations and thus force research teams to decide on an acceptable, albeit affordable, number of observations to conduct. Similar resource decisions were encountered in the current study. Relatedly, the limited number of observations conducted within each Fusion group may have impacted our capacity to minimize measurement error due to the instability or volatility of the EBPs observed (Doabler et al., 2018). This type of measurement error may explain the nonsignificant associations between rates of individual responses and student outcomes.

Implications

A growing body of research highlights the importance of engaging at-risk learners in explicitly designed instructional interactions (Doabler et al., 2015; Clements et al., 2013; Gersten et al., 2009). As such, the practical significance of supporting teachers' provision of frequent, high-quality instructional interactions may be paramount to increasing mathematics achievement among students with or at risk for MLD. However, equipping teachers with EBPs and having them use such practices with fidelity can be difficult. Implementation challenges, such as intervention feasibility or lack thereof, are common causes of the breaks in the implementation pipeline of EBPs (Onken et al., 2014). To fix the EBP pipeline in the field of education, experts recommend leveraging implementation science frameworks through all phases of intervention research (Smolkowski et al., 2019). This way researchers can address implementation challenges and, in turn, increase the probability of teachers' actual use of EBPs in today's classrooms. For example, to strengthen the fidelity of EBPs in the current study, we upped the implementation support (i.e., in-class coaching) to those interventionists who demonstrated further need.

Another implication is the use of a dual approach to simultaneously capture the quantity and quality of instructional interactions in real time. Little research has involved both observed quantity and quality of instructional interactions. A common feature of extant observational research is a singular focus on instructional interactions, having documented either quantity or quality, but not both. Yet, collecting evidence on both quantity and quality may provide the field with a more comprehensive picture of the enacted use of explicit instructional interactions.

Finally, our research may shine further light on the use of direct observation for capturing "practice-based evidence" (Green, 2008) of teachers' use of EBPs. Findings from the current study align with previous observation research (e.g., Clements et al., 2013; Smolkowski & Gunn, 2012) and thus lend further credence for the calls to continue the advancement and implementation of classroom observation instruments that offer actionable metrics of effective instruction (Pianta & Hamre, 2009). Standardized observation protocols, such as the CLASS (Pianta & Hamre, 2009), Instructional Content Emphasis (Edmonds & Briggs, 2003), Assessment-to-Instruction (Connor, 2019), and the COSTI-M (Doabler et al., 2015) have potential to improve the field's understanding of the connection between increased student achievement and instructional factors, such as the quantity and quality of EBPs.

Conclusion

Over the past several decades, the field of education science has come leaps and bounds in terms of establishing itself as an education science (Morris & Reardon, 2017). Showcasing this progress in education is the number of EBPs made available in classroom instruction. Although encouraging, additional research is needed to gain a deeper understanding of teachers' use of EBPs. One measurement tactic that can reliably and validly collect "practice-based evidence" (Green, 2008) of EBPs is direct observation. Through a continued program of direct observation research, the field may be able to unpack the "black box" of EBPs, gaining an understanding how and why these practices impact the achievement of students who are at risk for academic failure.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Drs. Ben Clarke and Christian Doabler are eligible to receive a portion of royalties from the University of Oregon's distribution and licensing of certain Fusion-based works. Potential conflicts of interest are managed through the University of Oregon's Research Compliance Services. In addition, the terms of this arrangement have been reviewed and approved by The University of Texas at Austin in accordance with its policy on objectivity in research.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education through Grant R324A160046.

References

Alfieri, L., Brooks, P., Aldrich, N., & Tenenbaum, H. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*, 1–18. https://doi.org/10.1037/a0021017

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. *Elementary School Journal*, *107*, 199–220. https://doi.org/10.1086/510655

Baldwin, S. A., Bauer, D. J., Stice, E., & Rhode, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods*, *16*(2), 149–165. https://doi.org/10.1037/a0023464

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). Macmillan.

Carnine, D. (1997). Bridging the research-to-practice gap. *Exceptional Children*, *63*, 513–521.

Clarke, B., Doabler, C. T., & Fien, H. (2016). *A randomized control trial of a tier 2 first-grade mathematics intervention*. Washington, DC: U.S. Department of Education; Institute of Education Sciences, National Center on Special Education Research.

Clarke, B., Doabler, C. T., Strand Cary, M., Kosty, D., Baker, S. K., Fien, H., & Smolkowski, K., & (2014). Preliminary evaluation of a tier-2 mathematics intervention for first-grade students: Using a theory of change to guide formative evaluation activities. *School Psychology Review*, *43*, 160–177.

Clarke, B., Gersten, R. M., Dimino, J., & Rolfhus, E. (2011). *Assessing student proficiency of number sense (ASPENS)*. Cambium Learning Group, Sopris Learning.

Clements, D. H., Agodini, R., & Harris, B. (2013). *Instructional practices and student math achievement: Correlations from a study of math curricula* (NCEE Evaluation Brief No. 2013–4020). National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences.

Connor, C. M. (2019). Using technology and assessment to personalize instruction: Preventing reading problems. *Prevention Science*, *20*, 89–99. https://doi.org/10.1007/s11121-017-0842-9

Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., & Test, D. W. (2015). CEC's standards for classifying the evidence base of practices in special education. *Remedial and Special Education*, *36*(4), 220–234.

Cook, B. G., Cook, L., & Landrum, T. J. (2013). Making research into practice: Can we make dissemination stick? *Exceptional Children*, *2*, 163–180.

Cook, B. G., & Cook, S. C. (2011). Unraveling evidence-based practices in special education. *Journal of Special Education*, *47*, 71–82

Dennis, M. S., Sharp, E., Chovanes, J., Thomas, A., Burns, R. M., Custer, B., & Park, J. (2016). A meta-analysis of empirical research on teaching students with mathematics learning difficulties. *Learning Disabilities Research & Practice*, *31*, 156–168. https://doi.org/10.1111/ldrp.12107

Deshler, D. D. (2015). Moving in the right direction but at what speed, and how smoothly? *Remedial and Special Education*, *36*(2), 72–76. doi:10.1177/0741932514558093

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

Doabler, C. T., & Clarke, B. (2012). *Quality of explicit mathematics instruction*. Unpublished observation instrument, Center on Teaching and Learning. University of Oregon, Eugene.

Doabler, C.T., Clarke, B., Stoolmiller, M., Kosty, D., Fien, F., Smolkowski, K., & Baker, S.K. (2017). Explicit instructional interactions: Exploring the black box of a tier 2 mathematics intervention. *Remedial and Special Education*, *32*, 98–110. doi:10.1177/0741932516654219

Doabler, C.T., Nelson, N. J., Kennedy, P., Stoolmiller, M., Fien, H., Clarke, B., Gearin, B., Smolkowski, K., & Baker, S. (2018). Investigating the longitudinal effects of a core kindergarten mathematics program on teachers' use of evidence-based teaching practices. *Learning Disability Quarterly*, *41*, 144–158. doi:10.1177/0731948718756040

Doabler, C. T., Clarke, B., Kosty, D., Fien, H., Smolkowski, K., Liu, M., & Baker, S. K. (2020). Measuring the quantity and quality of explicit instructional interactions in an empirically-validated tier 2 kindergarten mathematics intervention. *Learning Disability Quarterly*.

Edmonds, M. S., & Briggs, K. L. (2003). Instructional content emphasis instrument. In S. R. Vaughn & K. L. Briggs (Eds.), *Reading in the classroom: Systems for observing teaching and learning* (pp. 31–52). Baltimore, MD: Paul H. Brookes.

Fuchs, D., & Fuchs, L. S. (2019). On the importance of moderator analysis in intervention research: An introduction to the special issue. *Exceptional Children*, *85*, 126–128. https://doi.org/10.1177/0014402918811924

Gersten, R., Chard, D. J., 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. https://doi.org/10.3102/0034654309334431

Gersten, R., & Dimino, J. (2001). The realities of translating research into classroom practice. *Learning Disabilities Research & Practice*, *16*, 120–130.

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. https://doi.org/10.1177/07419325050260040201

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of Early Mathematics Ability–Third Edition* (TEMA-3). Pro-Ed.

Green, L. W. (2008). Making research relevant: If it is an evidence-based practice, where's the practice-based evidence? *Family Practice*, *25*, i20–i24.

Halpern, D., Aronson, J., Reimer, N., Simpkins, S., Star, J., & Wentzel, K. (2007). *Encouraging girls in math and science* (NCER 2007-2003). National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Hill, H. C., Blunk, M. L., Charalambos, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Loewenberg Ball, D. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*, 430–511.

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.

Hughes, C. A., Morris, J. R., Therrien, W. J., & Benson, S. K. (2017). Explicit instruction: Historical and contemporary contexts. *Learning Disabilities Research & Practice*, *32*(3), 140–148. https://doi.org/10.1111/ldrp.12142

Hughes, C. A., & Riccomini, P. J. (2019). Purposeful independent practice procedures: An introduction to the special issue. *TEACHING Exceptional Children*, *51*, 405–408.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

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.

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

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities*, *42*, 306–321. https://doi.org/10.1177/ 0022219408331037

Morris, P. A., & Reardon, S. F. (2017). Moving education science forward by leaps and bounds: The need for interdisciplinary approaches to improving children's educational trajectories. *Journal of Research on Educational Effectiveness*, *10*(1), 1–6.

Munter, C., Wilhelm, A. G., Cobb, P., & Cordray, D. S. (2014). Assessing fidelity of implementation of an unprescribed, diagnostic mathematics intervention. *Journal of Research on Educational Effectiveness*, *7*, 83–113. https://doi.org/10.1080/19345747.2013.809177

National Research Council. (2001). *Adding it up: Helping children learn mathematics*. Mathematics Learning Study Committee.

Nettina, S. M. (Eds.). (2019). *Lippincott manual of nursing practice* (11th ed.). Wolters Kluwer.

Onken, L. S., Carroll, K. M., Shoham, V., Cuthbert, B. N., & Riddle, M. (2014). Reenvisioning clinical science: Unifying the discipline to improve the public health. *Clinical Psychological Science*, *2*, 22–34. https://doi.org/10.1177/2167702613497932

Pellegrini, M., Lake, C., Inns, A., & Slavin, R. E. (2018). *Effective programs in elementary mathematics: A best-evidence synthesis*. Best Evidence Encyclopedia. www.bestevidence.org

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

Riccomini, P. J., & Morano, S. (2019). Guided practice for complex, multistep procedures in algebra: Scaffolding through worked solutions. *TEACHING Exceptional Children*, *51*, 445–454.

Rosnow, R. L., & Rosenthal, R. (2003). Effect sizes for experimenting psychologists. *Canadian Journal of Experimental Psychology*, *57*(3), 221–237.

SAS Institute. (2016). *Base SAS® 9.4 procedures guide: Statistical procedures* (5th ed.). SAS Institute. http://support.sas.com/documentation/index.html

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

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). Macmillan.

Simmons, D. (2015). Instructional engineering principles to frame the future of reading intervention research and practice. *Remedial and Special Education*, *36*, 45–51.

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.

Smolkowski, K., Crawford, L., Seeley, J. R., & Rochelle, J. (2019). Introduction to implementation science for research on learning disabilities. *Learning Disability Quarterly*, *42*, 192–203.

Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE.

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

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