Despina A. Stylianou, Rena Stroud, Michael Cassidy, Eric Knuth, Ana Stephens, Angela Gardiner & Lindsay Demers (2019) Putting early algebra in the hands of elementary school teachers: examining fidelity of implementation and its relation to student performance / *El álgebra temprana en manos del docente de primaria: un análisis de la fidelidad de ejecución y su relación con el rendimiento de los escolares*, Journal for the Study of Education and Development, 42:3, 523-569, DOI: 10.1080/02103702.2019.1604021

Putting Early Algebra in the Hands of Elementary School Teachers: Examining Fidelity of Implementation and its Relation to Student Performance

Despina A. Stylianou, The City College of New York
Rena Stroud, Merrimack College
Michael Cassidy, TERC
Eric Knuth, University of Texas at Austin
Ana Stephens, University of Wisconsin, Madison
Angela Gardiner, TERC
Lindsay Demers, Boston University

Guided by a theoretical framework emphasizing the importance of fidelity of implementation (FOI), this paper explores how Grades 3, 4 and 5 teachers implemented an early algebra intervention, and the relationship between FOI and student learning. The data shared in this paper come from a longitudinal experimental research project in which 3,208 students from 46 schools were followed for three years. Videotaped classroom observations, our primary source of FOI data, were coded to capture teachers' instructional practices, and an algebra assessment was given to assess student performance in response to the teachers' implementation of our instructional intervention. Results revealed a significant positive relationship between aspects of teachers' implementation and their students' performance.

Keywords: Algebra and Algebraic Thinking, Elementary School Education, Assessment and Evaluation, Fidelity of Implementation

Acknowledgements

The research reported here was supported by the US Department of Education under IES Award #R305A140092. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US Department of Education.

Putting Early Algebra in the Hands of Elementary School Teachers: Examining Fidelity of Implementation and its Relation to Student Performance

The position of algebra in school mathematics has changed dramatically over the past two decades. The traditional view of algebra as an isolated topic that is taught in secondary school has been challenged, as this approach creates numerous obstacles for students. In the United States (US) in particular, the well-known path of "first arithmetic then algebra" has been shown to be inadequate in preparing students for formal algebra (Hiebert, Carpenter, Fennema, Fuson, Wearne, Murray, & Human, 2005; National Center for Education Statistics [NCES], 1998; Stigler, Gonzales, Kawanaka, Knoll, & Serrano,1999), and has acted as a gatekeeper rather than a facilitator for college entrance (Schoenfeld, 1995). In response, reform documents in the US (e.g., National Council of Teachers of Mathematics [NCTM], 2000; National Governors Association Center for Best Practices [NGA] & Council of Chief State School Officers [CCSSO], 2010) have worked to instantiate the calls by several scholars (e.g., Kilpatrick, Swafford, & Findell, 2001; RAND Mathematics Study Panel, 2003) for algebraic reasoning to start as early as Kindergarten and span across the grades.

While broadening the scope of algebra instruction in this way is viewed as paramount to improving students' success in algebra, research is still needed to understand the impact such a significant shift in our approach might have on students' mathematical thinking. Promising research illustrating that elementary students are capable of engaging with algebraic ideas (e.g., Carpenter, Franke, & Levi, 2003; Carraher, Schliemann, Brizuela, & Earnest, 2006) has set the stage for exploring questions regarding what impact a sustained, comprehensive early algebra education would have on students' algebra understanding as compared to a traditional arithmetic-focused education.

In seeking to address the aforementioned questions regarding the effectiveness of early algebra, we initially designed a small-scale, quasi-experimental studies in which we examined the effectiveness of an early algebra intervention in Grades 3-5 (Blanton, Isler, et al., 2018; Blanton et al., 2015). We investigated the impact of our early algebra intervention by comparing the performance of students who participated in the intervention (as implemented by a member of the research team) with the performance of students who received their district's regularly-planned curriculum. We found that students who participated in the intervention outperformed their peers who received regular instruction with respect to their understanding of algebraic ideas. Given the positive results from the small-scale study, we implemented a large-scale, cluster randomized study of the impact of our early algebra intervention when taught by *regular* classroom teachers. We report on the results of that study here, focusing in particular on the fidelity of implementation (FOI) of Grades 3, 4, and 5 teachers (of students ages 8, 9, and 10 years old, respectively) as they implemented our early algebra intervention, as well as the relationship between FOI and student performance.

Rationale: Measuring Intervention Effectiveness

Historically, in the US algebra was reserved for successful secondary students. More recently, due to "Algebra for All" movements, the majority of US students take a minimum of one year of algebra by the time they graduate from secondary school, but not necessarily with success. During their elementary school years, students have typically only been exposed to mathematics that includes arithmetic, and algebra only appeared as a course upon their entry to secondary school. However, research has shown that this approach has been unsuccessful (Hiebert et al., 2005; NCES, 1998; Stigler et al., 1999), leading to the marginalization of students in school and society (Kaput 1999; Schoenfeld, 1995). Moses and Cobb (2001) argue that algebra functions as a gateway to academic and economic success, disenfranchising students in non-dominant groups even further. Calls for reform urged schools to consider a more longitudinal approach; to introduce algebraic ideas early in elementary school along with arithmetic (e.g., NCTM, 2000; NGA and CCSSO, 2010; RAND Mathematics Study Panel 2003).

Several studies have demonstrated that elementary students are capable of engaging with algebraic ideas (e.g., Carpenter et al., 2003; Carraher et al., 2006; and Kaput et al., 2008). However, systematic, instructional approaches for building children's algebraic reasoning are still lacking. Our work, which was informed by the pioneering work of Carraher et al. (2006) and Carpenter et al. (2003), is a step in that direction. In particular, using a learning progressions approach (described in Fonger et al., 2018), we designed an early algebra intervention for Grades 3-5 that focuses on developing students' abilities to generalize, represent, justify, and reason with mathematical structure and relationships across a variety of mathematical contexts. We then initiated a series of studies to examine the impact of this intervention in Grades 3-5 comparing the performance of students who participated in the intervention with the performance of students who received their district's regularly-planned curriculum. The first study in the series was a small-scale, cross-sectional *efficacy study* with about 120 students participating, in which the intervention was taught by a member of the study team (Blanton et al., 2015). This was followed by a small-scale *longitudinal* study in which approximately 170 intervention and comparison students were followed *across* Grades 3-5 (Blanton et al., 2018). As is the case with most efficacy studies, the main objective of these studies was to demonstrate that the program could, under the most favorable conditions, lead to the desired outcomes (O'Donnell, 2008; Raudenbush, 2007). In that study, we found that the intervention group significantly outperformed the comparison group on an assessment designed to measure understanding of core algebraic concepts and were more apt to use algebraic strategies in responding to the assessment items than were students in the comparison classrooms.

In our most recent study in the series, we have extended the prior studies by examining the *effectiveness* of our early algebra intervention at scale, with about 3200 students participating in a randomized study in which the intervention was implemented by classroom teachers. Effectiveness is defined by Dorland (1994) as "the ability of an intervention to produce the desired beneficial effect in actual use" (p. 531). As O'Donnell (2008) notes, effectiveness research complements efficacy research and refers to the degree to which the program achieves its outcome in actual field settings (e.g., Mihalic, 2002; Raudenbush, 2007; Summerfelt & Meltzer, 1998). Both efficacy (internal validity) and effectiveness (external validity - which can be generalized) are complementary, critical aspects of evaluating interventions (Summerfelt & Meltzer, 1998).

Because effectiveness studies are concerned with the ability of the program to produce the desired effect in actual use, it is important to study teachers' fidelity to that program in practice. Fidelity of implementation (FOI) is the commitment to using the intended instructional practices and procedures accurately and consistently when delivering a program or intervention (National Center on Response to Intervention, n.d., online). As such, FOI is closely related to effectiveness in that "no program—no matter how sound it is—can have impact if its essential elements are not used" (Yap, Aldersebaes, Railsback, Shaughnessy, & Speth, 2000). FOI studies can reveal important information about how likely an intervention can and will be implemented with fidelity in the classroom. Dusenbury and colleagues (2003) state, "If it is difficult to achieve fidelity of implementation in practice, a program has low feasibility. Programs that are implemented with high levels of fidelity but fail to produce desired effects may need to be redesigned" (p. 240)[1].

In this study, we examine the FOI of Grades 3, 4, and 5 teachers as they implemented our early algebra intervention from our first efficacy study, and the relationship between FOI and student performance. The data from this study are taken from the entire three years of our experimental study that tests the hypothesis that children who receive comprehensive, longitudinal early algebra instruction integrated in their curriculum and delivered by their teachers during the elementary grades are better prepared for algebra in middle school than children who have only "business-as-usual" (i.e., arithmetic-based) experiences during elementary grades.

Our goal was to measure the fidelity with which teachers in diverse demographic settings implemented the intervention, and how this intervention affected students' learning outcomes. This is a critical factor in evaluating an intervention's effectiveness (Mowbray, Holter, Teague, & Bybee, 2003; NRC, 2004; Summerfelt & Meltzer, 1998) and promoting external validity (O'Donnell, 2008). Thus, the questions guiding the study are the following:

1. To what extent are teachers implementing the intervention with fidelity?
2. To what extent does the fidelity with which teachers implement the intervention impact student learning?

Method

[1] FOI can be measured using different tools all of which involve tracking classroom instruction. Due to the size of this study. it was not feasible to be in all the classrooms day-to-day, hence we only observed a subset of the classrooms, and we further employed other tools as such as lesson logs. However, we acknowledged the limitations of self-report data so we did on-site visits to verify that the lessons were being implemented with fidelity as a means of triangulation with the lesson logs. Second, we selected the use of the MQI to complement our cognitive load measures because we wanted to use an established instrument (rather than just one that we'd developed) to ensure our data weren't too biased towards the curriculum. More detail is provided in the next section.

The data for this study are taken from a three-year longitudinal, cluster randomized trial (CRT) in which an early algebra intervention was implemented in Grades 3 –5. The longitudinal aspect concerned the students who were followed for 3 years. Pre- and post-intervention written assessments were administered to students in both the experimental and control groups. Implementation data in the form of monthly surveys and implementation logs were also collected from all participating experimental teachers. A subset of those teachers also participated in a more detailed round of data collection including videotaped classroom observations and interviews.

Participants

Participants included 3,208 students from 46 schools (23 experimental, 23 control) in three school districts in the Southeast US. In the first year of the study, the experimental schools included 98 participating Grade 3 teachers who taught the intervention, while the control schools included 108 Grade 3 teachers who taught only their regularly-planned curriculum[2]. The number of participating teachers and students changed each year for a variety of reasons (e.g., attrition, students moving between schools, changes in school structures such as departmentalization in some schools in 5th grade). Table 1 provides the numbers of teachers and students participating in the study each year. Note that these are the same students each year, with numbers of students decreasing each year due to attrition. In other words, students in Grade 5 are a subset of students in Grade 4, and students in Grade 4 are a subset of students in Grade 3.

Table 1. *Participating Teachers and Students Each Year*

Grade	Experimental			Control		
	Students	Schools	Teachers	Students	Schools	Teachers
3	1,637	23	98	1,448	23	108
4	1,341	23	82	1,245	23	75
5	1,087	23	49	1,079	23	52

[2] In the US, districts (and often individual schools within districts) can use a curriculum of their choice. The curriculum, however, must be aligned with state standards to prepare students for end-of-year state examinations. The three districts chose different curricula and provided teachers with recommended pacing and implementation guides. Our surveys and interviews indicated a great degree of similarity in addressing state standards across the three districts.

The three districts were diverse in terms of race and socioeconomic status, as well as location (rural/urban). Table 2 provides further detail on the demographics of the three districts.

Table 2. *Demographics of Participating Districts*

School district	Non-white students	Students qualifying for free/reduced lunch
A (mostly rural)	36%	20%
B (suburban)	40%	54%
C (urban)	82%	62%

Schools were comparable demographically within districts, but not between districts. Our cluster-randomized design, where randomization occurred at the school level, allowed us to account for these differences, as well as to avoid intervention spillover effects. In other words, the school-randomization ensured that the control classrooms remained in the control condition, in that they did not have any access to the intervention.

To study FOI, we examined the implementation practices of a sub-set of teachers and the performance on the assessments exhibited by their students. We observed about half of the teachers in the experimental condition at each grade level. The number of teachers and lessons to be observed was determined by considerations of statistical power and possible attrition. In other words, we calculated the number of teachers and sessions that we would need to perform quantitative analyses, while also considering that it was possible that we might lose access to certain teachers over time. In Grade 3, 50 teachers were observed, and we collected assessment data from a total of 670 students in those classrooms. In Grades 4 and 5, we observed 44 and 33 teachers, respectively, and collected assessment data from 826 and 777 of their students respectively (see Table 3).

Table 3. *Classroom Observations*

Grade	Number of teachers observed	Number of students whose teachers were observed
3	50	670
4	44	826
5	33	777

Intervention

Intervention Development. In our prior work, we designed a sequence of lessons across Grades 3-5. Each grade level sequence consisted of 18 one-hour lessons[3] that were designed to engage students in a range of critical algebraic thinking practices (generalizing, representing, justifying, and reasoning with mathematical structure and relationships) as they occur in the context of different big ideas regarding algebraic content (e.g., generalized arithmetic; equivalence, expressions, equations, and inequalities; and functional thinking). This approach to teaching and learning algebra, which has guided our prior work (see, e.g., Blanton et al., 2015; Fonger et al., 2018), is informed by Kaput's (2008) content analysis of algebra. Overall, these lessons used novel tasks as well as modified existing tasks from research that showed potential to facilitate students' construction of algebraic ideas (e.g., Brizuela & Earnest, 2008; Carpenter et al., 2003; Carraher, Schliemann, & Schwartz, 2008) and allowed for multiple points of entry so as to engage students at varying levels of understanding (see Appendix A for a sample lesson). This instructional sequence was refined through cycles of testing and revision during the earlier studies. The intervention lessons were implemented approximately weekly between the second and eighth month of the school year, following a common pacing guide.

Teacher Professional Development. As part of the work, we designed a series of professional development workshops for the teachers who were implementing the intervention in each of the three districts. Each year, participating teachers attended a one-day workshop before the school year started and subsequent half-day meetings once per month throughout the intervention. During professional development meetings, the intervention lessons were used to build teachers' content knowledge and pedagogical content knowledge around the teaching of early algebra. While focusing on the big ideas framing each lesson's content, teachers were consistently asked to engage in the algebraic thinking practices of generalizing, representing, justifying, and reasoning with relationships, and to think about how to engage their students in these practices across the different big ideas. Participating teachers were also provided with additional resources such as extended lesson plans, which offered teachers further information regarding

[3] These lessons were not added to students' mathematics instructional time. Rather, one hour of regular instructional time each week was replaced by an intervention lesson. Hence, the time spent on mathematics instruction remained the same across experimental and control groups.

expected student responses to tasks, common student misconceptions, and instructional questions that might help move students' thinking forward. Furthermore, teachers had at their disposal video-taped "model" lessons (taught by a team researcher in one of the previous small-scale studies) that they could access at any time, as well as on-line or phone access to project members for more help.

Data Collection

Student written assessment. Grade-level, one-hour written assessments were designed and validated in our prior work to align with the algebraic thinking practices and concepts (e.g., the relational view of the equal sign) targeted by our instructional intervention. Students in both the experimental and control conditions completed identical written assessments prior to the intervention in Grade 3 (as pre-test), as well as at the end of each year of the intervention (as post-test). The assessments were constructed using novel items as well as items that had performed well in previous studies, and, like those used in the instructional intervention, often allowed for multiple points of entry. Each assessment consisted of 12-14 items, almost all of which were open response. Many of the items contained multiple parts and were designed to be used longitudinally (i.e., across Grades 3–5). Nine of these items were used across the assessments in order to measure longitudinal growth. The remaining items were grade specific. See Appendix B for a copy of the assessment that was given as pre-test and post-test in Grade 3. Because most of the items were designed to be used longitudinally, they were rather challenging. Because of this, we were not expecting students to demonstrate mastery by the end of Grade 3. Rather, we were interested in the growth experimental students might show relative to control students (see Blanton et al., 2015, for a description of this instrument's development and validation process).

Instructional Data – Fidelity of Implementation (FOI). Each year, a randomly selected subset of teachers was observed and videotaped twice—once in the fall and once in the spring—while teaching intervention lessons. During the first year, 50 Grade 3 teachers were selected to be observed and videotaped, resulting in a total of 99[4] observations. During the second year, 29 Grade 4 teachers were observed twice and 16

[4] Due to a logistical complication, one teacher was only observed once at the beginning of the year

were observed once for a total of 74 observations. Finally, in the third year, 33 Grade 5 teachers were observed twice and 6 were observed once for a total of 72 observations. Across all grades we observed 245 lessons. One member of the research team completed all of the videotaping. One stationary camera was used at the back of the room, focusing on the teacher. An additional lapel microphone on the teacher ensured clear sound reception. Students were seated in the classroom such that only consenting students were captured on camera. Lesson videotaping took place throughout each school year with the exception of Grade 5, when all videotaping was completed during the first half of the year.

It is important to note that the longitudinal aspect of the study refers to students, not to teachers. As this was an effectiveness study, in which the intervention was implemented in natural school environments, teachers differed from year to year. Teachers were assigned to the classrooms at the schools' discretion[5]. Students, too, were moved to different classrooms with different teachers from year to year. However, our cluster-randomized design, where randomization occurred at the school level, allowed for experimental students to remain in the experimental condition for three years. Students who moved from an experimental school to a control school (or vice versa) during the course of the study were removed from subsequent analyses.

All participating experimental teachers were asked to complete an online lesson log corresponding to each intervention lesson taught. The log was designed to capture aspects of teachers' implementation with regard to the amount of time spent on each lesson, the kinds of lesson structures used, the use of practices that suggest a focus (or lack thereof) on students' algebraic thinking practices (e.g., whether they asked students to share their solution strategies with their peers or to justify their solutions), and which of the available resources they used, if any, to prepare to teach the lesson (e.g., whether they used the extended lesson plans, watched videos, searched for additional resources, or discussed the lessons with their colleagues). Additionally, all teachers were asked to complete a monthly survey regarding their regular instructional practices as a way to

[5] Only in two cases teachers followed the same group of students from 3rd to 5th grade as this was regular practice in their schools.

gauge potential similarities and differences in their regular instruction and the instruction that corresponded to the intervention.

Data Analysis

Written assessment data. Responses to the assessment items were coded for both correctness and strategy use, although for the purposes of this paper we focus only on the former[6]..

Instructional Data – Fidelity of Implementation (FOI). For FOI we collected two types of data – we conducted videotaped classroom observations and self-reported lesson logs.

Videotaped classroom observations were analyzed to examine the implementation of the two main objectives of the curriculum –the big ideas regarding algebraic content and the critical algebraic thinking practices. To examine whether the big ideas and the lesson structure were implemented we explored several factors related to an aspect of implementation: time spent on the lesson; lesson structures used such as whole class discussion, the use of group work or independent work; the use of worksheets; and presentation of tasks in oral or written form; whether tasks were read aloud; whether teachers demonstrated methods of presenting numerical information that might be unfamiliar to students (e.g., how to use a T-chart), and for those who did so, whether or not they used information from the lesson activity itself; and whether the teacher introduced new terms and definitions. By examining whether these factors were present, we examined whether the big ideas regarding algebraic content were introduced and explored as intended in the lesson plans.

We also coded whether or not students worked on their own (either independently or in groups) for most of the remaining class time as group work was suggested in the lessons teachers were asked to implement. Finally, during individual/group work, teachers were rated on whether they were active or passive. An active teacher would actively visit individuals/groups to help students with questions, and to challenge their thinking. A passive teacher would go around to groups only when asked, and would be

---

[6] Responses to assessment items were also coded for strategy use, though this is beyond the scope of this paper. For a description of this development and results, see Blanton et al., 2015, submitted)

largely reactive to students' needs rather than being proactive and challenging their mathematical thinking. Once again, this was related to the way lessons were designed.

To examine whether the critical algebraic thinking practices were implemented teachers were additionally rated on a five-point Likert scale on variables measuring the degree to which they engaged students in those practices that were the focus of the intervention. In particular, these ratings indicated the degree to which teachers asked students to engage in generalization, representation, and justification (briefly described below).

- *Generalize*: extent to which the teacher engages students in extrapolating a rule or property from a specific case to a more general case. Examples of this practice might include the teacher asking "will this work for all numbers?" or "Is it always true that if we add zero to a number we will always get the same number?";
- *Represent*: extent to which the teacher engages students in representing their thinking with variables and discusses with them the meaning of the representation[7], as for example explicitly asking students to use a variable to represent a quantity; and,
- *Justify*: extent to which a student's line of reasoning in the development of an argument to support a claim is probed and explored in its entirety. For example, the teacher requires the student to *explain* an answer with words or a drawing be it correct or incorrect, and follows up with additional questions.

The unit of analysis is the teacher. In other words, the analysis identified whether a teacher engaged with a practice during each lesson, and not the frequency with which he/she did so, and subsequently, the Likert scale ratings were averaged across the number of lessons observed for each practice. In general, a score of 1 on the five-point Likert scale indicated that teachers did not engage students in the thinking practices of generalizing, representing, or justifying. A score of 1 was rare. A score of 2 indicated that teachers followed the written lesson plan in terms of asking questions that prompted students to generalize, represent, and justify but did not pursue students' thinking. A score of 3 indicated at least minimal follow-up on students' attempts to generalize, represent, and justify (e.g., in the form of a follow-up question). A score of 4 indicated

[7] While we take representations to include a variety of forms, such as natural language, variable notation, tables, graphs, and pictures, we focus here on variable notation because of its seminal role in algebra (see, e.g., Kaput, 2008; Kline 1972)

that teachers pursued a line of questioning with individual students to really try to "get to the bottom" of their thinking. Finally, a score of 5 indicated that the lesson was characterized by these algebraic thinking practices and that these practices were pervasive, whether students' initial responses to questions were correct or incorrect.

Furthermore, the rating was done both at the whole class level (when teachers interacted with the whole class) and at a small group level (when teachers had discussions with smaller groups of students). The distinction is important: During whole class discussions, teachers were more likely to follow suggested questions and prompts (presented in the experimental curriculum), while in small groups teachers were more likely to follow students' work and the engagement in the practices of generalization, representation, and justification were more likely to be specific to the teacher and more spontaneous. In other words, group work discussions were less scripted by the experimental curriculum and were more often initiated by the teacher . Table 4 explores the time spent, on average, at each grade level in whole class and small group activities. Whole class time was recorded when the teacher was addressing all students simultaneously, while individual/group time referred to when students were working in small groups, independently, or when students were presenting their work to their teacher and classmates.

Table 4. *Average Time Spent on Whole Class[8] vs. Individual/Group Activities*

	Grade 3	Grade 4	Grade 5	Overall
Whole Class	30:44	29:16	30:35	30:15
Individual/Group	19:15	23:06	20:14	20:51

Finally, observations were coded and rated on a five-point Likert scale on six variables adapted[9] from the Mathematical Quality of Instruction (MQI) instrument (National Center for Teacher Effectiveness, 2014). The MQI was used as an externally designed and validated tool that provided a means of triangulation to our analysis:

[8] Class time varied. However, as the two categories (whole class vs. individual/group work) are mutually exclusive, they give us an average class time of 50 minutes across grades.

[9] We made a few adaptations of the MQI: we looked at fewer variables as they fit the focus of our study,and we also used a different segment length (the MQI only examines 20 minute sections, while we examined complete lessons).

- *Efficiency*: extent to which lesson time is used efficiently and class is on task;
- *Clarity*: extent to which the mathematics of the lesson is clear and not distorted;
- *Engagement*: extent to which the classroom environment is characterized by student engagement;
- *Attention to student difficulty*: extent to which teacher attends to student difficulty with the material;
- *Use of Student Ideas*: extent to which teacher uses student ideas and solutions to move the lesson forward; and,
- *Precision*: extent to which the mathematical language or notation used is precise (or not).

Once again, the unit of analysis is the teacher. With the exception of precision, all items were coded using a 5-point Likert scale. For these items, 1 is the most negative rating, indicating inefficient use of class time, severely distorted mathematics, total lack of student engagement with the lesson, student difficulty without any teacher remediation, and no substantive use of student ideas. Five is the most positive rating, indicating highly efficient use of class time, clarity in the math presented, high student engagement, high teacher attentiveness to student difficulty, and frequent inclusion of student ideas and solutions to help build the mathematics of the lesson. Precision was coded using a 4-point Likert scale[10]. Again, 1 is the most negative and indicates a level of imprecision that obscures the mathematics of the lesson, while 4 is the most positive and indicates high precision in notation or language.

In sum, what we attempted to capture in measuring FOI includes both the "nuts and bolts" details of lessons (e.g., the completion of required tasks, the amount of time students spent in various classroom configurations), as well as more nuanced characteristics that we believe are indications of effective teaching. Some of these characteristics are general to teaching mathematics (i.e., those measured by the adapted MQI variables), while others are more specific to the teaching of early algebra (i.e., those measured by the core algebraic practice variables).

Videos were coded independently by two members of the research team. Approximately 15% of videos were double coded in order to assess inter-rater reliability.

---

[10] As it was in original MQI.

The analysis suggested that raters had acceptable levels of agreement ($\kappa_w > .60$) for all MQI variables and for the three individual/group core algebraic practice variables (see Table 5). However, there was only moderate agreement for the whole class core algebraic practice variables at grades 3 and 4. Hence, while these variables are presented and explored here to a certain extent, subsequent factor analyses do not include the whole class core algebraic practice variables.

Table 5. *Weighted Cohen's Kappa for MQI and Core Algebraic Practice Variables*

	Grade 3	Grade 4	Grade 5	Overall
Generalize – Whole Class	0.49	0.59	0.87	0.65
Generalize – Ind/Group	0.88	0.80	1.00	0.90
Represent – Whole Class	0.53	0.54	0.82	0.61
Represent – Ind/Group	0.86	0.70	0.86	0.82
Justify – Whole Class	0.45	0.58	0.77	0.63
Justify – Ind/Group	0.68	0.93	0.95	0.86
Efficiency	0.65	0.73	0.88	0.73
Precision	0.67	0.77	1.00	0.78
Clarity	0.86	0.90	0.90	0.89
Engagement	0.75	0.70	0.78	0.74
Student Difficulty	0.70	0.63	0.73	0.69
Use of Student Ideas	0.71	0.60	0.90	0.70

Factor analysis was employed at each grade level to identify composite variables that could be used as teacher-level predictors of student outcomes (i.e., student performance on the early algebra assessments). After the latent factors (that is, underlying factors that are inferred from directly measured factors, in this case, the MQI variables and the core algebraic practice variables) were identified, multi-level regression analyses, with students nested within classrooms, was used to explore the potential impact of each composite variable on student performance.

Results

Student Performance

It is important to understand not only the fidelity with which teachers implemented the intervention, but also the relationship between FOI and student performance. As noted earlier, effective interventions are those that ultimately produce beneficial outcomes (Dorland, 1994). In an educational intervention such as ours, it is important to understand how improvements in students understanding of algebraic concepts and practices (i.e., beneficial outcomes) were influenced by the fidelity with which teachers implemented the intervention. In what follows, we describe results of analyses of student performance and FOI, and examine connections between the two.

Figure 1 illustrates students' overall performance, in terms of percent correct, across grades 3 through 5[11]. Using a 3-level longitudinal piecewise hierarchical linear model, we found the steepest increase in performance between the Grade 3 pre- and post-test, with students in the experimental condition gaining significantly faster than their peers in the control condition (Authors, submitted). Though there was a leveling off of the treatment effect in Grades 4 and 5, students who received the intervention maintained their advantage across the length of the study (see Table 6). It is also worth noting that there was no difference on students' performance on the pre-test across experimental and control groups.

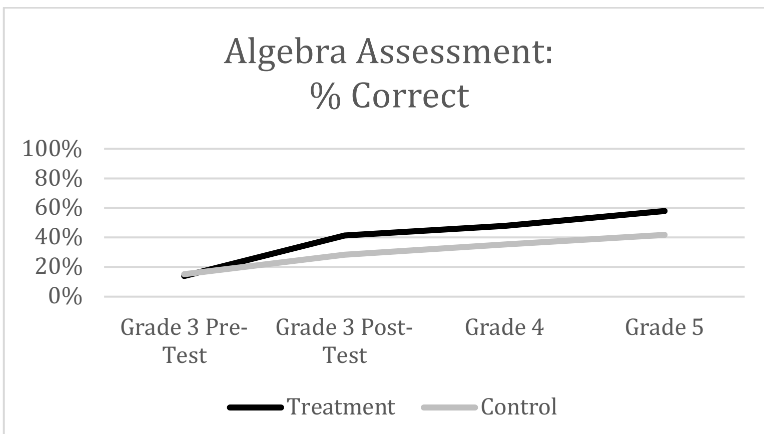


Figure 1. Overall percentage correct on assessment by testing time (Blanton, Stephens, et al., 2018).

Table 6. *Percent Correct on the Algebra Assessment by Year and Treatment Condition.*

	Treatment	Control
Grade 3 Pre-Test	13.96% (10.88%)	15.18% (11.41%)
Grade 3 Post-Test	41.34% (21.24%)	28.56% (16.18%)
Grade 4	48.00% (21.63%)	35.10% (17.39%)
Grade 5	57.88% (22.36%)	41.75% (19.43%)

**Standard deviation in parentheses.*

[11] It is beyond the scope of this paper to provide a full report on the student performance growth. For a complete report of the longitudinal analysis of student performance, see Authors (in preparation). Similarly, for a report on students' strategy use see Authors (submitted).

Videotaped Classroom Observations

In analyzing videotaped classroom observations, we were particularly interested in how teachers engaged with the intervention curriculum and the pedagogical strategies they chose to implement the lessons. Table 7 provides descriptive statistics by grade level as well as across grades regarding specific pedagogical strategies teachers used.

Table 7. *Summary of Instructional Activities.*

	Grade 3	Grade 4	Grade 5	Overall
Read aloud	90%	85%	79%	85%
Define	52%	41%	53%	49%
Demo	42%	28%	25%	33%
Demo info from lesson activity	98%	100%	94%	98%
Post-intro work	71%	78%	71%	73%
Active teacher	69%	77%	88%	77%

Overall, 85% of observed classroom teachers read the problem aloud or had a student read the problem aloud. However, the reading aloud of problems steadily decreased as students moved up the grades from 90% in Grade 3 to 79% in Grade 5, likely due to students' increasing reading ability and ability to engage with tasks autonomously. In about half of the classrooms, teachers tried to ensure that students understood any terms or concepts that might be unfamiliar to them. In about one third of the classrooms, teachers demonstrated methods of presenting numerical information that might be unfamiliar to students, and of those who did, 98% used information from the lesson itself.

After the lesson was introduced, in 73% of the observed classrooms students worked either independently or in groups on the remainder of the lesson activity– a number that stayed relatively consistent over the three grade levels. Finally, in 77% of observed classrooms, the teacher was coded as "active," that is, he or she actively visited individuals or groups to help students with questions and to challenge their thinking.

As described above, teachers were rated on a 5-point Likert scale on three variables intended to capture engagement with algebraic thinking practices in the classroom: *generalize* a mathematical relationship, *represent* with variables, and *justify* a generalization. Separate codes were given for whole class work and individual/group work, though we only report results for individual/group here due to reliability issues discussed previously. Table 8 shows the average ratings teachers received on these variables at each grade separately, and overall.

Table 8: *Mean Algebraic Thinking Practices Observation Ratings.*

		Grade 3	Grade 4	Grade 5	Overall
Generalize					
	Whole Class	3.51 (1.08)	3.54 (1.04)	3.51 (1.05)	3.52 (1.05)
	Individual/Group	2.55 (1.33)	2.32 (1.23)	2.76 (1.22)	2.54 (1.28)
Represent					
	Whole Class	3.00 (1.25)	3.12 (1.15)	3,29 (1.16)	3.40 (1.05)
	Individual/Group	2.10 (1.34)	1.93 (1.10)	2.62 (1.21)	2.35 (1.29)
Justify					
	Whole Class	3.75 (1.02)	3.62 (0.89)	3.69 (0.96)	3.69 (0.96)
	Individual/Group	2.91 (1.39)	2.66 (1.20)	3.11 (1.21)	2.89 (1.29)

Note. Standard deviation in parentheses.

The results suggest that teachers relatively faithfully met the expectations expressed in the lesson plans in terms of asking students to engage in generalization, representation, and justification though teachers did not go far beyond these written expectations.

Finally, Table 9 shows the average ratings teachers received for the six variables—*Efficiency, Clarity, Engagement, Student Difficulty, Use of Student Ideas,* and *Precision*—adapted from the MQI instrument (ratings range from 1 to 5). On average, teachers at each grade level scored fairly high across all six variables.

Table 9: *Mean Adapted MQI Classroom Observation Ratings.*

	Grade 3	Grade 4	Grade 5	Overall
Efficiency	3.33 (1.11)	3.43 (1.05)	3.86 (0.72)	3.52 (1.01)
Clarity/Distortion	4.24 (.96)	4.51 (0.81)	4.46 (0.82)	4.39 (0.88)
Engagement	3.80 (1.00)	3.61 (1.06)	3.90 (0.79)	3.77 (0.96)
Student Difficulty	3.56 (.99)	3.59 (1.11)	3.79 (0.89)	3.64 (1.00)
Uses Student Ideas	3.85 (.93)	4.05 (1.12)	4.31 (0.80)	4.04 (0.97)
Precision*	3.31 (.83)	3.51 (0.69)	3.61 (0.59)	3.46 (0.73)

Note. Standard deviation in parentheses. *Precision was coded on a 4-point scale.

Relationship Between Teaching Practices and Student Performance

In order to identify underlying latent variables, factor analysis using principal components analysis was utilized at each grade level. For teachers who were observed twice, codes were first averaged across observations to create one set of variables per teacher. The six MQI variables and the three individual/group algebraic thinking practice variables were entered. Solutions for two, three, and four factors were each examined at each grade level using varimax rotations of the factor loading matrix. Solutions were chosen based on: (a) the 'leveling off' of eigenvalues on the scree plot; and (b) the insufficient number of primary loadings and difficulty of interpreting the subsequent factors. The analysis resulted in a two-factor solution at grades 3 and 5, and a three-factor solution at grade 4 (see Table 8). In other words, in grades 3 and 5, the variations amongst the nine observed variables reflected two underlying factors: the six MQI variables loaded onto one factor, while the three algebraic thinking practice variables loaded onto another. In Grade 4, however, the six MQI variables loaded onto two distinct factors, while the three algebraic practices variables again loaded onto one, for a total of three factors at this grade level (see Table 10).

Table 10: *Factor Loadings Based on a Principal Components Analysis with Varimax Rotation.*

	Grade 3		Grade 4			Grade 5	
	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2
Generalize		0.86		0.88			0.92
Represent		0.82		0.81			0.91
Justify		0.81		0.82			0.90
Efficiency	0.67		0.77			0.65	
Clarity	0.85				0.87	0.81	
Engagement	0.64		0.86			0.74	
Student Difficulty	0.75		0.80			0.78	
Student Ideas	0.66		0.77			0.88	
Precision	0.84				0.90	0.82	

Note. Factor loadings < .4 are suppressed.

At grades 3 and 5, the two factors that emerged are referred to as "core algebraic practice" and "MQI." In grade 4, three factors emerged: 1) the three individual/group algebraic thinking practice codes (generalize, represent, justify) were averaged to create a composite variable ("core algebraic practice"), 2) "precision" and "clarity" were averaged to create a composite variable ("teacher precision"), and the remaining four MQI variables ("efficiency," "engagement," "student difficulty," and "use of student ideas") were averaged to create a composite variable ("student-focused") (see Table 11).

Table 11. *Descriptive Statistics for Composite Variables.*

	Mean (*SD*)	Minimum	Maximum
Grade 3			
Core algebraic practice	2.64 (1.01)	1.00	4.67
MQI	3.67 (0.68)	1.92	4.83
Grade 4			
Core algebraic practice	2.39 (1.00)	1.00	5.00
teacher precision	3.97 (0.67)	2.00	4.50
student-focused	3.62 (0.88)	1.00	5.00
Grade 5			
Core algebraic practice	2.90 (1.01)	1.00	5.00
MQI	3.98 (0.50)	2.75	4.67

To determine whether there was a relationship between teachers' practice, as measured by the composite variables, and student performance on the algebra assessment, we used Hierarchical Linear Modeling (HLM) with separate two-level, random intercept models at each grade level. HLM methods were utilized because of the hierarchical nature of the data, with students nested within classrooms (Raudenbush & Bryk, 2002). Given that students typically had different teachers at each grade level and different students were present in observed classrooms each year, it was not feasible to take a longitudinal approach to data analysis. Therefore, separate cross-sectional models were explored at each grade level.

Due to the hierarchical structure, the variance in the outcome measure of student scores on the algebra assessment at each grade level can be partitioned into two parts: variability due to individual differences and variability due to teacher characteristics. The intraclass correlation coefficient (ICC), which was estimated in an unconditional model (one with no predictor variables), indicated that, at each grade level, a significant proportion of the variance occurred at the teacher level: 39.5%, 29.7%, and 35.8% at grades 3, 4, and 5, respectively. These results suggest that student performance on the outcome measure is affected not only by individual differences, but also by characteristics of the classroom practice, and thus support the use of a two-level model for analysis at each grade level.

We present here the results from HLM analyses that describe the effects of teacher practice on each grade's post-test assessment performance. The models we used in our analyses are shown in Appendix C. First, in Grade 3 we explored the relationship between teachers' practice and student performance. Table 12 presents results from both the unconditional model (with no predictor variables) and the full model (with student- and teacher-level predictors included). In the full model, grade 3 pre-test performance was used as a level-1 (student-level) covariate, and core algebraic practice and MQI were included as level-2 (teacher-level) predictors.

Table 12. *Effects of Teacher Practice on Grade 3 Algebra Assessment Post-Test Performance.*

	Unstandardized β coefficients (with standard errors)	
	Model 1: Unconditional	Model 2: Full
Intercept	41.161 (2.107)***	41.399 (1.434)***
Student variable		
Grade 3 pre-test		0.862 (0.060)***
Teacher variables		
Core Algebraic Practice		4.552 (1.854)*
MQI		1.601 (2.780)
	Variance in algebra performance by source	
Variance estimates		
Student-level variance	299.623	234.708
Teacher-level variance	195.839	82.791

*p<.05; **p<.01; ***p<.001

As indicated in the full model, there was a statistically significant positive relationship (α = .02) between teachers' average score on the core algebraic practice composite variable and student performance on the algebra assessment, though the relationship between MQI and student performance was not significant. For each one-unit increase in a teacher's average core algebraic practice score, the estimated student score on the algebra assessment increased by 4.6%. Given the possible range of the core algebraic practice variable from 1 to 5, this equates to an 18.4% difference in average scores between students with teachers who received the lowest possible scores (core algebraic practice = 1) and those with teachers who received the highest possible score (core algebraic practice = 5). In other words, the average algebra score for a student in a classroom where the teacher was rated average on core algebraic practice was 41.4%. However, that score increased to 50.6% for students in a classroom where the teacher scored 2 points above average on core algebraic practices, and dropped down to 32.2% for students whose teacher scored two points below average on that measure.

Similarly, we continued this analysis with Grade 4. Table 13 presents the results of the unconditional and full models. Here, grade 3 post-test was used as the level-1 covariate, as it was the most recent prior measure of student performance, and the three teacher-level variables identified in the factor analysis were included at level 2.

Table 13. *Effects of Teacher Practice on Grade 4 Algebra Assessment Post-Test Performance.*

	Unstandardized β coefficients (with standard errors)	
	Model 1: Unconditional	Model 2: Full
Intercept	47.276 (1.951)***	48.534 (0.819)***
Student variable		
Grade 3 post-test		0.713 (0.024)***
Teacher variables		
Core Algebraic Practice		-0.650 (1.047)
Precision		0.800 (1.490)
Student-Focused		4.221 (1.205)***
	Variance in algebra performance by source	
Variance estimates		
Student-level variance	334.683	164.778
Teacher-level variance	141.313	17.839

**p*<.05; ***p*<.01; ****p*<.001

Contrary to what was found in Grade 3, there was no statistically significant relationship between teachers' average score on the core algebraic practice composite variable and student performance on the assessment, and also no relationship between precision and student performance. However, there was a significant relationship between the student-focused variable and student performance ($\alpha < .001$). For each one-unit increase in a teacher's student-focused score, the estimated student score on the assessment increased by 4.2%. Given the possible range of the student-focused variable from 1 to 5, this equates to a 16.8% difference in average scores between students with teachers who received the lowest possible scores (student-focused = 1) and those with teachers who received the highest possible score (student-focused = 5). In other words, the average algebra score for a student in a classroom where the teacher was rated average on the student-focused variable was 48.5%. However, that score increased to 56.9% for students in a classroom where the teacher scored 2 points above average on the student-focused variable, and dropped down to 40.1% for students whose teacher scored two points below average on that measure.

Finally, in Grade 5, the general structure of the model was identical to that of Grade 3, with the exception of the use of the grade 4 post-test as the level 1 covariate. Table 14 presents the results of the unconditional and full models.

Table 14. *Effects of Teacher Practice on Grade 5 Algebra Assessment Post-Test Performance.*

	Unstandardized β coefficients (with standard errors)	
	Model 1: Unconditional	Model 2: Full
Intercept	56.568 (2.678)***	58.097 (1.039)***
Student variable		
Grade 4 post-test		0.755 (0.024)***
Teacher variables		
Core Algebraic Practice		3.109 (1.315)*
MQI		1.473 (2.604)
	Variance in algebra performance by source	
Variance estimates		

Student-level variance	367.554	234.708
Teacher-level variance	205.210	82.791

**p*<.05; ***p*<.01; ****p*<.001

The Grade 5 results parallel the results found in Grade 3. There was a statistically significant positive relationship (α = .03) between teachers' average score on the core algebraic practice composite variable and student performance on the algebra assessment, though the relationship between MQI and student performance was not significant. For each one unit increase in a teacher's average core algebraic practice score, the estimated student score on the algebra assessment increased by 3.1%. Given the possible range of the core algebraic practice variable from 1 to 5, this equates to a 12.4% difference in average scores between students with teachers who received the lowest possible scores (core algebraic practice = 1) and those with teachers who received the highest possible score (core algebraic practice = 5). In other words, the average algebra score for a student in a classroom where the teacher was rated average on core algebraic practices was 58.1%. However, that score increased to 64.3% for students in a classroom where the teacher scored 2 points above average on core algebraic practices, and dropped down to 51.9% for students whose teacher scored two points below average on that measure.

Discussion

This study aimed to explore the relationship between the fidelity with which teachers implemented the early algebra intervention (as part of their regular classroom practice) and the development of students' understanding of algebraic concepts and practices. However, to do that, we needed to step out of the realm of the efficacy studies and move towards an effectiveness study that is more closely aligned to "regular" practice – taught by practicing teachers to a large number of students. While our smaller efficacy studies provided valuable information on how students engage with algebraic ideas and a glimpse of the possible affordances and gains early algebra instruction might offer, we needed to investigate the question of early algebra's feasibility and value to a larger number of students in regular classrooms. Here, we examined the *effectiveness* of an early algebra intervention at scale, aiming to understand how early algebra instruction impacts learning.

As we reported in depth elsewhere (Blanton, Stephens, et al., 2018) and briefly here, students in the experimental group who experienced the early algebra intervention over the three years of the study significantly outperformed their counterparts who received a more traditional arithmetic-based curriculum. These findings were encouraging and prompted us to more closely examine instruction during the intervention in experimental classes for distinctions that might help further explain our findings about experimental students' performance. This article reported on our findings.

In addressing the question of whether teachers implemented the intervention with fidelity, we found that teachers generally followed the intervention lessons as intended. On average, they dedicated the time and resources needed to follow the intervention lessons, and followed all the steps of the lessons as they were intended by the curriculum designers. Furthermore, teachers, for the most part, faithfully met the expectations expressed in the lesson plans in terms of asking students to engage in generalization, representation, and justification. We note, however, that participating teachers did not go far beyond these written expectations. In other words, most teachers did not ask other questions of an algebraic nature, nor did they explore algebraic tasks beyond what was expected of them. This is not surprising as research suggests that it is only through sustained professional development and implementation of a given curriculum over several years that teachers engage with more nuanced aspects of mathematical practices (see e.g., Jacobs, Lamb, & Philipp, 2010; Superfine, 2008).

Given such findings it was important for us to investigate what teacher-related instructional characteristics and approaches impacted and *predicted* student learning outcomes. A factor analysis revealed some important findings that were not necessarily consistent across the grades. Specifically, Grade 3 and 5 findings were aligned, while Grade 4 appeared to take a different direction.

In Grades 3 and 5, we found that the extent to which teachers engaged students in "core algebraic practices" of generalization, representation, and justification of mathematical structure and relationships significantly predicted students' post-test scores on the algebra test. In Grade 3 in particular, for each one-unit increase in a teacher's average core algebraic practice score, the estimated student score on the assessment increased by 4.6%. Similarly, in Grade 5 for each one-unit increase in a teacher's average core algebraic practice score, the estimated student score on the assessment increased by 3.1%. In other words, students whose teachers implemented the intervention with greater emphasis in engaging students in these core algebraic practices significantly outperformed students whose teachers did not engage them as deeply in algebraic practices. Thus, it seems clear that not only are the step-by-step following of the intended curriculum and the attention to the core algebraic ideas needed to increase student performance, but also teachers' engagement with core algebraic practices played a key role in strengthening and solidifying students' algebraic learning. The results suggest that teachers did not fall into two categories (those who engage students with core algebraic practices and those who fail to do so). Rather, the results clearly show a growth pattern – the more teachers engage students with core algebraic practices, the more the student performance grows. These findings support that a sustained and deeper engagement with core algebraic practices increases student learning and performance in early algebra.

Our findings in Grade 4 differed from those in Grades 3 and 5. In this case, the most important characteristic in predicting student learning outcomes was the teacher focus on student ideas. We found a significant relationship between the student-focused variable and student performance. For each one unit increase in a teacher's student-focused score, the estimated student score on the algebra assessment increased by about 4%. But, what exactly does it mean for a teacher to be "student-focused"? Teachers received a high score on the student-focused score when they pursued in more depth their students' ideas in algebraic meaning making. This did not mean that the teacher was only correcting or remediating student errors, but rather, the teacher attended to student contributions that emerged from the algebra lessons and used them to move instruction forward. Student contributions might include "student answers to questions (including one-word answers), comments, mathematical ideas, explanations, representations, generalizations, questions to the teacher, and student work" (quote from the MQI manual). It is then worth considering that it is not only engagement with algebraic ideas that predicted growth, but good teaching in general that attended to student needs and engaged with students' developmental and mathematical needs. Although we are not sure why the engagement with the actual core algebraic practices was not significant in Grade 4 as it was in Grades 3 and 5, it is still noteworthy that teachers' pursuit of students' ideas was a significant predictor of student learning outcomes.

Overall, emphasis on algebraic thinking practices (whether they originate from teachers or students) was the core predictor of improved student performance. This finding is aligned with our early algebra framework and its emphasis on algebraic practices (see Fonger et al., 2018 for an extensive discussion of this progression). However, our findings underscore the importance of further refinement of our observation instruments and our coding practices. As the study was not a laboratory-controlled study, but rather steeped in the reality of hundreds of teachers and thousands of students as any effectiveness study should be, lessons can be messy, and a more refined instrument might allow us to see beyond the haze of the realities of the classroom and better attend to the instructional practices. We acknowledge our limitations with respect to our observation instrument and coding.

Before closing, we briefly discuss how the results of our work can potentially impact teacher preparation programs as well as in-service education and professional development programs. Our professional development sessions focused explicitly on engaging participating teachers with algebraic practices and on modeling classroom instruction that promoted these practices. Elementary education teachers often receive limited instruction and training on early algebra concepts. Our work shows the importance of appropriate and in-depth use of core algebraic practices in the early mathematics classrooms as they impact student learning in significant ways. The stronger the presence of these core algebraic practices in the classroom, the more students benefit. As teacher educators, we are still learning how to incorporate these practices into our elementary preservice teacher education courses.

The early algebra field in mathematics education is currently in the process of navigating various standards and curricula across countries, and searching for ways to weave the big ideas and practices of early algebra into these curricula. Our study comes at this critical time to provide supporting evidence that, indeed, early algebra programs can be implemented with fidelity in actual field settings (not only in limited research environments) and can impact student learning in positive and rich ways. We will continue to explore how to strengthen these curricula, how to better support teachers, and how to measure what practicing teachers bring to our expanding field.

References

Blanton, M., Stephens, A., Knuth, E., Gardiner, A., Isler, I., & Kim, J. (2015). The Development of Children's Algebraic Thinking: The Impact of a Comprehensive Early Algebra Intervention in Third Grade. *Journal for Research in Mathematics Education, 46*(1), 39-87.

Blanton, M., Isler, I., Stroud, R., Stephens, A., Knuth, E., & Gardiner, A. (2018). Growth in children's understanding of generalizing and representing mathematical structure and relationships. Manuscript under review.

Blanton, M., Stephens, A., Stroud, R., Knuth, E., Stylianou, D., Gardiner, A. (2018). Does Early Algebra Matter? The Effectiveness of an Early Algebra Intervention in Grades 3–5. Manuscript under review.

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in the elementary school*. Portsmouth, NH: Heinemann.

Carraher, D. W., Schliemann, A. D., Brizuela, B. M., & Earnest, D. (2006). Arithmetic and algebra in early mathematics education. *Journal for Research in Mathematics Education, 37*(2), 87-115.

Common Core State Standards Initiative. (2010). National Governors Association Center for Best Practices and Council of Chief State School Officers.

Dorland, W. A. (1994). Borland's illustrated medical dictionary (28th ed.). Philadelphia: W. B. Saunder

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. Health Education Research Theory and Practice, 1

Fonger, N. L., Stephens, A., Blanton, M., Isler, I., Knuth, E., Gardiner, A. (2018). Developing a learning progression for curriculum, instruction, and student learning: An example from mathematics education. *Cognition and Instruction, 36*(1), 30-55.

Hiebert, J., Carpenter, T. P., Fennema, E., Fuson, K. C., Wearne, D., Murray, H., & Human, P. (1997). Making sense: Teaching and learning mathematics with understanding (Vol. 361, pp. 03801–3912). Portsmouth, NH: Heinemann.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, *26*(4), 430-511.

Jacobs, V. R., Lamb, L. L. C., and Philipp, R. A. (2010). Professional noticing of children's mathematical thinking. *Journal for Research in Mathematics Education, 41*(2), 169-202.

Kaput, J. (1999). Teaching and learning a new algebra. In E. Fennema & T. A. Romberg (Eds.), *Mathematical classrooms that promote understanding* (pp. 133–155). Mahwah, NJ: Lawrence Erlbaum Associates.

Kaput, J. (2008). What is algebra? What is algebraic reasoning? In J. Kaput, D. Carraher, & M. Blanton (Eds.), *Algebra in the early grades* (pp. 5–17). New York, NY: Lawrence Erlbaum Associates.

Kline, M. (1972). *Mathematical thought from ancient to modern times*. New York: Oxford University Press.

Moses, R. P., & Cobb, C. E. (2001). *Radical equations: Math literacy and civil rights.* Boston, MA: Beacon Press.

Mihalic, S. (2002, April). The importance of implementation fidelity. Boulder, CO: Center for the Study and Prevention of Violence

Mowbray, C., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, *24*(3), 315–340

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

National Governors Association Center for Best Practices & Council of Chief State School Officers (2010). *Common core state standards for mathematics*. Washington, DC: Author. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

National Research Council. (2004). *On curricular effectiveness: Judging the quality of K-12 mathematics evaluations.* Washington, DC: National Academy Press.

North Dakota Department of Public Instruction. (n.d.). *Fidelity of implementation.* Retrieved from www.dpi.state.nd.us/speced/personnel/fidelity.shtm

O'Donnell, C. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Review of Educational Research 78*(1), 33–84.

RAND Mathematics Study Panel Report. (2003). Mathematical proficiency for all students: A strategic research and development program in mathematics education. Washington, DC: U.S. Department of Education.

Raudenbush, S. W. (2007). Designing field trials of educational innovations. In B. Schneider & S.-K. McDonald (Eds.), Scale-up in education: Vol. II. Ideas in Practice (pp. 23-40). Lanham, MD: Rowman & Littlefiel

Stigler, J.W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States.* U.S. Department of Education National Center for Educational Statistics: NCES 99–074. Washington, DC: Government Printing Office.

Schoenfeld, A. (1995). Report of working group 1. In C. Lacampagne, W. Blair, & J. Kaput (Eds), *The Algebra Initiative Colloquium* (Vol. 2: Working Group papers, pp. 11-12). Washington, DC: US Department of Education, Office of Educational Research and Improvement.

Summerfelt, W. T., & Meltzer, H. Y. (1998). Efficacy vs. effectiveness in psychiatric research. *Psychiatric Services, 49*(6)*,* 834–835.

Superfine, A. S. (2008). Planning for mathematics instruction: A model of experienced teachers' planning processes in the context of a reform mathematics curriculum. *The Mathematics Educator, 18*(2), 11-22.

Yap, K., Aldersebaes, I., Railsback, J., Shaughnessy, J., & Speth, T. (2000). *Evaluating whole-school reform efforts: A guide for district and school staff* (2nd ed.). Portland, OR: Northwest Regional Educational Laboratory.

Appendix A:
Sample Lesson

Lesson 5:
Generalizations about Sums of Evens and Odds

CCSS.MP3
CCSS.MP5
CCSS.MP6
CCSS.MP7
CCSS.MP8
CCSS.MC.1.OA.7
CCSS.MC.2.OA.3
CCSS.MC.2.NBT.5
CCSS.MC.6.EE.4

Big Idea

Begin to develop arithmetic generalizations about sums of even numbers and odd numbers. Continue to develop an understanding of representation-based arguments for justifying a conjecture.

Talk About It!

Jumpstart

1. Find the missing value: 14 + ___ = 15 + 6
2. True or False: 34 + 10 = 44 + 9
3. What numbers make the equation $r + 0 = r$ true?
4. Find 19 + 52 17. Think about how you can use the properties you have learned.

Explore and Discuss
Generalizations about Sums of Evens and Odds

ACTIVITY

Teacher Notes

Allow students plenty of time to explore various numbers and come to think of numbers as combinations of pairs and leftovers.

For groups that need support, help them by using discussion building strategies (found in Teaching Tips section). Sit with groups and observe what they are saying. Use this to select students who may help you lead whole group discussions. Select students who exhibit different ways of thinking.

Talk About It!

Review and Discuss

Supporting Your Practice

Based on your observations today, develop

Lesson Materials:
Activity
Unifix cubes

Teaching Tips

- ✓ Spend sufficient time on the table portion of lesson activity. Exploration and group discussion are key to students' understanding of evens and odds. Students will refer back to what they learn in this activity when they begin building arguments about evens and odds.
- ✓ Discussion is essential in this lesson. Continue to encourage students to discuss their mathematical thinking. Encourage them to think about the ways their classmates are thinking, and discuss with each other the possible differences in their thinking.
- ✓ Questions that help build discussion:
    - Do you think your conjecture is true? How do you know?
    - For what numbers do you think your conjecture is true? Why?
    - Do you think your conjecture is true for *all* numbers? How do you know?
    - How would you argue to me (or to a classmate) that your conjecture is true?
    - Do your examples show that your conjecture is *always* true?
    - Can you think of another argument for your conjecture that does not look at specific numbers?
- ✓ Encourage students to explore representation-based reasoning. Ask students to show you what they know. Encourage the use of cubes to represent generic examples in their arguments.

a definition for "even numbers" and for "odd numbers."

Common Misconceptions

When asked to explore the sum of any two even numbers, students almost always turn immediately to examples. Some try one example (e.g., 2 + 4 = 6) and conclude that the sum of any two even numbers is always even. Some try a few examples. We want to begin moving them away from this strategy and help them build other types of arguments. Reminding them of the definitions of even numbers and odd numbers and the work they did in the first problem will get them thinking in a different way.

Targeted Student Thinking

"I noticed that any time there is an even number it breaks into pairs with no leftovers, but when we have an odd number like three I can only make one pair and then I have a leftover."

"I know that Jesse's answer is going to be even because two even numbers added together will always give us an even answer. I know this because if I take the first even number and break it into pairs and add those pairs with the pairs from the other even number, then I will still have all pairs with no leftovers. Like this:

4 + 2 = 6

VOCABULARY

Even Number

Odd Number

Notes:

Activity

Lesson 5: Generalizations about Sums of Evens and Odds

Name: ___________________________________ Date: _____________

Explore and Discuss with a partner:

Sums of Evens and Odds

How Many Pairs?

In this activity, you are going to “build” different numbers using cubes and think about how many pairs make up various numbers.

How many pairs of cubes are in the number 6? How many cubes are left over after you’ve made your pairs?

Use your cubes to complete the following table for the given numbers.

Number	Number of pairs created	Number of cubes left over
3		
4		
5		
6		
7		

What do you notice? What (kinds of) numbers have no cubes left over after all pairs are made? What (kinds of) numbers have a cube left over?
Write a sentence to describe each of your observations.

Exploring the sum of two even numbers

A. Jesse is adding two even numbers. Do you think his answer will be an even number or an odd number?

B. Develop a conjecture to describe what you found.

C. Is your conjecture true for any two even numbers you add together? How do you know? Use numbers, pictures, cubes, or words to explain your thinking.

Appendix B: Algebra Assessment

1. Fill in the blank with the value that makes the number sentence true.
   $7 + 3 = \_\_\_\_ + 4$
   Explain how you got your answer.

2. Circle True or False. Explain how you got your answer
   a) $12 + 3 = 10 + 5$
   b) $57 + 22 = 58 + 21$
   c) $39 + 121 = 121 + 39$

3. Marcy's teacher asks her to solve "23 + 15." She adds the two numbers and gets 38. The teacher then asks her to solve "15 + 23." Marcy already knows the answer is 38 because the numbers are just "turned around."
   a) Do you think Marcy's idea will work for any two numbers? Why or why not?
   b) Write an equation using variables (letters) to represent the idea that you can add two numbers in any order and get the same result.

4. Brian knows that if you add any three odd numbers, you will get an odd number. Explain why this is true.

5. Tim and Angela each have a piggy bank. They know that their piggy banks each contain the same number of pennies, but they don't know how many. Angela also has 8 pennies in her hand.
   a) How would you represent the number of pennies Tim has?
   b) How would you represent the total number of pennies Angela has?
   c) Angela and Tim combine all of their pennies. How would you represent the number of pennies they have altogether?
   Suppose Angela and Tim now count their pennies and find they have 16 all together. Write an equation that represents the relationship between this total and the expression you wrote above.

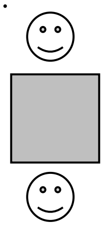
6. Circle the equation that best represents the following story:
   Owen had 10 stickers. Eric gave him some more. Now Owen has 27 stickers.
   a) $10 + 27 = c$ b) $10 + c = 27$ c) $10 – c = 27$ d) $27 – c = 10$

7. Nicole baked some cookies. She ate 3 cookies and now has only 9 cookies.
   a) Write an equation with a variable (letter) to represent this story.
   b) What does your variable (letter) in the equation in part a represent?

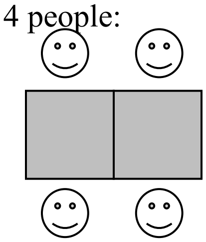
8. Find the value of n in the following equation. Show or explain how you got your answer.
   $5 \times n + 2 = 42$

9. Brady is celebrating his birthday at school. He wants to make sure he has a seat for everyone. He has square desks.

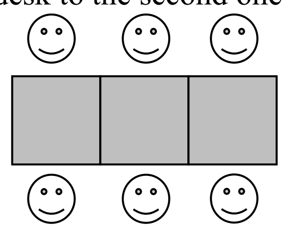
He can seat 2 people at one desk in the following way:



If he joins another desk to the first one, he can seat 4 people:



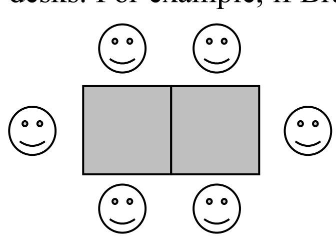
If he joins another desk to the second one, he can seat 6 people:



a) Fill in the table below to show how many people Brady can seat at different numbers of desks.

Number of desks	Number of people
1	2
2	4
3	
4	
5	
6	
7	

b) Do you see any patterns in the table from part a? If so, describe them.
c) Think about the relationship between the number of desks and the number of people.
Use words to write the rule that describes this relationship.
Use variables (letters) to write the rule that describes this relationship.
d) If Brady has 100 desks, how many people can he seat? Show how you got your answer.
e) Brady figured out he could seat more people if two people sat on the ends of the row of desks. For example, if Brady had 2 desks, he could seat 6 people.



How does this new information affect the rule you wrote in part c?
Use words to write your new rule:
Use variables (letters) to write your new rule:

10. The table below shows the relationship between two variables, k and p.
The rule $p = 2 \times k + 1$ describes their relationship.
    a) Some numbers in the table are missing. Use this rule to fill in the missing numbers.

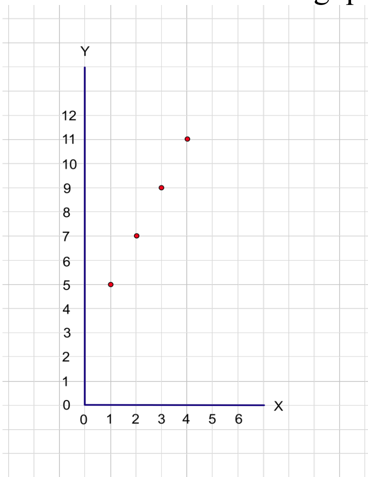
k	p
1	3
2	
	9

    b) What is the value of p when $k = 21$? Show how you got your answer.
    c) What is the value of k when $p = 61$? Show how you got your answer.

11. A fourth-grade class needs 5 leaves each day to feed its 2 caterpillars.

    a) How many leaves would they need each day for 12 caterpillars? Explain how you got your answer.
    b) How many leaves would they need each day for 15 caterpillars? Explain how you got your answer.

12. Answer the following question for the graph given below:



Which of the following tables could have been used to construct the graph? Circle the table and explain why you think it is the correct table.

x	y
1	2
2	4
3	6
4	8

x	y
1	5
2	7
3	9
4	11

Appendix C: HLM models

Grade 3

At grade 3, the student-level model included a grand-mean centered covariate, performance on the grade 3 pre-test. The level 1 model was:

$$Y_{ij} = \beta_{0j} + \beta_1(\text{Grade 3 pre-test})_{ij} + r_{ij}$$

In this model, the dependent variable Y_{ij} represents the Grade 3 post-test algebra assessment score of student *i* with teacher *j*; β_{0j} (intercept) represents the mean LEAP assessment score of the students taught by teacher *j*; β_1 is the Level 1 coefficient that measures the effects of Grade 3 pre-test score on an individual student's Grade 3 post-test score; and r_{ij} is the unique effect of student *i* on the assessment.

The full teacher-level model for analysis included the two composite variables that emerged from the factor analysis of grade 3 data: core algebraic practice and MQI. Both teacher-level variables were also grand-mean centered. The level 2 model was:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Core Algebraic Practice})_j + \gamma_{02}(\text{MQI})_j + \mu_{0j}$$

The dependent variable β_{0j} is the average LEAP assessment score of students taught by teacher *j*; γ_{00} is the average LEAP assessment score of students across teachers; γ_{01} and γ_{0s} are the Level 2 coefficients that measure the effects of composite variables for teacher practices; μ_{0j} is the residual (unexplained) variance at Level 2.

Grade 4:

The general structure of the models remained the same for Grade 4, with the exception that the teacher-level predictor variables differed based on the results of the factor analysis. Specifically, in grade 4, the teacher-level variables were core algebraic practice, teacher precision, and student focused. The Level 1 and Level 2 models were:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_1(\text{Grade 3 post-test})_{ij} + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Core Algebraic Practice})_j + \gamma_{02}(\text{Precision})_j + \gamma_{03}(\text{Student-Focused})_j + \mu_{0j}$$

At Level 1, students' Grade 3 post-test LEAP assessment score was used as the grand-mean centered covariate, as that was the most recent performance assessment prior to the Grade 4 assessment. All Level 2 variables were grand-mean centered.

Grade 5:

The general structure of the model for Grade 5 was identical to that of Grade 3. The Level 1 and Level 2 models were:

$$\text{Level 1: } Y_{ij} = \beta_{0j} + \beta_1(\text{Grade 4 post-test})_{ij} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Core Algebraic Practice})_j + \gamma_{02}(\text{MQI})_j + \mu_{0j}$$

At Level 1, students' Grade 4 post-test LEAP assessment score was used as the grand-mean centered covariate, as that was the most recent performance assessment prior to the Grade 5 assessment. All Level 2 variables were grand-mean centered.