Effectiveness of Mathematical Word Problem Solving Interventions for Students with Learning Disabilities and Mathematics Difficulties: A Meta-Analysis

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Amy E. Lein

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Dr. Asha Jitendra, Adviser

May, 2016

Acknowledgements

When I first applied to the Special Education doctoral program, Mary Farquhar thoughtfully matched me up with my adviser. I would like to thank my incredible adviser, Dr. Asha Jitendra for consistently providing me with guidance, support, and opportunities throughout my doctoral program. I would also like to thank my committee members: Dr. Sashank Varma, Dr. Michael Harwell, and Dr. Karen Storm, as well as Dr. Kristen McMaster. I consider myself very lucky to have learned from such amazing researcher-instructors.

Dedication

I dedicate my dissertation to my parents for their unwavering investment in my education beginning with Campus Lab School, through Carleton College, Lesley University, and now, the University of Minnesota. To my strong and intelligent sisters Jen, Emily, and Kari and my amazing nieces and nephews, Ellie, Adriaan, Abbey, Oskar, and Olive. I also dedicate it to my teacher friends and family (e.g., Auntie Mary & Kate, Ariadma B.), and the many students with special needs to whom I taught mathematics. I pledge to keep working to improve math education for you guys, integrating what I learned from you (e.g., Tyler's pop machine analogy for identifying functions), with special recognition of Sam Lane and Ashley Gardner who embody strength of spirit to which I can only aspire. Most of all, I dedicate this to Randy Donohue, for his love, support, and sacrifice. I can't possibly thank you enough.

Abstract

This meta-analysis synthesized the findings from 23 published and five unpublished experimental or quasi-experimental group design studies on word problem-solving instruction for K-12 students with learning disabilities (LD) and mathematics difficulties (MD). A secondary purpose of this meta-analysis was to analyze the relation between treatment effectiveness and various study features including (a) participant characteristics, (b) study design characteristics, (c) outcome measure characteristics, and (d) contextual characteristics of instruction. Results of a random effects model synthesizing the 31 independent effect sizes extracted from the 28 included studies showed an overall mean effect size of 1.03 (*SE* = 0.15). Grade level of participants, type of report (published vs. unpublished), assignment to conditions, reliability of outcome measure, instructional setting, interventionist, instructional arrangement, mathematics task, and intervention duration were found to moderate treatment effectiveness. Given that effect sizes in two studies (Fuchs, Fuchs, Finelli, et al., 2004; Fuchs, Fuchs, Prentice, Hamlett, et al., 2004) were over four standard deviations above the mean, analyses performed without these two influential points produced a lower mean effect size (*g* = 0.77, *SE* = 0.10) and impacted the results of moderator analyses. Specifically, only two of the six variables found to moderate intervention effectiveness when all 31 independent effect sizes were included remained significant after removing the two influential effect sizes. Those two variables were reliability of outcome measure and intervention duration. The results extend the findings of previous meta-analysis with regard to the effectiveness

of word problem solving interventions for students with LD and MD. Limitations and directions for future research are discussed.

Table of Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

Mathematical competence in school uniquely predicts student success in both higher education and future employment (Cavanagh, 2007; National Mathematics Advisory Panel [NMAP], 2008). Specifically, successful completion of higher-level mathematics courses (e.g., algebra) is associated with higher standardized test scores (National Center for Education Statistics [NCES], 2011), greater likelihood of college success, and access to better-paying jobs after college (Casner-Lotto & Barrington, 2006; OECD, 2010; Rivera-Batiz, 1992). Results from the National Assessment of Educational Progress (NAEP) mathematics assessment (NCES, 2013) indicate that less than half of students tested scored at or above the proficient level in mathematics. Further, among typically achieving students, 10% of fourth graders and 25% of eighth graders scored below the basic level. The results are even more dismal for students with disabilities. Approximately 75% of fourth grade students with disabilities scored below the proficient level, and almost half scored at or below the basic level. Ninety percent of eighth grade students with disabilities scored below proficient, and half scored below basic. These findings underscore the need for effective mathematics interventions for students with disabilities, who vary considerably in ability, achievement, and motivation, to develop the necessary mathematical knowledge to meet grade level benchmarks.

Mathematics involves conceptual knowledge, procedural fluency, and problem solving (Common Core State Standards Initiative, 2010; Jonassen, 2003; Van de Walle, Karp, & Bay-Williams, 2013; Voutsina, 2012). The present study focuses on problem

solving, a critical skill emphasized by educational research groups (e.g., Gonzales, Williams, Jocelyn, Kastberg, & Brenwald, 2008) and workforce development committees (e.g., OECD, 2010). Word problems constitute one of the most common types of problem solving (Jonassen, 2003) and serve as "a vehicle for developing students' general problem-solving skills" (Verschaffel, Greer, & De Corte, 2007, p.583).

Word problem solving is a complex process, which requires students to integrate cognitive and metacognitive processes to identify relevant information, determine what information is missing, and create an adequate representation of the problem, which leads to the selection and execution of appropriate solution strategies (Depaepe, De Corte, & Verschaffel, 2010; Desoete, Roeyers, & De Clercq, 2003; Mayer & Hegarty, 1996). Although many students with learning disabilities (LD) and students with mathematics difficulties (MD) struggle with mathematics in general, word problem solving is particularly challenging (e.g., Fuchs, Fuchs, & Prentice, 2004; Hanich, Jordan, Kaplan, & Dick, 2001; Jordan, Kaplan, & Hanich, 2002). These students evidence deficiencies in domain-general abilities such as working memory, language, and attentive behavior (Andersson & Lyxell, 2007; Fuchs et al., 2010; Geary, Hoard, Byrd-Craven, Nugent, & Numtee, 2007; Zheng, Swanson, & Marcoulides, 2011). Further, students with LD and MD also evidence deficits in domain-specific basic mathematics skills (e.g., calculation, fact fluency), and specific problem-solving skills (e.g., problem representation, metacognition, self-regulation) (Andersson, 2008; Desoete et al., 2003; Maccini & Ruhl, 2001).

Prior Research

Given the importance of problem solving skills and the well-documented inadequate performance of students with LD and MD, several meta-analyses of mathematics intervention research for students with LD and MD have been conducted (i.e., Gersten, Chard, Jayanthi, Baker, Morphy, & Flojo, 2009; Kroesbergen & Van Luit, 2003; Xin & Jitendra, 1999; Zhang & Xin, 2012; Zheng, Flynn, & Swanson, 2013). Three of these meta-analyses focused specifically on word problem solving interventions (Xin & Jitendra, 1999; Zhang & Xin, 2012; Zheng et al., 2013) and two meta-analyses addressed word problem solving interventions in addition to interventions in other domains (Gersten Chard, et al., 2009; Kroesbergen & Van Luit, 2003).

Xin and Jitendra (1999) conducted a meta-analysis on word problem solving interventions for Grade 1 – post-secondary students with high incidence disabilities (students with learning disabilities, mental retardation, and emotional disturbance) and students without disabilities who exhibited mathematics difficulties. The meta-analysis examined the effects of four instructional interventions: (1) representation techniques (i.e., diagrams, manipulatives, linguistic training and/or mapping instruction), (2) strategy training (i.e., direct instruction, explicit instruction in cognitive/metacognitive strategies, self-regulation in using heuristics), (3) computer-aided instruction (CAI) (i.e., computerized tutorials, interactive programming), and (4) "other" (i.e., no instruction, instruction not subsumed by the other three categories). A total of 14 group design studies and 12 single subject studies were included, which were analyzed using standardized mean change and percentage of non-overlapping data (PND), respectively.

Results of group design studies indicated a large, mean effect size (d = 0.89) for word problem solving interventions. Specifically CAI was found to be most effective, followed by representation techniques, and strategy training, which were superior to "other." One reservation with this meta-analysis is the inclusion of studies without control groups and the calculation of effect size in terms of gain scores for single groups, which does not "partial out the influence of pretest conditions" (Zheng et al., 2013, p.98).

The meta-analysis conducted by Kroesbergen and Van Luit (2003) focused on younger students (K-6) and included interventions aimed at preparatory mathematics, basic skills, and problem solving interventions. They included 34 group design studies and 24 single subject design studies and categorized studies by: (1) instructional approaches (direct instruction, self-instruction, and mediated or assisted instruction) and (2) teacher directed or computerized (CAI). Results indicated that problem solving interventions appeared to be less effective than interventions in the area of basic mathematics skills. Regarding instructional approaches, direct instruction and self-instruction produced similarly large, positive effect sizes (d = 1.45 and 0.91, respectively), which were more effective than mediated instruction (d = 0.34). Further, teacher directed instruction was found to be more effective than instruction delivered by computer. Several limitations of this study include: (a) exclusion of unpublished studies, which tends to bias effect sizes toward significant, positive results, (b) averaging of multiple dependent effect sizes within a single study without provision of evidence that the assumption of independence was not violated, and (c) decision to combine effect sizes

across group and single subject design studies, which tends to inflates effect sizes (Busse, Kratochwill, & Elliott, 1995).

Gersten, Chard, et al. (2009) synthesized interventions that covered a range of mathematics domains (i.e., operations, fractions, algebra, general math proficiency, word problem solving). The meta-analysis focused on students identified as having LD and included only randomized controlled studies and quasi-experimental design (QED) studies, "in which there was at least one treatment and one comparison group, evidence of pretest comparability for QEDs, and sufficient data with which to calculate effect size" (p. 1205). They examined the effects of: (1) instructional approaches or curriculum design (i.e., explicit instruction, use of heuristics, student verbalizations of mathematical reasoning, use of visual representations while solving problems, sequence and/or range of examples, other instructional and curricular variables) (2) providing formative assessment data and feedback to teachers on students' mathematics performance (e.g., student progress data, skill analysis, options for addressing instructional needs), (3) providing formative assessment data and feedback to students on their mathematics performance (e.g., general performance and effort feedback, progress toward a goal), and (4) peer-assisted instruction. With the exception of student feedback with goal setting and peer-assisted instruction, mean effects (range: g = 0.21 to 1.56) for all other instructional components were statistically significant. The largest effects were found for explicit instruction (g = 1.22) and use of heuristics (g = 1.56). One limitation of this meta-analysis is the significant overlap of instructional components, which hinders isolation of the unique contribution of specific components.

Zhang and Xin (2012) conducted a follow up to Xin and Jitendra's (1999) meta-analysis on word problem solving interventions for students with LD and MD. Zhang and Xin examined the impact of various education reforms (e.g., inclusion, standards based, response to intervention) on the effectiveness of mathematical word problem solving interventions. The research base included a total of 29 group-design and 10 single-subject design studies. Effect sizes for group-design studies ranged from d = -.054 to d = 11.82, with a mean of d = 1.58. A sub-category of the standards-based category (i.e., explicit instruction involving representation of the problem structure) produced the largest mean effect size (d = 2.64). Cognitive strategy instruction also yielded a large effect size (d = 1.86). Similar to Xin and Jitendra (1999), this meta-analysis is limited by the inclusion of studies without control groups and the calculation of effect size as the standardized difference between the posttest and pretest means for a single sample, which is known to inflate effect sizes and poses a threat to internal validity (Borman & D'Agostino, 1996; Borman, Hewes, Overman, & Brown, 2003; Cook & Campbell, 1979).

Most recently, Zheng et al., (2013) conducted a meta-analysis synthesizing word problem solving intervention research. They included seven group design and eight single subject design studies. This meta-analysis extended earlier meta-analyses by specifically examining "the role of sample characteristics within MD samples on treatment outcomes" (p.98). As such, they examined student characteristics in terms of presence and severity of mathematics difficulties (MD) with and without comorbid reading difficulties (MD+RD). In addition, the meta-analysis identified instructional components comprising the broader instructional categories examined in prior meta-analyses (e.g.,

Xin & Jitendra, 1999; Gersten, Chard, et al. (2009). Results showed that word problem solving interventions yielded a significant, positive mean effect (g = 0.76) for students with MD. In contrast, findings for students with MD+RD showed that the control group outperformed the treatment group on word problem solving (g = -0.45). While descriptive results suggested, "a great deal of commonality ... in the instructional components used in treatments that yield[ed] high ESs" (Zheng et al., 2013, p. 109), moderator analyses were not conducted due to the small number of studies included in the meta-analysis. In addition to small sample size, other limitations include exclusion of unpublished studies and calculation of multiple effect sizes per study without addressing the issue of dependency.

Although results of prior meta-analyses showed moderate to large effect sizes, indicating the overall effectiveness of mathematics interventions for students with LD and MD, findings related to study characteristics were mixed. For example, regarding the moderating effect of student age, one meta-analysis indicated that mathematics interventions are more effective for younger students (Gersten, Chard, et al., 2009), another found that they are more effective for older students (Kroesbergen & Van Luit, 2003), and a third found no significant difference by age (Xin & Jitendra, 1999). Findings across prior meta-analyses were also inconclusive regarding the performance of students with LD compared to students with MD. Specifically, findings from two meta-analyses showed that students with LD performed significantly lower than students with MD, while another found no significant difference between the two categories of students. The observed inconsistencies may be explained by variations in coding decisions in these

meta-analyses. Prior meta-analyses vary in terms of (a) coding decisions (i.e., sample characteristics, methodological characteristics, study characteristics), and (b) methodological decisions (i.e., inclusion of varying study designs, effect size extraction, statistical methods used to combine studies). The details of the relevant prior meta-analyses are examined in depth in chapter two.

Study Rationale

The purpose of the present meta-analysis was to extend the previous meta-analyses and reconcile discrepancies in their findings. The present meta-analysis provides precision in defining the population of studies and the potential moderator variables, including a clear coding rationale that includes operational definitions and illustrative examples. This type of transparency is critical so that readers can assess the credibility of meta-analytic findings (Harwell & Maeda, 2008).

First, I sought to identify the participant characteristics (e.g., LD vs. MD, grade level) that moderate mathematical word problem solving interventions. Previous meta-analyses either did not take into consideration the role of participant characteristics on treatment outcomes or were inconsistent in their operational definitions and coding decisions. In particular, clear description and operational definitions of criterion used to determine if a student has an LD or only MD is critical. Prior findings have been mixed when researchers compared the word problem solving performance between students with LD and those considered to be MD or "low-achieving" (LA). While some research suggests that students with LD perform similarly to those labeled MD or LA (e.g., Gonzalez and Espinel, 2002; Lackaye & Margalit, 2006; Montague, Enders, and Dietz,

2011), other research indicates that students with LD perform below the level of those labeled MD or LA in various aspects of mathematical knowledge and problem solving skills (e.g., Krawec, 2014; Mazzocco & Devlin, 2008). These discrepancies may be largely dependent on the criterion used to identify students as having LD or MD, which varies broadly across studies (see Fletcher & Vaughn, 2009; Geary, Hoard, Nugent, & Bailey, 2012; Murphy, Mazzocco, Hanich, & Early, 2007; Powell, Fuchs, Fuchs, Cirino, & Fletcher, 2009; Ysseldyke, Algozzine, Shinn, & McGue, 1982).

Second, I identified study design and outcome measure characteristics that may moderate intervention effectiveness. Gersten et al. (2005) proposed specific indicators of high-quality educational research that include the following: (a) random assignment of participants and interventionists to groups, (b) high reliability and validity of data yielded from outcome measures, (c) assessment of fidelity of implementation, and (d) documentation of attrition. As with participant characteristics, a majority of prior meta-analyses did not report sufficient detail regarding the coding of study design or outcome measure characteristics. Two prior meta-analyses (i.e., Gersten, Chard, et al., 2009; Xin & Jitendra, 1999) clearly defined how they coded the nature of assignment to conditions; one (Gersten, Chard, et al., 2009) reported coding the technical adequacy of measures; and in the three meta-analyses reporting FOI (i.e., Gersten, Chard, et al., 2009; Xin & Jitendra, 1999; Zheng et al., 2013), coverage was limited to a simple yes/no code. In the present meta-analysis, I provide operational definitions to explain the coding of relevant methodological characteristics. Further, unlike prior relevant meta-analyses, I evaluated

reporting of attrition and validity evidence associated with the outcome measure, and assessed reliability of outcome measure as a potential moderator variable.

Third, I sought to identify contextual characteristics of interventions (e.g., interventionist, duration) that may moderate effectiveness. Only Gersten, Chard, et al. (2009) considered the alignment between the intervention and the outcome measure, and the nature of instruction provided to the control group. Furthermore, only one study (Kroesbergen & Van Luit, 2003) assessed whether minutes of instruction per session and number of sessions impacted intervention effectiveness. Broad variations in coding procedures for categories such as interventionist, instructional setting and arrangement, and instructional components precluded comparison across meta-analyses in many cases. Where findings across meta-analyses could be directly compared, results were inconsistent. In the present meta-analysis, I provide detailed operational definitions to explain the coding of relevant contextual characteristics of interventions.

In addition to refining the coding procedures used in previous meta-analyses, I also employed meta-analytic best practices as recommended by Harwell and Maeda (2008). Specifically, I explicitly defined the population to which the results of the present meta-analysis will generalize, I examined the distribution of effect sizes, and I conducted all key analyses with and without potential outliers. Last, this meta-analysis presents a unique contribution to the literature in that, similar to Gersten, Chard, et al. (2009), the present meta-analysis: (a) explicitly stated the degree to which the critical assumption of independence of effect sizes was met and (b) avoided effect size calculations that are

vulnerable to inflation and pose a threat to internal validity. The present meta-analysis addressed the following research questions:

1. What is the effectiveness of word problem-solving interventions on the mathematical performance of school-aged (K-12) students with LD and/or MD?
2. Does intervention effectiveness vary as a function of participant characteristics (i.e., grade level, LD/MD status, race, socioeconomic status)?
3. Does intervention effectiveness vary as a function of study design characteristics (i.e., type of report, group assignment, type of comparison group, fidelity of implementation, attrition)?
4. Does intervention effectiveness vary as a function of outcome measure characteristics (i.e., type of measure, reliability, validity)?
5. Does intervention effectiveness vary as a function of contextual characteristics of interventions (i.e., interventionist, instructional arrangement, instructional setting, mathematics task, intervention duration)?

Chapter 2

LITERATURE REVIEW

In this chapter, I first discuss mathematical problem solving. Second I describe characteristics of students with mathematics learning disabilities (MLD) and mathematics difficulties (MD), as well as consider previous research on identifying students as having MLD or MD. Last, I review previous meta-analytic studies examining the effectiveness of mathematics interventions for students with MLD and MD.

Mathematical Problem Solving

Success in both daily life and employment depend on one's ability to quantify, calculate, and problem solve (Geary et al., 2012; Price & Ansari, 2013). Of these skills, problem solving may be the most important as "virtually everyone, in their everyday lives and professional lives, regularly solves problems" (Jonassen, 2000, p. 63). In this information age, even entry level jobs require mathematical competence, while success in higher level positions, which come with higher salaries and health benefits, require even stronger quantitative reasoning skills (Casner-Lotto & Barrington, 2006; OECD, 2010; Rivera-Batiz, 1992). Therefore, it is not surprising that the *Principles and Standards for School Mathematics* [National Council of Teachers of Mathematics (NCTM), 2000], the Common Core State Standards (CCSS, 2010), and international assessments like the Trends in International Mathematics and Science Study (TIMMS) emphasize the importance of problem solving.

Mathematical problem solving refers to "the cognitive process of figuring out how to solve a mathematics problem that one does not already know how to solve"

(Mayer & Hegarty, 1996, p. 31). In other words, problem solving involves the "process of moving from a given state to a goal state" (p. 31), with no clearly outlined solution path. Although mastery of concepts and specific mathematical procedures is a prerequisite for solution accuracy, problem solving involves processes that go above and beyond conceptual and procedural knowledge (Jonassen, 2000). Solving routine problems – those that students can solve using familiar methods in a step-by-step fashion (National Research Council, 2001; Polya, 1945) does not constitute problem solving (Woodward et al., 2012). For example, no-context problems such 78 + 65 = ? are routine problems, because the solution path is obvious (Mayer & Hegarty, 1996). Most children in upper elementary grades and adolescents can compute to solve this problem using prerequisite skills such as knowledge of number combinations and base ten operations. In contrast, nonroutine problems are those for "which there is not a predictable, well-rehearsed approach or pathway explicitly suggested by the task, task instructions, or a worked-out example" (Stein & Lane, 1996, p. 58). It is worth noting that whether a problem is routine or nonroutine is based on a student's prior experience solving those problems (Woodward et al., 2012).

Competence in problem solving (e.g., critical thinking, construction of arguments) is an important element of not only state and national standardized assessments, but also college entrance exams (Gonzales et al., 2008; Kirsch, Braun, Yamamoto, & Sum, 2007; Lesh, Hamilton, & Kaput, 2007; Levy & Murnane, 2004). Current emphasis in mathematics education is on solving complex, authentic problems situated in everyday contexts. However, in practice, the opportunity for experiential problem solving

instruction is limited and word problems that range from simple to complex represent "the most common form of problem solving" (Jonassen, 2003, p. 267) in school mathematics curricula. Moreover, learning how to solve word problems can help students develop their ability to mathematically model real-world problem situations (Depaepe, et al., 2010; Van de Walle, et al., 2013).

Word problems tend to be more challenging than no-context problems (Cummins, Kintsch, Reusser, & Weimer, 1988; Mayer, Lewis, & Hegarty, 1992) for several reasons (Jonassen, 2003; Lewis & Mayer, 1987; Lucangeli, Tressoldi, & Cendron, 1998; Schumacher & Fuchs, 2012; Schurter, 2002). Solving word problems requires the integration of several components – linguistic and factual knowledge, schematic knowledge (i.e., the mathematical relations among the various elements in the problem), strategic knowledge, and procedural knowledge (Mayer & Hegarty, 1996). Mayer (1998) describes the four phases involved in solving word problems. They include problem translation, problem integration, solution planning and monitoring, and solution execution.

Problem translation refers to the process of reading and mentally representing each statement in the word problem. This includes restating what is given in the problem and figuring out what one is being asked to solve (i.e., the problem goal). This first phase requires the application of linguistic and factual knowledge (Mayer, 1998). The second phase, problem integration, involves assembling the translated information into a coherent representation of the problem. This requires the application of schematic knowledge. Schemata refer to mental constructs that allow "problem solvers to group

problems into categories in which the problems in each category require similar solutions" (Cooper & Sweller, 1987, p. 348). The ability to differentiate between details that are relevant to the solution and irrelevant surface details facilitates identification of problem type based on underlying problem schemata, which is essential to mathematical problem solving (Quilici & Mayer, 1996; Sweller, Chandler, Tierney, & Cooper, 1990).

The third phase involves creating a solution plan and monitoring one's progress as that plan is implemented. This often includes breaking down the larger problem into sub-goals (e.g., identifying the problem type, comparing the current problem to previous problems, visually representing the problem) (Woodward et al., 2012). As students work through the problem solving process, it may be useful for them to refer to a list of these sub-goals to monitor their progress. Success with the planning and monitoring stage requires the application of strategic and meta-cognitive knowledge (Mayer, 1998). The final phase, problem execution requires the problem solver to apply procedural knowledge to calculate an answer. After the solution is executed, the problem solver must revisit the solution plan and the representation and assess the accuracy of the answer in the original context of the problem.

In sum, solving word problems is complex, because the multiple components comprising problem solving "do not necessarily follow a strictly linear model" (Depaepe, et al., 2010, p. 152). Depaepe et al. eloquently summarized problem solving as:

> understanding and defining the problem situation...constructing a mathematical model of the relevant elements, relations and conditions embedded in the situation; working through the model to derive some mathematical

> results...interpreting [those results]...in relation to the original problem situation; evaluating...if the ...mathematical outcome is appropriate and reasonable for its purpose; and, communicating the obtained solution of the original real-world problem (p.152).

The inherent complexity and coordination of multiple skills required for successful problem solving makes solving word problems challenging for many students, especially students with mathematics learning disabilities (MLD) and those with mathematics difficulties (MD).

Characteristics of Students with MLD

Students with MLD make up approximately 5-10% of the school-age population (Badian, 1983; Barbaresi, Katusic, Colligan, Weaver, & Jacobsen, 2005; Fuchs, Compton, Fuchs, Paulsen, Bryant, & Hamlett, 2005; Geary, 2004; Gross-Tsur, Manor, & Shalev, 1996; Lewis, Hitch, & Walker, 1994). Although these students' intellectual ability is within the average range, they exhibit cognitive and behavioral deficits that contribute to significantly lower achievement in mathematics than is expected (Johnson, Humphrey, Mellard, Woods, & Swanson, 2010). These deficits may include one or more of a variety of domain-general and domain-specific skills.

Domain-general skills include basic cognitive processes like working memory (e.g., Andersson & Lyxell, 2007; Geary et al., 2007), executive functioning, and attentive behavior (e.g., Andersson & Lyxell, 2007; Desoete et al., 2003; Geary et al., 2007; Hassinger-Das, Jordan, Glutting, Irwin, & Dyson, 2014). Many students with MLD exhibit deficits in the ability to store and retrieve arithmetic facts in both long term and

working memory, which negatively impacts problem solving. For example, when a student has to employ a counting strategy to compute an arithmetic fact instead of automatically retrieving it from memory, the problem solving process is disrupted. Many students with MLD also exhibit deficits in the area of executive functioning, which can negatively affect the ability to self-monitor and self-assess. Successful problem solvers tend to ask themselves questions and evaluate their performance as they systematically work through their solution plans. By contrast, students with MLD are often disorganized in their thinking and do not adequately monitor and regulate their problem solving performance (Desoete et al., 2003; Montague & Applegate, 1993).

In addition to possessing domain-general skill deficits, students with MLD are characterized by difficulties in one or more mathematics domains (e.g., number sense, arithmetic fluency, problem representation) (Fuchs, Fuchs, & Prentice, 2004; Jordan, Hanich, & Kaplan, 2003; Montague & Applegate, 1993). For example, these students may have difficulties with many aspects of basic number sense (e.g., reading numerals, judging magnitudes, understanding counting principles, number line concepts and estimation) (Hansen et al., 2015; Locuniak & Jordan, 2008; Mazzocco & Thompson, 2005). Such deficits in basic number sense may hinder their ability to make connections between knowledge about mathematical relations, underlying principles, and procedures (Gersten, Jordan, & Flojo, 2005). In addition, word problem solving is inherently influenced by decoding and comprehension skills, so students with deficits in these areas also tend to struggle with word problems (e.g., Hanich et al., 2001; Jordan & Hanich, 2000).

Research examining the relation between domain-general abilities (e.g., working memory, executive function), domain-specific mathematics skills (e.g., number sense, arithmetic fluency), and problem solving is somewhat unclear with regard to findings for students with MLD. While some studies suggest that the poor performance of students with MLD in basic mathematics skills can be explained primarily by working memory deficits (e.g., Geary et al., 2007), other research indicates that attention may be the primary contributing factor (Fuchs, et al., 2005; Hassinger-Das et al., 2014). At the same time, there is evidence that (a) number sense accounts for variance in mathematics outcomes beyond the contribution of basic cognitive abilities (Locuniak & Jordan, 2008; Mazzocco & Thompson, 2005) and (b) basic math skills mediate the impact of poor working memory on problem solving accuracy (Zheng et al., 2011).

Despite these mixed results, it is clear that the diminished problem solving performance of students with MLD cannot simply be explained by their deficits in working memory and associated lack of basic mathematics skills. Specifically, Andersson, (2008) reported that even after the effects of computation and arithmetic fact retrieval scores were partialled out, the differences between students with MLD and typically achieving students on word problem solving were statistically significant, favoring typically achieving students. In sum, students with MLD are an inherently heterogeneous group, varying in terms of basic cognitive processes, attitudes and behaviors, basic mathematics understanding, and problem solving strategies and skills.

Characteristics of Students with MD or Low Achievement (LA) in Mathematics

Although 5 to 10% of students have a learning disability in mathematics, many more students struggle to learn mathematics, and defining this larger population of students considered to be low achieving (LA) or to have mathematics difficulties (MD) is less clear than identifying students with MLD (Mazzocco, 2007). Unlike the term MLD, which includes a presumed biological cause (Mazzocco, 2007), MD or LA and MD refer to a broader category of students who exhibit below grade-level performance in mathematics. The factors that contribute to their low achievement may include: (a) inadequate instruction in previous years of schooling by teachers with limited pedagogical knowledge for mathematics (Sowder, Philipp, Armstrong, & Schappelle, 1998); (b) entering school with inadequate knowledge of number concepts and counting procedures (e.g., Griffin, Case, & Siegler, 1994; Price & Ansari, 2013; Vukovic & Siegel, 2010); (c) difficulties sustaining attention to academic tasks (Fuchs, et al., 2005; DiPerna, Lei, & Reid, 2007; Kolligian & Sternberg, 1987; Price & Ansari, 2013); and (d) issues associated with low motivation and maladaptive attribution style (Torgesen, 1994). Whatever the cause, research is unequivocal that poor scores on mathematics achievement tests at the beginning of formal schooling predict maladaptive behaviors (e.g., engaging in unrelated or distracting tasks, giving up easily), which subsequently predict continued poor mathematics achievement (DiPerna et al., 2007; Onatsu-Arvilommi & Nurmi, 2000). This persistent, iterative relation between mathematics performance and student behaviors may account for the low achievement of students considered to have MD.

Identifying students as MLD or MD. The inherent heterogeneity of the populations of students with MLD and MD has led to inconsistent operational definitions (Murphy et al., 2007; Price & Ansari, 2013). Across studies, inclusion criteria range from students identified as having LD based on state criteria (e.g., discrepancy between ability and achievement) to students identified by specific cut scores on various mathematics assessments, often ranging from the 10th percentile to the 35th percentile (Price & Ansari, 2013). Several studies have compared the performance of students grouped by varying criteria to assess the construct used to classify these students (e.g., Geary et al., 2012; Hanich et al., 2001; Jordan et al., 2002; Murphy et al., 2007).

Two studies (Geary et al., 2012; Murphy et al., 2007) compared the initial mathematics achievement and growth trajectories of three categories of students. Both Geary et al. and Murphy et al. operationally defined students scoring above the 25th percentile as typically achieving (TA), and both studies separated the remaining students into two groups: (a) those scoring at or below the 10th percentile on a mathematics achievement test and, (b) those scoring between the 11th-25th percentile on a mathematics achievement test; however, the authors labeled the two low-scoring groups differently. Namely, Geary et al. classified students scoring at or below the 10th percentile as having math learning disabilities (MLD), whereas students scoring between the 11th and 25th percentiles were considered to be low achieving (LA). In contrast, Murphy et al. (2007) defined both groups as having MLD, but differentiated in terms of the severity of MLD (i.e., MLD-10; MLD 11-25). The fact that the two studies classified students performing

at the same percentile into different categories underscores the lack of consensus in the field regarding operational definitions.

Regardless of the differences in criteria used to operationally define MLD or MD, students who struggle with mathematics problem solving are at risk for "long term difficulties in occupational and everyday activities that require basic mathematical knowledge" (Geary, et al., 2012, p. 206). Furthermore, research indicates the existence of a persistent achievement gap between students with disabilities and typically achieving students (Bandeira de Mello, Bohrnstedt, Blankenship, & Sherman, 2015; Fletcher & Vaughn, 2009; Fuchs & Fuchs, 2005). Therefore, it is of critical importance to identify and implement effective problem solving interventions for students with MLD and MD.

Review of Relevant Prior Meta-Analyses

In the following sections, I scrutinize the meta-analytic studies that examined the effectiveness of mathematics word problem solving interventions for students with LD or MD, detailing variations in coding schemes and methodological decisions (i.e., inclusion of various study designs, effect size calculation, and ways of handling potential outliers). I posit that these variations led to inconsistencies across findings, and propose that use of refined coding procedures and application of best practices regarding methodological decisions (see Harwell & Maeda, 2008; Cooper, Hedges, & Valentine, 2009) can reconcile discrepancies. Throughout this review of prior research, I will identify instances where the authors of previous meta-analyses have not reported sufficient information to allow the reader to assess the credibility of the reported findings (see Harwell & Maeda, 2008).

Three prior meta-analyses focused specifically on word problem solving interventions for students with LD and MD. Xin and Jitendra (1999) synthesized 14 group design word problem solving studies, Zhang and Xin (2012) synthesized 29 group design word problem solving studies, and Zheng et al. (2013) synthesized seven group design word problem solving studies. Kroesbergen and Van Luit (2003) and Gersten, Chard, et al. (2009) considered word problem solving as one of several mathematics topics covered in instructional interventions. Specifically, Kroesbergen and Van Luit (2003) included seven group design word problem-solving studies (21%), and the remaining 27 group design studies (79%) assessed basic mathematics skills interventions. Gersten, Chard, et al. (2009) included 13 group design studies of word problem solving interventions (31%). The remaining 29 studies (69%) were divided between operations, fractions, algebra, and general math proficiency interventions. Word problem solving intervention studies included in prior meta-analyses are listed in Appendix A.

Coding decisions. Across the five relevant prior meta-analyses, authors coded characteristics of included studies in different ways. In the following sections, I describe variations in coding of participant characteristics (i.e., grade level, LD/MD status), study design characteristics (i.e., fidelity of implementation, assignment to conditions, type of control group), outcome measure characteristics (i.e., psychometric properties associated with outcome measure, type of outcome measure), and contextual characteristics of intervention (i.e., duration, arrangement, setting, interventionist, instructional components).

Participant characteristics. Across the five relevant prior meta-analyses, definitions of the population of interest range from broad to narrow along at least two aspects of student characteristics (i.e., severity of disability, age/grade level). Four syntheses (Gersten, Chard, et al., 2009; Xin & Jitendra, 1999; Zhang & Xin, 2012; Zheng et al., 2013) examined mathematics intervention effectiveness for students in grades K-12, whereas one synthesis (Kroesbergen & Van Luit, 2003) narrowed their focus to students in grades K- 6. Findings regarding intervention effectiveness as a function of students' age or grade levels were inconsistent. For example, Gersten, Chard, et al. (2009) found smaller effects for older students as compared to effects for younger students. Specifically, they reported that for each grade level increase, the effect size decreased 0.07 standard deviations. In contrast, Kroesbergen and Van Luit (2003) found higher effects for older as opposed to younger elementary school aged students but the differences were not significant. It is important to note that the results reported in both Gersten, Chard, et al. (2009) and Kroesbergen and Van Luit (2003) were calculated for their entire samples, which included word problem solving and other interventions. Visual inspection of graphs of the effect sizes associated with word problem solving interventions only in each meta-analysis showed no obvious pattern by grade level (see Figures 1 and 2).

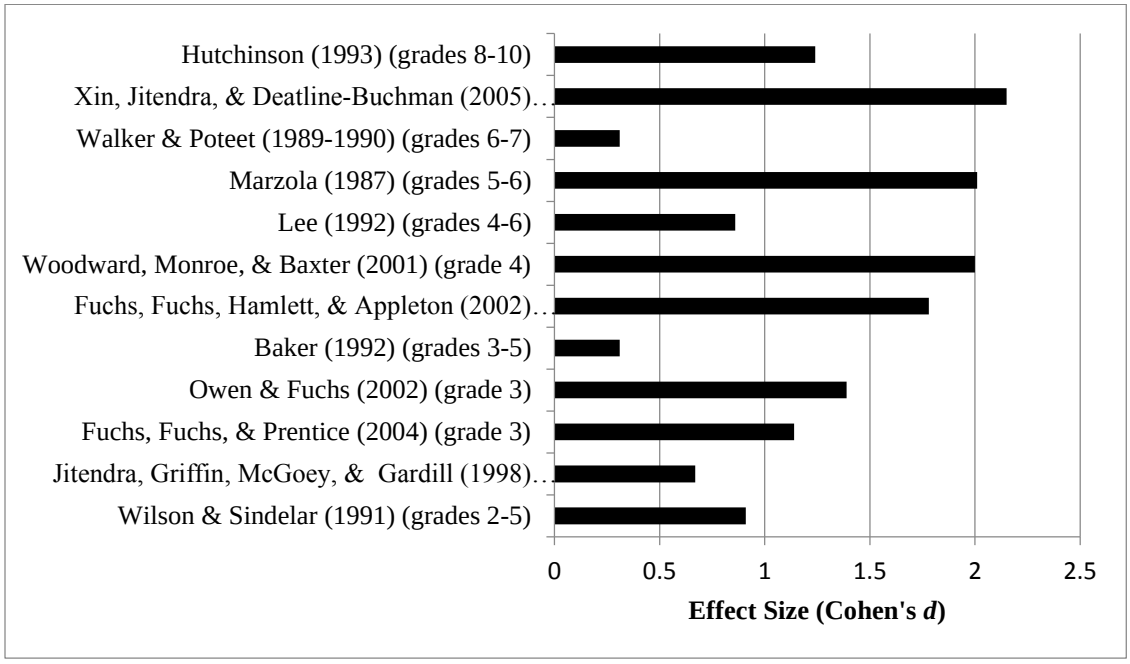


Figure 1. Effect sizes by grade level for word problem solving interventions reported in Gersten, Chard, et al. (2009).

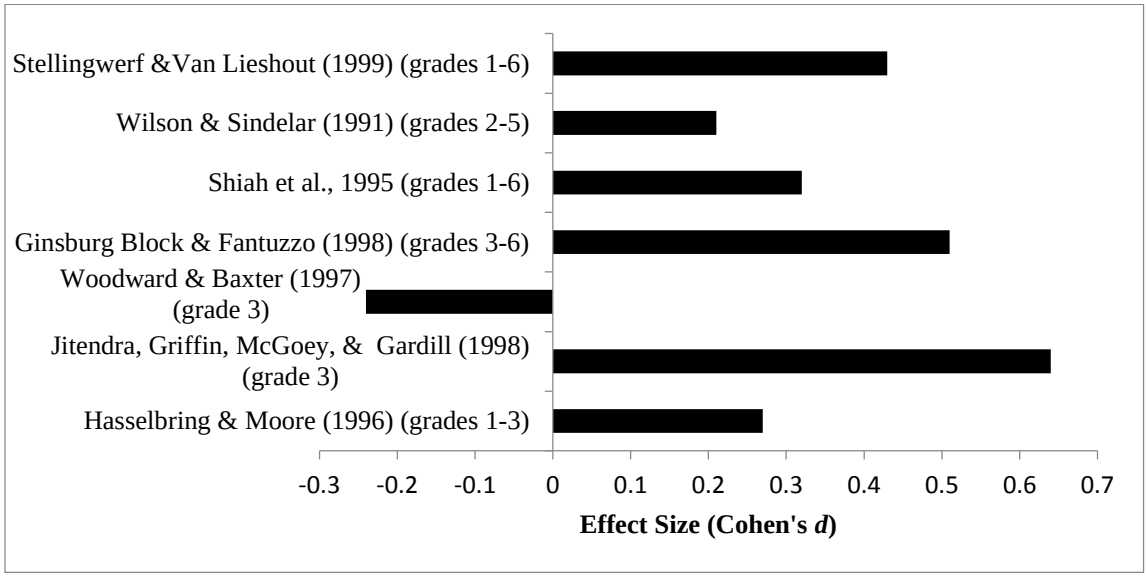


Figure 2. Effect sizes by grade level for word problem solving interventions reported in Kroesbergen and Van Luit (2003).

Xin and Jitendra (1999) found that effects at the secondary (grades 7-12) and elementary school (grades 1-6) levels did not differ significantly from each other. It is important to note that moderator analyses in meta-analysis are analogous to tests of interaction effects in primary studies; thus, they tend to be less powerful than tests for the overall average effect (Hedges & Pigott, 2004). By the same token, Xin and Jitendra's (1999) statistically non-significant finding for grade level does not "provide strong evidence for ruling out moderator effects" (Hedges & Pigott, 2004, p. 427). The current meta-analysis will provide additional evidence to aid in clarification of moderation of intervention effects as a function of student grade level.

Similar to the pattern of results for students at different grade levels, the findings of prior meta-analyses varied with regard to the performance of students with LD and students with MD (Kroesbergen & Van Luit, 2003; Xin & Jitendra, 1999; Zhang & Xin, 2012). For example, Xin and Jitendra (1999) reported that interventions had a significantly lower effect for students with LD than for either students with various disabilities (e.g., Attention Deficit Disorder, emotional/behavioral disability; mild cognitive impairment) or students with MD. Similarly, Kroesbergen and Van Luit (2003) reported that students with LD did not perform as well as those with mild mental retardation on problem solving tasks. Students with mixed disabilities (e.g., behavior and attention disorders) scored lower than those considered at-risk or low achieving. In contrast, Zhang and Xin (2012) reported no significant differences between students with LD and those with MD. However, because tests for moderator effects traditionally have

low power (Cooper et al., 2009; Hedges & Pigott, 2004; Mittlböck & Heinzl, 2006), this finding of no-significance must be interpreted cautiously.

One possible explanation for the discrepant findings with regard to sample characteristics across prior meta-analyses is due to widely varying definitions of LD and MD. For example, Kroesbergen and Van Luit (2003) defined students with special needs as any students who struggle more, perform more poorly, and/or require more specialized instruction than their peers. This includes students with "mild disabilities" (p.98) such as learning disabilities, mental retardation, and emotional/behavioral disorders. In contrast, Xin and Jitendra (1999) required that students score at or above 85 (i.e., one standard deviation below the mean) on a measure of IQ while still exhibiting a significant discrepancy between achievement and IQ score in order to be identified as having LD. This was done to "distinguish the population identified as learning disabled from other categories of disabilities" (Xin & Jitendra, 1999, p. 211). Zhang and Xin (2012) categorized students as having LD only if the study "explicitly define[d] students as having an LD due to discrepancy between IQ and low achievement scores" (Zhang & Xin, 2012, p. 311). When studies reported the presence of several students with LD and did not define LD using specific criteria, students in those studies were coded as having MD. As such, it is possible that the MD group in Zhang and Xin (2012) included students with LD.

Zheng et al. (2013) operationally defined math difficulties (MD) and reading difficulties (RD) using percentile cut off scores on standardized tests. They identified students who scored at or below the 25th percentile on a standardized math test, but above

the 25th on a standardized reading test as having MD only. Students scoring at or below the 25th percentile on standardized assessments of both reading and math were considered to have both MD and RD. Results indicated that, on average, interventions were effective at improving general word problem solving accuracy for students with MD, but not for students with MD and RD (Zheng et al., 2013). The present meta-analysis will provide additional evidence to aid in clarification of moderation of intervention effects as a function of students' LD/MD status.

Study design characteristics. Previous meta-analyses varied in regard to description of (a) fidelity of implementation (FOI), (b) nature of assignment of participants to conditions, and (c) type of control group. With regard to FOI, only three meta-analyses (Gersten, Chard, et al., 2009; Xin & Jitendra, 1999; Zheng et al., 2013) reported it and coded dichotomously (yes/no). In the present meta-analysis, each study was coded for FOI information, including percentage of essential instructional components that were met. In addition, unlike prior meta-analyses, I coded included studies for attrition.

Gersten, Chard et al. (2009), and Xin and Jitendra (1999) provided clear definitions for coding nature of assignment to conditions. Specifically, Xin and Jitendra (1999) stated:

> A procedure in which students were randomly assigned to treatment and comparison conditions was coded as random, a matching technique whereby students were matched on variables and assigned to the treatment

> and comparison conditions was coded as matched, and a study that used previously formed groups of students was coded as intact (p. 212).

Gersten, Chard, et al., (2009) explained:

> Quasi-experiments were included if students were pretested on a relevant mathematics measure and one of the following three conditions were met: (a) Researchers in the original study adjusted posttest performance using appropriate analysis of covariance (ANCOVA) techniques, (b) authors provided pretest data so that effect sizes could be calculated using the Wortman and Bryant (1985) procedure, or (c) if posttest scores could not be adjusted statistically for pretest performance differences, there was documentation showing that no significant differences (<.25 SD units) existed between groups at pretest on relevant measures of mathematics achievement (p. 1206).

On the other hand, Zheng et al. (2013) reported that the group design studies included in their meta-analysis were either randomized control trials (RCT) or quasi-experimental designs (QED). Kroesbergen and Van Luit (2003) categorized group design studies as: (a) randomly assigned, (b) randomly stratified, (c) restricted randomization, (d) randomized block design, or (e) matched groups.

Only one of the five prior meta-analyses conducted an in-depth examination of the instruction provided to the control group. In Gersten, Chard, et al. (2009), two raters coded all studies to "determine if the content covered was consistently relevant or minimally relevant to the purpose of the study" (p. 1207) and found no significant

difference between the two categories. By contrast, Kroesbergen and Van Luit (2003) noted only if the control condition received an intervention or not. The three remaining prior meta-analyses (i.e., Xin & Jitendra, 1999; Zhang & Xin 2012; Zheng et al., 2013) made no reference to the nature of instruction in the control group. I chose to code three categories for the nature of control variable in the present meta-analysis (see Method section for details).

Outcome measure characteristics. Previous meta-analyses varied in their coding of the psychometric properties associated with the primary outcome measures. Only one of the five relevant prior meta-analyses reported coding of the technical adequacy of the outcome measure (Gersten, Chard, et al., 2009). The remaining four meta-analyses did not report the reliability or validity of the outcome measures in the primary studies. In the present meta-analysis, each study was coded for information regarding reliability and validity of the outcome measure.

All five prior meta-analyses coded for the primary mathematics outcome measure as researcher developed or standardized/norm-referenced. However, only one of the meta-analyses (Gersten, Chard, et al., 2009) characterized the outcome measures according to the alignment between the intervention focus and the breadth of skills assessed by the measure. In the present meta-analysis, I coded all effect size comparisons as whether they were researcher developed or standardized, and examined the nature of the outcome measure, including its alignment with the intervention. Then, I selected the proximal, immediate posttest measures for inclusion (see Method section for effect size extraction details).

Contextual characteristics of interventions. The specific coding of contextual characteristics of interventions (i.e., duration, arrangement, setting, interventionist, instructional components) varied across prior meta-analyses, which may explain inconsistencies in findings across syntheses. Three of the relevant prior meta-analyses provided considerable detail on the duration of the intervention (i.e., Kroesbergen & Van Luit, 2003; Xin & Jitendra, 1999; Zheng et al., 2013). Specifically, Xin and Jitendra (1999) chose an arbitrary categorization of "short" (1-7 sessions total in one week), "intermediate" (8-30 sessions total in one month), and "long-term" (over 30 sessions, extended longer than one month). Results indicated that long-term interventions were more effective than short, but that short interventions were more effective than intermediate length interventions. Two meta-analyses coded the minutes per session and the number of sessions (Kroesbergen & Van Luit, 2003; Zheng et al., 2013). Kroesbergen and Van Luit found a negative correlation between instructional duration and effect size; Zheng et al. (2013) did not conduct moderator analyses due to the small number of studies. The remaining two studies (Gersten, Chard, et al., 2009; Zhang & Xin, 2012) did not code for duration of the intervention. The present meta-analysis recorded both minutes of instruction per session and the number of sessions, which were used to compute the total minutes of instruction.

Three of the five prior meta-analyses coded for instructional arrangement (Xin & Jitendra, 1999; Zhang & Xin, 2012; Zheng et al., 2013). However, differences in coding decisions and analysis techniques preclude direct comparisons of their findings. Both Xin and Jitendra and Zheng et al. coded instructional arrangement similarly (e.g., one-one-

one instruction, small group instruction). However, as described earlier, Zheng et al. (2013) did not conduct moderator analyses because of the small number of included studies. Xin and Jitendra (1999) reported that studies using a one-on-one instructional arrangement yielded a significantly higher mean effect size than studies using small group arrangement. No other prior meta-analyses coded for instructional arrangement.

With regard to instructional setting, Zhang and Xin (2012) coded for inclusive classrooms or special education classrooms. The latter encompassed pull-out settings as well as remedial classes, resource rooms, and self-contained classrooms. The authors reported that inclusive classes yielded a higher mean effect size than did special education classes. Xin and Jitendra (1999) grouped instructional setting somewhat differently than Zhang and Xin (2012), categorizing instructional setting into either instruction that took place in a special education classroom or instruction occurring outside of the classroom, which they labeled "pull-out." Xin and Jitendra (1999) reported inconsistent findings with and without the outliers. That is, there was no significant difference associated with variation in instructional setting when all effect sizes were included, however, the mean effect associated with special education classroom settings was significantly higher than for pull-out settings after trimming the outliers. Similar to Xin and Jitendra (1999), I coded for both instructional setting and arrangement in the present meta-analysis (see Method section for details).

The potential moderator variable of interventionist was coded in two different ways. Zhang and Xin (2012) and Kroesbergen and Van Luit (2003) compared computer-led instruction to instruction delivered by teachers, with findings indicating that the mean

effect size associated with teacher-led instruction was significantly higher than the mean effect size associated with computer-led instruction. Gersten, Chard, et al. (2009) and Xin and Jitendra (1999) coded as to whether teachers or researchers delivered the intervention. Gersten, Chard, et al. (2009) included a category for "other school personnel" as well (p. 1207); however they did not conduct moderator analyses on this variable. Xin and Jitendra (1999) included an additional category for "both teachers and researchers" (p. 216), which was associated with a significantly higher mean effect size than for either teachers or researchers. However, the criteria used to determine whether both teachers and researchers implemented the intervention were not explicitly reported. In the present meta-analysis, I provide operational definitions and illustrative examples to clarify coding decisions (see Method section).

Instructional components. Across prior meta-analyses, there is variability in definitions and coding of key instructional components. Despite differing operational definitions, there are some underlying similarities in the categories. Specifically, (a) explicit instruction, and (b) meta-cognitive strategies are key instructional components addressed in each meta-analysis, although the amount of overlap between categories varies by study. In addition, visual representation of the problem appears in four of the five meta-analyses (Gersten, Chard, et al. 2009; Xin & Jitendra, 1999; Zhang & Xin, 2012; Zheng et al., 2013). Researchers reported positive and significant effect sizes associated with each category described in the following section.

Kroesbergen and Van Luit (2003) defined direct instruction as explicit instruction sequenced to ensure mastery of each successive step, whereas Gersten, Chard et al.

(2009) separated explicit instruction (defined as a step-by-step plan for solving that is specific to a set of problems) and sequencing of examples into two different categories. Zheng et al. (2013) also coded sequencing separately from explicit instruction, and further categorized features of explicit instruction as skill modeling, explicit practice, fading probes or prompts, and elaborated explanation. Xin and Jitendra (1999) subsumed explicit instruction under a larger category referred to as strategy training (i.e., direct instruction, explicit instruction in cognitive/metacognitive strategies, self-regulation in using heuristics). Zhang and Xin (2012) also combined explicit instruction with additional components, although the specific components differed from those in Xin and Jitendra's (1999) meta-analyses. Specifically, Zhang and Xin (2012) conceptualized explicit instruction to include representation of the problem structure as a subcategory of the standards-based reform category.

Metacognitive and self-regulation strategies also varied across the meta-analyses. Kroesbergen and Van Luit (2003) referred to this category as self-instruction, which is defined as teaching students how to think-aloud while solving problems based on teacher modeling. Gersten, Chard, et al. (2009) refer to verbalization of mathematics reasoning, while Zheng et al. (2013) separately coded interventions for the presence of questioning and strategy cues (e.g., verbalization or thinking aloud). Zhang and Xin (2012) included the category cognitive strategy instruction and Xin and Jitendra's (1999) category of strategy training encompassed direct instruction as well as explicit instruction in cognitive/metacognitive strategies.

In four of five meta-analyses, representations were addressed as a separate instructional component category (Gersten, Chard et al., 2009) or part of a broader category (Xin & Jitendra, 1999; Zhang & Xin, 2012; Zheng et al., 2013). Gersten, Chard et al. (2009) viewed visual representation of the problem as a distinct instructional component. Representational approaches in Xin and Jitendra (1999) included pictorial (e.g., diagramming), concrete (e.g., manipulative materials), verbal (e.g., linguistic training) and mapping instruction (schema-based). Zhang and Xin (2012) combined explicit instruction involving representation of the problem structure as a subcategory within their standards-based reform category. Finally, Zheng et al. (2013) described the development of "pictorial representations, using specific material or computers" (p. 101) as part of the technology component. The current meta-analysis provides a clear rationale for the coding of instructional components that includes operational definitions and examples from coded studies.

Methodological decisions. In addition to variations in coding choices, previous meta-analyses varied in terms of methodological decisions. Specifically, authors' decisions regarding inclusion of different study designs varied broadly. Further, effect sizes were calculated in various ways, and decisions to combine effect size data and consideration of potential outliers in the data differed across meta-analyses.

Study designs and effect size calculation. Two meta-analyses (Xin & Jitendra, 1999; Zhang & Xin, 2012) included studies without control groups. In these studies, the effect size calculation was based on the standardized difference between the posttest and pretest means for a single sample. This practice tends to produce larger effect sizes than

those obtained from calculation of the standardized mean difference between treatment and control (Borman et al., 2003; Borman & D'Agostino, 1996). Further, threats to internal validity (e.g., history, maturation, regression-to-the-mean) are greater for the single group, pretest-posttest designs than for designs comparing treatment and control groups (Borman et al., 2003; Cook & Campbell, 1979). Four of the five of the prior meta-analyses included single subject design studies. Three of these calculated and reported effect sizes separately for single subject design and group design studies (Xin & Jitendra, 1999; Zhang & Xin, 2012; Zheng et al., 2013). Kroesbergen and Van Luit (2003) computed and combined the standardized mean difference (Cohen's *d*) for both single-subject and group design studies; a practice that produces greatly inflated effect sizes (Busse et al., 1995).

Zheng et al. (2013) calculated effect size using the differences within differences approach as outlined in *What Works Clearinghouse [WWC, 2011]*. This method involves finding the difference between the unadjusted pretest-posttest mean difference for the intervention group and the unadjusted pretest-posttest mean difference for the comparison group and dividing it by the weighted variance of the posttest scores. Although Zheng et al. noted that one of the limitations of this method, is the failure to account for covariance between pre-test and post-test, their use of this method is based on the fact that "no study reported the correlation between pretest and posttests measures or reported adjusted posttest means" (Zheng et al., 2013, p. 101). However, two additional limitations to this method of effect size calculation were not considered. First, Zheng et al. (2013) did not report whether the assumption was met that the pretest and the posttest are the same test.

If they are not the same test, calculating pre-post-test difference for each group is not necessarily appropriate. Second, they do not address the likelihood of over- or under-estimating "the adjusted group mean difference, depending on which group performed better on the pretest (WWC, 2011, p.12)." The present meta-analysis synthesized studies comparing a treatment to an equivalent control group. Effect size was calculated as standardized mean difference corrected for sample size using the Hedges' *g* correction, which aligns with the effect size calculations performed by Gersten, Chard, et al. (2009).

Combining effect size data. Three of the relevant prior meta-analyses (Gersten, Chard, et al., 2009; Kroesbergen & Van Luit, 2003; Zheng et al., 2013) calculated multiple effect sizes per study. This is a concern because when multiple measures are used within a single sample or when multiple subsamples within a single study are used to produce effect sizes, the assumption of independence is violated, and the dependency among effect sizes must be accounted for in the meta-analysis (Cooper et al., 2009; Harwell & Maeda, 2008).

Gersten, Chard, et al. (2009) provide a rationale and explicitly describe their effect size calculations as well as their procedures for averaging across multiple effect sizes calculated for a single study. Though they note the limitations of simply averaging across effect sizes, they reported a finding of no systematic biasing based on the results of a sensitivity analysis "regressing effect size onto the number of groups, time points, and measures aggregated in a fixed-weighted analysis" (p.1214). In addition, Gersten, Chard, et al. (2009) state that, as a result of applying their systematic selection and estimation procedures, only independent effect sizes were used in their meta-analysis.

In contrast, Kroesbergen and Van Luit (2003) and Zheng et al. (2013) do not "provide clear evidence of the independence of the effect sizes being analyzed" (Harwell & Maeda, 2008, p. 422). Kroesbergen and Van Luit (2003) simply report that "when more than one test or subtest was used to measure mathematics performance, we calculated the effect sizes for all tests and then used the mean effect size in the meta-analysis" (p.100). They explain that they made this decision to avoid having multiple outcome measures for some studies resulting in unequal weightings, but they do not provide evidence that the effect sizes involved in the meta-analyses are independent. Similarly, Zheng et al. (2013) calculated multiple effect sizes per study and used them all to conduct inferential analyses without providing evidence that the assumption of independence was met. In the present meta-analysis, I provide evidence as to "the extent to which the statistical analyses ... appeared to satisfy the assumption of independence" (Harwell & Maeda, 2008, p. 423).

Potential outliers. Harwell and Maeda (2008) recommend identifying potential outliers and "performing key analyses after temporarily excluding suspect studies and examining the similarity of findings with and without these studies" (p. 424). Xin and Jitendra (1999) performed such sensitivity analyses, reporting and comparing the trimmed and untrimmed data sets. Zheng et al. (2013) reported removing outliers, which they defined as "effect sizes lying beyond the first gap of at least one standard deviation between adjacent effect size values in a positive direction" (p.101). They did not report results of sensitivity analyses comparing findings with and without outliers. In the remaining three meta-analyses, potential outliers, extreme values, or influential points

were not considered and is disconcerting when extremely high effect sizes are included. For example, Zhang and Xin (2012) note that the highest effect size, $d = 11.82$, came from Fuchs, Fuchs, Prentice, Hamlett, et al. (2004). In the present meta-analysis, I examined the distribution of effect sizes, identified potential outliers, or influential points, and conducted all key analyses with and without influential points.

Chapter 3

METHOD

In this chapter, I first describe the population of studies to which the results of this meta-analysis would generalize. Second, I detail the search and screening procedures, and the procedure for extracting independent effect sizes. Third, I describe the procedure for coding the studies selected for review, and assessment of inter-rater agreement. Finally, I describe the data analysis procedures used in this meta-analysis.

Populations

Gersten et al. (2005) suggested that researchers "provide enough information about participants so that readers can identify the population of participants to which results may be generalized" (p. 155). As such, meta-analysts must provide sufficient information about the sample of studies. The sample of studies included in the present meta-analysis is presumed to be a random sample from a hypothetical "universe of possible studies—studies that realistically could have been conducted or might be conducted in the future" (Cooper et al., 2009, p. 297). As such, the results should generalize reasonably well to the population of studies that serve to answer the following question: what is the expected effect of word problem-solving interventions on the mathematical performance of students in grades K-12 with LD or MD? Although geographical limitations were not part of the search criteria, only one of the 28 included studies was conducted outside of the United States (Hutchinson, 1993; conducted in Vancouver, Canada). Therefore, these results may not generalize well to studies conducted in countries other than the United States and Canada. Further, it would be

erroneous to infer that "study-level variables found to be associated with effect sizes are also descriptive of relationships at the level of individual participants" (Cooper & Patall, 2009, p.170); therefore, inferences about the behavior of individual participants within each study in the sample should not be inferred from aggregate data alone.

A random effects model is more appropriate for the current meta-analysis than a fixed effects model. Within a random effects model, "studies under synthesis can be viewed as representative of a larger population or universe of implementations of a treatment" (Cooper et al., 2009, p. 306). The primary assumption of the fixed effects model is that "one true effect size underlies all the studies in the analysis, and that all differences in observed effects are due to sampling error" (Borenstein, Hedges, Higgins, & Rothstein, 2010, p.97). This assumption seems implausible for studies in the social sciences in general and in education in particular, as they tend to vary along multiple dimensions (Borman et al., 2003; Borenstein, et al., 2010; Cooper et al, 2009). Given that the included studies represent different samples, methodological features, and study characteristics, we could expect a distribution of effect sizes that is due to more than sampling error alone. As such, the total variance of the distribution of effect sizes consists of both within- and between-studies error variance:

$$v_i* = v_i + \tau^2$$

where $v_{i\,is}$ the within-studies error variance, which is unique to each study, and τ^2 is the between-studies variance, which is common to all studies (Borenstein et al., 2010).

Search and Screening Procedures

The current study aims to synthesize data obtained from available studies (both published and unpublished) through 2014 on word problem-solving interventions for students in grades K-12 with LD or MD. The Education Source (previously Education Full Text), ERIC, PsycINFO, and Digital Dissertations online databases were searched using a combination of the following search terms: *math(ematics) and word problem(s) or problem solving, and* instruction *or instructional or intervention or teach(ing).* The resulting subset of articles was searched using the following terms: learning disability/disabilities, special education, at risk, learning difficulties, low performing, low achieving. In addition, the references of published meta-analytic studies (i.e., Gersten, Chard, et al., 2009; Kroesbergen & Van Luit, 2005; Xin & Jitendra, 1999; Zhang & Xin, 2012; Zheng et al., 2013) and the articles reviewed therein were examined. Further, a hand search was conducted of major journals in special and elementary education (i.e., *Elementary School Journal, Exceptional Children, Journal of Educational Psychology, Journal of Learning Disabilities, Journal of Special Education, Learning Disabilities Quarterly, Remedial & Special Education, and Learning Disabilities Research & Practice*). Citations and abstracts identified using these search processes were imported into Refworks where duplicates were eliminated.

Title and abstract screening. The titles and abstracts retrieved using the process detailed above were examined for further review using the following criteria:

1. The study assessed the effectiveness of a mathematics word problem-solving intervention. Studies employing word problem-solving measures, but that did

not include interventions focusing specifically on word problem solving were excluded.

2. The study had a pretest-posttest control group design that included (a) a randomized controlled trial or (b) a quasi-experimental design (i.e., students not randomly assigned to groups; intact classrooms). Studies that reported only pre-post gains within a single group were excluded. Single case design (SCD) studies were excluded "because there is no known statistical procedure for valid combination of single-subject and group design studies" (Gersten, Chard et al., 2009, p. 1204).
3. The study focused on school-aged students (grades K-12) with learning disabilities (LD) or mathematics difficulties (MD).
4. The study was written in, or translated into, English.

Throughout this process, the number of articles excluded, detailed information on these articles, and the reason for exclusion were recorded.

Full-text screening. The 140 studies remaining after the title and abstract screening were retrieved for in-depth examination. In addition to the criteria used in the title and abstract screening procedure, the studies were assessed according to the following criteria:

1. Studies reported data needed to calculate effect size.
2. If random assignment at the student level was not used, then equivalence on key measures at pretest was established.

3. Treatment condition was compared to an equivalent control group. Studies comparing the treatment condition to a sample of typically achieving students (i.e., non-equivalent control group) were excluded.
4. The control group provided either: (a) no intervention (attention only), (b) business-as-usual instruction, or (c) an alternate intervention. Studies examining the relative effectiveness of different instructional components were excluded if the alternate intervention was identical to the treatment with the exception of one dimension of the intervention.
5. Studies examining a wider range of ages (beyond K-12) without disaggregating the results by age were excluded.
6. The study focused on students with learning disabilities (LD) or mathematics difficulties (MD). Studies were excluded if less than 50% of participants had LD or MD, and data for the subgroup of students with LD or MD were not disaggregated. We relied on the authors' documentation of LD and recorded the reported criteria used for this determination (e.g., state criteria, screening test). Documentation of MD had to include a cut score at or below the 35th percentile on a mathematics measure.

Examples of exclusions. A total of 28 group design studies met the above criteria for inclusion in the present meta-analysis; 21 (75%) were previously included in one or more prior meta-analyses. There were 65 studies included in prior meta-analyses, which were excluded from the present meta-analysis during the full text screening. Of these, 41 studies (63%) did not include a WPS intervention. The remaining 24 studies (37%) were

excluded from the present meta-analysis on the basis of issues with the control group or sample characteristics (i.e., age, LD/MD status) (See Table 1 for details).

Table 1
Studies from prior meta-analyses excluded from present meta-analysis

Reason for Exclusion	**Study**	**Prior meta-analyses**
Sample not K-12	Noll (1983)	Xin & Jitendra (1999)
	Toppel (1996)	Zhang & Xin (2012)
	Zawaiza & Gerber (1993)	Xin & Jitendra (1999)
No control group/inappropriate control group	Bennett (1981)	Xin & Jitendra (1999); Gersten, Chard, et al. (2009)
	Bottge Heinrichs, Chan, & Serlin (2001)	Zhang & Xin (2012)
	Bottge, Heinrichs, Mehta, and Hung (2002)	Gersten, Chard, et al. (2009); Zhang & Xin (2012)
	Bottge, Rueda, LaRoque, Serlin, & Kwon (2007)	Zhang & Xin (2012)
	Bottge, Rueda, Serlin, Hung, & Kwon (2007)	Zhang & Xin (2012)
	Gleason, Carnine, & Boriero (1990)	Xin & Jitendra (1999)
	Jitendra, Griffin, Deatline-Buchman, & Sczeniak (2007)	Zhang & Xin (2012)
	Lang (2001)	Zhang & Xin (2012)
	Montague, Applegate, & Marquard (1993)	Xin & Jitendra (1999)
	Troff (2004)	Zhang & Xin (2012)

Table 1 (continued)

Reason for Exclusion	Study	Prior meta-analyses
Data were not disaggregated for LD or MD sample/MD determination not based on cut score at or below the 35th percentile	Bottge (1999)	Zhang & Xin (2012)
	Bottge, Heinrichs, Chan, Mehta, & Watson (2003)	Zhang & Xin (2012)
	Fuchs, Fuchs, Prentice, Burch, Hamlett, Owen, Hosp, Janacek (2003)	Zhang & Xin (2012)
	Fuchs, Fuchs, Prentice, Burch, Hamlett, Owen, Schroeter (2003)	Zhang & Xin (2012)
	Ginsburg-Block & Fantuzzo (1998)	Kroesbergen & Van Luit (2003); Zhang & Xin (2012)
	Griffin & Jitendra (2009)	Zhang & Xin (2012)
	Hasselbring & Moore (1996)	Kroesbergen & Van Luit (2003); Zhang & Xin (2012)
	Jitendra et al. (2009)	Zhang & Xin (2012)
	Jitendra, Griffin, Haria, Leh, Adams, & Kaduvettoor, (2007)	Zhang & Xin (2012)
	Moore & Carnine (1989)	Xin & Jitendra (1999)
	Stellingwerf & Van Lieshout (1999)	Xin & Jitendra (1999); Kroesbergen & Van Luit (2003)

Issues with control group. Four studies (Bottge, Rueda, LaRoque, Serlin, & Kwon 2007; Bottge, Rueda, Serlin, Hung, Kwon, 2007; Jitendra, Griffin, Deatline-Buchman, & Sczeniak, 2007; Troff, 2004) did not include a control group. The effect sizes reported from these studies illustrated pre- to post-test improvement within a single sample. Research suggests that analyses of pre-post gains within a single group tend to produce significant positive biases in results as compared to designs employing matched control-group designs (Borman et al., 2003; Borman & D'Agostino, 1996). Four studies (Bennett, 1981; Bottge, Heinrichs, Chan, & Serlin, 2001; Bottge, Heinrichs, Mehta, & Hung, 2002; Lang, 2001) included a non-equivalent control group (i.e., the comparison group was composed of typically achieving students). One study (Gleason, Carnine, & Boriero 1990) used a control condition that did not assess the effectiveness of word problem solving but examined the effects of medium of instruction (computer- vs. teacher-delivered). Specifically, Gleason et al. (1990) compared the same curriculum for both conditions; with the only difference being that in one condition instruction was provided by a teacher, and in the other, a computer instructed students. In another study (Montague, Applegate, & Marquard, 1993) the comparison conditions were identical to the treatment except for one aspect. Montague et al. (1993) focused on "the issue of separability of cognitive and metacognitive components of instruction" (p. 223) by comparing three conditions: metacognitive strategy instruction (MSI), cognitive strategy instruction (CSI), and MSI + CSI.

Sample characteristics. Among studies that included less than 50% students with LD or MD, two studies (Griffin & Jitendra, 2009; Hasselbring & Moore, 1996) were

excluded because they did not disaggregate the data for students with LD. Nine studies (Bottge, 1999; Bottge, Heinrichs, Chan, Mehta, & Watson, 2003; Fuchs, Fuchs, Prentice, Burch, Hamlett, Owen, Schroeter, 2003; Fuchs, Fuchs, Prentice, Burch, Hamlett, Owen, Hosp, Janacek, 2003; Ginsburg-Block & Fantuzzo, 1998; Jitendra, Griffin, et al., 2007; Jitendra, et al., 2009; Moore & Carnine 1989; Stellingwerf & Van Lieshout, 1999) were not included because the determination of MD status was based on either a cut score between the 36th and 50th percentile or teacher referral for supplemental instruction, and/or history of low achievement. Furthermore, three studies (5%) (Noll, 1983; Toppel, 1996; Zawaiza & Gerber, 1993) were excluded because their sample comprised adults with LD or MD in post-secondary education.

Effect Size Extraction

Only independent effect sizes were included in this meta-analysis because dependence among effect sizes can bias statistical inferences (Cooper et al., 2009). When studies used multiple outcome measures, the primary, proximal measure of mathematics WPS performance aligned with the intervention, often referred to as the immediate posttest, was coded as the independent effect size. Comparisons of the performance of the same student sample on additional measures (e.g., transfer measures, maintenance measures) were recorded as dependent effect sizes. When authors reported immediate posttest results adjusted for pretest performance in addition to unadjusted posttest performance, the adjusted posttest means and standard deviations were extracted for effect size calculation.

In the case of multiple treatment studies, each of which was compared to the same control group, I coded the comparison containing the simplest version of the treatment of interest as the independent effect size. For example, Fuchs, Fuchs, Finelli, et al. (2004) compared two different treatment groups to a single control. One of the treatments was called Schema-Based Transfer Instruction (SBTI). The other treatment group, Expanded SBTI, contained additional treatment features. I selected the SBTI treatment without the additional features as the group to compare with the control to obtain an independent effect size. Similarly, Fuchs, Fuchs, Prentice, Hamlett, et al. (2004) compared SBTI to control and SBTI with sorting to control. I used the comparison containing the simpler treatment, SBTI, to calculate the independent effect size. This criteria was used to select the comparison for three additional studies. Owen & Fuchs, (2002) compared the same control group to (a) "acquisition", (b) "low dose acquisition + transfer", and (c) "full dose acquisition + transfer". Powell & Fuchs (2010) compared both "word-problem tutoring" and "word-problem tutoring plus equal-sign instruction" to the same BAU control. Shiah et al. (1994) compared two variants of a computer-assisted intervention involving an explicit problem solving strategy to the same control (i.e., "explicit problem solving strategy + animated pictures" vs. "no strategy + static pictures"; "explicit problem solving strategy + static pictures" vs. "no strategy + static pictures"). In each case, I chose the simplest treatment ("acquisition"; "word-problem tutoring"; and "explicit problem solving strategy + static pictures", respectively) for the independent comparison.

In the multiple treatment studies described above, the simplest treatment seemed to best capture the intervention. In the following multiple treatment studies, the

intervention consisted of two major components, and the simpler variants of the treatment served as a type of control. For example, Swanson, Lussier, and Orosco (2013) tested a treatment composed of general heuristic instruction (GHI) and visual schematic instruction. To assess the differential contribution of each of the two major components, they also provided one group with GHI only and another with visual schematic instruction only. I reasoned that selecting one of the two simpler components was not an accurate representation of the intended intervention, so I chose the dual-component treatment (i.e., GHI + visual schematic) for the independent comparison. Similarly, Moran, Swanson, Gerber, and Fung (2014) investigated the effectiveness of a paraphrasing intervention, and included two additional groups receiving only one subset of the paraphrasing strategies, so I coded the comparison between the complete treatment and the control as the independent effect size. Both Swanson et al. (2013) and Moran et al. (2014) compared their treatments to a BAU control group, specifying that their interventions were a supplement to BAU instruction. Wilson and Sindelar (1991) did not include a BAU control, but rather compared increasingly complex variants of the intervention to each other (i.e., "sequence only", "strategy only", and "sequence +strategy"). In this case, I reasoned that the "sequence only" group was similar to BAU in that students worked on one type of problem each class session, but did not receive specific strategy instruction. I compared the "sequence only" control group with the "sequence + strategy" treatment in keeping with the BAU + supplement conditions identified in Moran et al. (2014) and Swanson et al. (2013).

When multiple-treatment studies provided disaggregated data for mutually exclusive sub-samples (i.e., no subjects shared across sub-samples) then the effect sizes were coded as independent. Although "any feature shared in common by subsamples within a study—for example, being studied by the same investigator—can introduce statistical dependences into the effect sizes" (Cooper et al., 2009, p. 149), the primary threat to statistical independence is averted by comparing mutually exclusive sub-samples. For example, Fuchs, Fuchs, Craddock, Hollenbeck, Hamlett, and Schatschneider (2008), and Woodward et al. (2001) both compared WPS instruction with and without supplementary tutoring to control group instruction with and without supplementary tutoring. The two by two comparisons provided a total of four contrasts (i.e., treatment vs. control; treatment + supplement vs. control + supplement; treatment + supplement vs. control; and treatment vs. control + supplement). The results of the two mutually exclusive comparisons (i.e., treatment vs. control and treatment + supplement vs. control + supplement) were coded as two independent effect sizes.

In addition, when single treatment studies reported disaggregated results for separate subgroups of students receiving the same intervention, then the results of each mutually exclusive subsample comparison were coded as independent. For example, Fuchs, Fuchs, and Prentice (2004) provided whole-class instruction to a sample consisting of students identified as having MD (i.e., < 25th percentile in computation and > 40th percentile in reading comprehension) and students identified as having both math and reading difficulties (MDRD) (i.e., < 25th percentile in computation and reading

comprehension). Results were reported separately for the MD subsample and the MDRD subsample, and they were coded as two independent effect sizes.

Coding of Studies

Once determinations were made regarding independent effect size data extraction from each of the included studies, coding began. The coding process was iterative and involved reading the articles, coding them, reviewing the codes with an expert in the field, revising the coding scheme, and re-coding the articles until the grouping categories seemed to best capture the data at hand. Each included article was coded for the following: (a) participant characteristics (e.g., LD/MD status, grade level), (b) study design characteristics (e.g., random assignment, type comparison group), (c) outcome measure characteristics (e.g., reliability) and (d) contextual characteristics of interventions (e.g., instructional arrangement, interventionist). The final coding scheme used is presented in Appendix B.

Participant characteristics. Student demographic information (e.g., age/grade level, sex, ethnicity, socio-economic status, LD/MD status) was identified and coded from each article. I recorded the grade level of the sample for each included article, then categorized them into two levels: elementary (grades K-6), and secondary (grades 7-12). With regard to ethnicity, I first coded whether or not the authors reported the ethnicity of the sample, then I recorded the percentage of White students and the percentage of Minority students. Minority was defined as African American, Asian, Hispanic, or categories other than strictly Caucasian (i.e., biracial). Although there are multiple possible measures of the multifaceted construct of socio-economic status (SES) (Dalton,

2011; Harwell & LeBeau 2010), eligibility for Free or Reduced Price Lunch (FRL) was the only estimate of SES reported in a subsample of the reviewed studies. This variable was coded in two phases: first, I coded whether or not the authors reported eligibility for FRL of the sample, then I recorded the percentage of students eligible for FRL.

In terms of LD/MD status, studies were coded as "LD" if 50% or more of the sample was identified as having a learning disability and documented this using state criteria (e.g., IQ/achievement discrepancy). Studies were coded as "MD" when less than half the sample had a documented learning disability, but more than half the sample met the cut score criteria of 35^{th} percentile or below. First, I identified and coded the number of student participants in each study identified as having a learning disability (LD), and the number of students identified as having MD. Second, I identified and recorded the type of criteria used to establish LD/MD status (i.e., state LD criteria, specific cut score on a standardized or researcher developed screening measure).

Study design characteristics. Each study was coded on five study design characteristics: (a) type of report (publication status), (b) assignment to groups, (c) fidelity of implementation (FOI), (d) attrition rate, and (e) type of comparison group. Studies were coded as either "published" or "not published" (e.g., dissertations). Assignment to groups was dichotomously coded as either random assignment at the student level, or random assignment at the classroom or school level, with group equivalence established at pretest. Fidelity of implementation (FOI) was coded in two phases. First I recorded whether or not the authors reported FOI. When studies reported FOI, I then recorded the percentage of essential instructional components addressed.

Assessment of attrition is important to the internal validity of a study. When information on attrition is not provided, one cannot be sure that establishment of equivalency between groups at the outset has been maintained through the conclusion of the study (Gersten et al., 2005). In the present meta-analysis, attrition data were coded in two stages. The first was a dichotomous code of "reported" or "not reported," and the second was the specific attrition rate. In the sample of included studies, attrition was rarely explicitly reported. In fact, the word "attrition" only appeared in two of the included studies. One study (Jitendra et al., 2013) provided a detailed attrition analysis section. The other (Fuchs et al., 2008) provided a brief report noting comparability between students who did and those who did not complete the study and a lack of "significant interactions between AR students' tutoring condition and attrition status" (p. 496). To assess attrition rate in the remaining studies, I searched for data facilitating calculation of attrition. For example, if a study reported that two students moved to a different school before the end of the study, then I divided two by the initial sample size to calculate attrition rate. In studies lacking a discussion of the number of students who did not complete the study, I identified the initial sample size, which was often reported as part of the sample demographic data. Then I identified the final sample size associated with results of the intervention and compared the two sample sizes using the formula below:

$$\text{Attrition rate} = \frac{n_1 - n_2}{n_1}$$

Where n_1 is the initial sample size and n_2 is the final sample size.

Studies were classified with respect to the type of comparison group and coded as (a) "business as usual" (regular classroom instruction), (b) alternate intervention or (c) no intervention/attention only. Business as usual control was defined as regular classroom instruction based on the textbook and supplemental materials used by the school, with little, if any, researcher adaptation of these materials. The type of comparison group was considered to be an "alternate intervention" if: (a) the control group received an intervention that did not come from a standard textbook and supplemental materials, but rather was developed by the researcher; (b) the control group received BAU supplemented in part by researcher developed instruction, or (c) the intervention was based on the classroom text and materials, but the researcher adapted them to provide a more relevant control than would be provided by business-as-usual classroom instruction.

Six studies met the first condition of providing the control group with alternative, researcher-developed instruction. Bottge and Hasselbring (1993) reported that the comparison group intervention was specifically designed to parallel the treatment in all but one key way: embedding the problems within a realistic, connected context, which was the primary independent variable. Fuchs, Fuchs, et al. (2008) compared a treatment group receiving the researcher-developed schema broadening classroom intervention plus supplemental schema broadening tutoring to a comparison group that received only the schema broadening classroom intervention. Shiah et al. (1994) developed two computerized interventions that both taught students to solve the same word problems. The control group received a simplified version of the same program without certain key features (e.g., explicit instruction, general problem solving strategy, animated pictures).

Similarly, Baker (1992), Konold (2004) and Wilson and Sindelar (1991) were coded as having an alternate treatment control group because students in the control group received a simplified version of the treatment intervention.

One study was coded as having an alternate control group because it met the second condition: the control group received BAU supplemented with researcher-developed instruction. Specifically, Fuchs, Fuchs, Prentice, Hamlett et al. (2004) employed a combination of "teacher designed and implemented instruction … as well as a three-week researcher-designed and implemented unit on general problem-solving strategies" (Fuchs, Fuchs, Prentice, Hamlett et al., 2004, p. 637). Three studies (Jitendra et al., 2013; Xin et al., 2005, 2011) were coded as alternate control group because they met the third condition: the intervention was based on the classroom text and materials, but was adapted by the researcher to provide a more relevant control than would be provided by business-as-usual classroom instruction. In addition to differentiating between "BAU" and "alternate intervention," a third type of control group was coded as "no instruction/attention only." One study, (Marzola, 1985) reported that control students received the same problems as the treatment group, but did not receive any specific instruction. Specifically, after the students completed the problems, the researcher marked them as correct or incorrect, but gave no guidance on how to fix the incorrectly answered problems.

Outcome measure characteristics. The nature of the primary outcome measure was coded as standardized norm-referenced or researcher developed. The reliability of the primary outcome measure was initially coded as inadequate ($\alpha < 0.60$) (see Gersten,

et al., 2005), adequate ($0.60 \leq \alpha < .80$), high ($\alpha \geq 0.80$) (see Thorndike & Thorndike-Christ, 2010), or not reported. However, only one study (Powell & Fuchs, 2010) reported inadequate reliability ($\alpha = 0.54$) for the outcome measure, and three studies (Jitendra et al., 2013; Lambert, 1996; Moran et al., 2014) reported adequate reliability ($\alpha = 0.75$, $\alpha = 0.78$, and $\alpha = 0.80$, respectively); therefore, I selected $\alpha \leq 0.86$ vs. $\alpha > 0.86$ for categorizing reported reliability estimates. Validity of the primary outcome measure was coded first as "reported" or "not reported." Then, I descriptively noted the type of validity evidence provided.

Contextual characteristics of intervention. The sample of studies included in the present meta-analysis varied across six contextual characteristics of intervention (i.e., instructional setting, instructional arrangement, duration of intervention, interventionist, mathematics task, instructional components). The instructional setting in each study was described as (a) general education classroom, (b) special education classroom, (c) other (e.g., library, computer lab, hallway), or (d) not reported. The primary instructional arrangement in each study was categorized as: whole class, small group (generally 2-7 students), one-on-one, or not reported.

The duration of the intervention was calculated as total hours of instruction based on data reported by the authors. When authors did not report total instructional time, I reread the article to locate additional information needed to calculate the total time. For example, Fuchs et al. (2009) reported the following: "each lesson lasts 20–30 min ... standard protocol runs 16 weeks, with three sessions per week. These 48 lessons are divided into four units" (p.567). Given this data, I calculated average instructional

duration by multiplying the average time per lesson (25 min) by the number of lessons (48), for a total of 1200 min or 20 hours. Ultimately, hours of instruction was assigned three levels: less than 10 hours, between 10 and 16 hours, and greater than 16 hours of instruction.

The interventionist category was coded as follows: (1) classroom teacher, (2) researcher, (3) teacher and researcher, (4) other (e.g., volunteers from the community, computer), or (5) not reported (NR). I coded for the primary interventionist; for example, if the researcher implemented the intervention and the classroom teacher was present to help with classroom management, I coded the interventionist as "researcher." However, in cases where implementation of treatment was divided relatively equally, I coded the study as being implemented by both teachers and researchers. For example, in three studies (Bottge & Hasselbring, 1993; Xin et al., 2005; Xin et al., 2011) there were multiple interventionists—half of them researchers, half of them classroom teachers, counterbalanced across conditions to control for instructor effects. One study (Fuchs, Fuchs, & Prentice, 2004) described a uniquely collaborative design with regard to implementation. Specifically, within a six-lesson unit, the first of two lesson types (i.e., "problem solution", and "transfer") was taught by the researcher with the classroom teacher present to observe and assist. Then the classroom teacher, usually with a research assistant present, taught the remaining lessons.

The "mathematics task" category was coded as: (1) arithmetic word problems involving only addition and/or subtraction (but not multiplication and/or division), (2) arithmetic word problems involving multiplication/division or all four operations, and (3)

higher level mathematics including fractions, ratio, proportion, geometric and/or algebraic word problems. I coded selective instructional components of mathematics interventions found to be effective for struggling learners (see Gersten, Chard et al., 2009; Jitendra & Xin, 1997; Kroesbergen & Van Luit, 2003; Xin & Jitendra, 1999). Table 2 provides (a) a summary of the instructional components, (b) specific wording found in studies that indicated the presence of these components, and (c) other meta-analyses that also addressed the component. It is worth noting that the instructional component categories are not mutually exclusive; rather, there is extensive overlap.

Table 2
Instructional components

Category Name	**Description**	**Indicators***	**Noted in prior meta-analyses**
Representations	Visual & verbal representations	Visual – Manipulative materials (i.e., concrete objects such as base-10 blocks, unifix cubes); schematic diagrams that illustrate the relationships described in the problem; pictorial diagrams that focus on surface features (images depicting the visual appearance of objects or persons described in the problem), a combination of different types of representations (e.g., concrete-representational-abstract); mnemonic illustrations. Verbal – paraphrasing; mental imagery.	Gersten, Chard, et al. (2009) Xin & Jitendra (1999) Zhang & Xin (2012) Zheng, et al. (2013)
Metacognition	Thinking aloud, self-questioning, self-regulation, self-monitoring	Students use think-alouds or ask themselves questions to regulate and monitor the problem solving process, which generally involves identifying the goal (e.g., restating the problem in own words to identify question), monitoring strategy use, and/or evaluating the outcome of a word problem.	Gersten, Chard, et al. (2009) Xin & Jitendra (1999) Zhang & Xin (2012) Zheng, et al. (2013)
Prerequisite/ Foundational Skills Instruction	Knowledge that is foundational to new content taught	Prior knowledge needed to learn the new content or skill (e.g., representing numbers in base-ten, place value, basic calculation skills, number combinations for solving addition/subtraction word problems).	none
Explicit Instruction	Instruction includes "models of proficient problem solving" (Gersten, Beckman, et al., 2009, p. 65) specific to certain problems. Extensive modeling, thinking aloud procedures, and scaffolded practice and review with corrective feedback.	Worked examples are used to illustrate a specific step-by-step method to solve certain types of problems; specific problem solving strategy instruction in which teacher models, demonstrates, and uses think aloud procedures to solve problems before students work on their own.	Kroesbergen & Van Luit (2003) Gersten, Chard, et al. (2009) Xin & Jitendra (1999) Zheng, et al. (2013)

Table 2 (continued)

Category Name	Description	Indicators*	Noted in prior meta-analyses
Teach for transfer	Instruction explicitly focuses on the concept of transfer (i.e., generalizing knowledge to novel situations) (e.g., Fuchs et al., 2008).	Instruction focuses on the meaning of the word *transfer* and promotes transfer (i.e., recognizing that novel problems, even though different in certain features, are related to previously solved problems) by highlighting superficial-problem features (e.g., different format, different key word, additional or different question, and problem scope – problem is placed within a larger problem-solving context).	Zhang & Xin (2012)
Contextualized Approach	Emphasis is on "real-world applications of mathematical principles" (Baker, Gersten, & Lee, 2002, p. 63). Instruction is embedded in a realistic context or integrated with other disciplines.	Enhanced anchored instruction that includes contextualized problem solving; interactive videodisc instruction	Kroesbergen & Van Luit (2003) Zhang & Xin (2012)
Problem types or problem structure	Instruction focuses on identifying specific problem types that share underlying common problem features (e.g., Jitendra et al., 2013; in press).	Schema based instruction; schema broadening instruction; number families instruction Instructional emphasis is on different problem types: a. Group/Total/Part-part-whole – two or more amounts being combined. b. Compare/Difference – two amounts being compared. c. Change – initial amount that increases or decreases. d. Proportion/ Equal Groups; Shopping List (e.g., how much money is needed to buy varying amounts of items given a constant unit price); Buying Bag (e.g., determine number of groups needed to obtain a certain amount, given a constant number within each group). e. Multiplicative compare/Ratio; Half (e.g., dividing various amounts into two equal groups) f. Percent/Percent of Change	Gersten, Chard, et al. (2009) Xin & Jitendra (1999) Zhang & Xin (2012)

Note. *Key words and phrases reported in the reviewed studies served to indicate the presence of specific instructional components.

Inter-rater agreement. The author served as the first rater, coding all studies for sample, methodological, and study characteristics. A special education doctoral student served as the second rater, and independently coded a randomly selected sample of seven studies (25%). I calculated average interrater agreement (IRA) using the following formula:

$$\text{IRA} = \frac{\text{agreements}}{\text{agreements} + \text{disagreements}} \times 100$$

Training for inter-rater agreement coding progressed through three stages: (1) co-coding, (2) individual practice coding, and (3) formal coding. An iterative process for clarification and rewriting of codes was part of the process at each stage. In stages 1 and 2, coders used a bank of studies excluded from the meta-analysis for various reasons. Three of the articles used for coder training (Toppel, 1996; Yadrick, Regian, Connolly-Gomez, Robertson-Schule, 1997; Zawaiza & Gerber, 1993) were excluded because the sample was not within the grade levels K-12 or assessed the effects of teacher delivered versus computer-delivered word problem solving instruction (Leh & Jitendra, 2013).

In stage 1, co-coding, the second coder was provided with a copy of the coding manual. Then a training meeting was scheduled wherein the first author clarified questions about the coding manual and provided examples. Next, the two raters coded one study together (Yadrick et al., 1997) using the coding manual. After additional discussion and clarification of the coding manual, stage 2 was initiated wherein each rater independently coded three of the excluded articles (Leh & Jitendra, 2013; Toppel, 1996; Zawaiza & Gerber, 1993) and IRA was assessed. Average IRA was 82% for this first

round of practice coding. Codes with low IRA were discussed until consensus was reached. Then the coding manual was clarified and rewritten in greater detail. For example, coders disagreed on determination of "instructional setting" and "mathematics task" in all three practice-coded studies. After discussion about instructional setting, an additional coding option was added: "author did not specify" and space was provided on the coding sheet for qualitative details extracted from the study, supporting this determination. Here is an example of one of the coders' reasoning for the decision to code instructional setting as, "author does not specify." Although the authors say "whole class" instruction, we are not given information telling us definitively where the class took place (e.g., a general education classroom, special education classroom, computer lab) as is shown in this excerpt from the study:

> The teacher conducted whole-group instruction by initially projecting the computer program on a large screen in the classroom and discussed the problem solution process as a group before students worked on their individual computers using headphones (Leh & Jitendra, 2013, p.71).

The "mathematics task" category required rewording for increased clarity. Specifically, the coding manual initially described the types of mathematics task as follows (1) arithmetic (add & subtract), (2) arithmetic (all four operations), and (3) higher order. These categories were not explicitly mutually exclusive, so the coding manual was rewritten for greater clarity as follows: (1) arithmetic word problems involving only addition and/or subtraction, (2) arithmetic word problems involving multiplication/division or all four operations, and (3) fraction, ratio, proportion,

geometric and/or algebraic word problems. Stage 3, formal coding, began with random selection of seven studies, representing 25% of the total sample of studies. All 28 studies were numbered and then an online random number generator found at https://www.random.org/ was used to select the random sample for coding. The second rater coded the randomly selected sample and IRA was assessed. Average IRA for this sample was 86%. All disagreements between raters were discussed and resolved by consensus before analyzing coded data.

Data Analysis

All analyses were conducted using the software, Comprehensive Meta-Analysis (CMA; Borenstein, Hedges, Higgins & Rothstein, 2015) and R Software (R Core Team, 2013). Graphs from CMA and R were used to visually inspect the data for patterns in effect size magnitude (i.e., distribution of effect sizes; forest plot with 95% confidence intervals for each study) and likelihood of publication bias (i.e., funnel plot depicting the relation between effect size and study size).

Effect size calculation. In meta-analyses, "effect sizes can be viewed as the dependent (or criterion) variables and the features of the study designs as independent (or predictor) variables (Cooper et al., 2009, p. 13). To facilitate comparison across studies, the same effect size estimate was calculated for all findings in this meta-analysis. Specifically, the standardized mean difference, or Cohen's *d*, was calculated using the formula below:

$$\bar{d} = \frac{\overline{X_1} - \overline{X_2}}{S_{pooled}}$$

where X_1 and X_2 are posttest means for the treatment and control groups and S_{pooled} is the standard pooled deviation defined as:

$$S_{pooled} = \sqrt{\frac{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

Where s_1^2 is the standard deviation for group 1, s_2^2 is the standard deviation for group 2, n_1 is the number of subjects in group 1, and n_2 is the number of subjects in group 2. Because Cohen's d "has a slight bias, tending to overestimate the absolute value of δ in small samples" (Cooper et al., 2009, p.226), the Hedges' g correction was applied to produce an unbiased estimate:

$$1 - \frac{3}{4N - 9}$$

where N is the total sample size. Each study was weighted by the inverse of its variance before combination to determine the grand mean effect size of the population under the random effects model:

$$W_i^* = \frac{1}{V_{Y_i} + T^2}$$

where V_{Y_i} is the within-studies variance for study i and T^2 is the between-studies variance (Borenstein, Hedges, Higgins, & Rothstein, 2009).

The precision of all effect size estimates was established by calculating the standard error (SE) of the mean (i.e., square root of the sum of the inverse variance weights) and using it to create a 95% confidence interval around the mean (see Cooper et al., 2009):

$$\text{Lower limit of 95\% } CI = \bar{g} - 1.96 * SE$$

$$\text{Upper limit of 95\% } CI = \bar{g} + 1.96 * SE$$

Estimation of heterogeneity. To assess consistency across studies, heterogeneity of the distribution of effect sizes was assessed using the *Q* statistic and the I^2 statistic. The *Q* statistic has an approximate chi-square distribution with *k-1* degrees of freedom, where *k* is the number of effect sizes (Cooper et al., 2009). Significant Q statistics indicate the existence of heterogeneity, while "I^2 describes the percentage of total variation across studies that is due to heterogeneity rather than chance" (Higgins, Thompson, Deeks, & Altman, 2003, p.557). I^2 was calculated using the formula below:

$$I^2 = 100\% * \frac{Q-(k-1)}{Q}$$

Where *Q* is the homogeneity statistic and *k* is the number of independent effect sizes. Suggested guidelines for interpreting the value of I^2 follow: I^2 = 25% suggests a small amount of heterogeneity, I^2 = 50% suggests medium heterogeneity and I^2 = 75% suggests large heterogeneity (Cooper et al., 2009; Higgins et al., 2003).

Moderator analyses. The presence of heterogeneity in the grand mean effect size estimate indicates the possible existence of variables that serve to moderate the effect (e.g., sample, methodological, and study features) (Borenstein et al., 2010; Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006). As Cooper et al. (2009) noted, I conducted moderator analyses by disaggregating study effect size estimates and grouping them into appropriate categories. The chi-square test of homogeneity (*Q* statistic) of effect sizes and the associated I^2 statistics were used to determine the significance and magnitude of between (Q_B) and within (Q_w) group differences in mean effect size for each potential moderator variable. Because these tests traditionally have

low power to detect departures from homogeneity (Cooper et al., 2009; Mittlböck & Heinzl, 2006), some researchers (e.g., Petitti, 2001) recommend selecting a significance level of α = .10. However, I chose to use the traditional significance level of α = .05 to avoid increasing Type I error (risk of a false positive) (Higgins et al., 2003).

A significant Q_B statistic suggests that the observed differences between subgroups of the moderator category are significantly different from each other. A significant Q_W statistic suggests the existence of additional heterogeneity yet to be explained, whereas a non-significant Q_W statistic suggests that any variation among effect sizes for a given level of a moderator is attributable to sampling error. In the case of a significant Q_B statistic found for a moderator variable with more than two sub-groups, post-hoc pairwise comparisons were conducted. The Dunn-Bonferroni method (see Howell, 2010) was used to adjust the alpha level by taking the traditional α = .05 and dividing it equally among the number of comparisons. For example, a variable with three levels yields three comparisons (1 vs. 2; 1 vs. 3; 2 vs. 3) producing $\alpha = \frac{.05}{3} = 0.01\bar{6}$. A variable with four levels yields six comparisons (1 vs. 2; 1 vs. 3; 1 vs. 4; 2 vs. 3; 2 vs. 4; 3 vs. 4) producing $\alpha = \frac{.05}{6} = .008\bar{3}$.

Addressing influential points. If we consider the percent of data located in a normal distribution, only the top 99.7%-100% of the data would be located three standard deviations above the mean. As such, an effect size of three standard deviation units or higher is highly unlikely, regardless of the myriad possible causes for such error (Cooper

et al., 2009). Therefore, all statistical analyses were conducted with and without any such influential effect size estimates and both are reported.

Chapter 4

RESULTS

In this chapter, I first report descriptive statistics by study, noting instances of missing data across coded variables. Next, I discuss findings related to the grand mean effect of mathematics word-problem solving instruction on immediate posttest performance of students with LD or MD, and consider the impact of two influential effect size estimates. Last, I report the results of analyses of potential moderator variables.

Descriptive Statistics

A total of 31 independent Hedges' *g* effect sizes were extracted from 28 studies. Table 3 provides a summary of included studies. The mean sample size of the 28 studies was 44.74 (*SD* = 30.22) and ranged from 11 to 164. The studies were conducted from 1987 to 2014, with slightly over half (54%) conducted after 2004. Five of the 28 studies were unpublished dissertations, and the remaining studies were published in peer-reviewed journals. To assess possible publication bias, a funnel plot (see Figure 3) was used to illustrate the relation between effect size and study size. Figure 3 shows that larger studies are distributed on the top and in the middle of the funnel plot, and there is a small gap near the left bottom corner of the funnel plot. This gap where the small-scale studies would have been if they could be located suggests possible publication bias against studies with small sample size and non-significant effect sizes (Cooper et al., 2009). Although "these studies may be less influential on the meta-analytic results because they tend to provide a small weight in the weighted average effect size

computation" (Pai, Sears, & Maeda, 2015, p. 86), publication status was analyzed as a potential moderator variable. These results are reported later in this chapter.

Table 3
Summary of Included Studies

Study	Participants	Assignment to groups	Interventionist, Arr., & Setting	Mean Hrs.	Intervention	Control	Measure	WPS Content
Baker (1992)*	*N* = 46; grades 3-5; 100% LD; 48% Minority; FRL=NR Attrition = NR	RA at student level	Researcher; small groups (2-6); other	1.5	4 step heuristic + self-generated drawings	Alt. Int. (4 step heuristic)	Researcher-developed, immediate posttest (change problems) (α = 0.82)	Arithmetic (+, -, x, ÷)
Bottge & Hasselbring (1993)	*N* = 29; grades 9-10; 48% LD/52% MD; Minority= NR; FRL= NR; Attrition = 0%	RA at student level (matched pairs)	Teachers & researchers; small groups (7-8); NR	13.3	Contextualized problem solving with videodisc	Alt. Int.	Researcher-developed, contextualized problem solving test (α = NR)	Fractions
Fede et al. (2013)	*N* =32; grade 5; 19% LD/81% MD 3% Minority; FRL= NR Attrition = 0%	RA at student level	NR; NR; NR	13.5	Computer-assisted, schema-based instruction	BAU	Researcher-developed, adapted from MCAS (α = NR)	Arithmetic (+, -, x, ÷)
Fuchs, Fuchs, Craddock, et al. (2008)	*N* =84; grade 3; 100% MD; 75% Minority; 76% FRL; Attrition = 16%	RA at classroom level	Researchers; whole class; gen. ed.	33.4	BAU + Schema-broadening instruction tutoring	BAU	Researcher-developed, immediate posttest (α = 0.95)	Arithmetic (+, -, x, ÷)
Fuchs, Fuchs, Craddock, et al. (2008)	*N*= 159; grade 3; 100% MD; 71% Minority; 73% FRL; Attrition = 16%	RA at classroom level	Researchers; whole class + small groups (2-4) for tutoring; gen. ed.	33.4	Schema-broadening instruction + tutoring	Schema broadening instruction	Researcher-developed, immediate posttest (α = 0.95)	Arithmetic (+, -, x, ÷)
Fuchs, Fuchs, Finelli, et al. (2004)	*N* =61; grade 3; 25%LD/ % MD; 73% Minority; FRL=NR; Attrition = NR	RA at classroom level	Researchers; whole class; gen. ed.	18.4	SBTI	BAU	Researcher-developed, immediate posttest (α = 0.97)	Arithmetic (+, -, x, ÷)

Table 3 (continued)

Study	Participants	Assignment to groups	Interventionist Arr., & Setting	Mean Hrs.	Intervention	Control	Measure	WPS Content
Fuchs et al. (2002)	*N* =20; grade 4; 100% LD; 60% Minority; 40-65% FRL; Attrition (trt) = 0%; Attrition (cntrl) = 5%	RA at student level	Researchers; small groups (2-4); spec. ed.	15.9	SBTI	BAU	Researcher-developed, immediate posttest ($\alpha = 0.95$)	Arithmetic (+, -, x, ÷)
Fuchs, Fuchs, Prentice, Hamlett, et al. (2004)	*N* =52; grade 3; 100% MD; 69% Minority; 81% FRL; Attrition = NR	RA at classroom level	Teachers & Researchers; whole class; gen. ed.	16.8	SBTI	Alt. Int.	Researcher-developed, immediate posttest ($\alpha = 0.95$)	Arithmetic (+, -, x, ÷)
Fuchs, Fuchs, Prentice (2004)	*N* = 13; grade 3; 100% MD; 60% Minority; 48% FRL; Attrition = NR	RA at classroom level	Teachers & Researchers; whole class; gen. ed.	15.8	SBTI + self-regulation strategy instruction	BAU	Researcher-developed, immediate posttest ($\alpha = 0.95$)	Arithmetic (+, -, x, ÷)
Fuchs, Fuchs, Prentice (2004)	*N* = 32; grade 3; 100% MDRD; 83% Minority; 75% FRL; Attrition = NR	RA at classroom level	Teachers & Researchers; whole class; gen. ed.	15.8	SBTI + self-regulation strategy instruction	BAU	Researcher-developed, immediate posttest ($\alpha = 0.95$)	Arithmetic (+, -, x, ÷)
Fuchs et al., (2009)	*N* = 89; grade 3; 17% LD/83% MD; 82% Minority; 78% FRL; Attrition = NR	RA at student level	Researchers; small groups; other	20.0	Schema-broadening instruction tutoring	BAU	Researcher-developed, Vanderbilt Story Problems test ($\alpha = 0.86$)	Arithmetic (+, -)

Table 3 (continued)

Study	Participants	Assignment to groups	Interventionist Arr., & Setting	Mean Hrs.	Intervention	Control	Measure	WPS Content
Fuchs, Seethaler, et al. (2008)	*N* = 35; grade 3; 9% LD/91% MD; 77% Minority; 88% FRL; Attrition (trt) = 24%; Attrition (cntrl) = 10%	RA at student level	Researchers; 1-on-1; other	15.0	Schema-broadening instruction	BAU	Researcher-developed, Jordan's Story Problems test (α = 0.86)	Arithmetic (+, -)
Hutchinson (1993)	*N* = 20; grades 8-10; 100% LD; Minority= NR; FRL= NR; Attrition = NR	RA at student level	Researchers; 1-on-1; spec. ed.	26.7	Cognitive Strategy Instruction	BAU	Standardized, BCA (α = NR)	Algebra
Jitendra et al. (1998)	*N* = 34; grade 2-5; 50% LD/50% MD; 18% Minority; FRL= NR Attrition = 0%	RA at student level	Researchers; small groups (3-6); other	13.9	Schema-based instruction	BAU	Researcher-developed, immediate posttest (α = 0.83)	Arithmetic (+, -)
Jitendra, et al. (2013)	*N* = 55; grade 3; 9% LD/91% MD; Minority = NR; FRL= NR; Attrition = 13%	RA at student level	Volunteer tutors from the community; small groups; other	30.0	Schema-based instruction	Alt. Int.	Researcher-developed, immediate posttest (α = 0.75)	Arithmetic (+, -)
Konold (2004)*	*N* = 61; grades 6-12; 100% LD; Minority = NR; FRL = NR; Attrition = NR	RA at classroom level	Teachers; whole class; spec. ed.	5.50	C-R-A; strategy instruction	Alt. Int. (Abstract only)	Researcher-developed, immediate posttest (α = NR)	Algebra

Table 3 (continued)

Study	Participants	Assignment to groups	Interventionist Arr., & Setting	Mean Hrs.	Intervention	Control	Measure	WPS Content
Lambert (1996)*	*N* = 76; grades 9-12; 100% LD; Minority = NR; FRL = NR; Attrition = NR	Cluster sampling at class level	Teachers; whole class; spec. ed.	7.3	Cognitive Strategy Instruction	BAU	Researcher-developed, fractions word problems posttest (α = 0.78)	Fractions, decimals, percent
Lee (1992)*	*N* = 32; grades 4-6; 100% LD; Minority = NR; 80% FRL; Attrition (trt) = 0% Attrition (cntrl) = 13%	RA at classroom level	Researchers; whole class; spec. ed.	6.8	Number families arrow model	BAU	Researcher-developed, answer accuracy (α = 0.89)	Arithmetic (+, -)
Marzola (1987)*	*N* = 60; grades 5-6; 100% LD; Minority= NR; FRL = NR; Attrition = NR	RA at school level	Researchers: whole class; spec. ed.	6.0	Explicit problem solving with problem typing	No intervention (worksheet completion)	Researcher-developed, addition word problems posttest (α = NR)	Arithmetic (+, -)
Moran et al. (2014)	N = 37; grade 3; 100% MD; 89% Minority; FRL=NR; Attrition = NR	RA at student level	NR (Trained tutors); small group (3-5); NR	9.2	Supplemental paraphrasing intervention	BAU	Standardized, composite score of TOMA, CMAT, KeyMath (α = 0.80)	Arithmetic (+, -)
Owen & Fuchs (2002)	*N* = 12; grade 3; 50% LD/50% MD; 50% Minority; 100% FRL; Attrition = 0%	RA at classroom level	Researchers; NR, gen ed.	NR	Problem solving heuristic acquisition	BAU	Researcher-developed, product measure (α = 0.89)	Arithmetic (+, -, x, ÷)

Table 3 (continued)

Study	Participants	Assignment to groups	Interventionist Arr., & Setting	Mean Hrs.	Intervention	Control	Measure	WPS Content
Powell & Fuchs (2010)	*N* = 56; grade 3; 33%LD/67%MD; 93% Minority; 74% FRL; Attrition (trt) = 10%; Attrition (cntrl) = 3%	RA at student level	Teachers & Researchers; small group (3 students); other	6.9	Schema-broadening tutoring	BAU	Researcher-developed; missing information before equal sign (α = 0.54)	Arithmetic (+, -)
Shiah et al. (1994)	*N* = 20; grades 1-6; 100% LD; Minority= NR; FRL=NR; Attrition = 0%	RA at student level	Computer; one-on-one; NR	0.7	CAI with explicit problem solving heuristic	Alt, Int. (Simplified CAI)	Researcher-developed, immediate online posttest (α = NR)	Arithmetic (+, -)
Swanson et al. (2013)	*N* = 38; grade 3; 100% MD; Minority = NR; FRL=NR; Attrition = NR	RA at student level	Research assistants & paraprofessionals; small groups (2-4); gen. ed.	10.0	Supplemental GHI & visual schematic	BAU	Standardized, composite score of CMAT, WRAT-3, WIAT (α = 0.90)	Arithmetic (+, -)
Walker & Poteet (1989)	*N* = 70; grades 6-8; 100% LD; Minority= NR; FRL=NR; Attrition = NR	RA at classroom level	Teachers; small group (3-18); spec. ed.	8.5	Draw a diagram method	Alt. Int. (keyword method)	Researcher-developed, solution accuracy on 1-step word problems (α = 0.84)	Arithmetic (+, -)
Wilson & Sindelar (1991)	*N* = 41; grades 2-5; 100% LD; 65% Minority; FRL=NR; Attrition = NR	RA at small group level	Researchers; small groups (3-5); other	7.0	Explicit problem solving strategy instruction + sequencing	Alt. Int. (Sequencing only)	Researcher-developed, immediate posttest (α = 0.88)	Arithmetic (+, -)

Table 3 (continued)

Study	Participants	Assignment to groups	Interventionist Arr., & Setting	Mean Hrs.	Intervention	Control	Measure	WPS Content
Woodward & Baxter (1997)	*N* = 38; grade 3; 32% LD/ 68% MD; Minority =NR; FRL= NR; Attrition =NR	RA at classroom level	Teachers; small group; gen. ed.	NR	Student-centered problem solving with C-R-A	BAU	Standardized, ITBS problem solving subtest (α = NR)	Arithmetic (+, -, x, ÷)
Woodward et al. (2001)	*N*= 25; grade 4; 100% MD; Minority = NR; FRL= NR; Attrition = NR	RA at classroom level	Teachers; whole class; gen. ed.	15.8	Performance assessment tasks	BAU	Researcher-developed, immediate posttest (α = 0.85)	Arithmetic (+, -, x, ÷) & geom.
Woodward et al. (2001)	*N* = 11; grade 4; 100% LD; Minority = NR; FRL= NR; Attrition = NR	RA at classroom level	Teachers; whole class + small group tutoring; gen. ed.	55.3	Performance assessment tasks & ad-hoc tutoring	BAU	Researcher-developed, immediate posttest (α = 0.85)	Arithmetic (+, -, x, ÷) & geom.
Xin, et al. (2005)	*N* = 22; grades 6 - 8; 82% LD/18% MD; 70% Minority; FRL= NR; Attrition =NR	RA at student level	Teachers & Researchers; small groups (4-7); NR	12.0	Schema-based instruction	Alt. Int. (GHI)	Researcher-developed, immediate posttest (α = 0.88)	Arithmetic (+, -, x, ÷)
Xin, et al. (2011)	*N* = 27; grades 4-5; 35% LD/ 65% MD; 58% minority; FRL= NR; Attrition= 7%	RA at student level	Teachers & Researchers; small groups (6-7); NR	11.3	Schematic diagram instruction	Alt. Int. (GHI)	Researcher-developed, immediate posttest (α = 0.86)	Arithmetic (+, -, x, ÷)

Note. *Unpublished dissertation; Alt. int. = alternate intervention; Arr. = arrangement; BAU = business-as-usual; BCA = British Columbia Applications; CAI = computer assisted instruction; CMAT = Comprehensive Mathematical Abilities Test; cntrl = control; C-R-A=concrete-representational-abstract; FRL= eligible for free or reduced lunch; gen. ed.= general education; geom.=geometry; GHI = general problem solving heuristic instruction; Hrs. = hours; ITBS = Iowa Test of Basic Skills; LD = learning disabilities; MCAS = Massachusetts Comprehensive Assessment System; MD = mathematics difficulties; NR = not

reported; SBTI = schema broadening + transfer instruction; spec. ed. = special education; TOMA = Test of Mathematical Abilities; trt = treatment; WRAT-3 = Wide Range Achievement Test; WIAT= Wechsler Individual Achievement Test

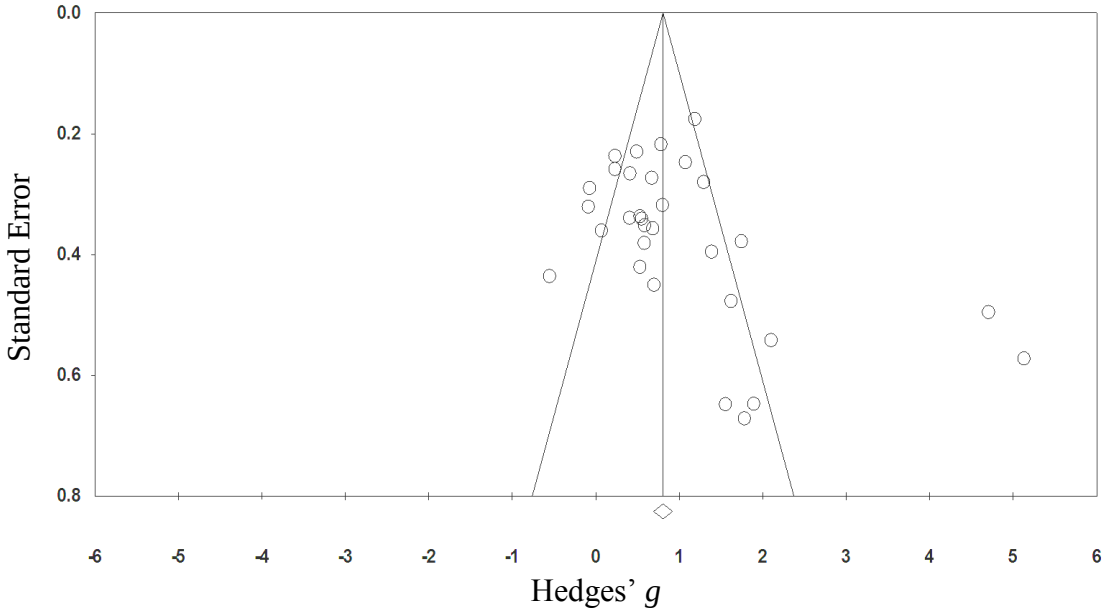


Figure 3. Funnel plot of standard error by Hedges' g

Instructional components. Studies were coded for the presence of seven instructional components (See Table 2 for description of components). Most studies reported indicators of multiple instructional components. Specifically, studies included between one and six of the selected instructional components, with a median of three and a mode of four components. Figure 4 lists the instructional components by study. The most commonly reported instructional components were explicit instruction (93% of studies), followed by representations (75%) and problem types or problem structure (57%). Metacognitive strategy instruction was present in 50% of studies, prerequisite/foundational skills instruction was a component in 32% of studies, and teaching for transfer was part of instruction in 25% of studies. Lastly, the contextualized approach was present in only two studies (7%).

Study	*g*	RP	PT	MT	EI	TT	PQ	CX	Total
Baker (1992)	-0.07	x		x	x				3
Bottge & Hasselbring (1993)	1.07*							x	1
Fede et al. (2013)	0.59	x	x		x			x	4
Fuchs et al. (2002)	2.11*		x		x	x	x		4
Fuchs, Fuchs, Craddock, et al. (2008)	1.15* 1.12*	x	x	x	x	x	x		6
Fuchs, Fuchs, Finelli, et al. (2004)	4.71*		x		x	x	x		4
Fuchs, Fuchs, & Prentice (2004)	1.40* 1.90*		x	x	x	x	x		5
Fuchs, Fuchs, Prentice, et al. (2004)	5.14*		x		x	x	x		4
Fuchs, Powell, et al. (2009)	0.79	x	x	x	x	x	x		6
Fuchs, Seethaler et al. (2008)	0.54	x	x	x	x	x	x		6
Hutchinson (1993)	0.71	x	x	x	x				4
Jitendra et al. (1998)	0.56	x	x		x				3
Jitendra et al. (2013)	0.68*	x	x	x	x		x		5
Konold (2004)	0.24	x		x	x				3
Lambert (1996)	0.50*	x		x	x				3
Lee (1992)	0.69*	x	x		x				3
Marzola (1985)	1.30*		x	x	x				3
Moran et al. (2014)	0.41	x			x				2
Owen & Fuchs (2002)	1.56*	x			x				2
Powell & Fuchs (2010)	0.42		x	x	x		x		4
Shiah et al. (1994)	-0.26	x		x	x				3
Swanson et al. (2013)	1.75*	x		x	x				3
Walker & Poteet (1989)	0.24	x			x				2
Wilson & Sindelar (1991)	0.81*	x			x				2
Woodward & Baxter (1997)	-0.08	x	x						2
Woodward et al. (2001)	1.79* 0.54	x	x		x				3
Xin et al. (2005)	1.63*	x	x	x	x				4
Xin et al. (2011)	0.59	x	x	x	x				4

Note. REP = representations; PT/S = problem types or problem structure; MET = metacognition; EI = explicit instruction; TT = teach for transfer; PRQ = prerequisite/foundational skills; CTX= contextualized instruction

Figure 4. Instructional Components by Study

Missing data. In several studies, authors did not provide complete information on potential moderator variables (i.e., participant characteristics, outcome measure characteristics, contextual characteristics of intervention, study design characteristics). With regard to participant characteristics, authors did not report the (a) socioeconomic status (SES) of participants in 68% of the included studies, (b) ethnicity of participants in 46% of studies, and (c) sex of participants in 25% of studies. In the nine studies that reported SES, 72% of the sample, on average, was eligible for free or reduced price lunch (FRL). Of the 15 studies that provided information about ethnicity of the sample, 34% of the participants, on average, were White. In the 21 studies reporting sex of the sample, 43% of the participants, on average, were female.

In terms of outcome measure characteristics, seven studies (25%) did not provide reliability estimates for the primary outcome measure, and 22 (86%) did not present evidence of validity for the data set associated with the primary outcome measure. Of the six studies (14%) that provided evidence of validity, four studies (Fuchs, Fuchs, Hamlett, & Appleton, 2002; Fuchs, Fuchs, Craddock, et al., 2008; Fuchs, Fuchs, Finelli, et al., 2004; Fuchs, Fuchs, Prentice, Hamlett, et al., 2004) presented concurrent validity with standardized measures (i.e., Woodcock Johnson III, Terra Nova). One study (Walker & Poteet, 1989) reported content validity of the outcome measure, and the remaining study (Woodward & Baxter, 1997) used a standardized measure (Iowa Test of Basic Skills), and merely noted that it had "well documented reliability and validity" (p. 377).

With regard to contextual characteristics of intervention, two of the studies (Owen & Fuchs, 2002; Woodward & Baxter, 1997) did not provide information needed to

calculate the total minutes of instruction; six did not report the instructional setting (Bottge & Hasselbring, 1993; Fede et al., 2013; Moran et al. (2014); Shiah et al., 1994; Xin et al., 2005, 2011); two did not report the instructional arrangement (Fede et al., 2013; Owen & Fuchs, 2002); and two did not indicate who delivered the intervention (Fede et al., 2013; Moran et al., 2014). Of the studies that reported duration of the intervention, it was less than 10 hours (M = 5.93, SD = 2.63) in 38% of the studies, between 10 and 16 hours (M = 13.84, SD = 1.96) in 35%, and more than 16 hours (M = 27.30, SD = 12.04) in 27% of the studies. In the studies that described the instructional setting, most interventions occurred in a general education or special education classroom (35% each), and the remaining (30%) took place outside the classroom (e.g., in the library). Instructional arrangement was described in 26 studies. In 42% of these studies, instruction was implemented in small groups (generally ranging from 2-7 students per group), followed by whole-class instruction (37%) and one-on-one instruction (15%). Of the 26 studies that reported who delivered the intervention, the majority of the studies (48%) were delivered by researchers, followed by teachers (18.5%), and both researchers and teachers (15%). In the remaining studies (18.5%), others (e.g., computer, volunteers from the community) delivered the intervention.

In terms of study design characteristics, 20 studies (71%) did not report complete information on FOI, and 17 (61%) did not report information about sample attrition. In 12 studies (43%), fidelity of implementation (FOI) was reported for neither the treatment group nor the control group; and in eight studies (29%) FOI was reported for the treatment, but not the control group. For the 17 studies reporting FOI for the treatment

group, the mean percent of essential instructional components addressed in the treatment group was 96%. For the nine studies reporting FOI for the comparison group, the mean percent of essential instructional components addressed was 97%. In eight of the 12 studies that reported attrition, there was less than 10% attrition. The highest attrition rate was 24%, which was reported by Fuchs, Seethaler, et al. (2008). Attrition rates are reported by study in Table 3.

Effect of Mathematics Word-Problem-Solving Instruction

Figure 5 presents a distribution of the 31 independent Hedges' *g* effect sizes extracted from the 28 included studies. The graph reveals a negatively skewed distribution with many effect sizes clustering between 0.0 and 2.0. The skewness and kurtosis estimates for the data set were 2.25 (0.42) and 5.84 (0.82), respectively. After removal of the two influential effect size estimates, skewness and kurtosis statistics were reduced to 0.33 (0.43) and -0.63 (0.85), respectively.

Figure 6 shows the distribution of the 31 effect sizes with their 95% confidence intervals. The mid-point of the square represents the point estimate of each effect size, and the width of the line shows the 95% chance that the true effect will be within that range. Variation in the width of the confidence intervals represents variation in the precision of effect size estimates. There is considerable variation of the 31 effect sizes with relatively large confidence intervals. Both Figures 5 and 6 show that two of the effect size point estimates are at least four times larger than the overall mean effect size. As such, all calculations were performed both with and without these two influential points.

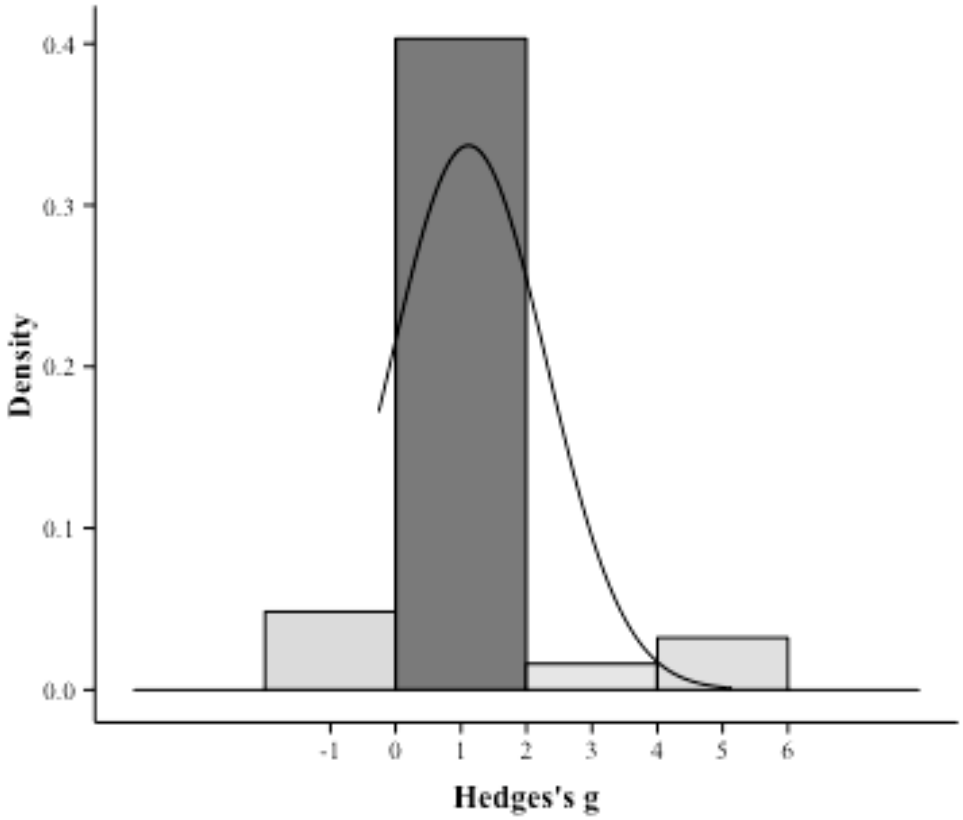


Figure 5. Distribution of 31 Hedges' *g* Effect Sizes

Study Name	*g*	CI
Baker (1992)	-0.07	[-0.64, 0.50]
Bottge & Hasselbring (1993)	1.07*	[0.31, 1.83]
Fede et al. (2013)	0.59	[-0.10, 1.28]
Fuchs et al. (2002)	2.11*	[1.04, 3.17]
Fuchs, et al. (2008) BAU + SBI T vs. BAU	1.15*	[0.67, 1.63
Fuchs, et al. (2008) SBI + T vs. SBI	1.12*	[0.76, 1.47]
Fuchs, Fuchs, Finelli, et al. (2004)	4.71*	[3.74, 5.69]
Fuchs, Fuchs, Prentice (2004) MD	1.40*	[0.62, 2.17]
Fuchs, Fuchs, Prentice (2004) MDRD	1.90*	[0.63, 3.17]
Fuchs, Fuchs, Prentice, et al. (2004)	5.14*	[4.01, 6.26]
Fuchs, Powell, et al. (2009)	0.79	[0.36, 1.22]
Fuchs, Seethaler, et al. (2008)	0.54	[-0.13, 1.20]
Hutchinson (1993)	0.71	[-0.18, 1.59]
Jitendra et al. (1998)	0.56	[-0.11, 1.23]
Jitendra, Dupuis, et al. (2013)	0.68*	[0.14, 1.21]
Konold (2004)	0.24	[-0.27, 0.75]
Lambert (1996)	0.50*	[0.05, 0.95]
Lee (1992)	0.69*	[-0.01, 1.39]
Marzola (1985)	1.30*	[0.75, 1.85]
Moran et al. (2014)	0.41	[-0.25, 1.08]
Owen & Fuchs (2002)	1.56*	[0.29, 2.83]
Powell & Fuchs (2010)	0.42	[-0.11, 0.94]
Shiah et al. (1994)	-0.26	[-1.10, 0.59]
Swanson et al. (2013)	1.75*	[1.01, 2.49]
Walker & Poteet (1989)	0.24	[-0.23, 0.70]
Wilson & Sindelar (1991)	0.81*	[0.18, 1.43]
Woodward & Baxter (1997)	-0.08	[-0.71, 0.55]
Woodward, et al. (2001) LD	1.79*	[0.47, 3.11]
Woodward, et al. (2001) MD	0.54	[-0.29, 1.36]
Xin et al. (2005)	1.63*	[0.69, 2.56]
Xin et al. (2011)	0.59	[-0.16, 1.34]
Mean	1.03*	[0.73, 1.32]

LD = learning disabilities; MD = mathematics difficulties; MDRD = reading and mathematics difficulties; SBI = schema broadening instruction; T= tutoring;
*p < .05

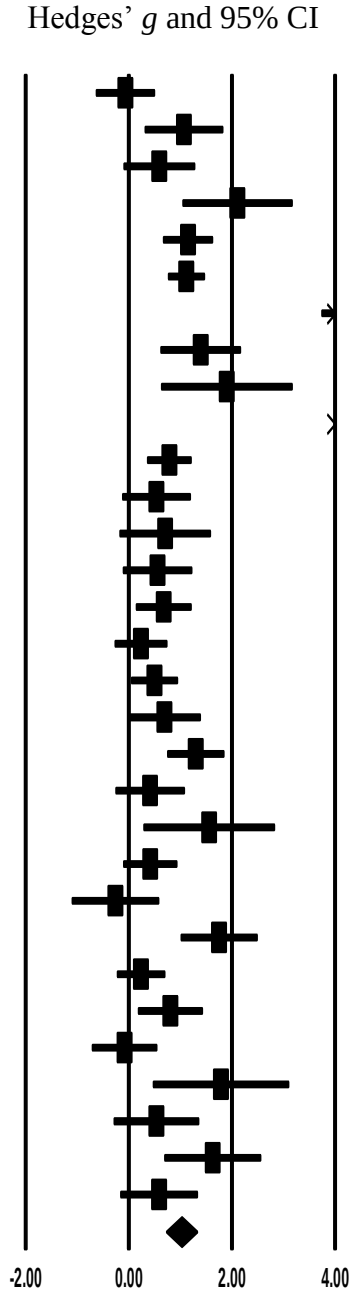


Figure 6. Hedges' g Effect Sizes with 95% CI by Study

A random effects model was selected for the current meta-analysis in light of the fact that the effect sizes involved were extracted from studies varying in sample, methodological, and study characteristics, "each of which could contribute to heterogeneity among the population effects" (Pai et al., 2015, p.89). The grand mean effect size (Hedges' *g*) for mathematics word problem solving instruction on immediate posttest performance was 1.03 (*SE* = 0.15; 95% CI = 0.73 to 1.33). When the two influential point estimates were excluded, the grand mean effect size was 0.77 (*SE* = 0.10; 95% CI = 0.57 to 0.96). These findings indicate that, on average, students with LD or MD who received WPS treatment outperformed their peers in the control condition at immediate posttest by 0.57 to 1.33 of a standard deviation unit, which translates to an improvement index of 22 to 41 percentile points (WWC, 2014).

Homogeneity statistics were calculated to "determine whether there are genuine differences underlying the results of the studies (heterogeneity), or whether the variation in findings is compatible with chance alone (homogeneity)" (Higgins et al., 2003, p. 557). When all 31 independent effect sizes were included, the chi-square test of homogeneity was significant ($Q_B = 51.32$, $p < .01$) and the I^2 statistic was 41.54%, which suggests that a moderate percentage of the total variation across studies can be accounted for by heterogeneity as opposed to chance (Higgins & Thompson, 2002; Higgins et al., 2003). By contrast, when the two influential estimates were excluded, the chi-square test of homogeneity was no longer significant ($Q_B = 31.29$, $p > .25$), and the magnitude of heterogeneity was reduced to $I^2 = 10.52\%$.

Relation of Potential Moderator Variables to Effect Size

The presence of low to moderate heterogeneity in the grand mean effect size estimate warrants moderator analyses to account for that variance (Borenstein et al., 2010; Huedo-Medina et al., 2006). Based on Cooper et al.'s (2009) guidelines, moderator analyses were conducted by disaggregating study effect size estimates and grouping them into appropriate categories. The chi-square test of homogeneity (*Q* statistic) of effect sizes and the associated I^2 statistics were used to determine the significance and magnitude of between (Q_B) and within (Q_w) group differences in mean effect sizes for each potential moderator variable (i.e., participant characteristics, study design characteristics, outcome measure characteristics, and contextual characteristics of intervention). As with the grand mean effect size estimate for mathematics word problem solving instruction, moderator analyses were conducted both with and without the two influential effect size estimates that were over four times larger than the overall average effect size.

Participant characteristics. Tables 4 and 5 present the results of moderator analyses by participants with and without the two influential points. LD/MD status did not moderate effectiveness of instruction either with or without the two influential effect size estimates. In contrast, the difference between elementary and secondary grade levels (favoring elementary) was significant with all 31 effect sizes but not with the trimmed data set.

Table 4
Summary of Effect Sizes by Participants

Variable	Q_B	*k*	*g*	*SE*	95% CI
LD/MD Status	3.47				
LD		14	0.72*	0.16	[0.40, 1.04]
MD		17	1.25*	0.24	[0.79, 1.71]
Grade Level	3.94*				
Elementary (K-6)		25	1.12*	0.18	[0.76, 1.48]
Secondary (7-12)		6	0.61*	0.18	[0.25, 0.97]

Note. CI confidence interval; *k* = number of effect sizes; LD = learning disability; MD = mathematics difficulty; Q_B= *Q Between*; Q_W = *Q Within*
*p < .05; **p < 0.01

Table 5
Summary of Effect Sizes by Participants After Removing Influential Points

Variable	Q_B	*k*	*g*	*SE*	95% CI
LD/MD Status	0.22				
LD		14	0.72*	0.16	[0.40, 1.04]
MD		15	0.81*	0.12	[0.58, 1.05]
Grade Level	0.82				
Elementary (K-6)		23	0.80*	0.11	[0.58, 1.02]
Secondary (7-12)		6	0.61*	0.18	[0.25, 0.97]

Note. CI confidence interval; *k* = number of effect sizes; LD = learning disability; MD = mathematics difficulty; Q_B= *Q Between*; Q_W = *Q Within*
*p < .05; **p < 0.01

Study design. Tables 6 and 7 present the results of moderator analyses by study design with and without the two influential points. For type of report, published studies yielded a significantly larger effect size than unpublished studies. However, removal of the two influential effect size estimates reduced the mean effect associated with published studies such that the difference was no longer significant. Similarly, studies employing random assignment at the classroom or school level yielded a significantly larger effect

size than studies employing student-level random assignment when the two influential effect sizes were included, but not after their removal. The remaining variable, type of comparison group, did not moderate effectiveness of instruction either with or without the two influential effect size estimates.

Table 6
Summary of Effect Sizes by Study Design

Variable	Q_B	*k*	*g*	*SE*	95% CI
Type of Report	4.55*				
Published		26	1.14*	0.18	[0.79, 1.49]
Unpublished		5	0.53*	0.23	[0.08, 0.97]
Random Assignment	4.63*				
Student level		16	0.71*	0.14	[0.44, 0.98]
Classroom or school level		15	1.34*	0.26	[0.83, 1.85]
Type of Comparison Group	0.86				
No intervention/attention only		1	1.30*	0.28	[0.75, 1.85]
Business-as-usual		19	1.07*	0.19	[0.70, 1.45]
Alternate intervention		11	0.93*	0.28	[0.38, 1.47]

Note. CI confidence interval; *k* = number of effect sizes; Q_B= *Q Between*; Q_W = *Q Within*
*p < .05; **p < 0.01

Table 7
Summary of Effect Sizes by Study Design After Removing Influential Points

Variable	Q_B	*k*	*g*	*SE*	95% CI
Type of Study	1.39				
Published		24	0.82*	0.11	[0.61, 1.04]
Unpublished		5	0.53*	0.23	[0.08, 0.97]
Random Assignment	0.33				
Student level		15	0.71*	0.14	[0.44, 0.98]
Classroom or school level		14	0.83*	0.14	[0.54, 1.11]
Type of Comparison Group	4.51				
No intervention/attention only		1	1.30*	0.28	[0.75, 1.85]
Business-as-usual		18	0.83*	0.12	[0.59, 1.08]
Alternate intervention		10	0.59*	0.17	[0.25, 0.92]

Note. CI confidence interval; *k* = number of effect sizes; Q_B= *Q Between*; Q_W = *Q Within*
*p < .05; **p < 0.01

Outcome measure characteristics. Tables 8 and 9 present the results of moderator analyses by outcome measure characteristics with and without the two influential points. Assessment of between-studies homogeneity indicated that the type of outcome measure (i.e., researcher-developed or standardized) did not moderate effectiveness of instruction either with or without the two influential effect size estimates. In contrast, reliability of outcome measure moderated word problem solving effectiveness both with and without the two influential effect sizes. Specifically, studies using mathematics measures with reliability estimates at or below 0.86 (k = 12) yielded a significantly lower mean effect size than studies using measures with reliability estimates higher than 0.86 (k = 13); a finding that held when the two influential effect size estimates were removed. A scatterplot of reliability coefficient as a continuous variable by Hedges' g suggests that a linear model is appropriate to represent the relationship between the dependent and predictor variable (see Appendix C). Specifically, as the reliability coefficient associated with the primary outcome measure goes up, there is a proportional increase in effect size.

Table 8

Summary of Effect Sizes by Outcome Measure Characteristics

Variable	Q_B	*k*	*g*	*SE*	95% CI
Type of Outcome Measure	0.85				
Standardized		4	0.68*	0.40	[-0.10, 1.46]
Researcher-developed		27	1.08*	0.17	[0.75, 1.41]
Reliability of Outcome Measure[a]	19.40**				
≤ 0.86		12	0.50*	0.09	[0.33, 0.66]
> 0.86		12	1.93*	0.33	[1.29, 2.58]
NR		7	0.52*	0.22	[0.08, 0.96]

Note. CI confidence interval; *k* = number of effect sizes; NR = not reported; Q_B= *Q Between*; Q_W = *Q Within;* [a] = statistic calculated without the NR category.

*p < .05; **p < 0.01

Table 9

Summary of Effect Sizes by Outcome Measure Characteristics After Removing Influential Points

Variable	Q_B	*k*	*g*	*SE*	95% CI
Type of Outcome Measure	0.05				
Standardized		4	0.68	0.40	[-0.10, 1.46]
Researcher-developed		25	0.77*	0.10	[0.58, 0.97]
Reliability of Outcome Measure[a]	25.12**				
≤ 0.86		12	0.50*	0.09	[0.33, 0.66]
> 0.86		10	1.25*	0.12	[1.00, 1.49]
NR		7	0.52*	0.22	[0.08, 0.96]

Note. CI confidence interval; *k* = number of effect sizes; NR = not reported; Q_B= *Q Between*; Q_W = *Q Within;* [a] = statistic calculated without the NR category.

*p < .05; **p < 0.01

Contextual characteristics of intervention. Tables 10 and 11 present the results of moderator analyses by contextual characteristics of intervention with and without the two influential points. The effectiveness of instruction varied as a function of interventionist both with (Q_B = 14.52, p = 0.006) and without the two influential effect size estimates (Q_B = 9.46, p = 0.050) (see Tables 8 and 9). Specifically, interventions implemented by researchers (g = 1.42, SE = 0.28) yielded significantly higher effect sizes than teacher-led interventions (g = 0.35, SE = 0.15) (Q_B = 11.31, p = 0.001). However, when this comparison was conducted with the trimmed data set, the difference was not significant at the adjusted level (α = 0.008) (Q_B = 6.43, p = 0.011).

Table 10
Summary of Effect Sizes by Contextual Characteristics of Interventions

Variable	Q_B	*k*	*g*	*SE*	95% CI
Interventionist[a]	14.52**				
Teacher		6	0.35*	0.15	[0.05, 0.66]
Researcher		14	1.42*	0.28	[0.88, 1.97]
Teacher & Researcher		6	1.04*	0.24	[0.58, 1.51]
Other (e.g., computer)		3	0.74	0.52	[-0.27, 1.75]
NR		2	0.50*	0.24	[0.02, 0.98]
Instructional Arrangement[a]	9.61*				
Whole class		12	1.62*	0.32	[0.99, 2.25]
Small group		14	0.70*	0.15	[0.41, 0.98]
One-on-one		3	0.34	0.28	[-0.21, 0.89]
NR		2	0.92*	0.46	[0.02, 1.82]
Instructional Setting[a]	10.40*				
General education classroom		11	1.85*	0.39	[1.09, 2.62]
Special education classroom		7	0.73*	0.21	[0.32, 1.13]
Other (outside of classroom)		7	0.55*	0.11	[0.32, 0.77]
NR		6	0.65*	0.23	[0.20, 1.10]
Mathematics Task	7.00*				
Arithmetic WPs (+, -)		12	0.67*	0.13	[0.41, 0.92]
Arithmetic WPs (x, ÷; all four operations)		13	1.61*	0.35	[0.93, 2.29]
Fraction, ratio and proportion, & algebra WPs		6	0.62*	0.17	[0.29, 0.95]
Intervention Duration[a]	16.08**				
< 10 hours (40-550 min)		10	0.44*	0.13	[0.19, 0.70]
≥10 to ≤ 16 Hours (600-955 min)		11	1.06*	0.17	[0.72, 1.41]
> 16 Hours (1,000-3,315 min)		8	1.92*	0.43	[1.08, 2.76]
NR		2	0.64	0.82	[-0.96, 2.24]

Note. CI confidence interval; *k* = number of effect sizes; NR = not reported; Q_B= *Q Between*; Q_W = *Q Within;* WPs = word problems; [a] = statistic calculated without the NR category.
*p < .05; **p < 0.01

Table 11
Summary of Effect Sizes by Contextual Characteristics of Interventions After Removing Influential Points

Variable	Q_B	*k*	*g*	*SE*	95% CI
Interventionist[a]	9.46*				
Teacher		6	0.35*	0.15	[0.05, 0.66]
Researcher		12	0.88*	0.14	[0.61, 1.14]
Teacher & Researcher		6	1.04*	0.24	[0.58, 1.51]
Other (e.g., computer)		3	0.74	0.52	[-0.27, 1.75]
NR		2	0.50*	0.24	[0.02, 0.98]
Instructional Arrangement[a]	4.04				
Whole class		10	0.95*	0.15	[0.65, 1.26]
Small group		14	0.70*	0.15	[0.41, 0.98]
One-on-one		3	0.34	0.28	[-0.21, 0.89]
NR		2	0.92*	0.46	[0.02, 1.82]
Instructional Setting[a]	6.23				
General education classroom		9	1.13*	0.21	[0.73, 1.53]
Special education classroom		7	0.73*	0.21	[0.32, 1.13]
Other (outside of classroom)		7	0.55*	0.11	[0.32, 0.77]
NR		6	0.65*	0.23	[0.20,1.10]
Mathematics Task	2.07				
Arithmetic WPs (+, -)		12	0.67*	0.13	[0.41, 0.92]
Arithmetic WPs (x, ÷; all four operations)		11	0.97*	0.21	[0.57, 1.38]
Fraction, ratio and proportion, & algebra WPs		6	0.62*	0.17	[0.29, 0.95]
Intervention Duration[a]	12.28**				
< 10 hours (40-550 min)		10	0.44**	0.13	[0.19, 0.70]
$\geq$10 to $\leq$ 16 Hours (600-955 min)		11	1.06**	0.17	[0.72, 1.41]
> 16 Hours (1,000-3,315 min)		6	0.97**	0.11	[0.76, 1.18]
NR		2	0.64	0.82	[-0.96, 2.24]

Note. CI confidence interval; *k* = number of effect sizes; NR = not reported; Q_B= *Q Between*; Q_W = *Q Within;* WPs = word problems; [a] = statistic calculated without the NR category.
*p < .05; **p < 0.01

For the variable, instructional arrangement, analysis with the complete data set indicated significant differences between the three levels of the intervention arrangement variable (whole class, small group, one-on-one). Post-hoc pairwise comparisons showed that the effect for "whole class" (*g* = 1.62, *SE* = 0.32) was significantly larger than either

"small group" (g = 0.70, SE =0.15) or "one-on-one" instruction (g =0.34, SE=0.28). After removal of the two influential effect size estimates, instructional arrangement no longer moderated effectiveness.

Instructional setting and mathematics task variables moderated effectiveness only when the two influential effect size estimates were included, and not after removal of the influential effect size estimates. Interventions implemented in general education classrooms yielded significantly larger effects (g = 1.85, SE = 0.39) than interventions conducted in either special education (g = 0.73, SE = 0.21) or "other" settings (g = 0.55, SE = 0.11). However, none of the comparisons were significant following removal of the two influential effect sizes. Similarly, with the complete data set, arithmetic word problem solving interventions involving all four basic operations (+, -, x, ÷) produced a significantly higher effect size (g =1.61, SE = 0.35) than did interventions involving only addition and subtraction word problems (g = 0.67, SE = 0.13) and interventions involving higher level mathematics (e.g., fractions, algebra) (g = 0.62, SE =0.17). However, the removal of the two influential effect size estimates reduced the effect size associated with "arithmetic word problems (+, -, x, ÷)" (g = 0.98, SE= 0.21) such that significant differences were no longer detected.

Intervention duration moderated word problem solving effectiveness both with and without the two influential effect sizes. Post hoc pairwise comparisons indicated that interventions lasting under 10 hours yielded a significantly lower effect size (g = 0.44, SE = 0.13) than interventions lasting between 10-16 hours (g = 1.06, SE = 0.17) – a comparison that remained unchanged after removal of the two influential effect sizes.

Further, studies lasting less than 10 hours yielded a significantly lower effect size than interventions that were over 16 hours in duration (g = 1.92, SE = 0.43)(Q_B = 10.87 p < 0.01). This finding held after the two influential effect sizes were removed (g = 0.98, SE = 0.11) (Q_B = 9.79, p < 0.01). No significant difference was found between studies lasting 10-16 hours and studies lasting over 16 hours (see Tables 10 and 11 for additional details). A scatterplot of average duration of each intervention as a continuous variable by Hedges' g suggests that a linear model is appropriate to represent the relationship between the dependent and predictor variable (see Appendix C). Specifically as the intervention duration increases, there is a proportional increase in effect size. In addition to analyzing the results of the Q_B statistics, which indicate when observed differences between subgroups of the moderator category are statistically significant, analyses of Q_W and I^2 statistics were performed to identify the presence of any additional variation within a moderator variable. As with all prior calculations, these were conducted both with and without the two influential effect size estimates, and results were compared (see Table 12). When all effect sizes were included, nine variables yielded significant Q_W statistics, and I^2 statistics indicating the existence of moderate heterogeneity (Higgins & Thompson, 2002). These variables are (1) the "published" level of the Type of Report variable, (2) the "classroom or school" level of the Random Assignment variable, (3) the "alternate intervention" level of the Type of Comparison Group variable, (4) the "elementary (k-6)" level of the Grade Level variable, (5) the "MD" level of the LD/MD Status variable, (6) the "whole class" level of the Instructional Arrangement variable, (7) the "researcher" level of the Interventionist variable, (8) the " > 16 hours" level of the

Intervention Duration variable and (9) the "researcher-developed test" level of the Type of Outcome Measure variable. These nine variables produced I^2 statistics between 41.15%-54.33%. The removal of the influential effect size estimates reduced the Q_W statistics of each of these moderator variable levels such that they were no longer significant and the associated I^2 statistics ranged from 0% -13.7%, which suggests that any variation among effect sizes for a given level of a moderator is attributable to sampling error.

Table 12
Q_W and I^2 Statistics by Moderator Variable

	Full Data Set (k=31)		Trimmed Data Set (k=29)	
Variable	Q_W	I^2 (%)	Q_W	I^2 (%)
LD/MD Status				
LD	15.71	17.27	15.71	17.27
MD	31.60*	49.38	15.07	7.07
Grade Level				
Elementary (K-6)	42.06*	42.94	25.49	13.70
Secondary (7-12)	5.88	15.01	5.88	15.01
Type of Report				
Published	42.48*	41.15	26.36	12.73
Unpublished	4.11	2.60	4.11	2.60
Random Assignment				
Student level	16.67	16.02	16.67	16.02
Classroom or school level	28.21*	46.83	13.93	6.68
Type of Comparison Group				
No intervention	0.00	0.00	0.00	0.00
Business-as-usual	26.41	31.84	17.00	13.12
Alternate intervention	21.52*	53.52	9.00	3.71
Interventionist				
Teacher	6.05	17.40	6.05	17.40
Researcher	26.14*	50.26	12.20	9.83
Teacher & Researcher	4.79	0.00	4.79	0.00
Other (e.g., computer)	2.43	17.79	2.43	17.79
NR	0.13	0.00	0.13	0.00

Table 12 (continued)

	Full Data Set (k = 31)		Trimmed Data Set (k=29)	
Variable	Q_W	I^2 (%)	Q_W	I^2 (%)
Instructional Arrangement				
Whole class	21.16*	48.01	9.46	4.89
Small group	15.94	18.45	15.94	18.45
One-on-one	2.06	2.75	2.06	2.75
NR	1.00	0.00	1.00	0.00
Instructional Setting				
General education classroom	15.11	33.81	8.73	8.34
Special education classroom	7.29	17.65	7.29	17.65
Other (outside of classroom)	5.94	0.00	5.94	0.00
NR	5.62	11.02	5.62	11.02
Mathematics Task				
Arithmetic WPs (+, -)	12.17	9.60	12.17	9.60
Arithmetic WPs (x, ÷; all four operations)	18.88	36.45	11.28	11.34
Fraction, ratio, proportion, algebra WPs	5.32	5.95	5.32	5.95
Intervention Duration				
< 10 hours (40-550 min)	9.92	9.28	9.40	4.26
≥10 to ≤ 16 Hours (600-955 min)	10.49	4.63	10.49	4.63
> 16 Hours (1,000-3,315 min)	15.33*	54.33	4.84	0.00
NR	1.00	0.00	1.00	0.00
Type of Outcome Measure				
Standardized test	2.87	0.00	2.87	0.00
Researcher-developed test	46.66*	46.27	27.17	11.65
Reliability of Outcome Measure				
Reliability ≤ 0.86	11.00	0.04	11.00	0.04
Reliability > 0.86	16.90	34.91	9.40	4.30
NR	5.70	0.00	5.70	0.00

Note. Q_W = *Q Within*

*p < .05; **p < 0.01

Chapter 5

DISCUSSION

This present meta-analysis (a) assessed the effectiveness of word-problem solving interventions on the mathematical performance of school-aged (K-12) students with LD and/or MD, (b) examined the relationship between intervention effectiveness and participant characteristics (e.g., grade level, LD/MD status), and (c) analyzed the relationship between intervention effectiveness and characteristics of study design, contextual characteristics of interventions, and outcome measure characteristics. A total of 28 studies conducted from 1987 – 2014 were located that met the criteria for inclusion. In this discussion, I summarize the findings, discuss implications of the findings, consider study limitations, and provide directions for future research.

What is the effectiveness of word problem-solving interventions on the mathematical performance of school-aged (K-12) students with LD and/or MD?

The results indicated a large grand mean effect for word problem solving interventions on the performance of students with LD or MD. Before removing the two influential effect size estimates, which were over four times the size of the mean effect size, the mean effect size was 1.03 [CI_{95} 0.73 1.32]. That is, on average, students in the treatment group scored 35 percentile points above students in the control group (WWC, 2014). After removing the two influential effect size estimates, the mean effect size was 0.77 [CI_{95} 0.57, 0.96] suggesting that students in the treatment group scored an average of 28 percentile points above those in the control group. Another interpretation of the improvement index is that, on average, 78% of those in the treatment group performed

above the mean of the control group (WWC, 2014). This finding is consistent with the results of prior meta-analyses (Gersten, Chard, et al., 2009; Kroesbergen & van Luit, 2003; Xin & Jitendra, 1999; Zhang & Xin, 2012; Zheng et al., 2013), which reported that (a) treatment outperformed control groups by 22 to 47 percentile points and (b) 72-97% of the treatment group performed above the mean of the control group.

Does intervention effectiveness vary as a function of participant characteristics?

Findings of moderator variable analyses suggested that studies with elementary school students yielded significantly higher effect sizes than studies with middle and/or high school students. However, after removing the two influential effect sizes, the difference was no longer significant; a finding that aligns with Xin and Jitendra (1999) who also did not detect a significant difference between elementary and secondary students. These results must be interpreted with caution considering the low power of moderator tests to detect departures from homogeneity (Cooper et al., 2009; Mittlböck & Heinzl, 2006). As such, failure to reject the null hypothesis of homogeneity does not provide strong evidence that there are not significant differences between sub-categories. That is, there were not sufficient studies at each grade level to examine the significant role of grade level. Further research with a larger database is needed to examine observed differences in mathematics performance by grade level.

Results of moderator analyses both with and without the two influential effect size estimates showed that LD/MD status did not moderate intervention effectiveness. This finding is consistent with that of Zhang and Xin (2012) but contradicts findings from the meta-analyses by Xin and Jitendra (1999) and Kroesbergen and Van Luit (2003), who

reported higher outcomes for students with MD than for students with LD. Although the finding of no differences between intervention effectiveness for students with LD and MD is encouraging, these results must be interpreted cautiously due to the small sample of studies. Eleven studies (see Table 1) from the research data base were not included in the present meta-analysis because they either (a) did not disaggregate data on students with LD or MD, or (b) reported that students were low performing but did not delineate the criteria for labeling students as MD (i.e., scored $\leq 35^{th}$ percentile on mathematics measures).

In addition to the issue of low power for moderator analyses, the continually evolving criteria for determining LD eligibility make it difficult to interpret these results. In the present meta-analysis, approximately one-third of the studies used only the state criteria to determine the presence of a learning disability, which, until the 2004 reauthorization of the Individuals with Disabilities Education Improvement Act (IDEA), depended solely on the IQ-achievement discrepancy model (Bradley, Danielson, & Doolittle, 2005). Since 2004, states may include response to intervention (RTI) in their eligibility determinations (Berkeley, Bender, Peaster, & Saunders, 2009). Yet, states vary in the degree to which they combine the traditional discrepancy model with RTI and in their implementation of RTI models (Ahearn, 2009; Berkeley et al., 2009; Fletcher & Vaughn, 2009; Zirkel, 2010). Thus, additional research is needed to determine the factors that affect research findings for students with LD and MD. These factors may include: (a) use of varying cutoff scores, (b) the degree to which RTI is used to determine eligibility, and (c) variation in RTI implementation.

With regard to sample demographics, there were not sufficient data to assess the potential moderating effect of SES or race on intervention effectiveness. Prior research suggests that studies with a high percentage of White students and students with high SES in the sample tend to be associated with higher effect sizes than studies with larger percentages of minority students and students with low SES (Dalton, 2011; Gonzales et al., 2008; National Center for Education Statistics, 2011). However, the moderating effects of SES and race were not tested in the present meta-analysis because 68% of the studies did not report SES data and 46% were missing data on race. Across the studies that reported race, there was extensive variation in the definition of categories. None of the studies reported racial categories fully aligned with census categories – White, Black or African American, Hispanic, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander (US Department of Health & Human Services, 2011). Four studies included in this meta-analysis coded race in terms of White, African-American, Hispanic, or Other (Fuchs, et al., 2008; Fuchs, Seethaler, et al., 2008; Fuchs et al., 2009; Xin et al., 2011). The remaining studies that reported race were increasingly less specific in their categories (i.e., White, African-American or Hispanic; White or African American; White or non-White). Such variations made it difficult to assess the extent to which race moderated intervention effectiveness.

The prevalence of using eligibility for FRL as a proxy for SES in educational research has been criticized (e.g., Harwell & LeBeau, 2010). Socioeconomic status (SES) is a complex index combining "educational attainment, occupational status, and family income" (Dalton, 2011, p. 11). As such, eligibility for FRL is not likely a valid indicator

of the economic resources available to the student (Harwell & LeBeau, 2010). All of the nine studies (32%) in the present meta-analysis that reported SES data used eligibility for FRL as a proxy for SES. Therefore, due to the small sample size and issues regarding use of SES measures that are not validated, I did not analyze the impact of participant SES on intervention effectiveness.

Does intervention effectiveness vary as a function of study design characteristics?

Results regarding the moderating impact of study design characteristics were mixed. For two study design characteristics, FOI and attrition, most studies did not report data to allow for moderator variable analyses. Two other design characteristics, type of report and level of random assignment, moderated effectiveness when the two influential effect sizes were included but not after their removal. The remaining study design characteristic, type of comparison group, did not moderate intervention effectiveness.

The fact that FOI data were only reported in 12 of the included studies is disconcerting. Specifics regarding instruction in both treatment and control groups provide critical information needed to establish that the measured outcomes "are the direct result of implementing a specified intervention" (Gersten et al., 2005, p.157). Toward this end, it is essential that studies include evidence regarding the reliability of implementation. Considerable research examining the research to practice gap in education has focused on the importance of faithfully implementing evidence-based practices in the classroom (e.g., Cook & Odom, 2013; Espin & Deno, 2001; Fixsen, Naoom, Blase, Friedman, & Wallace, 2005; Gersten & Dimino, 2001; Sindelar & Brown, 2001); and their findings show that differences in implementation fidelity impact the

magnitude of treatment effect. Similar to FOI, attrition data were reported in only 12 studies. It is important to report overall and differential attrition to ensure that the initial comparability between groups is maintained throughout the study (Gersten et al., 2005; WWC, 2014). If attrition rates are large overall, or differ substantially between groups, it is necessary to provide evidence that those who withdrew from the study were comparable at pretest to those who completed the study to ensure that the attrition is random and does not contribute to potential bias of the estimated effect (Trochim, 2006; WWC, 2014).

Results regarding the moderating effect of type of report (published vs. unpublished) were mixed. When all 31 independent effect sizes were included, published studies yielded significantly larger effect sizes than unpublished dissertations. However, after removing the two influential effect sizes, the difference between published and unpublished studies was no longer significant; a result that aligns with the subset of findings in education and special education literature, where effect sizes did not vary as a function of publication status (e.g., Swanson, 1999). Although I did not consistently find a significant difference between the effect sizes associated with published and unpublished studies, this finding must be interpreted with caution given the small sample of unpublished studies (k =5). As such, it was not possible to detect publication bias, which refers to the finding that positive and statistically significant findings tend to be published more frequently than negative or nonsignificant findings (Banks, Kepes & Banks, 2012; Ferguson & Heene, 2012; Rosenthal, 1979). Interestingly, an examination of the funnel plot (see Figure 3) from the present meta-analysis suggested possible

publication bias against studies with small sample size and non-significant effect sizes. I tried to minimize publication bias by conducting a thorough search for relevant unpublished studies. Following the recommendation of Harwell and Maeda (2008), I provided explicit details on the results of my search procedures to allow readers to assess the thoroughness of my search and facilitate replication.

Regarding the level of random assignment, results of analyses with 31 effect sizes indicated that estimated effects were significantly larger when random assignment occurred at the classroom/teacher level than at the student level. However, this finding was not robust following the removal of the two influential effect sizes. This suggests similar effect sizes are obtained whether studies employ random assignment at either the individual or teacher/classroom level; which aligns with the WWC (2014) research guidelines. For the remaining study design variable, type of comparison group (i.e., no intervention, BAU, alternate intervention), I did not detect significant differences with either the complete or trimmed data sets. This result is in contrast to research suggesting that when the comparison group shares features of high quality instruction with the treatment group, effect sizes tend to be lower (e.g., Gersten, Chard et al., 2009; Lemons, Fuchs, Gilbert, & Fuchs, 2014).

Does intervention effectiveness vary as a function of outcome measure characteristics?

As with study design characteristics (e.g., assignment to groups, attrition rate, FOI), the internal validity of a study is affected by the "quality of the measures selected or developed to evaluate intervention effects**"** (Gersten et al., 2005, p. 158). No

significant differences were found between standardized and researcher-developed measures. This finding does not align with prior research (e.g., Swanson & Hoskyn, 1998) indicating, "treatment effects were stronger on experimenter-developed measures than on standardized measures of the same construct" (Gersten, Baker, & Lloyd, 2000, p. 13). This discrepancy may be explained by not only the inherent low power of moderator tests, but also the small sample of studies using standardized assessments as the primary dependent outcome measure ($k = 5$).

It is considered best practice to provide details on the dependent measures used in intervention research, including calculation and reporting of psychometric properties (i.e., reliability, validity) (e.g., Gersten et al., 2005; Green, Chen, Helms, & Henze, 2011). Studies using measures with higher reliability estimates ($\alpha > 0.86$) were found to yield significantly larger effect sizes than those using measures with lower reliability estimates ($\alpha \leq 0.86$). Specifically, when the reliability estimate was $\alpha \leq 0.86$, between 63% and 75% of the treatment group outperformed the control group; whereas, in studies using measures with higher reliability estimates ($\alpha > 0.86$), between 84% and 99% of treatment students outperformed those in the control group. This finding was robust following the removal of the two influential effect sizes, indicating that studies with highly reliable measures are associated with strong treatment effects. This finding aligns with the notion that the use of measures with high reliability estimates "increases the power of a study" (Gersten et al., 2005, p. 159).

In addition to assessing the reliability of data obtained from dependent measures, evidence of the validity of that data is also essential (Gersten et al., 2005; Green et al.,

2011). Validity evidence of the primary outcome measure was only reported in six studies (Fuchs et al., 2008; Fuchs, Fuchs, Finelli, Courey & Hamlett, 2004; Fuchs, Fuchs, Hamlett & Appleton, 2002; Fuchs, Fuchs, Prentice, Hamlett, Finelli, & Courey, 2004; Walker & Poteet, 1989-90; Woodward & Baxter, 1997). As such, there were insufficient data to assess validity of measures as a potential moderator.

Does intervention effectiveness vary as a function of contextual characteristics of intervention?

Results regarding the moderating impact of contextual characteristics of intervention were mixed. Intervention duration moderated intervention effectiveness both with and without the two influential effect size estimates. The remaining four characteristics (interventionist, instructional arrangement, setting, and type of mathematics task) moderated effectiveness only when the two influential effect sizes were included but not after their removal. These mixed findings are not surprising given that the broader literature base on instruction for struggling students suggests difficulty disentangling the influence of instructional duration, setting, arrangement, and interventionist on student outcomes (Richards-Tutor, Baker, Gersten, Baker, & Smith, 2016).

Intervention duration moderated intervention effectiveness, both with and without the two influential effect sizes. Specifically, interventions lasting less than 10 hours were associated with an increase of eight to 26 percentile points; interventions lasting over 16 hours yielded improvements of 28 to 50 percentile points. The effectiveness of interventions ranging from 10-16 hours did not differ significantly from those over 16

hours in duration. These findings partially align with Xin and Jitendra (1999), who found that interventions lasting more than 30 sessions (longer than one month) were more effective than interventions lasting 1-7 sessions (one week or less). However, given that Xin and Jitendra (1999) did not calculate total minutes of instruction, their findings are not a direct comparison with those of the present meta-analysis.

It is important to note that the impact of duration of instruction depends on the quality of instruction or how instructional time is spent. If instruction is not effective, then more of the same instruction is not likely to predict better student performance. On the other hand, many students with LD and low achieving students benefit from explicit and systematic instruction, repetition of key material, frequent opportunities to respond, and corrective feedback (Gersten, Beckman, et al., 2009), all of which may be impacted by instructional setting and arrangement. For example, duration of instruction for struggling students receiving Tier 2 support would be longer as they typically receive supplemental instruction in addition to core mathematics instruction. Further, Tier 2 support often occurs in small group instructional arrangement outside of the general education classroom.

In terms of interventionist, an analysis with the complete data set showed that when researchers implemented, between 82% and 98% of students in the treatment group outperformed students in the control group. By contrast, when teachers implemented, between 2% and 75% of treatment students outperformed the control group. After removing the two influential effect sizes, the improvement index associated with researcher-led interventions suggested that between 73% and 87% of treatment students

outperformed the control. After adjusting the significance level for the six post-hoc comparisons, this difference between researcher- and teacher-led interventions was not statistically significant at the adjusted criterion of $p = 0.008$. That is, interventions tend to be equally effective when implemented by researchers, teachers, teachers and researchers combined, or others (e.g., volunteers. This finding, however, is in direct contrast to the substantial literature base on the research to practice gap in education, which suggests significant differences in implementation and associated student outcomes when interventions are implemented by researchers as compared to classroom teachers (e.g., Cook & Odom, 2013; Espin & Deno, 2001; Fixsen et al. 2005; Gersten & Dimino, 2001; Sindelar & Brown, 2001). Perhaps, the Dunn-Bonferroni method to adjust the critical *p*-value for multiple post-hoc comparisons provided an overly conservative estimate.

In terms of intervention setting, instruction that occurred in the general education classroom was associated with a greater improvement index (47 percentile points) than instruction provided in an alternate setting (e.g., library, computer lab, hallway), which resulted in an improvement index of 21 percentile points. This finding was not robust following the removal of the two influential effect size estimates. Similarly, the difference between general education classroom and special education classroom (favoring general education classes) was significant only when the two influential effect sizes were included, but not after their removal. These mixed results converge with the broader knowledge base examining the differential benefits of various instructional settings for students with disabilities – "efficacy research…which spans more than 3 decades, provides no compelling research evidence that place is the critical factor in the

academic or social progress of students with mild/moderate disabilities" (Zigmond, 2003, p.195).

Findings regarding instructional arrangement indicated that differences between whole-class arrangement and small-group or one-on-one instruction (favoring whole-class) were statistically significant when the two influential effect sizes were included, but not after their removal. Both of the influential effect size estimates came from studies having the same instructional arrangement (i.e., whole-class). Their removal from the data set lowered the mean effect size associated with whole-class instruction from g = 1.62 (SE = 0.32) to g = 0.95 (SE = 0.15). As such, the difference between the groups was no longer significant. Similar to the results for instructional setting, these mixed findings are not surprising. It may be that there is no one arrangement that is best for all students with LD or MD, because effectiveness of interventions may depend on several factors such as student characteristics (severity of academic difficulties, mastery of prerequisite skills), ease or difficulty of the to be learned content.

With regard to content domain, there is evidence that low-level tasks (e.g., problem solving involving whole numbers) serve as the foundation for high-level mathematics tasks involving fractions, proportions, and/or algebra (e.g., Berk, Taber, Gorowara, & Poetzl, 2009; Boyer, Levine, & Huttenlocher, 2008; NMAP, 2008). As such, it was expected that effect sizes associated with high-level mathematics tasks would be significantly lower than low level tasks. This expectation was met when all 31 independent effect sizes were included in the analysis but not after their removal. Specifically, effect sizes were significantly larger when the mathematics task involved

mathematics word problems using all four arithmetic operations but did not address higher-level concepts (i.e., fractions, proportions, algebraic skills). However, both studies yielding influential effect sizes were coded into the same category of mathematics task: arithmetic word problems involving all four operations. The removal of these two effect sizes reduced the mean for this category from g = 1.61 (SE = 0.35) to g= 0.98 (SE = 0.21) making between group differences nonsignificant. This finding must be interpreted with caution as only five studies addressed higher-level mathematics content.

In summary, the moderating impacts of contextual characteristics of interventions involve multiple variables. The mixed results of the moderator analyses for contextual characteristics on intervention aligns with the notion that these variables are intricately intertwined (Richards-Tutor, et al., 2016), and variation in one may lead to variation in others. Specifically, students with more severe deficits in mathematics may not benefit from the potentially positive aspects of inclusion (e.g., subject matter specialists teaching higher level content, access to typically achieving peers), at least not without intensive supplemental instruction or tutoring. Furthermore, the cumulative and multidimensional nature of mathematics content leads to increased variation regarding what material is prohibitively difficult for students to learn effectively in the general education classroom. For example, when learning fractions, a student with LD or MD might exhibit adequate progress in skills when provided with typical, Tier 1 instruction. However, when the content shifts to algebraic equations of increasing complexity, the same student may require a different approach, instructional setting, arrangement, and/or duration of instruction.

Instructional components. Perhaps the most important contextual characteristic of mathematics instruction is the nature and quality of instruction (Slavin & Lake, 2008; Slavin, Lake, & Groff, 2009). I did not attempt to statistically assess whether or not specific instructional components served to moderate intervention effectiveness due to extensive overlap of instructional components across studies. However, I provide tentative conclusions based on the frequency of certain instructional components across studies as well as the combination of multiple components within studies. Specifically, explicit instruction, representations, and problem types/problem structure were present in a majority of the included studies. This suggests that the following are key components of effective word problem instruction for students with LD or MD: (a) extensive modeling, think-aloud procedures, scaffolded instruction, and review with corrective feedback, (b) using visual and verbal representations, and (c) identifying specific problem types that share underlying common problem features. These conclusions align with the IES practice guide recommendations for improving mathematical problem solving: select "visual representations that are appropriate for students and the problems they are solving; Use think-alouds and discussions to teach students how to represent problems visually" (Woodward et al., 2012, p.45). Similarly, the findings of the present meta-analysis align with the following recommendations found in the Common Core Mathematical Practice Standards (2010): (a) model how to monitor and reflect on the problem-solving process, (b) identify important quantities in a situation and map their relationship onto diagrams, and (c) look closely at a problem to discern a pattern or a structure.

Limitations and Directions for Future Research

Several limitations of the present meta-analysis suggest caution in interpreting the findings. First, the relatively small sample size of 28 included studies hindered investigation of moderator variables. Despite an increase in the number of word problem solving intervention studies conducted over the years, there are still not sufficient studies that focus on students with LD and MD. I explicitly specified the population of students with MD by excluding studies that did not provide evidence of student performance below the 35th percentile on standardized, norm-referenced mathematics measures. Use of this criterion was important in specifying the population to which the results generalize, but also resulted in exclusion of a number of potentially important studies. In addition, across the studies targeting students with LD, there was considerable variation in criteria used to identify students. For example, only five studies (Baker, 1992; Fede et al., 2013; Fuchs, Seethaler, et al., 2008; Marzola, 1985; Powell & Fuchs, 2010) specified criteria for identifying the presence of mathematics disability using a standardized, norm-referenced mathematics measure (see Appendix D for details). This is problematic in generalizing the findings, because students with LD are a heterogeneous group and often vary in terms of their academic achievement across content area skills (e.g., decoding skills, reading comprehension, written expression, mathematics computation, mathematics reasoning). Future research should investigate word problem solving interventions for students with LD, who exhibit specific mathematics deficits.

A second limitation involves the prevalence of missing, incomplete, and/or inconsistent data reporting on critical student demographic information (e.g., race, SES),

which impeded the examination of potential moderating variables. This is problematic given that differences in educational achievement as a function of student race and SES dominate national news and drive educational reform. Further, the omission of attrition data, FOI details, and detailed evidence of the validity and reliability of outcome measures in many studies made it difficult to examine the potential impact of these variables on intervention effectiveness. Despite repeated calls for improved reporting practices in educational research (e.g., Gersten et al., 2000, 2005; Harwell & Maeda, 2008; Lemons et al., 2014; WWC, 2014), it appears that there is still room for improvement. Further research is warranted that includes (a) description of the population to which the study generalizes based on the sample by including specific criteria for inclusion and demographic details, (b) assessment of the reliability and validity of the data obtained from the outcome measures, (c) detailed description of FOI, and (d) assessment of overall and differential attrition. Also, future research should describe in greater detail the contextual characteristics of the intervention (i.e., duration, setting, arrangement, interventionist, instructional components) as these variables may interact in multiple ways to differentially influence student outcomes.

Third, due to the extensive overlap of instructional components, I did not attempt to isolate the unique amount of variance in mathematics performance associated with each individual instructional component. In addition, I did not statistically assess the impact of various clusters of instructional components to determine if their contribution as a group outweighs the contribution of the sum of each individual component. Considering the continually increasing amount of mathematics content that students are

exposed to in school within a finite time frame, future research regarding which instructional components or clusters of components confer the greatest educational benefit for most students would be beneficial to practitioners.

Fourth, two of the included studies yielded mean effect sizes over four times the grand mean effect size. When these two effect sizes were included in the analyses, the critical assumption of normality was violated, and there was significant heterogeneity within several of the potential moderator variables; both of which threaten the validity of conclusions (Cooper et al., 2009; Shadish, Cook, & Campbell, 2002). Therefore analyses were conducted with and without these two influential points. After removal of the two influential effect sizes, the data were more normally distributed, and homogeneous findings for potential moderating variables emerged, increasing the certainty of the empirical relation (Cooper et al., 2009). However, the cause for these anomalously large effect sizes remains unclear. I calculated the effect sizes for these two studies in several different ways (e.g., using the control *SD* instead of the pooled *SD* in the Cohen's *d* formula; using the differences-within-differences method) and the resultant effect sizes were still over four times the grand mean effect size. As such, it is critical to investigate studies for influential artifacts so these can be controlled for in future research. More importantly, if there is no identifiable flaw explaining the large effect sizes, then further examination into what makes these interventions work so well is certainly warranted. Considering that the highest effect sizes in the present meta-analysis came from the same research team, one might hypothesize that the researchers were especially adept at building rapport, engaging students, and clearly and succinctly explaining key concepts.

In addition, both researchers and teachers delivered the intervention in one of the two studies yielding influential effect size estimates. As such, it would be important to (a) understand the methods used by the researchers for positively impacting teacher implementers to change their instructional practices, and (b) consider the possible impact of the quality of the relationship between researchers and teachers.

Another key limitation is that the funnel plot (see Figure 3) suggests publication bias against studies with small sample sizes and non-significant effect sizes; therefore, there is likely a slight upward bias in the effect sizes in the present meta-analysis. Though I tried to minimize publication bias by searching for relevant unpublished studies, I undoubtedly omitted some of the grey literature by failing to augment my literature search by searching conference proceedings, searching research registers, and contacting researchers in the field (Cooper, et al., 2009; Ferguson & Heene, 2012). Nonetheless, location and inclusion of grey literature meeting inclusion criteria would not solve the larger issue of publication bias because "unless one knows the number of studies carried out for a particular intervention…regardless of final publication status, one will not know how many studies have been missed" (Cooper et al., 2009, p. 122).

As with recommendations for improved reporting practices (i.e., Gersten et al., 2000, 2005; Harwell & Maeda, 2008; Lemons et al., 2014; WWC, 2014), there have also been recommendations for possible solutions to the issue of publication bias. Namely, Cooper et al. (2009) recommended the establishment of research registries where researchers register their studies before they are completed so that synthesists could more accurately gauge the number of studies conducted on a topic. They noted that they hoped

that by the publication of the next edition of the text, they would be able to "report on the creation and growth of these structures" (p. 122). However, a recent meta-analysis (Suggate, 2016), reiterated the call for the creation of such a registry.

Finally, in addition to an overall dearth of rigorous and relevant intervention studies that are representative of the entire literature base, two areas were noticeably underrepresented in the literature: secondary grade level, and high-level mathematics tasks (e.g., proportions, algebra). Only six studies (21%) included in the present meta-analysis involved secondary students (grades 7-12). Of those, only four (Bottge & Hasslebring 1993; Hutchinson 1993; Lambert 1996; Konold 2004) addressed high-level mathematics content. Further, only one included study implemented with elementary level students included high-level mathematics tasks (Woodward et al., 2001). As such, there is a need for additional research with secondary students with LD and MD, and a need for more studies examining high-level mathematics instruction for struggling students in both elementary and secondary levels.

Conclusions

A key finding from this meta-analysis is that mathematics word problem-solving interventions are generally effective at improving WPS performance of school-aged students with LD or MD. Specifically, the interventions reviewed here predict, on average, a 28% improvement in percentile rank on mathematics WPS, and that 78% of students in the treatment group will outperform the control group. Second, the reliability of outcome measures and duration of intervention moderated intervention effectiveness. Intervention effectiveness did not appear to be moderated by the following variables: (a)

grade level of students, (b) LD/MD status of students, (c) type of report (published vs. unpublished), (d) level of random assignment to groups, (e) type of control group, (f) interventionist, (g) instructional arrangement (e.g., whole class, small group), (h) instructional setting (e.g., general education classroom, special education classroom), (i) type of mathematics word problems (e.g., arithmetic; fractions; algebra), or (j) type of measure (standardized vs. researcher developed). However the non-significant results of the moderator analyses must be interpreted cautiously as the relatively small number of included studies, coupled with missing data from many studies on key variables limited the investigation of moderator variables. Yet, the results serve to reinforce the need for additional, high-quality research on WPS interventions for students with LD and MD (i.e., Gersten et al., 2000, 2005; Lemons et al., 2014; WWC, 2014).

References

*included in meta-analysis

Ahearn, E. (2009). State eligibility requirements for specific learning disabilities. *Communication Disorders Quarterly, 30*(2), 120-128. doi: 10.1177/1525740108325221

Andersson, U. (2008). Mathematical competencies in children with different types of learning difficulties. *Journal of Educational Psychology*, *100*(1), 48–66. doi: 10.1037/0022-0663.100.1.48

Andersson, U., & Lyxell, B. (2007). Working memory deficit in children with mathematical difficulties: A general or specific deficit? *Journal of Experimental Child Psychology*, *96*(3), 197-228. doi: 10.1016/j.jecp.2006.10.001

Badian, N. A. (1983). Dyscalculia and nonverbal disorders of learning. In H. R. Myklebust (Ed.), *Progress In Learning Disabilities* (Vol. 5, pp. 235–264). New York: Stratton

Baker, D.E. (1992). *The effect of self-generated drawings on the ability of students with learning disabilities to solve mathematical word problem*s. (Doctoral dissertation).

Bandeira de Mello, V., Bohrnstedt, G., Blankenship, C., and Sherman, D. (2015). *Mapping State Proficiency Standards Onto NAEP Scales: Results From the 2013 NAEP Reading and Mathematics Assessments* (NCES 2015-046). U.S.

Department of Education, Washington, DC: National Center for Education Statistics. Retrieved [date] from http://nces.ed.gov/pubsearch

Banks, G. C., Kepes, S., & McDaniel, M. A. (2012). Publication Bias: A call for improved meta-analytic practice in the organizational sciences. *International Journal of Selection and Assessment*, *20*(2), 182-196. doi: 10.1111/j.1468-2389.2012.00591.x

Barbaresi, W.J., Katusic, S.K., Colligan, R.C., Weaver, A.L., & Jacobsen, S.J. (2005). Math learning disorder: Incidence in a population-based birth cohort, 1976–82, Rochester, Minn. *Ambulatory Pediatrics, 5*, 281–289. doi: 10.1367/A04-209R.1

Bennett, K. K. (1981). *The Effect of Syntax And Verbal Mediation on Learning Disabled Students' Verbal Mathematical Problem Scores.* Unpublished doctoral dissertation, Northern Arizona University, Flagstaff.

Berk, D., Taber, S. B., Gorowara, C. C., & Poetzl, C. (2009). Developing prospective elementary teachers' flexibility in the domain of proportional reasoning. *Mathematical Thinking and Learning, 11*, 113-135. doi:10.1080/10986060903022714

Berkeley, S., Bender, W. N., Peaster, L. G., & Saunders, L. (2009). Implementation of Response to Intervention A Snapshot of Progress. *Journal of Learning Disabilities*, *42*(1), 85-95. doi: 10.1177/0022219408326214

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis.* John Wiley & Sons. Ltd, Chichester, UK.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97-111. doi: 10.1002/jrsm.12

Borenstein, Hedges, Higgins & Rothstein (2015). Comprehensive Meta-Analysis (CMA) software package.

Borman, G. D., & D'Agostino, J. V. (1996). Title I and student achievement: A meta-analysis of federal evaluation results. *Educational Evaluation and Policy Analysis, 18,* 309-326. doi: 10.2307/1164335

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform and achievement: A meta-analysis. *Review of Educational Research, 73,* 125–230. doi: 10.3102/00346543073002125

Bottge, B., Heinrichs, M., Chan, S., & Serlin, R. (2001). Anchoring adolescents' understanding of math concepts in rich problem-solving environments. *Remedial and Special Education, 22,* 299–314. doi: 10.1177/074193250102200505

Bottge, B., Heinrichs, M., Chan, S., Mehta, Z., & Watson, E. (2003). Effects of video-based and applied problems on the procedural math skills of average and low-achieving adolescences. *Journal of Special Education Technology, 18,* 5–22. http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=eue&AN=507822535&site=ehost-live

Bottge, B., Heinrichs, M., Mehta, Z., & Hung, Y. (2002). Weighing the benefits of

anchored math instruction for students with disabilities in general education classes. *The Journal of Special Education, 35*, 186–200. doi: 10.1177/002246690203500401

Bottge, B., Rueda, E., LaRoque, P. T., Serlin, R. C., & Kwon, J. (2007). Integrating reform-oriented math instruction in special education settings. *Learning Disabilities Research & Practice, 22*, 96–109. doi: 10.1111/j.1540-5826.2007.00234.x

Bottge, B., Rueda, E., Serlin, R. Hung, Y., & Kwon, J. (2007). Shrinking achievement differences with anchored math problems: Challenges and Possibilities. *The Journal of Special Education, 41*, 31–49. doi: 10.1177/00224669070410010301

*Bottge, B.A. & Hasselbring, T.S. (1993). A comparison of two approaches for teaching complex, authentic mathematics problems to adolescents in remedial math classes. *Exceptional Children 59*(6), 556-566. http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=eue&AN=508476533&site=ehost-live

Bottge, B.A. (1999). Effects of contextualized math instruction on problem solving of average and below-average achieving students. *The Journal of Special Education 33*(2), 81-92. doi: 10.1177/002246699903300202

Boyer, T. W., Levine, S. C., & Huttenlocher, J. (2008). Development of proportional reasoning: Where young children go wrong. *Developmental Psychology, 44*, 1478-1490. doi:10.1037/a0013110

Bradley, R., Danielson, L. Doolittle, J. (2007). Responsiveness to intervention: 1997-2007. *Teaching Exceptional Children, 39*(5), 8-12. http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=eue&AN=507976006&site=ehost-live

Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meta-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology, 33*(4), 269-285. doi: 10.1016/0022-4405(95)00014-D

Casner-Lotto, J., & Barrington, L. (2006). *Are They Really Ready to Work? Employers' Perspectives on the Basic Knowledge and Applied Skills of New Entrants to the 21st Century US Workforce.* Partnership for 21st Century Skills. 1 Massachusetts Avenue NW Suite 700, Washington, DC 20001.

Cavanagh, S. (2007, June 12). What kind of math matters? *Education Week, 26*(40), 21-23. Retrieved from: http://www.edweek.org/ew/articles/2007/06/12/40math.h26.html

Common Core State Standards Initiative. (2010). *Common Core State Standards for mathematics*. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf

Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Chicago: Rand McNally, 1979.

Cook, B. G., & Odom, S. L. (2013). Evidence-based practices and implementation science in special education. *Exceptional Children, 79*(2), 135-144.

http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=eue&AN=84513025&site=ehost-live

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14*(2), 165-176. doi: 10.1037/a0015565

Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of educational psychology, 79*(4), 347-362. doi:10.1037/0022-0663.79.4.347

Cooper, H. Hedges, L.V., Valentine, J.C. (2009). *The Handbook of Research Synthesis and Meta-Analysis, 2nd ed.* New York: Russell Sage Foundation.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive psychology, 20*(4), 405-438. doi:10.1016/0010-0285(88)90011-4

Dalton, B. (2011). *US Educational Achievement on International Assessments: The Role of Race and Ethnicity.* RTI Press publication No. OP-0005-1105. Research Triangle Park, NC: RTI Press. Retrieved [11/27/13] from http://www.rti.org/rtipress

Depaepe, F., De Corte, E., & Verschaffel, L. (2010). Teachers' approaches towards word problem solving: Elaborating or restricting the problem context. *Teaching and Teacher Education: An International Journal of Research and Studies, 26*, 152-160. doi:10.1016/j.tate.2009.03.016

Desoete, A., Roeyers, H., & De Clercq, A. (2003). Can offline metacognition enhance mathematical problem solving? *Journal of Educational Psychology, 95*(1), 188-200. doi: 10.1037/0022-0663.95.1.188

DiPerna, J. C., Lei, P. W., & Reid, E. E. (2007). Kindergarten predictors of mathematical growth in the primary grades: An investigation using the Early Childhood Longitudinal Study--Kindergarten cohort. *Journal of Educational psychology, 99*(2), 369-379. doi: 10.1037/0022-0663.99.2.369

Espin, C.A., & Deno, S.L. (2000). Introduction to the special issue of learning disabilities research & practice: Research to practice: Views from researchers and practitioners. *Learning Disabilities Research & Practice, 15*(2), 67-68. doi: 10.1207/SLDRP1502_1

*Fede, J.L., Pierce, M.E., & Matthews, W.J. (2013). The effects of a computer-assisted, schema-based instruction intervention on word problem-solving skills of low-performing fifth grade students. *Journal of Special Education Technology, 28*(1), 9-21. doi: 10.1177/016264341302800102

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*(6), 555-561. doi: 10.1177/1745691612459059

Fixsen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation Research: A Synthesis of the Literature.* Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The

National Implementation Research Network (FMHI Publication #231). http://nirn.fmhi.usf.edu/resources/publications/Monograph/pdf/monograph_full.pdf

Fletcher, J. M., & Vaughn, S. (2009). Response to intervention: Preventing and remediating academic difficulties. *Child Development Perspectives*, *3*(1), 30-37. doi: 10.1111/j.1750-8606.2008.00072.x

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology, 97*(3), 493-513. doi: 10.1037/0022-0663.97.3.493

Fuchs, L. S., Fuchs, D. (2005). Enhancing mathematical problem solving for students with disabilities. *The Journal of Special Education, 39,* 45-57. doi: 10.1177/00224669050390010501

*Fuchs, L. S., Fuchs, D., Craddock, C., Hollenbeck, K. N., Hamlett, C. L., & Schatschneider, C. (2008). Effects of small-group tutoring with and without validated classroom instruction on at-risk students' math problem solving: Are two tiers of prevention better than one? *Journal of Educational Psychology, 100*(3), 491-509. doi: 10.1037/0022-0663.100.3.491

*Fuchs, L. S., Fuchs, D., Finelli, R., Courey, S. J., & Hamlett, C. L. (2004). Expanding schema-based transfer instruction to help third graders solve real-life mathematical problems. *American Educational Research Journal, 41*(2), 419-

445. doi: 10.3102/00028312041002419

*Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Appleton, A. C. (2002). Explicitly teaching for transfer: Effects on the mathematical Problem-Solving performance of students with mathematics disabilities. *Learning Disabilities Research & Practice, 17*(2), 90-106. doi: 10.1111/1540-5826.00036

Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., & Schroeter, K. (2003). Enhancing third-grade students' mathematical problem solving with self-regulated learning strategies. *Journal of Educational Psychology, 95*(2), 306-315. doi: 10.1037/0022-0663.95.2.306

Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., . . . Janacek, D. (2003). Explicitly teaching for transfer: Effects on third-grade students' mathematical problem solving. *Journal of Educational Psychology, 95*(2), 293-304. doi:10.1037/0022-0663.95.2.293

*Fuchs, L. S., Fuchs, D., Prentice, K., Hamlett, C. L., Finelli, R., & Courey, S. J. (2004). Enhancing mathematical problem solving among third-grade students with schema-based instruction. *Journal of Educational Psychology, 96*(4), 635-647. doi: 10.1037/0022-0663.96.4.635

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Hamlett, C. L., Seethaler, P. M., . . . Schatschneider, C. (2010). Do different types of school mathematics development depend on different constellations of numerical versus general cognitive abilities? *Developmental Psychology, 46*(6), 1731-1746. doi:

10.1037/a0020662

*Fuchs, L. S., Powell, S. R., Seethaler, P. M., Cirino, P. T., Fletcher, J. M., Fuchs, D., . . . Zumeta, R. O. (2009). Remediating number combination and word problem deficits among students with mathematics difficulties: A randomized control trial. *Journal of Educational Psychology, 101*(3), 561-576. doi:10.1037/a0014701

*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. doi: 10.1177/001440290807400202

Fuchs, L.S., Fuchs, D., & Compton, D.L. (2010). Rethinking response to intervention at middle and high school. *School Psychology Review, 39*(1), 22-28. http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=eue&AN=508142499&site=ehost-live

*Fuchs, L.S., Fuchs, D., & Prentice, K. (2004). Responsiveness to mathematical problem-solving instruction: Comparing students at risk of mathematics disability with and without risk of reading disability. *Journal of Learning Disabilities, 37*(4), 293-306. doi: 10.1177/00222194040370040201

Geary, D. C, Hoard, M. K., Byrd-Craven, J., Nugent, L, & Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematics learning disability. *Child Development, 78*(4), 1343–1359. doi:

10.1111/j.1467-8624.2007.01069.x

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities, 37*, 4–15. doi: 10.1177/00222194040370010201

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. doi: 10.1037/a0025398

Gersten, R., Baker, S., & Lloyd, J. W. (2000). Designing high-quality research in special education group experimental design. *The Journal of Special Education, 34*(1), 2-18. doi: 10.1177/002246690003400101

Gersten, R., Beckman, S., Clarke, B., Foegen, A., Marsh, L., Star, J., & Witzel, B. (2009). Assisting students struggling with mathematics: Response to intervention (RtI) for elementary & middle school. Institute of education science national center of educational evaluation & regional assistance. *Washington, DC: US Department of Education.*

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research, 79*(3), 1202-1242. doi: 10.3102/0034654309334431

Gersten, R. & Dimino, J. (2001). The realities of translating research into classroom practice. *Learning Disabilities Research & Practice, 16*(2), 120-130. doi:

10.1111/0938-8982.00013

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children, 71*(2), 149-164. doi: 10.1177/001440290507100202

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities, 38*(4), 293-304. doi: 10.1177/00222194050380040301

Ginsburg-Block, M., & Fantuzzo, J. (1998). An evaluation of the relative effectiveness of NVTM standards- based interventions for low-achieving urban elementary students. *Journal of Educational Psychology, 90*, 560–569. doi: 10.1037/0022-0663.90.3.560

Gleason, M., Carnine, D., & Boriero, D. (1990). Improving CAI effectiveness with attention to instructional design in teaching story problems to mildly handicapped students. *Journal of Special Education Technology, 10*(3), 129-36. http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=eue&AN=508383926&site=ehost-live

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2008). *Highlights from TIMSS 2007: Mathematics and science achievement of US fourth- and eighth-grade students in an international context* (NCES 2009-001). Washington, DC: National Center for Education Statistics, Institute of

Education Sciences, US Department of Education.

Gonzalez, J. E., & Espinel, A. I. (2002). Strategy choice in solving arithmetic word problems: Are there differences between students with learning disabilities, G-V poor performance, and typical achievement students? *Learning Disabilities Quarterly, 25*, 113–122. doi: 10.2307/1511278

Green, C.E., Chen, C.E., Helms, J.E., & Henze, K.T. (2011). Recent reliability reporting practices in psychological assessment: Recognizing the people behind the data. *Psychological Assessment, 23* (3), 656-669. doi: 10.1037/a0023089

Griffin, C. C., & Jitendra, A. K. (2009).Word problem-solving instruction in inclusive third-grade mathematics classrooms. *Journal of Educational Research, 102*, 187–202. doi: 10.3200/JOER.102.3.187-202

Griffin, S. A., Case, R., & Siegler, R. S. (1994). Rightstart: Providing the central conceptual prerequisites for first formal learning of arithmetic to students at risk for school failure. *In McGilly, K. (Ed.) Classroom Lessons: Integrating Cognitive Theory and Classroom Practice, pp. 25–49, Cambridge, MA:MIT Press.*

Gross-Tsur, V., Manor, O., & Shalev, R. S. (1996). Developmental dyscalculia: Prevalence and demographic features. *Developmental Medicine and Child Neurology, 38*, 25–33. doi: 10.1111/j.1469-8749.1996.tb15029.x

Hanich, L. B., Jordan, N. C., Kaplan, D., & Dick, J. (2001). Performance across different areas of mathematical cognition in children with learning difficulties.

Journal of Educational Psychology, 93(3), 615-626. doi: 10.1037/0022-0663.93.3.615

Hansen, N., Jordan, N. C., Fernandez, E., Siegler, R. S., Fuchs, L., Gersten, R., & Micklos, D. (2015). General and math-specific predictors of sixth-graders' knowledge of fractions. *Cognitive Development, 35*, 34-49. doi: 10.1016/j.cogdev.2015.02.001

Harwell, M. Maeda, Y. (2008). Deficiencies of reporting in meta-analyses and some remedies. *Journal of Experimental Education, 76* (4), 403-428. doi: 10.3200/JEXE.76.4.403-430

Harwell, M., & LeBeau, B. (2010). Student eligibility for a free lunch as an SES measure in education research. *Educational Researcher, 39*(2), 120-131. doi: 10.3102/0013189X10362578

Hasselbring, T., & Moore, P. (1996). Developing mathematical literacy through the use of contextualized learning environments. *Journal of Computing in Childhood Education, 7*, 199–222. http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=eue&AN=508582152&site=ehost-live

Hassinger-Das, B., Jordan, N. C., Glutting, J., Irwin, C., & Dyson, N. (2014). Domain-general mediators of the relation between kindergarten number sense and first-grade mathematics achievement. *Journal of Experimental Child Psychology, 118*, 78-92. doi:10.1016/j.jecp.2013.09.008

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6*(3), 203-217. doi: 10.1037/1082-989X.6.3.203

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9*(4), 426-445. doi: 10.1037/1082-989X.9.4.426

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*(7414), 557-560. doi: 10.1136/bmj.327.7414.557

Higgins, J.P.T., & Thompson, S.G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539-1558. doi: 10.1002/sim.1186

Howell, D. C. (2010). *Statistical Methods for Psychology.* Belmont, CA: Cengage Wadsworth.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods, 11*(2), 193-206. doi: 10.1037/1082-989X.11.2.193

*Hutchinson (1993). Effects of cognitive strategy instruction on algebra problem solving of adolescents with learning disabilities. *Learning Disability Quarterly, 16*(1), 34-63. doi: 10.2307/1511158

*Jitendra, A. K., Dupuis, D. N., Rodriguez, M. C., Zaslofsky, A. F., Slater, S., Cozine-Corroy, K., & Church, C. (2013). A randomized controlled trial of the impact of schema-based instruction on mathematical outcomes for third-grade students

with mathematics difficulties. *The Elementary School Journal*, *114*(2), 252-276. doi: 10.1086/673199

Jitendra, A. K., Griffin, C. C., Haria, P., Leh, J., Adams, A., & Kaduvettoor, A. (2007). A comparison of single and multiple strategy instruction on third-grade students' mathematical problem solving. *Journal of Educational Psychology*, *99*(1), 115-127. doi: 10.1037/0022-0663.99.1.115

*Jitendra, A. K., Griffin, C. C., McGoey, K., Gardill, M. C., Bhat, P., & Riley, T. (1998). Effects of mathematical word problem solving by students at risk or with mild disabilities. *The Journal of Educational Research, 91,* 345-355. doi: 10.1080/00220679809597564

Jitendra, A. K., Griffin, C., Haria, P., Leh, J., Adams, A., & Kaduvettoor, A. (2007). A comparison of single and multiple strategy instruction on third grade students' mathematical problem solving. *Journal of Educational Psychology, 99,* 115–127. doi: 10.1037/0022-0663.99.1.115

Jitendra, A.K., Griffin, C., Deatline-Buchman, A., & Sczesniak, E. (2007). Mathematical word problem solving in third-grade classrooms. *The Journal of Educational Research, 100,* 283–302. doi: 10.3200/JOER.100.5.283-302

Jitendra, A. K., Star, J. R., Starosta, K., Leh, J. M., Sood, S., Caskie, G., ... & Mack, T. R. (2009). Improving seventh grade students' learning of ratio and proportion: The role of schema-based instruction. *Contemporary Educational Psychology*, *34*(3), 250-264. doi: 10.1016/j.cedpsych.2009.06.001

Johnson, E. S., Humphrey, M., Mellard, D. F., Woods, K., & Swanson, H. L. (2010). Cognitive processing deficits and students with specific learning disabilities: A selective meta-analysis of the literature. *Learning Disability Quarterly*, *33*(1), 3-18. http://www.jstor.org.ezp2.lib.umn.edu/stable/25701427

Jonassen, D. H. (2003). Designing research-based instruction for story problems. *Educational Psychology Review, 15*(3), 267-296. doi: 10.1023/A:1024648217919

Jonassen, D. H. (2000). Toward a design theory of problem solving. *Educational technology research and development*, *48*(4), 63-85. doi: 10.1007/BF02300500

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

Jordan, N. G., Hanich, L. B., & Kaplan, D. (2003). Arithmetic fact mastery in young children: A longitudinal investigation. *Journal of Experimental Child Psychology*, *85,* 103–119. doi: 10.1016/S0022-0965(03)00032-8

Jordon, N. C., Kaplan, D., & Hanich, L. B. (2002). Achievement growth in children with learning difficulties in mathematics: Findings of a two-year longitudinal study. *Journal of Educational Psychology*, *94*(3), 586-597. doi: 10.1037/0022-0663.94.3.586

Kirsch, I., Braun, H., Yamamoto, K., & Sum, A. (2007). *America's Perfect Storm: Three Forces Changing Our Nation's Future*. Educational Testing Service.

Kolligian, J., & Sternberg, R. J. (1987). Intelligence, information processing, and specific learning disabilities: A triarchic synthesis. *Journal of Learning Disabilities*, *20*(1), 8-17. doi: 10.1177/002221948702000103

*Konold, K. B. (2004). *Using the Concrete-Representational-Abstract Teaching Sequence to Increase Algebra Problem-Solving Skills* (Doctoral dissertation).

Krawec, J. L. (2014). Problem representation and mathematical problem solving of students of varying math ability. *Journal of Learning Disabilities, 47*(2), 103-115. doi: 10.1177/0022219412436976

Kroesbergen, E. H., & Van Luit, J. E. H. (2003). Mathematics interventions for children with special educational needs. *Remedial and Special Education, 24*(2), 97-114. doi: 10.1177/07419325030240020501

Lackaye, T.D., & Margalit, M. (2006). Comparisons of achievement, effort, and self-perceptions among students with learning disabilities and their peers from different achievement groups. *Journal of Learning Disabilities*, *39*(5), 432-446. doi: 10.1177/00222194060390050501

*Lambert, M. (1996). *Teaching students with learning disabilities to solve word-problems: A comparison of a cognitive strategy and a traditional textbook method* (Doctoral dissertation).

Lang, C. (2001). *The effects of self-instructional strategies on problem solving in algebra I for students with special needs* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 30000279)

*Lee, J.W. (1992) *The effectiveness of a novel direct instructional approach on math word problem solving skills of elementary students with learning disabilities.* (Doctoral dissertation).

Leh, J. M., & Jitendra, A. K. (2013). Effects of computer-mediated versus teacher-mediated instruction on the mathematical word problem-solving performance of third-grade students with mathematical difficulties. *Learning Disability Quarterly, 36*(2), 68-79. doi: 10.1177/0731948712461447

Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world reconsidering the counterfactual in education research. *Educational Researcher, 43*(5), 242-252. doi: 0013189X14539189

Lesh, R., Hamilton, E., & Kaput, J. (Eds.) (2007). *Foundations for the Future in Mathematics Education.* Mahwah, NJ: Lawrence Erlbaum Associates

Levy, F., & Murnane, R. J. (2004). Education and the changing job market. *Educational Leadership, 62*(2), 80-84. http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=eue&AN=507935789&site=ehost-live

Lewis, C., Hitch, G. I., & Walker, P. (1994). The prevalence of specific arithmetic difficulties and specific reading difficulties in 9 to 10 year old boys and girls. *Journal of Child Psychology and Psychiatry, 35*, 283-292. doi: 10.1111/j.1469-7610.1994.tb01162.x

Lewis, A. B., & Mayer, R. E. (1987). Students' miscomprehension of relational

statements in arithmetic word problems. *Journal of Educational Psychology, 79*(4), 363-371. doi: 10.1037/0022-0663.79.4.363

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. doi: 10.1177/0022219408321126

Lucangeli, D., Tressoldi, P. E., & Cendron, M. (1998). Cognitive and metacognitive abilities involved in the solution of mathematical word problems: Validation of a comprehensive model. *Contemporary Educational Psychology, 23*(3), 257-275. doi: 10.1006/ceps.1997.0962

Maccini, P., & Ruhl, K. L. (2001). Effects of a graduated instructional sequence on the algebraic subtraction of integers by secondary students with learning disabilities. *Education and Treatment of Children, 23,* 465 – 489.

*Marzola, E. (1985). *An arithmetic problem solving model based on a plan for steps to solution, mastery learning, and calculator use in a resource room setting for learning disabled students*. (Doctoral dissertation).

Mayer, R. E., Lewis, A. B., & Hegarty, M. (1992). Mathematical misunderstandings: Qualitative reasoning about quantitative problems. *Advances in Psychology, 91*, 137-153. doi: 10.1016/S0166-4115(08)60886-9

Mayer, R. E. & Hegarty, M. (1996). The process of understanding mathematical problem solving. In R. J. Sternberg & T. Ben-Zeev (Eds.), *The nature of mathematical thinking* (pp. 29–54). Hillsdale, NJ, England: Lawrence Erlbaum

Associates, Inc.

Mayer, R. E. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instructional Science, 26,* 49–63. doi: 10.1023/A:1003088013286

Mazzocco, M. M., & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research & Practice, 20*(3), 142-155. doi: 10.1111/j.1540-5826.2005.00129.x

Mazzocco, M.M.M. (2007). Defining and differentiating mathematical learning disabilities and difficulties. In Berch, D.B., & Mazzocco, M.M.M. (Eds). *Why is Math So Hard for Some Children? The Nature and Origins of Mathematical Learning Difficulties and Disabilities*, pp. 29-47, Baltimore, MD: Paul H Brookes Publishing.

Mazzocco, M. M., & Devlin, K. T. (2008). Parts and 'holes': Gaps in rational number sense among children with vs. without mathematical learning disabilities. *Developmental Science, 11*(5), 681-691. doi: 10.1111/j.1467-7687.2008.00717.x

Mittlböck, M., & Heinzl, H. (2006). A simulation study comparing properties of heterogeneity measures in meta-analyses. *Statistics in Medicine, 25*(24), 4321-4333. doi: 10.1002/sim.2692

Montague, M., & Applegate, B. (1993). Mathematical problem solving characteristics of middle school students with learning disabilities. *Journal of Special Education, 27,* 175–201. doi: 10.1177/002246699302700203

Montague, M., Applegate, B., & Marquard, K. (1993). Cognitive strategy instruction and mathematical problem-solving performance of students with learning disabilities. *Learning Disabilities Research & Practice, 8*(4), 223-232.

Montague, M., Enders, C., & Dietz, S. (2011). Effects of cognitive strategy instruction on math problem solving of middle school students with learning disabilities. *Learning Disabilities Quarterly, 34*(4), 262-272. http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=eue&AN=525908204&site=ehost-live

Moore, L. J., & Carnine, D. (1989). Evaluating curriculum design in the context of active teaching. *Remedial and Special Education, 10*(4), 28-37. doi: 10.1177/074193258901000406

*Moran, A. S., Swanson, H. L., Gerber, M. M., & Fung, W. (2014). The effects of paraphrasing interventions on problem-solving accuracy for children at risk for math disabilities. *Learning Disabilities Research & Practice*, 29(3), 97-105. doi: 10.1111/ldrp.12035

Murphy, M. M., Mazzocco, M. M., Hanich, L. B., & Early, M. C. (2007). Cognitive characteristics of children with mathematics learning disability (MLD) vary as a function of the cutoff criterion used to define MLD. *Journal of Learning Disabilities, 40*(5), 458-478. doi: 10.1177/00222194070400050901

National Center for Education Statistics (2011). *The Nation's Report Card: Mathematics 2011*(NCES 2012–458). Institute of Education Sciences, U.S.

Department of Education, Washington, D.C.

National Center for Education Statistics (2013). *The Nation's Report Card: A First Look: 2013 Mathematics and Reading* (NCES 2014-451). Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics*. Reston, VA: Author.

National Mathematics Advisory Panel. (2008). *Foundation for success: The final report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

National Research Council. (2001). *Adding It Up: Helping Children Learn Mathematics.* J. Kilpatrick, J. Swafford, and B. Findell (Eds.). Mathematics Learning Study Committee, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Noll, R. S. (1983). *Effects of verbal cueing and a visual representation on percent problem-solving performance of remedial adults*. Unpublished doctoral dissertation, Fordham University, New York.

OECD (2010). *The High Cost of Low Education Performance: The Long-Run Economic Impact of Improving Educational Outcomes*. Paris.

Onatsu-Arvilommi, T., & Nurmi, J.-E. (2000). The role of task-avoidant and task-focused behaviors in the development of reading and mathematical skills during the first school year: A cross-lagged longitudinal study. *Journal of Educational*

Psychology, 92(3), 478-491. doi:10.1037/0022-0663.92.3.478

*Owen, R. L., & Fuchs, L. S. (2002). Mathematical problem-solving strategy instruction for third-grade students with learning disabilities. *Remedial and Special Education, 23*(5), 268-278. doi: 10.1177/07419325020230050201

Pai, H. H., Sears, D. A., & Maeda, Y. (2014). Effects of small-group learning on transfer: A meta-analysis. *Educational Psychology Review, 27*(1), 79-102. doi: 10.1007/s10648-014-9260-8

Petitti, D. B. (2001). Approaches to heterogeneity in meta-analysis. *Statistics in Medicine, 20*(23), 3625-3633. doi: 10.1002/sim.1091

Pólya, G. 1945. *How to Solve It: A New Aspect of Mathematical Model*. Princeton University Press, Princeton, NJ.

*Powell, S. R., & Fuchs, L. S. (2010). Contribution of equal-sign instruction beyond word-problem tutoring for third-grade students with mathematics difficulty. *Journal of Educational Psychology, 102*(2), 381-394. doi: 10.1037/a0018447

Powell, S. R., Fuchs, L. S., Fuchs, D., Cirino, P. T., & Fletcher, J. M. (2009). Effects of Fact Retrieval Tutoring on Third-Grade Students with Math Difficulties with and without Reading Difficulties. *Learning Disabilities Research & Practice, 24*(1), 1-11. doi: 10.1111/j.1540-5826.2008.01272.x

Price, G. R., & Ansari, D. (2013). Dyscalculia: Characteristics, causes, and treatments. *Numeracy, 6*(1), 2-16. doi: 10.5038/1936-4660.6.1.2

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to

categorize statistics word problems. *Journal of Educational Psychology, 88*(1), 144-161. doi: 10.1037/0022-0663.88.1.144

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Richards-Tutor, C., Baker, D.L., Gersten, R., Baker, S.K., & Smith, J.M. (2016). The effectiveness of reading interventions for English language learners: A research synthesis. *Exceptional Children, 82*(2), 144-169. doi: 10.1177/0014402915585483

Rivera-Batiz, F. L. (1992). Quantitative literacy and the likelihood of employment among young adults in the United States. *The Journal of Human Resources, 27*, 313–328. doi: 10.2307/145737

Rosenthal, R. (1979). The ‘‘file drawer problem’’ and tolerance for null results. *Psychological Bulletin, 86*, 638–641. doi: 10.1037/0033-2909.86.3.638

Schumacher, R.F., & Fuchs, L.S. (2012). Does understanding relational terminology mediate effects of intervention on compare word problems? *Journal of Experimental Child Psychology, 111*, 607-628. doi: 10.1016/j.jecp.2011.12.001

Schurter, W. A. (2002). Comprehension monitoring: An aid to mathematical problem solving. *Journal of Developmental Education, 26*(2), 22-29. http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=eue&AN=507794430&site=ehost-live

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

*Shiah, R. L., Mastropieri, M. A., Scruggs, T. E., & Fulk, B. J. M. (1994). The effects of computer-assisted instruction on the mathematical problem solving of students with learning disabilities. *Exceptionality, 5*(3), 131-161. doi: 10.1207/s15327035ex0503_2

Sindelar, P.T., & Brown, M.T. (2001). Research to practice dissemination, scale and context: We can do it, but can we afford it? *Teacher Education and Special Education, 24* (4), 348-355. doi: 10.1177/088840640102400408

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research, 78*(3), 427-515. http://www.jstor.org/stable/40071135

Slavin, R. E., Lake, C., & Groff, C. (2009). Effective programs in middle and high school mathematics: A best-evidence synthesis. *Review of Educational Research*, *79*(2), 839–911. http://www.jstor.org.ezp2.lib.umn.edu/stable/40469058

Sowder, J. T., Philipp, R. A., Armstrong, B. E., & Schappelle, B. P. (1998). *Middle grade teachers' mathematical knowledge and structuring content in mathematics: A research monograph*. New York: State University of New York Press.

Stein, M. K., & Lane, S. (1996). Instructional tasks and the development of student capacity to think and reason: An analysis of the relationship between teaching and learning in a reform mathematics project. *Educational Research and Evaluation, 2*(1), 50-80. doi: 10.1080/1380361960020103

Stellingwerf, B.P., & Van Lieshout, E. (1999). Manipulatives and number sentences in computer aided arithmetic word problem solving. *Instructional Science, 27*, 459-476. doi: 10.1007/BF00891974

Suggate, S.P. (2016). A meta-analysis of the long-term effects of phonemic awareness, phonics, fluency, and reading comprehension interventions. *Journal of Learning Disabilities, 49*(1), 77-96. doi: 10.1177/0022219414528540

Swanson, H. L. (1999). *Interventions for students with learning disabilities: A meta-analysis of treatment outcomes*. NY: Guilford Press.

Swanson, H. L., & Hoskyn, M. (1998). Experimental intervention research on students with learning disabilities: A meta-analysis of treatment outcomes. *Review of Educational Research, 68*(3), 277-321. doi: 10.3102/00346543068003277

*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 & Practice, 28*(4), 170-183. doi: 10.1111/ldrp.12019

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology:*

General, 119(2), 176-192. doi: 10.1037/0096-3445.119.2.176

Thorndike, R.M., & Thorndike-Christ, T. (2010). *Measurement and Evaluation in Psychology and Education* (8th ed.) Boston, MA: Pearson Education Inc., publishing as Merrill.

Toppel, C. D. (1996). *The effect of a diagramming instructional strategy on error patterns, problem resolution, and self-perception of mathematical ability of community college students with learning disabilities.* (Ed.., University of San Francisco). ProQuest Dissertations and Theses, Retrieved from http://search.proquest.com.ezp3.lib.umn.edu/docview/304351021?accountid=14586

Torgesen, J. K. (1994). Issues in the assessment of executive function: An information-processing perspective. In G. R. Lyon (Ed.), *Frames of Reference for the Assessment of Learning Disabilities: New Views on Measurement Issues* (pp. 143–162). Baltimore, MD: Paul H. Brookes.

Trochim, William M. (2006). The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: http://www.socialresearchmethods.net/kb/

Troff, D. (2004). *An explicit instruction design approach for teaching students with learning disabilities to solve mathematical problems involving proportions* (Master's thesis).Available from ProQuest Dissertations and Theses database. (UMI No. 1422331)

Van de Walle, J.A., Karp, K.S & Bay-Williams, J.M. (2013). *Elementary and Middle*

school Mathematics Teaching Developmentally (8th ed.). New Jersey: Pearson Education, Inc.

Verschaffel, L., Greer, B., & De Corte, E. (2007). Whole number concepts and operation. In F. K. Lester Jr. (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 557–628). National Council of Teachers of Mathematics. Charlotte, NC: Information Age.

Voutsina, C. (2012). Procedural and conceptual changes in young children's problem solving. *Educational Studies in Mathematics, 79*, 193-214. doi: 10.1007/s10649-011-9334-1

Vukovic, R. K., & Siegel, L. S. (2010). Academic and cognitive characteristics of persistent mathematics difficulty from first through fourth grade. *Learning Disabilities Research & Practice, 25*, 25–38. doi: 10.1111/j.1540-5826.2009.00298.x

*Walker, D.W. & Poteet, J.A. (1989). A comparison of two methods of teaching mathematics story problem–solving with learning disabled students. *National Forum of Special Education Journal, 1*(1), 44-51.

WWC (2011). *What Works Clearinghouse Procedures and Standards Handbook, Version 2.1.*

WWC (2014). *What Works Clearinghouse Procedures and Standards Handbook, Version 3.0.*

Whetzel, D. (1992). *The Secretary of Labor's Commission on Achieving Necessary*

Skills. Retrieved from ERIC database (ED339749). Washington DC: US Department of Labor.

*Wilson, C. L., & Sindelar, P. T. (1991). Direct instruction in math word problems: Students with learning disabilities. *Exceptional Children, 57*(6), 512-519. http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=eue&AN=508354942&site=ehost-live

*Woodward, J. & Baxter, J. (1997). The effects of an innovative approach to mathematics on academically low achieving students in mainstreamed settings. *Exceptional Children, 63*(3), 373-388. http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=eue&AN=507573537&site=ehost-live

Woodward, J., Beckmann, S., Driscoll, M., Franke, M., Herzig, P., Jitendra, A., … Ogbuehi, P. (2012). *Improving Mathematical Problem Solving In Grades 4 Through 8: A Practice Guide* (NCEE 2012-4055). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications_reviews.aspx#pubsearch/

*Woodward, J., Monroe, K., & Baxter, J. (2001). Enhancing student achievement on performance assessments in mathematics. *Learning Disabilities Quarterly, 24*(1), 33-46. doi: 10.2307/1511294

Wortman, P. M., & Bryant, F. B. (1985). School desegregation and Black achievement:

An integrative review. *Sociological Methods and Research, 13*, 289–324. doi: 10.1177/0049124185013003002

Xin, Y. P., & Jitendra, A. K. (1999). The effects of instruction in solving mathematical word problems for students with learning problems: A meta-analysis. *The Journal of Special Education, 32*(4), 207-225. doi: 10.1177/002246699903200402

*Xin, Y.P., Jitendra, A.K., & Deatline-Buchman, A. (2005). Effects of mathematical word problem-solving instruction on middle school students with learning problems. *The Journal of Special Education, 39*(3), 181-192. doi: 10.1177/00224669050390030501

*Xin, Y. P., Zhang, D., Park, J. Y., Tom, K., Whipple, A., & Si, L. (2011). A comparison of two mathematics problem-solving strategies: Facilitate algebra-readiness. *The Journal of Educational Research, 104*(6), 381-395. doi: 10.1080/00220671.2010.487080

Yadrick, R.M., Regian, J.W., Connolly-Gomez, C., Robertson-Schule, L. (1997). Dyadic vs. individual practice with exploratory and directive mathematics tutors. *Journal of Educational Computing Research, 17*(2), 165-185. doi: 10.2190/1LVL-N5PK-M9QM-RXFF

Ysseldyke, J. E., Algozzine, B., Shinn, M. R., & McGue, M. (1982). Similarities and differences between low achievers and students classified learning disabled. *The Journal of Special Education, 16*(1), 73-85. doi:

10.1177/002246698201600108

Zawaiza, T. R. W., & Gerber, M. M. (1993). Effects of explicit instruction on math word-problem solving by community college students with learning disabilities. *Learning Disabilities Quarterly, 16*(1), 64-79. doi: 10.2307/1511159

Zhang, D., & Xin, Y.P. (2012). A follow-up meta-analysis for word-problem-solving interventions for students with mathematics difficulties. *The Journal of Educational Research, 105*(5), 303-318. doi: 10.1080/00220671.2011.627397

Zheng, X., Flynn, L. J., & Swanson, H. L. (2013). Experimental intervention studies on word problem solving and math disabilities: A selective analysis of the literature. *Learning Disability Quarterly, 36*(2), 97-111. doi: 10.1177/0731948712444277

Zheng, X., Swanson, H. L., & Marcoulides, G. A. (2011). Working memory components as predictors of children's mathematical word problem solving. *Journal of Experimental Child Psychology*, *110*(4), 481-498. doi: 10.1016/j.jecp.2011.06.001

Zigmond, N. (2003). Where should students with disabilities receive special education services? Is one place better than another? *The Journal of Special Education, 37*(3), 193-199. doi: 10.1177/00224669030370030901

Zirkel, P. A. (2010). The legal meaning of specific learning disability for special education eligibility. *Teaching Exceptional Children*, *42*(5), 62-68. http://login.ezproxy.lib.umn.edu/login?url=http://search.ebscohost.com/login.as

px?direct=true&AuthType=ip,uid&db=eue&AN=508160420&site=ehost-live

Appendix A

WPS intervention studies included in prior meta-analyses

WPS Intervention Studies	Kroesbergen & Van Luit (2003)	Gersten, Chard, et al. (2009)	Xin & Jitendra (1999)	Zhang & Xin (2012)	Zheng, et al. (2013)	Present meta-analysis
Baker (1992)		x	x			x
Bottge & Hasselbring (1993)			x			x
Bottge (1999)				x		
Fuchs, Fuchs, Hamlett, & Appleton (2002)		x		x		x
Fuchs, Fuchs, Finelli, Courey, & Hamlett (2004)				x		x
Fuchs, Fuchs, Craddock, et al. (2008)				x		x
Fuchs, Fuchs, & Prentice (2004)		x			x	x
Fuchs, Fuchs, Prentice, Burch, Hamlett, Owen, & Schroeter (2003)				x		
Fuchs, Fuchs, Prentice, Burch, Hamlett, Owen, Hosp, & Janacek (2003)				x		
Fuchs, Fuchs, Prentice, Hamlett, Finelli, & Courey (2004)				x		x
Fuchs, Seethaler, et al. (2008)				x		x
Gleason, Carnine, & Boriero (1990)			x			
Hutchinson (1993)		x	x			x
Jitendra, et al. (2009)				x		
Jitendra, Griffin, Deatline-Buchman & Sczeniak (2007)				x		

Jitendra, Griffin, Haria, Leh, Adams, & Kaduvettoor (2007)				x		
Jitendra, Griffin, McGoey, & Gardill (1998)	x	x		x		x
Konold (2004)				x		x
Lambert (1996)				x		x
Lang (2001)				x		
Lee (1992)		x	x			x
Marzola (1987)		x	x			x
Montague, Applegate, & Marquard (1993)			x			
Moore & Carnine (1989)			x			
Owen & Fuchs (2002)		x		x		x
Shiah, Mastropieri, Scruggs & Fulk (1994-1995)	x		x		x	x
Walker & Poteet (1989-1990)		x	x		x	x
Wilson & Sindelar (1991)	x	x	x		x	x
Woodward & Baxter (1997)	x				x	x
Woodward, Monroe, & Baxter (2001)		x			x	x
Xin, Jitendra, & Deatline-Buchman (2005)		x		x	x	x

Appendix B

Coding Sheet

Name of Article:

Title of Groups:	***Treatment***		***Control***	
N				
	M	**SD**	**M**	**SD**
PRETEST				
Posttest				
Description of intervention & control:				

Variable name	**Variable description**	**Variable codes/categories**	**Source Data**	**Page #(s)**
EISD	**Effect size ID**	5 digit ID		
Doc Yr	**Year Document appeared**	4 digit year		
Doc Type	**Type of Document**	1= journal article 2= conference paper 3= gov't/project report 4= dissertation/master's thesis 5 = other		
DocSource	**Source of document**	1=PsycINFO 2=ERIC/EBSCO 3=More than one of above 4=Ancestral search		
Dependent variable information				**Page #(s)**
Variable	**Variable description**	**Variable codes/categories**	**Source Data**	
MAType	**Type of math achievement measure**	**1=** standardized test **2=** teacher/researcher created test **3=** both **4=** author did not specify (include qualitative notes/details below)		
MArelYN	**Reliability reported?**	0= no 1= yes		
MArel	**Math achievement measure reliability**	#		
MAval	**Math achievement validity reported?**	0= no 1= yes		
Student Sample Info				**Page #(s)**
Variable	**Variable description**	**Variable codes/categories**	**Source Data**	
FemaleN	**Number female**	**#(%)**		
LD N	**# (%) students LD**	**#(%)**		

At risk N	**# (%) students at risk for MD**	**#(%)**			
LDCrit	**LD**	**0=** NA, no LD students **1=** cut score on screening test/math achievement measure AND state criteria **2=** cut score on screener only (authors specified students were LD but only provided screening data) **3=** state criteria only (authors specified students were LD—may or may not cite specific state criteria) **4=** Other (if other, include qualitative notes/details below)			
	LD cut score details: type of measure used	**0=** NA, no LD students OR no screening measure **1=** researcher developed measure **2=**standardized measure **3=** other (if other, include qualitative notes/details below)			
	LD cut score details: cut score used	**0=** NA, no LD students OR no screening measure **1=** percent correct/incorrect **2=** percentile **3=** other (if other, include qualitative notes/details below)			
	LD cut score:	List #, %, percentile or NR for not reported			
ARCrit	**At risk/MD criteria/ definition**	**0=** NA, no AR students **1=** cut score on screening test/math achievement measure **2=** low achievement in math (e.g., teacher referral, recommended for remedial classes, poor grades in math previously)			
	At risk/MD cut score details: type of measure used	**0=** NA, no AR students OR no screening measure cut score reported **1=** researcher developed measure **2=** standardized measure **3=** other (if other, include qualitative notes/details below)			
	At risk/MD details: cut score used	**0=** NA, no AR students OR no screening measure cut score reported **1=** percent correct/incorrect **2=** percentile **3=** other (if other, include qualitative notes/details below)			

	AR cut score:	NR for not reported List #, %, percentile		
White	**% Students White**	**#**		
Asian	**% Students Asian**	**#**		
Black	**% Students Black**	**#**		
Hispanic	**% Students Hispanic**	**#**		
Other	**% students other ethnicities**	**#**		
NonWhite	**% Students NonWhite**	**#(%)**		
StudGrd	**Student grade level**	1 =grades K-5 2 = grades 6-8 3 = grades 9-12 4 = mixed grades 5= author did not specify		
SES rep	**SES status reported?**	0= no 1= yes		
SampSES	**SES of sample**	% FRL		
Attrition	**Attrition reported?**	0= no 1= yes		
SampAttr	**Attrition**	% attrition		
Methodological Information				**Page #(s)**
Variable name	**Variable description**	**Variable codes/categories**	**notes**	
Assign	**Assignment to groups**	**1=** RA at student level (includes matched pairs) **2=** RA at classrm/teacher lvl **3=** RA at school level **4=** No RA (convenience sample; intact groups) **5=** author did not specify (include details below)		
Pretest	**Affirmed group equivalence on pretest scores (OR Pretest differences controlled for)**	0= no 1= yes		
Independent Variable Information				**Page #(s)**
Variable name	**Variable description**	**Variable codes/categories**	**notes**	
IMPL	**Implementer/ interventionist**	**0=**teacher **1=**researcher **2=** mix (teachers and researchers) **3=** computer **4=** other (volunteers, paraprofessionals) **5=** author did not specify (include qualitative notes/details below)		

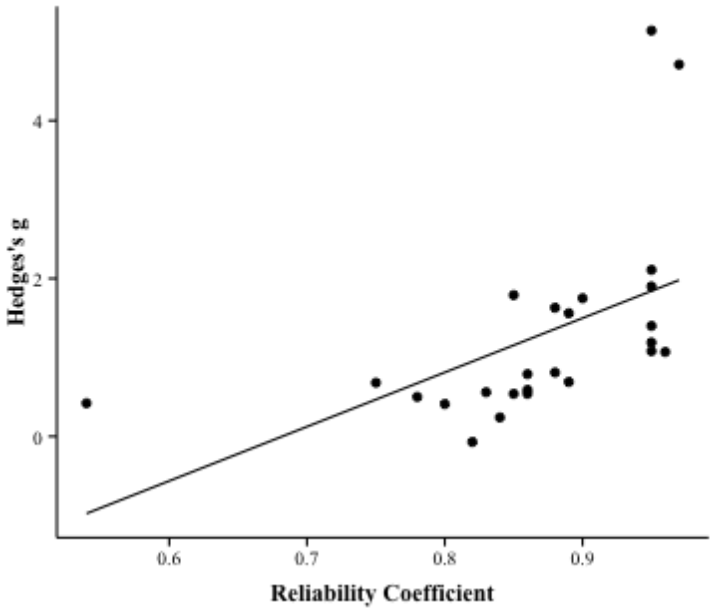
intSett	**Setting of intervention**	**1=** general education class **2=** special education class/resource room **3=** other (e.g., in library, outside of classroom) **4=** author did not specify (include qualitative notes/details below)		
insArr	**Instructional arrangement**	**1=** whole class **2=** small group **3=**one-on-one **4=** more than one of above (include qualitative notes/details below) **5=** author did not specify (include qualitative notes/details below)		
control	**Type of control/comparison**	**0=** no intervention (e.g., attention only) **1=** business as usual control group BAU (textbook instruction as per schools guidelines) **2=** alternate intervention (researcher created; may be based on textbooks, but is more controlled than BAU) **3=** author did not specify (include qualitative notes/details below)		
minINS	**Minutes of instruction**	NR if author does not report **#** (days x min per day spent on intervention—not counting pre and posttest)		
<u>Calculations used to determine min of instruction:</u>				
Type prob	**Type of math problems**	**1=** only addition, subtraction, or addition & subtraction (does not include x, ÷) (e.g., addition and subtraction one-step and two-step problems involving *Group* or *Parts and Total*, *Change*, and *Compare* or *Comparison* problems) **2=** more than just add and/or subtract, but still arithmetic only (+, -, x, ÷) (e.g., cover stories for addition and subtraction and cover stories for multiplication and division problems) **3=** higher level than simple arithmetic (e.g., algebra story problems, word problems involving geometry equations like perimeter = 4s; word problems involving ratios, proportions, fractions) **4=** other (include qualitative notes/details below) **5 =** author did not specify (include qualitative notes/details below)		

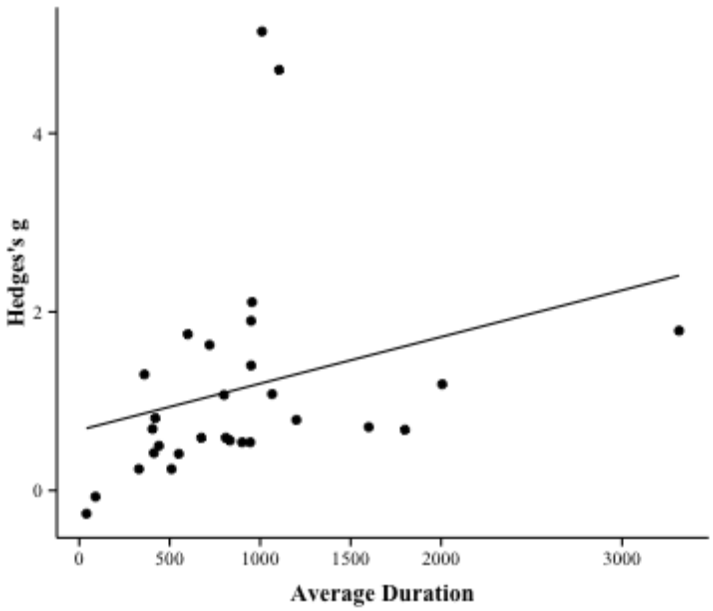
FOI	**Fidelity of Implementation Reported?**	0= no 1= yes		
FOI#	**Fidelity of Implementation**	#(%)		

InstrComp	**Instructional components**	1. Representations 2. Metacognition 3. General problem solving procedures 4. Explicit instruction 5. Prerequisite/foundational skills 6. Underlying problem structure 7. Teach for transfer 8. Contextualized approach 9. Student centered instruction 10. Peer-assisted instruction 11. Other (include qualitative notes/details below)		

Appendix C

Scatterplots of reliability and duration of intervention by Hedges' *g*





Appendix D

LD criteria by study

Studies including students with LD	LD criteria reported
Baker (1992)	State criteria: (a) IQ ≥ average, (b) discrepancy ≥ 19 standard score points between IQ and achievement (c) evidence of one or more significant disorders in the essential learning processes which are manifested by reading, writing, spelling, or mathematics disabilities (4) the elimination of other handicapping condition(s) as the cause of the learning disorder (Educational Standards for New Mexico Schools, 1990). "In addition…each student demonstrated a discrepancy of 15 or more points between his/her expectancy score, as measured on the Wechsler Intelligence Test for Children - Revised (WISC-R) (Wechsler, 1974), and his/her standard score on a test of math achievement" (Baker, 1992, p. 33). Cut scores: On grade level in simple computation skills as measured by Diagnostic Screening Test: Math The Basic Processes Section ≤ 75% correct on researcher-developed WPS pretest.
Bottge & Hasselbring (1993)	State criteria and teacher recommendation
Fede et al. (2013)	State criteria; mathematics goal in IEP Cut score: ≤ 30th percentile on GMADE Process and Applications subtest
Fuchs et al. (2002)	State criteria Cut score: 1.5 *SD* below regional normative sample on a researcher-developed test of computational fluency
Fuchs, Fuchs, Finelli, et al. (2004)	State criteria Cut score: ≤ 25th percentile on researcher-developed WPS measure

Fuchs, Seethaler, et al. (2008)	State criteria Cut score: < 26th percentile on Wide Range Achievement Test–Revised arithmetic and reading
Hutchinson (1993)	State criteria: discrepancy of more than three years on a standardized mathematics achievement test Cut scores: ≥ 80% correct on researcher-developed computation test ≥ 60% correct on researcher-developed test of one-step word problems involving multiplication < 40% correct on researcher-developed test of algebra word problems
Jitendra et al. (1998)	State criteria Cut scores: ≥ 90% on researcher-developed simple computation measure ≥ 90% researcher-developed simple action problems measure ≤ 60% on researcher-developed one-step WPS pretest
Konold (2004)	State criteria and history of low achievement in mathematics
Lambert (1996)	State criteria: "Specific learning disability is defined by the State of Florida as: 1. A disorder in one or more of the basic psychological processes involved in understanding or in using spoken or written language. Disorders may be manifested in listening, thinking, reading, talking, writing, spelling, or arithmetic. Such disorders do not include learning problems, which are due primarily to visual, hearing, or motor handicaps, to mental retardation, to emotional disturbance, or to an environmental deprivation. 2. A student is eligible for special programs for specific learning disabilities if the student meets all of the following criteria: (a) evidence of a disorder in one or more of the basic psychological processes which include visual, auditory, motor, and language processes. (b) Evidence of academic achievement, which is significantly below the student's level of intellectual functioning. (c) Evidence that learning problems are not due primarily to other handicapping conditions. (d) Documented evidence which indicates that general educational alternatives have been attempted and found to be ineffective in meeting the student's educational needs" (Lambert, 1996, p. 11).

	Cut score: ≥ 80% correct on researcher-developed computation pretest (with use of calculator)
Lee (1992)	State criteria only
Marzola (1985)	State criteria and recommendation of special education teacher Cut scores: ≤ 4 months below grade level in computation on Key Math Test ≥1 year below grade level in arithmetic WPS on Key Math Test
Powell & Fuchs (2010)	State criteria: "school-identified disability" (Powell & Fuchs, 2010, p. 385) Cut scores: < 26th percentile on the Arithmetic subtest of the Wide Range Achievement Test–Revised < 36th percentile and the Math Problem Solving and Data Interpretation subtest of the Iowa Test of Basic Skills
Shiah et al. (1994)	State criteria Cut score: ≤ 33% on researcher-developed WPS pretest
Walker & Poteet (1989)	State criteria Cut scores: ≥ 90% on researcher-developed computation screening test ≤ 60% on researcher-developed two-step WPS pretest
Wilson & Sindelar (1991)	State criteria only
Woodward & Baxter (1997)	State criteria and receiving special education services in mathematics
Woodward, et al. (2001)	State criteria and receiving special education services in mathematics

Xin et al. (2005)	State criteria and teacher recommendation Cut score: ≤ 70% on researcher-developed WPS pretest (multiplication and division word problems)
Xin et al. (2011)	State criteria Cut score: ≤ 70% on researcher-developed WPS pretest (multiplication and division word problems)