Does Word-Problem Performance Maintain?

Follow-Up One Year After Implementation of a Word-Problem Intervention

Sarah R. Powell[1], Katherine A. Berry[1], Anna-Maria Fall[1], Greg Roberts[1], Marcia A. Barnes[2], Lynn S. Fuchs[2], Amanda Martinez-Lincoln[2], Suzanne R. Forsyth[1], Rebecca K. Vinsonhaler[1], Sarah A. Benz[3], Brenda L. Zaparolli[1], and Xin Lin[1]

[1]The University of Texas at Austin

[2]Vanderbilt University

[3]American Institutes for Research

Citation: Powell, S. R., Berry, K. A., Fall, A.-M., Roberts, G., Barnes, M. A., Fuchs, L. S., Martinez-Lincoln, A., Forsyth, S. R., Vinsonhaler, R. K., Benz, S. A., Zaparolli, B., & Lin, X. (2022). Does word-problem performance maintain? Follow-up one year after implementation of a word-problem intervention. *Journal of Research on Educational Effectiveness, 15*(1), 52–77. https://doi.org/10.1080/19345747.2021.1961332

Author Note

Sarah R. Powell	https://orcid.org/0000-0002-6424-6160
Katherine A. Berry	https://orcid.org/0000-0001-8340-3259
Anna-Maria Fall	https://orcid.org/0000-0002-6257-6684
Greg Roberts	
Marcia A. Barnes	https://orcid.org/0000-0002-9446-3000
Lynn S. Fuchs	https://orcid.org/0000-0003-2099-5247
Amanda Martinez-Lincoln	https://orcid.org/0000-0002-6271-0353
Suzanne R. Forsyth	https://orcid.org/0000-0001-8082-9975
Rebecca K. Vinsonhaler	https://orcid.org/0000-0002-9024-6586
Sarah A. Benz	https://orcid.org/0000-0002-9729-5583
Brenda L. Zaparolli	
Xin Lin	

This research was supported in part by Grant R324A150078 from the Institute of Education Sciences in the U.S. Department of Education to the University of Texas at Austin. The content is solely the responsibility of the authors and does not necessarily represent the official views of the U.S. Department of Education.

Correspondence concerning this article should be addressed to Sarah R. Powell, 1 University Station D5300, Austin, TX 78712. E-mail: srpowell@austin.utexas.edu

Abstract

Powell, Berry, et al. (in press) conducted a randomized control trial assessing the effects of two variants of word-problem intervention with third graders (n = 304) experiencing mathematics difficulty. Students were assigned to a business-as-usual condition (BaU) or one of two variants of word-problem intervention. One variant included a pre-algebraic reasoning component (relational understanding of the equal sign as well as standard and nonstandard equation solving); the other included word-problem intervention without pre-algebraic reasoning. Students in both interventions significantly outperformed students in the BaU with large effect sizes (2.66 and 2.44), but there were no significant differences between the two intervention conditions. The purpose of the present analysis was to assess maintenance of effects 6 to 12 months after intervention with the students (n = 229), now in fourth grade. At follow-up, only students in the word-problem intervention with pre-algebraic reasoning significantly outperformed the BaU on a measure of word-problem solving. This finding suggests an advantage for the pre-algebraic reasoning component, yet the follow-up effect between intervention conditions was not significant. The ESs of 0.43 and 0.31 and persistence rates of 16% and 13% reveal substantial forgetting for both conditions, which suggests the dose of intervention may not be adequate for many students.

Keywords: follow-up; learning difficulty; maintenance; mathematics; pre-algebra; word problems

Does Word-Problem Performance Maintain?

Follow-Up One Year After Implementation of a Word-Problem Intervention

While many researchers conduct randomized controlled trials to determine efficacious ways to improve students' academic outcomes, fewer conduct follow-up testing to determine the long-term impacts of such interventions (Watts et al., 2019). When follow-up does occur, significant differences at posttest often fade (Bailey et al., 2017). Understanding the long-term impact of interventions is critical for determining the level of ongoing support students may require for success in and beyond school (Bailey et al., 2020). Long-term impact data also can help educators select cost effective options for use in classroom or intervention settings (Crowley et al., 2018).

As described by Bailey et al. (2017), interventions are more likely to produce long-term benefits if they meet three criteria. The outcome (i.e., skills) addressed via the intervention must be (1) malleable throughout the intervention, (2) fundamental for school or society success, and (3) unlikely to develop in a comparison condition that did not receive intervention. Unfortunately, even an intervention that addresses this trifecta is not guaranteed to produce long-term success. Instead, the benefit of an intervention often fades as intervention students lose skills learned during intervention (Kang et al., 2018) or as the control group catches up (Bailey et al., 2017). Despite the demonstrated difficulty in maintaining intervention results, the search for interventions that yield long-lasting effects continues.

To address this aim, we investigated follow-up effects of two variants of word-problem intervention that meet Bailey and colleagues' three criteria. First, our intervention focused on word-problem solving, which has proved malleable in prior intervention research (Freeman-Green et al., 2015; Fuchs et al, in press; Jitendra et al., 2017; Peltier et al., 2020). Second, the

focus on word-problem outcomes is essential for school success, especially in mathematics classrooms. The overwhelming majority of mathematics items on high-stakes tests involve word problems (Powell, Namkung, et al., in press) and performance on such tests opens or closes opportunities in mathematics and beyond (Byun et al., 2015). Third, word-problem skill is slow to develop under business-as-usual (BaU) circumstances because word-problem instruction in typical school programs is not practiced daily and often relies on ineffective strategies, such as linking keywords to operations (Powell, Berry, et al., 2020).

The purpose of the present analysis was to assess whether word-problem intervention affords lasting word-problem advantages. The base study was Powell, Berry, et al. (in press), a randomized controlled trial (RCT) evaluating the effects of word-problem schema intervention among third graders experiencing mathematics difficulty. For the present analysis, we examined effects 6 to 12 months later in fourth grade. In this Introduction, we explain the rationale for targeting students who experience mathematics difficulty. Next, we describe previous efforts to improve word-problem solving and our two intervention variants. Then, we review the follow-up literature involving mathematics interventions in the elementary and middle school grades for students experiencing mathematics difficulty. Finally, we summarize the purpose of the present analysis.

Students Experiencing Mathematics Difficulty

Approximately 5 to 7% of students receive special education services in mathematics (Devine et al., 2018; Mann Koepke & Miller, 2013), with an even higher percentage of students also demonstrating persistent difficulty with mathematics. This broader group of students, often referred to as those experiencing *mathematics difficulty* (MD) was the focus of the Powell, Berry, et al. (in press) RCT. As described in the literature, the achievement differences of students with

MD compared to students without MD are pervasive and span addition and subtraction (Andersson, 2010), multiplication (Bartelet et al., 2014), division (Landerl et al., 2009), and word problems (Cirino et al., 2015, Reikerås, 2009, Swanson et al., 2013; Tolar et al., 2016). Further, the low mathematics performance of students with MD is persistent (Nelson & Powell, 2018; Vukovic, 2012). For example, Geary et al. (2012) demonstrated that students with MD performed consistently lower on tests of counting, quantity discrimination, addition, subtraction, multiplication, division, and rational numbers than their peers without MD from first through fifth grade. Similar patterns of low performance for students with MD have been identified on early numeracy items across two grades (Navarro et al., 2012), on fluency with arithmetic facts across three grades (Vanbinst et al., 2015), and on fractions across five grades (Mazzocco et al., 2013).

Word-Problems Intervention for Students Experiencing MD

To address the pervasive and persistent difficulty with word problems (Kingsdorf & Krawec, 2014), researchers have developed interventions for improving the word-problem performance of students experiencing MD. These interventions have involved various approaches: schema instruction (Fuchs et al., 2008, in press; Griffin et al., 2018; Peltier et al., 2020; Powell et al., 2015), cognitive strategies (Krawec et al., 2012; Swanson et al., 2014), or drawings or graphic organizers (Flores et al., 2016; van Garderen et al., 2014). Intervention studies found word-problem improvement for each of these approaches.

In Powell, Berry, et al.'s (in press) RCT, the focus was schema instruction with use of cognitive strategies and graphic organizers. We investigated the efficacy of and the paths by which word-problem intervention, when conducted with versus without an embedded pre-algebraic reasoning component, improved the word-problem performance of students with MD.

We defined pre-algebraic reasoning as understanding the equal sign as a relational symbol and solving equations with a single unknown (Pillay et al., 1998). Our focus on pre-algebraic reasoning stemmed from research demonstrating that students often misinterpret the equal sign as an operational symbol (Matthews et al., 2012; Vincent et al., 2015), which leads to difficulty solving different types of equations (Driver & Powell, 2015).

To assess the added value of an embedded pre-algebraic reasoning component on word-problem outcomes and to deepen insight into the connection between equation solving and word problems, we randomly assigned third-grade students with MD to one of three conditions: word-problem intervention with the pre-algebraic reasoning component (Pirate Math Equation Quest; PMEQ), word-problem intervention without the pre-algebraic reasoning (Pirate Math-alone; PM-alone), or business-as-usual (BaU). Students in the two active treatment conditions participated in 45 to 51 individual sessions (30 min each session) about the additive word-problem schemas (i.e., Total, Difference, Change) and learned how to set up and solve word problems using schema knowledge. Students in PMEQ received 2 to 5 min of practice each session on interpreting the equal sign as relational and solving standard (e.g., 3 + __ = 12) and nonstandard (e.g., 8 = 13 – __ or 7 + __ = 9 + 5) equations. In lieu of pre-algebraic reasoning tasks, students in PM-alone completed 2 to 5 min of review activities each session on telling time, money, geometry, perimeter, area, place value, and fractions.

We administered pre- and posttests about interpreting the equal sign (*equal sign*), solving equations (*open equations*), and solving word problems (*word problems*). At posttest, PMEQ students and PM-alone students outperformed students in BaU on *word problems*, with ESs of 2.66 and 2.44, respectively. The *word problems* performance between PMEQ and PM-alone was comparable. Even so, multilevel path analytic analysis revealed a significant indirect effect for

PMEQ versus BaU, in which (a) PMEQ intervention (when contrasted to BaU) improved student performance on *equal sign*, which in turn improved *open equations*, which in turn improved *word problems* performance. This indirect effect for the contrast between PM-alone versus BaU was not significant.

In this way, the RCT of Powell, Berry, et al. (in press) revealed comparable efficacy for the two variants of word-problem schema intervention as indexed at the end of intervention. At the same time, Powell, Berry, et al. (in press) suggested an additional route to word-problem competence, which accrued due to stronger equal sign and open equation skill. A central question in the present analysis is: Does embedding pre-algebraic reasoning in word-problem schema intervention (i.e., PMEQ) afford a more lasting advantage over PM-alone on word problems?

Follow-Up Effects for Mathematics Interventions

In this section, we highlight recent studies in which the authors conducted follow-up testing after implementation of a mathematics intervention in the elementary or middle school grades for students with MD. We defined *follow-up* as testing delayed at least 1 month after intervention ends. We discuss follow-up studies in two categories: those occurring within the same school year in which researchers conducted intervention and those occurring in subsequent school years. In the spirit of Bailey et al. (2018), we calculated a persistence rate of effect sizes (ES) from posttest to follow-up by dividing the follow-up ES by the posttest ES. We interpreted this quotient proportionally as a *persistence rate* (i.e., 82% of the effect persisted from posttest to follow-up). Table 1 presents the posttest and follow-up ESs as well as the persistence rates from this review of recent studies.

Same School Year

In our search of the literature from the last decade, we identified four mathematics intervention studies conducted with students with MD with same-year follow-up data collection. Dyson et al. (2013) administered follow-up testing six weeks after posttesting with a sample of kindergartners with MD. The mean posttest ES comparing a number sense intervention and BaU was 0.42 on a proximal number sense measure and 0.13 on a distal mathematics measure, with both comparisons significant. Intervention students continued to show significantly higher number sense scores at follow-up over students in the comparison group, with an ES of 0.44 (persistence rate of 104%) with no difference between conditions on the distal mathematics measure (ES = 0.01).

Dyson et al. (2015) identified significant effects at posttest favoring a kindergarten mathematics intervention with addition and subtraction flashcard practice compared to a BaU condition on measures of number sense (ES = 0.82), mathematics facts (ES = 0.78), and computation (ES = 0.60). At follow-up 8 weeks after posttest, students in the intervention condition continued to significantly outperform students in the BaU, but with moderate fading, on number sense (ES = 0.56; persistence = 68%) and mathematics facts (ES = 0.58; persistence = 74%), with a marginally significant difference on computation (ES = 0.49; persistence = 82%). For a second mathematics intervention featuring addition and subtraction game practice compared to a BaU, the authors identified significant gains at posttest on mathematics facts (ES = 0.69) and computation (ES = 0.58) but these effects faded at 8-week follow-up with ESs of 0.27 and 0.12, neither of which demonstrated significance.

With a focus on fractions and students with MD at Grade 6, Dyson et al. (2020) reported posttest effects comparing students who did and did not receive intervention of 0.90 for placing fractions on a number line, 0.99 on fraction concepts, and 0.48 on fraction computation. At 7-

week follow-up testing, the authors noted significant differences for placing fractions on a number line (ES = 1.02; persistence = 113%) and fraction concepts (ES = 0.63; persistence = 64%). The authors did not identify a significant difference at follow-up on fraction computation (ES = 0.35).

With a similar 7-week follow-up span with a sixth-grade fraction intervention, Barbieri et al. (2020) identified significant effects favoring the fraction intervention over BaU at posttest on placing fractions on a number line (ES = 0.85), fraction concepts (ES = 1.09), and fraction comparison (ES = 0.82) but no significant posttest effect on fraction computation (ES = 0.17). The authors reported lower scores at follow-up than at posttest, but still identified a significant difference favoring intervention students over control, with ESs on three number line, fraction concepts, and fraction comparison of 0.60 (persistence = 71%), 0.66 (persistence = 61%), and 0.61, respectively (persistence = 74%).

To summarize, for follow-up testing conducted within the same school year as intervention, we noted that, of the 13 significant effects at posttest, 11 of these faded from posttest to follow-up. Six of the 11 faded effects remained significant at follow-up, but five were insignificant. Of the six faded yet significant effects at follow-up, we noted the posttest ES of all was greater than 0.75. Persistence rates for these six studies ranged from 61% to 74%, suggesting that students retained two-thirds to three-fourths of posttest performance gains at follow-up. The other two effects increased slightly from posttest to follow-up with Dyson et al.'s (2013) effect size on a proximal number sense measure increasing from 0.42 at posttest to 0.44 at follow-up, with a persistence rate of 104%. Dyson et al. (2020) demonstrated no apparent decrease over time with an ES of 0.90 at posttest and 1.02 at follow-up on a task about placing fractions on a number line. This led to a persistence rate of 113%. These two persistence rates greater than

100% should be interpreted with caution. Our interpretation is that intervention students maintained all of their performance advantage at follow-up compared to students in a BaU. We understand that BaU students may have declined in performance more than intervention students, so we do not interpret these persistence rates greater than 100% as meaning that intervention students continued to grow compared to students in the BaU.

Subsequent School Years After Intervention

From the literature over the last decade, we located six follow-up studies conducted in subsequent school years after intervention. In these studies, researchers identified few to no significant differences between intervention and control conditions. Hallstedt et al. (2018) examined the effects of an intervention that used a mathematics game about addition and subtraction facts played on a tablet. The authors randomly assigned second-grade students with reading difficulty or MD to one of two variants of a mathematics game, a reading game, or a comparison condition. Students in the reading game and comparison condition acted as the BaU group. Students played the mathematics game for anywhere from 10.6 to 37.4 hours over the span of their second-grade year. At posttest, the authors noted significantly higher scores on addition 0-12 (ES = 0.67), subtraction 0-12 (ES = 0.53), and subtraction 0-18 (ES = 0.50). The authors identified no significant difference on addition 0-18 (ES = 0.13). At a 6-month follow-up session, the authors identified only one significant difference on subtraction 0-12 (ES = 0.28; persistence = 52%) with ESs ranging from -0.11 to 0.18 on the three other measures. At a 12-month follow-up, the authors identified no significant differences.

Clarke et al. (2016) also collected data one year following intervention. In their kindergarten year, students with MD participated in a number sense intervention or a BaU condition. At posttest, the authors demonstrated significant effects favoring the intervention

students on four of five measures (ES = 0.28 – 0.75). At follow-up in first grade, the authors noted no significant differences on a follow-up measure different from the measures used in kindergarten. In similar studies with kindergarten students with MD followed into first grade, both Clarke et al. (2017) and Doabler et al. (2016) identified no significant differences at follow-up after identifying significant differences at posttest. We did not calculate persistence rates for these studies because they used a different follow-up measure from the posttest measures.

Smith et al. (2013) collected follow-up data in second grade, one year after implementation of a first-grade mathematics intervention. Similar to other studies, the authors identified significant differences between intervention and BaU at posttest on mathematics facts (ES = 0.15), applied problems (ES = 0.28), and quantitative concepts (ES = 0.24). At follow-up, the authors identified no significant differences between intervention and BaU on any measures (ESs = 0.09, 0.00, and 0.06, respectively).

Conducting follow-up at 1 and 2 years past intervention implementation, Bailey et al. (2020) followed first-grade students with MD who received one of two variants of a number sense intervention or who participated in a comparison condition. At posttest in first grade, the authors noted students in the number sense intervention with speeded practice outperformed students in the comparison on four of five measures (ESs = 0.14 – 0.42). For the other intervention with nonspeeded practice, students receiving intervention outperformed comparison students on three of five measures (ESs = 0.20 – 0.29). At follow-up in second grade, the authors identified no significant differences on any of the five measures for students in either intervention condition with ESs ranging from -0.03 to 0.16. This trend maintained at third grade with ESs ranging from -0.02 to 0.12.

To summarize, for this set of studies in which the authors collected follow-up data in subsequent school years after implementation of the intervention, the long-term results are limited. In only one study (Hallstedt et al., 2018) with only one measure, follow-up results showed a continued significant difference between intervention at BaU, but this effect faded (ES of 0.53 to 0.28) and demonstrated a persistence rate of 52%. For all other studies in which the authors administered the same measures at posttest and then at follow-up one year later, all significant effects from posttest faded to nonsignificant at follow-up. In Table 1, we calculated persistence rates for the insignificant follow-up effects with a significant posttest effect, but we do not interpret these because of the insignificance of the follow-up data. In the three studies by Clarke, Doabler, and colleagues (Clarke et al., 2016, 2017; Doabler et al., 2016), a new measure administered at follow-up failed to demonstrate significant differences between the intervention and BaU conditions. As the only study to follow-up two years after intervention, Bailey et al. (2020) noted no significant differences, which was expected given no significant differences at follow-up one year after intervention.

Purpose of the Present Analysis

Watts et al. (2019) described a need to conduct follow-up testing to assess fade-out effects within interventions evaluated in RCTs. Current reporting of long-term impacts of efficacious interventions remains limited, which creates a dearth in the research literature (Watts et al., 2019). Limited information about long-term impacts impedes learning how a specific intervention can impact a student's success across the elementary and secondary grades, as well as into college or career.

To address this need and contribute to the research base on follow-up effects, we tested students from the RCT of Powell, Berry, et al. (in press) 6 to 12 months after word-problem

solving intervention ended in the academic year subsequent to when intervention had occurred. We focused on word-problem schema intervention, conducted with third graders with MD. Because Powell, Berry, et al. (in press) included two variants of intervention, we also examined whether embedding pre-algebraic reasoning within the word-problem schema intervention (i.e., PMEQ) afforded a more lasting advantage over PM-alone on word problems. Based on the results of Powell, Berry, et al. (in press) in which PMEQ students demonstrated a word-problem advantage when we conducted a sequential mediation analysis, we hypothesized PMEQ students may continue to have a practical advantage over PM-alone students on *word problems* at fourth grade.

Method

Before describing the details of the present analysis, we provide an overview of the Powell, Berry, et al. (in press) RCT. We recruited third-grade teachers at the beginning of the school year. We screened their classrooms and identified students with MD. We randomly assigned these students to 1 of 3 conditions. Students in two of the conditions received word-problem intervention for 16 weeks during their third-grade year. We posttested all students with MD at the end of their third-grade year. After a summer school break, we located and conducted follow-up testing of students with MD (now in fourth grade) who completed posttesting in the previous school year. Therefore, the sequence of the study was: pretesting in third grade, intervention (for two-thirds of students with MD), posttesting, summer break, follow-up testing in fourth grade.

Context and Setting

After receiving approval from our university's Institutional Review Board, we recruited third-grade classroom teachers from a large urban school district in the Southwest of the U.S.

This public school district served over 80,000 students. On average, the district reported 55.5% of students as Hispanic, 29.6% as Caucasian, 7.1% as African American, and 7.7% as belonging to another race or ethnic category. In the school district, 27.1% of students identified as dual-language learners, 52.4% qualified as economically disadvantaged, and 12.1% received special education services. The graduation rate was 90.7%. Table 2 presents the demographic information for the students in our sample.

Participants

We recruited two cohorts of third-grade students for project participation across two years. During the 2016-2017 school year, for cohort 1, we recruited 37 third-grade teachers from 13 elementary schools. These 37 third-grade teachers taught 52 separate mathematics classes. From these 52 classrooms, we screened 916 third-grade students. During the 2017-2018 school year, for cohort 2, we recruited 44 teachers from 13 schools who taught 51 classrooms of students. We screened 818 third-grade students in the second cohort. In this study, we combined the data from cohorts 1 and 2 for a total of 1,734 third-grade students who participated in screening.

In third grade, we screened all students using a measure of *Single-Digit Word Problems* (Jordan & Hanich, 2000). We used this measure to screen for mathematics difficulty (MD) in the area of word problems because word-problem solving was the primary focus of the intervention. For study eligibility, we identified students scoring at or below the 25th percentile, a common cut-off score in research related to MD (Geary et al., 2012; Hecht & Vagi, 2010; Locuniak & Jordan, 2008). After completion of the pretest battery, we identified 304 third-grade students with MD across the two cohorts. During the third-grade year, we randomly assigned the 304 students to one of three conditions, blocking by classroom teachers: Pirate Math Equation Quest

(PMEQ) intervention; Pirate Math without Equation Quest (PM-alone) intervention; and business-as-usual (BaU) comparison. In our analyses, in which we compared PMEQ, PM-alone, and BaU students, we determined each of these conditions at the beginning of each student's third-grade year. After completion of posttesting in third grade, we provided no further intervention to any student.

In their fourth-grade year (the 2017-2018 school year for Cohort 1 and 2018-2019 school year for Cohort 2), we contacted the original schools of each of the 304 students. Of the 304, we found and tested 196 students in their original schools. We identified and tested 32 additional students who had moved to 25 other schools in the area. Of these, 15 students were in 12 schools in the same school district as the original study. The other 17 students were in 13 different schools in 9 different school districts in the surrounding area. The furthest we drove for follow-up testing was 45 miles one-way. We also identified and tested 1 student who was homeschooled. The reasons for not conducting follow-up testing with a student included the following: could not locate student (n = 27), student moved out of Texas (n = 15), principal refused follow-up testing in her school (n = 15), student moved out of school district (n = 10), student moved out of country (n = 2), no principal response from original school (n = 2), caregiver revoked consent (n = 2), student homeschooled and unable to contact parent (n = 1), and student moved schools after the first follow-up session (n = 1).

Therefore, we completed follow-up testing at fourth grade with 229 students across the two cohorts. Attrition rates for follow-up testing varied across treatment conditions. In PMEQ, 31% of students had missing follow-up data, whereas 20% of the students in PM-alone failed to complete the follow-up battery. In BaU, 22% of students did not complete the follow-up battery. An overall attrition rate of 25% combined with differential attrition of 1.5% (PM-alone versus

BaU) and 9.7% (PMEQ versus BaU) represents tolerable threat of bias under optimistic assumptions set by What Works Clearinghouse (What Works Clearinghouse, 2017). Table 2 displays the demographics of the 229 students who completed follow-up testing in fourth grade with a comparison to the demographics of the 304 who participated at third grade.

Word-Problem Intervention

See Powell, Berry, et al. (in press) for a full description of the two variants of the word-problem intervention. In this section, we provide a brief overview. For the students assigned to the PMEQ and PM-alone conditions, interventionists completed 45 to 51 individual sessions, 3 times per week, for 30 min each session. PMEQ and PM-alone students participated in five activities for each session: (1) Math Fact Flashcards, (2) Equation Quest or Pirate Crunch, (3) Buccaneer Problems, (4) Shipshape Sorting, and (5) Jolly Roger Review. Only one activity (i.e., Equation Quest or Pirate Crunch) differed for students in the two intervention conditions. Interventionists provided intervention to students in both word-problem intervention conditions. Students also received general education mathematics instruction from their classroom teacher.

During Math Fact Flashcards, all students answered as many addition and subtraction fact flashcards as they could during two, 1-min timing. At the end of the second 1-min timing, students graphed the highest score from the two trials.

During Equation Quest (for PMEQ students only), students learned to interpret the equal sign as a relational symbol. Students also solved standard and nonstandard equations by balancing both sides of the equation with concrete manipulatives (e.g., balance scale and blocks), drawing pictures, or solving equations presented with numbers and symbols. Students learned a set of steps to balance equations with a variable (i.e., “X”), which involved isolating the variable with both standard and nonstandard equations. In Pirate Crunch (for PM-alone students only),

students participated in paper and pencil tasks about concepts of telling time, money, geometry, perimeter, area, place value, and fractions.

In the Buccaneer Problems, all students participated in interventionist-led schema instruction through a series of three Buccaneer Problems. Students first learned to approach any word problem using the RUN attack strategy: *R*ead the problem, *U*nderline the label and cross out irrelevant information, and *N*ame the problem type (i.e., choose the correct schema to use). For each of the three schemas, Total, Difference, and Change, students learned to use an equation to represent the problem and to mark "X" to represent the missing information.

During Shipshape Sorting, students sorted word problems by schema during a 1-min timing. During the Jolly Roger Review, students worked independently for 1 min to answer math facts, solve computation problems, or write appropriate equations for the three word-problem schemas. Then, students worked independently for 2 min to solve a word problem using the schema steps taught during the Buccaneer Problems.

All lesson guides and student materials for PMEQ are available for free. See www.piratemathequationquest.com for information and videos about implementation of the intervention and to download all PMEQ materials.

Business-as-Usual Comparison

Students in the BaU condition did not receive any intervention from our research team. These students received regular classroom mathematics instruction. Classroom word-problem instruction for BaU students (as well as PMEQ and PM-alone students) incorporated general mnemonic devices (e.g., RICE: Read and restate, Illustrate, Calculate, Explain and edit), keyword clues (e.g., *altogether* means add), and practice in applying problem-solution rules, as self-reported by participating teachers. Notably, none of the core mathematics classroom

practices included schema instruction or explicit discussions about the equal sign as a relational symbol.

Examiners/Interventionists

We recruited 28 research assistants to serve as examiners for pre-, post, and follow-up testing and interventionists during tutoring. All research staff were pursuing or had obtained a Master's or doctoral degree in an education-related field. During the 2016-2017 school year (cohort 1), research staff (n = 15) were predominately female (n = 13), with 53% identifying as Caucasian (n = 8), 27% as Hispanic (n = 4), 13% as Asian American (n = 2), and 7% as African American (n = 1). During the 2017-2018 school year (cohort 2), all research staff were female (n = 16), with 69% (n = 11) identifying as Caucasian, 13% percent as Hispanic (n = 2), and 6% as American Indian (n = 1), African American (n = 1), and Asian American (n = 1), respectively. Only 10% (n = 3) of research staff were the same from cohort 1 to cohort 2.

Throughout the implementation of the interventions, research staff participated in trainings to ensure strong preparation of all aspects of the intervention. For fourth-grade follow-up, follow-up testers (n = 5) met with the research staff in late August during 2, 1-hr meetings to develop a testing schedule, review the scheduling protocol, and practice implementing the testing battery. The Project Manager conducted weekly check-ins with the follow-up testers to assess testing progress, assist in locating students who moved schools, and update the testing schedule. For both cohorts, follow-up testing of fourth-grade students began in October and ended in March.

Measures

Screening Measure in Grade 3

As mentioned, we used *Single-Digit Word Problems* as the primary measure for

screening and identifying students with MD in third grade (Jordan & Hanich, 2000). This measure included 14 one-step word problems involving sums or minuends of 9 or less categorized into the Total, Difference, and Change schemas. We only used this screening measure (*Single-Digit Word Problems*) for identifying MD. We did not administer this measure at subsequent timepoints.

Pretest in Grade 3, Posttest in Grade 3, and Follow-Up in Grade 4

Word Problems. *Texas Word Problems-Brief* included eight word problems requiring double-digit computation, with one Total, three Difference, and four Change problems. For each problem, examiners read the problem aloud and provided approximately 1 min for students to solve the problem and write an answer. Examiners could re-read each problem up to one time upon student request. We scored this measure as the number of correct numerical and label responses for a maximum score of 16. At Grade 3, Cronbach's α was .83. At follow-up in Grade 4, α was .80.

On *Texas Word Problems-Part 1*, students solved nine double-digit word problems: two Total problems, one Difference problem, four Change problems, and two multi-schema problems (i.e., Difference and Change; Total and Difference). Two problems featured the interpretation of graphs. Examiners read each problem aloud and provided students time to solve the problem and write an answer. This measure had a maximum score of 18. In Grade 3, Cronbach's α was 84. At follow-up, α was .81.

For *Texas Word Problems-Part 2*, students solved nine double-digit word problems: two Total problems, two Difference problems, three Change problems, one multi-schema problem (i.e., Total and Change), and one multiplicative problem (i.e., Equal Groups schema). Three problems featured the interpretation of graphs, and one problem included irrelevant information.

The maximum score was 18. We measured Cronbach's α in Grade 3 at .85. In Grade 4 at follow-up, α was .81.

Open Equations. For *Open Equations*, students solved 10 equations in a standard (e.g., 3 + __ = 8) format. Students also solved equations in nonstandard formats, including two identity statements (e.g., __ = 4), 10 nonstandard equations with an operator symbol on the right side (e.g., 5 = 9 – __), and eight nonstandard equations with operator symbols on both sides (e.g., 9 – 6 = 7 – __). Excluding the identity statements, 14 of the equations included addition operator symbols and 14 included subtraction operator symbols. Students completed as many problems as possible within the 6-min timing. We scored this measure as the number of correct answers, with a maximum score of 30. At pretest in Grade 3, α was .86. At follow-up in Grade 4, Cronbach's α was .88.

Equal Sign Tasks. *Equal-Sign Tasks* assessed students' understanding of the equal sign and equivalence in written format. First, examiners asked students to write a definition of the equal sign. Next, students decided if the equal sign was used correctly in nonstandard, closed equations. Then, students read statements of equivalence and decided whether each statement was always true, sometimes true, or never true. Finally, students viewed a closed equation with addends on both sides, broke the equation into two parts, and defined the meaning of the equal sign in the equation. The maximum score was 14. At pretest in Grade 3, α was .64. Cronbach's α at follow-up was .75.

Scoring

At pre- and posttest in Grade 3, we combined the three word-problem measures (*Texas Word Problems-Brief*, *Texas Word Problems-Part 1*, and *Texas Word Problems-Part* 2) for a composite score named *word problems*. Cronbach's α at third grade was .94. Similar to our

analyses at third grade, we created a composite *word problems* score in which we combined the *Texas Word Problems-Brief*, *Texas Word Problems-Part 1*, and *Texas Word Problems-Part 2.* At follow-up, the Cronbach's α for this composite score was .92.

Two examiners independently entered scores on 100% on the test protocols for each outcome measure on an item-by-item basis into an electronic database, resulting in two separate databases. We removed student names from all tests to ensure examiners did not know the identity or condition of any student. We compared the discrepancies between the two databases across each outcome measure and rectified any inconsistencies to reflect the original response. Two examiners and the Project Manager resolved all discrepancies. Then, we converted students' responses to correct (1) and incorrect (0) scores using spreadsheet commands, which ensured 100% accuracy of scoring. Original scoring reliability was 99.9% for pretesting, 99.8% for posttesting, and 97.8% for follow-up.

Procedure

In third grade, during the first week of September, examiners administered the whole-class screening in one, 55-min session. Identification of students with MD occurred shortly thereafter, with four weeks of individual pretesting during the last two weeks of September and the first two weeks of October. During the third week of October, approximately 4 to 6 days after pretesting, intervention began and occurred three times per week for 16 weeks, concluding the third week in March. During this time, examiners filled the roles of interventionists. Approximately 4 to 6 days after the last intervention session, posttesting occurred in five, 45-min small group sessions with four students or fewer. We administered posttesting over three weeks, beginning the last week of March and ending the second week of April.

We contacted school principals for permission to conduct follow-up testing in August and

September of the following school year. School staff helped to identify the students who attended the same school and the students who had moved. We conducted follow-up testing by school to ensure that all students in one school, regardless of condition, participated in follow-up testing in the same time frame. Follow-up testing at fourth grade occurred in two, 45-min sessions with individual students. In the first session, examiners administered *Texas Word Problems-Brief* and *Texas Word Problems-Part 2*. In the second session, examiners administered *Open Equations, Texas Word Problems-Part 1,* and *Equal-Sign Tasks*. As with pretest and posttest, examiners at follow-up did not know the condition of any of the students. We administered follow-up testing over a 6-month span each school year, beginning in early October and ending in late March.

Research Design and Nesting

Powell, Berry, et al. (in press) randomly assigned students to one of three conditions (i.e., PMEQ, PM-alone, or BaU). Random assignment occurred at the student level. An interventionist conducted the intervention individually with students in the two active intervention conditions. Students randomized to BaU were not assigned to interventionists because BaU students did not receive supplemental intervention from the research team. This arrangement, where only a subset of a multilevel sample is nested (here in interventionists), is commonly described as a partially nested randomized design (Lohr et al., 2014). The data structure also was cross-classified because cases from different levels of the model were not completely nested. Here, teacher and interventionist were crossed because students from the same teacher may have had different interventionists and because students from different teachers may have had the same interventionist. We accounted for different variance structures as described in Luo et al. (2015).

Results

Preliminary Analyses

Table 3 displays the means and standard deviations for each measure. All variables distributed normally based on estimates of skewness and kurtosis. We identified no outlying values. For our analysis, we refer to the composite score of three word-problem measures as *word problems.* We refer to the score on *Open Equations* as *open equations* and the score from *Equal-Sign Tasks* as *equal sign.*

Posttest Effects at Grade 3

See Powell, Berry, et al. (in press) for the complete results from third grade. On *equal sign*, students in PMEQ outperformed students in PM-alone (ES = 1.01) and BaU (ES = 0.73), but the difference between PM-alone intervention and BaU was not significant (ES = - 0.26). On *open equations*, students in PM-alone outperformed students in BaU (ES = 0.30). The contrast of PMEQ and BaU did not differ (ES = 0.14) and PMEQ and PM-alone did not differ (ES = -0.11). On *word problems*, students in the PMEQ and PM-alone conditions significantly outperformed students in the BaU at posttest with ESs of 2.66 for PMEQ and 2.44 for the PM-alone. Students in the PMEQ and PM-alone conditions did not score significantly different from each other (ES = 0.22).

We also conducted a sequential mediation analysis (Powell, Berry, et al., in press) in which we modeled the mediating effect of intervention via improvement on *equal sign* and *open equations* on the word-problem outcome using multilevel path analytic framework (Bauer et al., 2006). This indirect effect was significant for the PMEQ versus the BaU condition, but not for PM-alone versus the BaU. Thus, the effects of PMEQ on *word problems* accrued via a combination of direct and indirect effects by directly improving *word problems* and indirectly strengthening *word problems* through *equal sign* and *open equations*. By contrast, the effects of

PM-alone on *word problems* accrued entirely from the *direct effects of word-problem schema intervention*.

Follow-Up Effects at Grade 4

To determine the long-term impact of the two variants of the intervention, we administered follow-up assessments during each student's fourth-grade year (i.e., approximately 6 to 12 months after completion of posttesting). We estimated main effect contrasts as blocked partially nested cross-classified data, with students nested in teachers in all three conditions, students nested in interventionists in the two treatment conditions (PMEQ and PM-alone), and interventionists and teachers crossed in the PMEQ and PM-alone groups. We estimated average treatment effects (ATE) as multilevel, partially nested, and cross-classified in R using the *lme4* package (Luo et al., 2015) under intent-to-treat assumptions. We included scores at pretest for each outcome as covariates in the respective models. We calculated treatment effects as the across-arm mean difference divided by the *SD* within the control arm only (Lai & Kwok, 2016). Intraclass correlations (ICCs) at the teacher level were .11, .08, and .10 for *equal sign, open equations,* and *word problems*, respectively. Interventionist-level ICCs in the treatment groups were .07, .13, and .12 for *equal sign, open equations,* and *word problems*.

Results, which are displayed in Table 4, indicate that on *word problems*, students in the PMEQ condition maintained their learning over time and continued to outperform students in the BaU at follow-up ($\beta = 3.76$, $p = .01$). The effect size was 0.43. We calculated a persistence rate of 16% by dividing 0.43 by 2.66. By contrast, the difference between PM-alone and BaU was not statistically significant ($\beta = 2.67$, $p = .08$, ES = 0.31). We calculated a persistence rate of 13% by dividing 0.31 by 2.44. On *equal sign* and *open equations,* students receiving pre-algebraic instruction embedded within word-problem schema intervention (PMEQ) did not maintain their

advantage over BAU or PM-alone for *equal sign* nor did students in PM-alone maintain their advantage over BAU for *open equations*.

Discussion

In this analysis, we explored whether the effects of third-grade word-problem schema intervention persisted into fourth grade or, as often demonstrated for mathematics intervention, effects faded, particularly when follow-up occurs in the school year(s) following the intervention. As described in Powell, Berry, et al. (in press), both word-problem interventions demonstrated strong significant direct effects at the end of third grade upon completion of the intervention. Students in PMEQ outperformed BaU students on *word problems* outcomes with an ES of 2.66. Similarly, students in PM-alone outperformed students in the BaU with an ES of 2.44, and there was no significant difference between the two word-problem intervention variants. The posttest ESs contrasting each word-problem intervention condition against BaU were large and similar to posttest ESs in other RCTs evaluating similar word-problem schema interventions for students with MD (Fuchs et al., 2020; Jitendra et al., 2013; Powell et al., 2015).

At follow-up, which was conducted in fourth grade, 6 to 12 months after third-grade posttesting, we identified one significant difference between PMEQ and BaU students on the *word problems* outcome (ES = 0.43). We calculated a persistence rate of 16% meaning PMEQ students retained approximately 16% of the effect from posttest at follow-up. This significant difference, which continued to favor students in the PMEQ intervention condition over BaU, differed from previous subsequent-year follow-up research for students with MD who participated in mathematics interventions (Bailey et al., 2020; Clarke et al., 2016; Doabler et al., 2016; Smith et al., 2013). These earlier studies identified no or minimal significant effects at follow-up. Hallstedt et al. (2018) identified one significant effect on 1 of 4 measures at a 6-

month follow-up (ES = 0.28) with a persistence rate of 52%, but no significant effects at a 12-month follow-up. Therefore, our persistence rate is commendable compared to other research with subsequent-year follow-up, but it also shows the substantial loss of word-problem knowledge in the year following implementation of the intervention.

In one important sense, the significant follow-up effect favoring students in PMEQ demonstrates the long-term impact of the PMEQ intervention. This result may corroborate Bailey and colleagues' (2017) proposed trifecta of essential components for persistent intervention benefits to accrue. First, in PMEQ, students learned a skill (i.e., word-problem solving) that was malleable via the intervention. As demonstrated in many other studies with students with MD, word-problem solving can be taught to the population of students with MD (Fuchs et al., 2014; Krawec et al., 2012; Morin et al., 2017; Swanson et al., 2013). Second, in PMEQ, we focused on word problems, an essential skill that is fundamental for success in school. In an analysis of the problems on high-stakes mathematics tests, Powell, Namkung, et al. (in press) determined the majority of mathematics problems involved reading and solving word problems. Because students in the elementary grades must set up and solve word problems to demonstrate mathematics competency, word-problem solving is necessary for success in the elementary mathematics classroom, in later grade levels, and beyond school. In PMEQ, we also focused on helping students understand the equal sign as relational and practice solving different types of equations through the pre-algebraic reasoning component of Equation Quest. Both of these skills are necessary for pre-algebraic reasoning, as essential connector between arithmetic and algebra (Pillay et al., 1998). Third, because the word-problem instruction provided in classrooms often is limited in scope and reliant on ineffective word-problem strategies (Karp et al., 2019; Powell, Berry, et al., 2020), it is unlikely for BaU students to develop strong word-

problem skills without the PMEQ intervention.

Notably, however, PMEQ's significant follow-up effect was not documented for our two other outcomes of interest: *equal sign* or *open equations*. We did not expect to see differences on *open equations* because there were not significant differences between PMEQ and BaU on this measure at posttest in third grade. We did expect to see a significant difference on *equal sign* between PMEQ and BaU at follow-up, given the significant difference at third grade (ES = 0.73) and the repetitive modeling and practice about the equal sign as relational as provided through Equation Quest in third grade. Perhaps interpreting the equal sign as relational is a skill that may eventually develop with BaU (Bailey et al., 2017) or a concept that students may forget about over time (Kang et al., 2018). The persistence rate on *equal sign* for PMEQ students was only 8%, which indicates high loss of equal-sign knowledge from posttest to follow-up.

Unlike the long-term impact on the *word problems* outcome for PMEQ students, the follow-up contrast for the same word-problem intervention without the pre-algebraic reasoning component (i.e., PM-alone) was not significant. The ES comparing word-problem intervention without algebraic reasoning versus BaU was 0.31, and we calculated a persistence rate of 13%. This is in contrast to the ES of 0.43 and persistence rate of 16% comparing word-problem intervention with algebraic reasoning versus BaU. At third-grade posttest, PM-alone students demonstrated a slightly lower ES on *word problems* than students in the PMEQ, which was nonetheless a large effect (i.e., 2.44 for PM-alone compared to 2.66 for PMEQ). Although not significant at fourth grade, PM-alone students continued to demonstrate a small advantage on *word problems* over students in the BaU. This result, favoring PMEQ over PM-alone when compared to the BaU, may be attributed to the alternate indirect path to word-problem performance provided by the PMEQ intervention with the embedded pre-algebraic reasoning

component (Powell, Berry, et al., in press).

This differing pattern of results at follow-up on the *word problems* outcome is important but worthy of more research. At the end of third grade, students in both word-problem interventions demonstrated significantly higher *word problems* scores than students in the BaU; at fourth-grade follow-up, only one word-problem intervention repeated this pattern. We would like to claim that PMEQ, with its embedded pre-algebraic reasoning component, is the superior word-problem intervention variant when comparing PMEQ and PM-alone. But such a claim warrants further investigation given the nonsignificant, but trending positive, performance on *word problems* for PM-alone compared to the BaU and the similar persistence rates of 16% and 13% between PMEQ and PM-alone. Further research is necessary because the PM-alone intervention also meets the proposed trifecta of essential components for persistent intervention benefits as outlined by Bailey et al. (2017). Similar to PMEQ, PM-alone students learned about word-problem solving, which is a malleable skill. In PM-alone, we focused on an essential skill fundamental success in school (i.e., word-problem solving), just as we did with PMEQ. Finally, and similar to PMEQ, we compared PM-alone students to students in a BaU, who were unlikely to develop strong word-problem skills without the intervention.

As explained by Kang et al. (2018), fadeout occurs because students do not learn what they already know. From this point of view, the BaU students began to learn more about the equal sign, solving equations, and solving word problems, and they catch up to their peers. The average means for BaU students from pretest to posttest to follow-up corroborate this hypothesis. For example, on *open equations* from posttest to follow-up, we noted an approximately 3-point gain for both PMEQ and BaU and a 1-point gain for students in PM-alone. This indicates all students improve on *open equations* over time, and students in the BaU start to catch up PM-

alone students on solving standard and nonstandard equations (Bailey et al., 2017). PMEQ and PM-alone students may not progress in their learning because they retained these skills during the intervention provided in third grade.

Forgetting occurs because students who participate in intervention – especially a "rich intensive intervention" (Kang et al., 2018, p. 591) – may forget content after the intervention ceases. We hypothesize our results on *equal sign* and *word problems* may be contributed to forgetting, and our persistence rates corroborate this hypothesis. The decreases in performance from posttest to follow-up indicate PMEQ and PM-alone students forgot some of what they learned during intervention whereas the BaU students continued to catch up with their increases in performance.

The difference in *word problems* ESs between posttest and follow-up was substantial in both conditions: from 2.66 to 0.43 for PMEQ (16% persistence) and from 2.44 to 0.31 for PM-alone (13% persistence). Together with prior studies, the present analysis suggests that, for many students, although intervention is necessary, it is not sufficient for many students. Instead, sustained intervention or booster sessions are likely required for many students with MD with a focus on simultaneous development of conceptual (i.e., schemas) and procedural knowledge (i.e., attack strategy; Rittle-Johnson et al., 2001). We envision booster sessions could be implemented once a month for 30 min with small groups of participating students to review the attack strategy (RUN) and the schemas (Total, Difference, and Change). In these sessions, students could practice setting up and solving word problems similar to those featured during intervention with appropriate feedback from the teacher.

Limitations

Four limitations emerged during follow-up testing and should be considered when

interpreting these findings. First, we did not collect follow-up data from all 304 students who finished posttesting in their third-grade year. We only were able to locate and conduct follow-up testing with 75% (n = 229) of fourth-grade students who participated in the study during third grade. Also, we experienced differential attrition for students in PMEQ, though this attrition was within tolerable limits. Future research should determine the most effective methods for contacting students who move to another school district, state, or country and determine whether virtual follow-up testing offers a feasible alternative that yields accurate results.

Second, we did not conduct follow-up testing beyond fourth grade. Future research should follow-up two or more years upon completion of the posttest, as recommended by Bailey et al. (2020), to understand when interventions effects cease to be significant. Third, we only asked students to respond to pre-, post-, and follow-up tests in written format. We assumed students' written responses represented their mathematics knowledge; however, asking students to respond orally or with pictorial representations may have allowed students to answer a greater number of problems correctly. Future research should include measures that capture different methods of response to the questions. Future research also should consider collecting process data from students as they solve word problems to understand which strategies from PMEQ maintained over time and which faded or were forgotten.

Fourth, we administered a limited number of measures at follow-up. Given time constraints in schools for teachers and students, we conducted two 45-min follow-up sessions. After we scheduled the *equal sign*, *open equations,* and *word problems* measures into the follow-up sessions, we did not have time to administer other mathematics measures (e.g., fact fluency, computation, mathematics anxiety) or different word problems than those administered at third grade. As suggested by Watts et al. (2019), our follow-up battery should have included a

standardized test of word-problem solving at fourth grade that aligned more closely with long-term success in school (e.g., released fourth-grade word problems from the National Assessment of Educational Progress). Future research should consider adding a third or fourth follow-up session to collect more robust data on students to examine long-term impacts beyond the core measures of the study. As such, we should also collect data from high-stakes mathematics tests administered at the state level to understand the transfer of word-problem performance to holistic measures of mathematics performance.

Conclusion

For students who participated in the PMEQ intervention in third grade, we identified a long-term direct effect on *word problems* at fourth grade, 6 to 12 months after the completion of the intervention. The PMEQ intervention, with the practical advantage of the pre-algebraic reasoning component, led to improved word-problem outcomes at third grade and fourth grade, although the effect at fourth grade was substantially smaller than at the end of third grade (*word problems* ES of 2.66 vs. 0.43 for intervention vs. BaU) with a persistence rate of only 16%. Although the effect at posttest was significant and of similar magnitude (ES = 2.44) for the word-problem intervention without the pre-algebraic reasoning component, the follow-up effect did not achieve signficance (ES = 0.31) with a persistence rate of only 13%. Still forgetting in both word-problem conditions was similar and substantial. This suggests that the dose of intervention for the MD population, although necessary, it is not sufficient for many students. Instead, sustained intervention and/or continual booster sessions about word-problem solving likely are required.

References

Andersson, U. (2010). Skill development in different components of arithmetic and basic cognitive functions: Findings from a 3-year longitudinal study of children with different types of learning difficulties. *Journal of Educational Psychology, 102*(1), 115–134. https://doi.org/10.1037/a0016838

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness, 10*(1), 7–39. https://doi.org/10.1080/19345747.2016.1232459

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist, 73*(1), 81–94. https://doi.org/10.1037/amp0000146

Bailey, D. H., Fuchs, L. S., Gilbert, J. K., Geary, D. C., & Fuchs, D. (2020). Prevention: Necessary but insufficient? A 2-year follow-up of an effective first-grade mathematics intervention. *Child Development, 91*(2), 382–400. https://doi.org/10.1111/cdev.13175

Barbieri, C. A., Rodrigues., J., Dyson, N. & Jordan, N. C. (2020). Improving fraction understanding in sixth graders with mathematics difficulties: Effects of a number line approach combined with cognitive learning strategies. *Journal of Educational Psychology, 112*(3), 628–648. https://doi.org/10.1037/edu9999384

Bartelet, D., Ansari, D., Vaessen, A., & Blomert, L. (2014). Cognitive subtypes of mathematics learning difficulties in primary education. *Research in Developmental Disabilities, 35*, 657–670. https://doi.org/10.1016/j.ridd.2013.12.010

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and

recommendations. *Psychological Methods, 11*(2), 142–163. https://doi.org/10.1037/1082-989X.11.2.142

Byun, S.-Y., Irvin, M. J., & Bell, B. A. (2015). Advanced math course taking: Effects on math achievement and college enrollment. *The Journal of Experimental Education, 83*(4), 439–468. https://doi.org/10.1080/00220973.2014.979570

Cirino, P. T., Fuchs, L. S., Elias, J. T., Powell, S. R., & Schumacher, R. F. (2015). Cognitive and mathematical profiles for different forms of learning difficulties. *Journal of Learning Disabilities, 48*(2), 156–175. https://doi.org/10.1177/0022219413494239

Clarke, B., Doabler, C. T., Kosty, D., Nelson, E. K., Smolkowshi, K., Fien, H., & Turtura, J. (2017). Testing the efficacy of a kindergarten mathematics intervention by small group size. *AERA Open, 3*(2), 1–16. https://doi.org/10.1177/2332858417706899

Clarke, B., Doabler, C., Smolkowski, K., Nelson, E. K., Fien, H., Baker, S. K., & Kosty, D. (2016). Testing the immediate and long-term efficacy of a Tier 2 kindergarten mathematics intervention. *Journal of Research on Educational Effectiveness, 9*(4), 607–634. https://doi.org/10.1080/19345747.2015.1116034

Crowley, D. M., Dodge, K. A., Barnett, W. S., Corso, P., Duffy, S., Graham, P., Greenberg, M., Haskins, R., Hill, L., Jones, D. E., Karoly, L. A., Kuklinski, M. R., & Plotnick, R. (2018). Standards of evidence for conducting and reporting economic evaluations in prevention science. *Prevention Science, 19*, 366–290. https://doi.org/10.1007/s11121-017-0858-1

Devine, A., Hill, F., Carey, E., & Szücs, D. (2018). Cognitive and emotional math problems largely dissociate: Prevalence of developmental dyscalculia and mathematics anxiety. *Journal of Educational Psychology, 110*(3), 431–444. https://doi.org/10.1037/edu0000222

Doabler, C. T., Clarke, B., Kosty, D. B., Kurtz-Nelson, E., Fien, H., Smolkowski, K., & Baker, S. K. (2016). Testing the efficacy of a Tier 2 mathematics intervention: A conceptual replication study. *Exceptional Children, 83*(1), 92–110. https://doi.org/10.1177/0014029166600084

Driver, M. K., & Powell, S. R. (2015). Symbolic and nonsymbolic equivalence tasks: The influence of symbols on students with mathematics difficulty. *Learning Disabilities Research and Practice, 30*(3), 127–134. https://doi.org/10.1111/ldrp.12059

Dyson, N., Jordan, N. C., Beliakoff, A., & Hassinger-Das, B. (2015). A kindergarten number-sense intervention with contrasting practice conditions for low-achieving children. *Journal for Research in Mathematics Education, 46*(3), 331–370.

Dyson, N. I., Jordan, N. C., & Glutting, J. (2013). A number sense intervention for low-income kindergartners at risk for mathematics difficulties. *Journal of Learning Disabilities, 46*(2), 166–181. https://doi.org/10.1177/0022219411410233

Dyson, N. I., Jordan, N. C., Rodrigues, J., Barbieri, C., & Rinne, L. (2020). A fraction sense intervention for sixth graders with or at risk for mathematics difficulties. *Remedial and Special Education, 41*(4), 244–254. https://doi.org/10.1177/0741932518807139

Flores, M. M., Hinton, V. M., & Burton, M. E. (2016). Teaching problem solving to students receiving tiered interventions using the concrete-representational-abstract sequence and schema-based instruction. *Prevention School Failure: Alternative Education for Children and Youth, 60*(4), 345–355. https://doi.org/10.1080/1045988X.2016.1164117

Freeman-Green, S. M., O'Brien, C., Wood, C. L., & Hitt, S. B. (2015). Effects of the SOLVE strategy on the mathematical problem solving skills of secondary students with learning disabilities. *Learning Disabilities Research and Practice, 30*(2), 76–90.

Fuchs, L. S., Seethaler, P. M., Powell, S. R., Fuchs, D., Hamlett, C. L., & Fletcher, J. M. (2008). Effects of preventative tutoring on the mathematical problem solving of third-grade students with math and reading difficulties. *Exceptional Children, 74*(2), 155–173. https://doi.org/10.1177/001440290807400202

Fuchs, L. S., Seethaler, P. M., Sterba, S. K., Craddock, C., Fuchs, D., Compton, D. L., Geary, D. C., & Changas, P. (in press). Closing the word-problem achievement gap in first grade: Schema-based word-problem intervention with embedded language comprehension instruction. *Journal of Educational Psychology.* https://doi.org/10.1037/edu0000467

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2012). Mathematical cognition deficits in children with learning disabilities and persistent low achievement: A five-year prospective study. *Journal of Educational Psychology*, *104*(1), 206–223. https://doi.org/10.1037/a0025398

Griffin, C. C., Gagnon, J. C., Jossi, M. H., Ulrich, T. G., & Myers, J. A. (2018). Priming mathematics word problem structures in a rural elementary classroom. *Rural Special Education Quarterly, 37*(3), 150–163. https://doi.org/10.1177/8756870518772164

Hallstedt, M. H., Klingberg, T., & Ghaderi, A. (2018). Short and long-term effects of a mathematics tablet intervention for low performing second graders. *Journal of Educational Psychology, 110*(8), 1127–1148. https://doi.org/10.1037/edu0000264

Hecht, S. A., & Vagi, K. J. (2010). Sources of group and individual differences in emerging fraction skills. *Journal of Educational Psychology, 102*(4), 843–859. https://doi.org/10.1037/a0019824

Jitendra, A. K., Harwell, M. R., Dupuis, D. N., & Karl, S. R. (2017). A randomized trial of the effects of schema-based instruction on proportional problem-solving for students with

mathematics problem-solving difficulties. *Journal of Learning Disabilities, 50*(3), 322–336. https://doi.org/10.1177/0022219416629646

Jitendra, A. K., Rodriguez, M., Kanive, R., Huang, J.-P., Church, C., Corroy, K. A., & Zaslofsky, A. (2013). Impact of small-group tutoring interventions on the mathematical problem solving and achievement of third-grade students with mathematics difficulties. *Learning Disability Quarterly, 36*(1), 21–35. https://doi.org/10.1177/0731948712457561

Jordan, N. C., & Hanich, L. B. (2000). Mathematical thinking in second-grade children with different forms of LD. *Journal of Learning Disabilities, 33*(6), 567–578. https://doi.org/10.1177/002221940003300605

Kang, C. Y., Duncan, G. J., Clements, D. H., Sarama, J., & Bailey, D. H. (2018). The roles of transfer of learning and forgetting in the persistence and fadeout of early childhood mathematics interventions. *Journal of Educational Psychology, 111*(4), 590–603. https://doi.org/10.1037/edu0000297

Karp, K. S., Bush, S. B., & Dougherty, B. J. (2019). Avoiding the ineffective keyword strategy. *Teaching Children Mathematics, 25*(7), 429–435.

Kingsdorf, S., & Krawec, J. (2014). Error analysis of mathematical word problem solving across students with and without learning disabilities. *Learning Disabilities Research and Practice, 29*(2), 66–74.

Krawec, J., Huang, J., Montague, M., Kressler, B., & Melia de Alba, A. (2012). The effects of cognitive strategy instruction on knowledge of math problem-solving processes of middle school students with learning disabilities. *Learning Disability Quarterly, 36*(2), 80–92. https://doi.org/10.1177/0731948712463368

Lai, M. H. C., & Kwok, O.-M. (2016). Estimating standardized effect sizes for two- and three-

level partially nested data. *Multivariate Behavioral Research, 51*(6), 740–756. https://doi.org/10.1080/00273171.2016.1231606

Landerl, K., Fussenegger, B., Moll, K., & Willburger, E. (2009). Dyslexia and dyscalculia: Two learning disorders with different cognitive profiles. *Journal of Experiemental Child Psychology, 103*, 309–324. https://doi.org/10.1016/j.jecp.2009.03.005

Locuniak, M. N., & Jordan, N. C. (2008). Using kindergarten number sense to predict calculation fluency in second grade. *Journal of Learning Disabilities*, *41*(5), 451–459. https://doi.org/10.1177/0022219408321126

Lohr, S., Schochet, P.Z., and Sanders, E. (2014). *Partially nested randomized controlled trials in education research: A guide to design and analysis.* (NCER 2014-2000) Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Luo, W., Cappaert, K. J., & Ning, L. (2015). Modelling partially cross-classified multilevel data. *British Journal of Mathematical and Statistical Psychology, 68*(2), 342–362. https://doi.org/10.1111/bmsp.12050

Mann Koepke, K., & Miller, B. (2013). At the intersection of math and reading disabilities: Introduction to the special issue. *Journal of Learning Disabilities, 46*(6), 483–489. https://doi.org/10.1177/0022219413498200

Matthews, P., & Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of Experimental Child Psychology, 104*(1), 1–21. https://doi.org/10.1016/j.jecp.2008.08.004

Matthews, P., Rittle-Johnson, B., McEldoon, K., & Taylor, R. (2012). Measure for measure: What combining diverse measures reveals about children's understanding of the equal

sign as an indicator of mathematical equality. *Journal for Research in Mathematics Education, 43*(3), 316–350.

Mazzocco, M. M. M., Myers, G. F., Lewis, K. E., Hanich, L. B., & Murphy, M. M. (2013). Limited knowledge of fraction representations differentiates middle school students with mathematics learning disability (dyscalculia) versus low mathematics achievement. *Journal of Experimental Child Psychology, 115*, 371–387. http://doi.org/10.1016/j.jecp.2013.01.005

Morin, L. L., Watson, S. M. R., Hester, P., & Raver, S. (2017). The use of a bar model drawing to teach word problem solving to students with mathematics difficulties. *Learning Disability Quarterly, 40*(2), 91–104. https://doi.org/10.1177=0731948717690116

Navarro, J. I., Aguilar, M., Marchena, E., Ruiz, G., Menacho, I., & Van Luit, J. E. H. (2012). Longitudinal study of low and high achievers in early mathematics. *British Journal of Educational Psychology, 82*, 28–41. https://doi.org/10.1111/j.2044-8279.2011.02043.x

Nelson, G., & Powell, S. R. (2018). A systematic review of longitudinal studies of mathematics difficulty. *Journal of Learning Disabilities, 51*(6), 523–539. https://doi.org/10.1177/0022219417714883

Peltier, C., Sincalir, T. E., Pulos, J. M., & Suk, A. (2020). Effects of schema-based instruction on immediate, generalized, and combined structured word problems. *The Journal of Special Education, 54*(2), 101–112. https://doi.org/10.1177/0022466919883397

Pillay, H., Wilss, L., & Boulton-Lewis, G. (1998). Sequential development of algebra knowledge: A cognitive analysis. *Mathematics Education Research Journal, 10*(2), 87–102. https://doi.org/10.1007/bf03217344

Powell, S. R. (2007). *Open Equations.* Available from S. R. Powell, 1912 Speedway, D5300,

Austin, TX 78712.

Powell, S. R., & Berry, K. A. (2015). *Texas Word Problems.* Available from S. R. Powell, 1912 Speedway, D5300, Austin, TX 78712.

Powell, S. R., Berry, K. A., & Benz, S. A. (2020). Analyzing the word-problem performance and strategies of students experiencing mathematics difficulty. *Journal of Mathematical Behavior, 58*(100759), 1–16. https://doi.org/10.1016/j.jmathb.2020.100759

Powell, S. R., Berry, K. A., Fall, A.-M., Roberts, G., Fuchs, L. S., & Barnes, M. A. (in press). Alternative paths to improved word-problem performance: An advantage for embedding pre-algebraic reasoning instruction within word-problem intervention. *Journal of Educational Psychology.* https://doi.org.10.1037/edu0000513

Powell, S. R., Fuchs, L. S., Cirino, P. T., Fuchs, D., Compton, D. L., & Changas, P. C. (2015). Effects of a multitier support system on calculation, word problem, and pre-algebraic learning among at-risk learners. *Exceptional Children, 81*(4), 443–470. https://doi.org/10.1177/0014402914563702

Powell, S. R., Namkung, J. M., & Lin, X. (in press). An investigation of using keywords to solve word problems. *The Elementary School Journal.*

Reikerås, E. K. L. (2009). A comparison of performance in solving arithmetical word problems by children with different levels of achievement in mathematics and reading. *Investigations in Mathematics Learning, 1*(3), 49–72.

Rittle-Johsnon, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346–362. https://doi.org/10.1037//0022-0663.93.2.346

Smith, T. M., Cobb, P., Farran, D. C., Cordray, D. S., & Munter, C. (2013). Evaluating math

recovery: Assessing the causal impact of a diagnostic tutoring program on student achievement. *American Educational Research Journal, 50*(2), 397–428. https://doi.org/10.3102/0002831212469045

Swanson, H. L., Lussier, C., & Orosco, M. (2013). Effects of cognitive strategy interventions and cognitive moderators on word problem solving in children at risk for problem solving difficulties. *Learning Disabilities Research and Practice, 28*(4), 170–183.

Swanson, H. L., Orosco, M. J., & Lussier, C. M. (2014). The effects of mathematics strategy instruction for children with serious problem-solving difficulties. *Exceptional Children, 80*(2), 149–168. https://doi.org/10.1177/001440291408000202

Tolar, T. D., Fuchs, L., Fletcher, J. M., Fuchs, D., & Hamlett, C. L. (2016). Cognitive profiles of mathematical problem solving learning disability for different definitions of disability. *Journal of Learning Disabilities, 49*(3), 240–256. https://doi.org/10.1177/0022219414538520

Vanbinst, K., Ceulemans, E., Ghesquière, P., & De Smedt, B. (2015). Profiles of children's arithmetic fact development: A model-based clustering approach. *Journal of Experimental Child Psychology, 133*, 29–46. https://doi.org/10.1016/j.jecp.2015.01.003

van Garderen, D., Scheuermann, A., & Poch, A. (2014). Challenges students identified with a learning disability and as high-achieving experience when using diagrams as a visualization tool to solve mathematics word problems. *ZDM Mathematics Education, 46*, 135–149. https://doi.org/10.1007/s11858-013-0519-1

Vincent, J., Bardini, C., Pierce, R., & Pearn, C. (2015). Misuse of the equals sign: An entrenched practice from early primary years to tertiary mathematics. *Australian Senior Mathematics Journal, 29*(2), 31–39.

Vukovic, R. K. (2012). Mathematics difficulty with and without reading difficulty: Findings and implications from a four-year longitudinal study. *Exceptional Children, 78*(3), 280–300.

Watts, T. W., Bailey, D. H., & Li, C. (2019). Aiming further: Addressing the need for high-quality longitudinal research in education. *Journal of Research on Educational Effectiveness, 12*(4), 648–658. https://doi.org/10.1080/19345747.2019.1644692

What Works Clearinghouse (2017). *What Works Clearinghouse: Procedures and Standards Handbook, Version 4.0.* Washington, DC: Author. Retrieved from: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Table 1

Persistence Rates from Previous Studies

Author	Intervention	Grade	Measure	Posttest ES	Follow-up ES	Persistence rate[a]	Weeks from posttest to follow-up
Same School Year							
Dyson et al. (2013)	Number sense	K	Number sense	0.42*	0.44*	104%	6
	Number sense	K	Mathematics	0.13	0.01		6
Dyson et al. (2015)	Add/subtract flashcard practice	K	Number sense	0.82*	0.56*	68%	8
	Add/subtract flashcard practice	K	Mathematics facts	0.78*	0.58*	74%	8
	Add/subtract flashcard practice	K	Computation	0.60*	0.49	82%	8
	Add/subtract game practice	K	Number sense	0.32	0.26		8
	Add/subtract game practice	K	Mathematics facts	0.69*	0.27	39%	8
	Add/subtract game practice	K	Computation	0.58*	0.12	21%	8
Dyson et al. (2020)	Fraction sense	6	Fraction number line	0.90*	1.02*	113%	7
	Fraction sense	6	Fraction concepts	0.99*	0.63*	64%	7
	Fraction sense	6	Fraction computation	0.48*	0.35	73%	7
Barbieri et al. (2020)	Fraction sense	6	Fraction number line	0.85*	0.60*	71%	7
	Fraction sense	6	Fraction concepts	1.09*	0.66*	61%	7
	Fraction sense	6	Fraction computation	0.17	0.11		7
	Fraction sense	6	Fraction comparison	0.82*	0.61*	74%	7
Subsequent School Years							
Hallstedt et al. (2018)	Add/subtract tablet game	2	Addition 0-12	0.67*	0.18	27%	26
	Add/subtract tablet game	2	Subtraction 0-12	0.53*	0.28*	52%	26
	Add/subtract tablet game	2	Addition 0-18	0.13	-0.11		26
	Add/subtract tablet game	2	Subtraction 0-18	0.50*	0.04	8%	26
	Add/subtract tablet game	2	Addition 0-12	0.67*	0.03	5%	52
	Add/subtract tablet game	2	Subtraction 0-12	0.53*	0.13	24%	52
	Add/subtract tablet game	2	Addition 0-18	0.13	0.02		52

	Add/subtract tablet game	2	Subtraction 0-18	0.50*	0.07	14%	52
Smith et al. (2013)	Math recovery	1	Mathematics facts	0.15*	0.09	60%	52
	Math recovery	1	Applied problems	0.28*	0.00	0%	52
	Math recovery	1	Quantitative concepts	0.24*	0.06	25%	52
Bailey et al. (2020)	Number sense speeded practice	1	Mathematics facts	0.42*	0.16	38%	52
	Number sense speeded practice	1	Number sets	0.33*	0.09	27%	52
	Number sense speeded practice	1	Computation	0.30*	0.08	27%	52
	Number sense speeded practice	1	Number line	0.14	0.11		52
	Number sense speeded practice	1	Numeration	0.14*	0.00	0%	52
	Number sense speeded practice	1	Mathematics facts	0.42*	-0.01	0%	104
	Number sense speeded practice	1	Number sets	0.33*	0.12	36%	104
	Number sense speeded practice	1	Computation	0.30*	0.03	10%	104
	Number sense speeded practice	1	Number line	0.14	-0.02		104
	Number sense speeded practice	1	Numeration	0.14*	0.08	57%	104
	Number sense nonspeeded practice	1	Mathematics facts	0.24*	0.09	38%	52
	Number sense nonspeeded practice	1	Number sets	0.20*	-0.03	0%	52
	Number sense nonspeeded practice	1	Computation	0.29*	0.01	3%	52
	Number sense nonspeeded practice	1	Number line	0.06	0.12		52
	Number sense nonspeeded practice	1	Numeration	0.06	0.00		52
	Number sense nonspeeded practice	1	Mathematics facts	0.24*	0.04	17%	104
	Number sense nonspeeded practice	1	Number sets	0.20*	0.12	60%	104
	Number sense nonspeeded practice	1	Computation	0.29*	0.08	28%	104
	Number sense nonspeeded practice	1	Number line	0.06	0.02		104
	Number sense nonspeeded practice	1	Numeration	0.06	0.08		104

[a]We calculated persistence rates for measures in which the posttest ES was significant. We calculated persistence rate as: follow-up ES divided by posttest ES. When a follow-up ES was negative, we interpreted the persistence rate as 0%.
* = significant based on author report

Table 2

Participant Demographics

	Grade 3	Grade 4 at follow-up			
	Overall	Overall	PMEQ	PM-alone	BaU
	(*N* = 304)	(*N* = 229)	(*n* = 72)	(*n* = 67)	(*n* = 90)
Gender (% female)	56.3	56.8	58.3	50.7	61.1
Race/ethnicity					
African American	11.5	11.8	16.7	7.5	11.1
Hispanic/Latinx	69.4	70.7	62.5	80.6	70.0
Caucasian	5.3	5.7	6.9	3.0	6.7
Asian	2.6	1.7	1.4	3.0	1.1
Multi-racial	6.6	7.0	9.7	6.0	5.6
Other	2.3	2.6	2.8	0.0	4.4
Students in special education	12.2	11.8	9.7	16.4	10.0
Dual-language learners	60.5	61.6	63.9	65.7	56.7

Table 3

Means and Standard Deviations For Outcomes Measures

		Pretest			Posttest			Follow-up		
		n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Equal sign	PMEQ	72	5.88	2.73	72	10.25	3.07	72	9.51	4.13
	PM-alone	67	5.69	2.97	67	7.25	2.87	67	8.70	3.63
	BaU	90	5.76	2.76	90	8.07	2.84	90	9.13	3.55
Open equations	PMEQ	72	5.65	3.88	72	11.71	6.28	72	14.81	6.24
	PM-alone	67	5.25	4.57	67	12.25	4.75	67	13.46	6.15
	BaU	90	5.69	4.10	90	10.77	5.94	90	13.70	5.53
Word problems	PMEQ	72	6.74	4.76	72	26.64	10.85	72	17.31	10.56
	PM-alone	67	6.64	5.20	67	25.57	10.40	67	15.91	9.54
	BaU	90	6.60	4.89	90	10.49	6.06	90	13.42	8.75

Note. BaU = Business as usual; PM-alone = Pirate Math without Equation Quest; PMEQ = Pirate Math Equation Quest.

Table 4

Results from Cross-Classified Models Testing Effects of Intervention in Grade 4 Controlling for Differences in Grade 3

	Word problems				*Open equations*				*Equal sign*			
Fixed effects	Estimate	*SE*	*p*-value	ES	Estimate	*SE*	*p*-value	ES	Estimate	*SE*	*p*-value	ES
Posttest effects												
Intercept	10.69	0.68	0.00		10.52	0.51	0.00		7.94	0.26	0.00	
Pretest	0.62	0.09	0.00		0.55	0.07	0.00		0.37	0.06	0.00	
PM-alone vs BaU	14.86	1.53	0.00	**2.44**	1.74	0.72	0.02	**0.30**	-0.75	0.41	0.07	-0.26
PMEQ vs BaU	16.17	1.61	0.00	**2.66**	0.85	0.86	0.33	0.14	2.11	0.45	0.00	**0.73**
Follow-up effects												
Intercept	13.25	1.03	0.00		13.66	0.6	0.00		9.16	0.41	0.00	
Pretest	0.60	0.12	0.00		0.55	0.09	0.00		0.29	0.09	0.00	
PM-alone vs BaU	2.67	1.50	0.08	0.31	-0.09	0.89	0.92	-0.02	-0.37	0.60	0.54	-0.10
PMEQ vs BaU	3.76	1.48	0.01	**0.43**	1.01	0.88	0.25	0.18	0.23	0.59	0.70	0.06
Random effects	Variance	ICC			Variance	ICC			Variance	ICC		
Posttest effects												
Student-level												
Intercept	38.55	0.44			20.93	0.61			5.48	0.73		
Tutor-level												
Intercept	2.47	0.03			6.20	0.18			0.54	0.07		

PM-alone vs BaU	18.74	0.21	6.06	0.18	0.79	0.11
PMEQ vs BaU	23.96	0.27	0.00	.00	0.18	0.02
Teacher-level						
Intercept	3.50	0.04	0.95	0.03	0.48	0.06
Follow-up effects						
Student-level	66.96	0.81	26.52	0.90	11.29	0.84
Tutor-level	7.18	0.09	1.55	0.05	1.00	0.07
Teacher-level	8.90	0.11	1.51	0.05	1.20	0.09

Note. BaU = Business as usual; PM-alone = Pirate Math without Equation Quest; PMEQ = Pirate Math Equation Quest.