Journal for Research in Mathematics Education
2023, Vol. 54, No. 1, 24–42

Word Problem Performance of U.S. First Graders in the 20th Century and Common Core Era

Robert Schoen and Ian Whitacre
Florida State University

Zachary Champagne
The Discovery School

Changes in U.S. textbooks indicate that U.S. first-grade students in the Common Core era were exposed to a wider variety of word problem types than students in previous generations were. We compared the performance of U.S. first graders in the Common Core era with that of previous generations in solving 11 types of additive word problems to investigate a decades-long debate—whether certain types of word problems are inherently more difficult than others or whether relative difficulty is influenced by exposure. We found that overall patterns of relative difficulty persist; however, U.S. first graders in the Common Core era outperformed their historical counterparts when solving the types of problems that rarely appeared in textbooks used in the 1980s.

Keywords: Word problems; Cognitively guided instruction; Problem solving; Mathematics; Curriculum

Curriculum research employs a wide variety of traditions and approaches. Some studies compare international curricula, and the scope of those international comparison studies range from big-picture perspectives to sharply focused ones (Fan & Zhu, 2007; Fuson et al., 1988; Hong & Choi, 2014; Mayer et al., 1995; Schmidt et al., 2002; Schoen et al., 2011; Son, 2012; Stigler et al., 1986; Tarr et al., 2008). Some studies examine readily observable features of curricula, whereas others explore more abstract or subtle features (Cai et al., 2002, 2010; Herbel-Eisenmann, 2007; Hsu & Silver, 2014; Mesa, 2004; Smith et al., 2016; White & Mesa, 2014). Other studies compare the efficacy of different approaches to the design and implementation of textbooks or other curriculum resources with respect to their effects on key outcomes, such as student learning (Agodini & Harris, 2010; Grouws et al., 2013; Remillard et al., 2014; Sarama et al., 2008; Schoen & Koon, 2021; Tarr et al., 2013). Still others have contributed theoretical frameworks, such as the conceptualization by Remillard and Heck (2014) of the many interacting facets of curriculum. Regardless of differences in focus or approach taken in the research, scholars widely accept that curriculum resources exert a major influence on teaching and learning.

In this article, we focus attention on additive word problems in first-grade textbooks in the United States and student performance on those word problems. First, we discuss the increased exposure to word problems in the Common Core era (CCE), summarize some of the relevant literature about types of word problems, and provide a synthesis of empirical findings related to the relative difficulty of various types of word problems for U.S. students during the latter part of the 20th century. We then describe an intranational comparison study wherein we examine whether those same patterns of relative problem difficulty continue to be observed during the CCE. The findings provide new empirical insights about theoretical models concerning the inherent difficulty of various types of word problems and how exposure to different types of word problems may affect their difficulty. Because the changes in textbooks were promoted, at least in part, by policy decisions, these new empirical findings have implications about policy decisions and their subsequent effects on instructional materials and student learning.

Background

During the latter half of the 20th century, scholars identified consistent patterns in relative difficulty of various types of word problems through empirical study of children's performance solving those problems and generated theory to explain the differences in difficulty on the basis of inherent features of those problem types (Fuson, 1992; Riley & Greeno, 1988). Stigler et al. (1986) observed that the inherent features of the problems—which were thought to exert a major influence on their relative difficulty—were conflated with exposure, because the three easiest types of problems for students to solve were also the three predominant types in U.S. textbooks of the 1980s. More recently, scholars have presented results from observational studies, primarily using international comparisons, that appear to support the interpretation that variation in exposure to the different types of problems could be the dominant factor in determining their relative difficulty.

In the 2000s, mathematics curriculum in the U.S. was frequently criticized for being "a mile wide and an inch deep" (Schmidt et al., 1997, p. 120). This prompted policy recommendations in favor of narrowing the focus in the K–12 mathematics curriculum to enable a more thorough treatment of a fewer number of topics in a more coherent system (National Council of Teachers of Mathematics, 2006). Subsequent revisions of state curriculum standards in the late 2000s and early 2010s resulted in removal or reduction of topics in the state curriculum standards in the U.S. in many grade levels (Schmidt & Houang, 2012; Schoen et al., 2011). For example, with the adoption of the Common Core State Standards for Mathematics (CCSSM; National Governors Association Center for Best Practices [NGA Center] & Council of Chief State School Officers [CCSSO], 2010),

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

most states removed probability from their K–2 curriculum standards. The removal of some topics from the intended curriculum created more space for the remaining topics, thereby enabling the inclusion of aspects of mathematics that had not traditionally been included. For example, CCSSM include a taxonomy of additive word problems, including the three traditional types identified by Stigler et al. (1986) and eight (or more) other types of additive word problems that were not historically included in the U.S. curriculum. Critics of the curricular changes attributed to CCSSM expressed concern that the emphasis on nontraditional tasks and a higher priority on mathematical practices such as reasoning and communication would result in decreased learning of traditionally emphasized mathematics topics (Otten & de Araujo, 2015).

Widespread adoption of CCSSM occurred in the U.S. in the early 2010s (LaVenia et al., 2015). Recent research has established that CCE first-grade textbooks included a greater number and more balanced distribution of additive word problems than did U.S. first-grade textbooks of the 1980s (Schoen et al., 2021). These curricular differences present an opportunity to investigate whether, or to what extent, previously established patterns of word problem difficulty still hold when differences in exposure exist. Specifically, the textbooks of the 1980s were dominated by three traditional problem types, which were also found to be the easiest types for first graders to solve (relative to eight other problem types on which first graders' performance was studied extensively). Thus, the disparity in difficulty between the traditional and nontraditional problem types (i.e., those problem types that featured prominently in the textbooks of the 1980s vs. those that did not) is of particular interest.

Factors That Determine the Relative Difficulty of Word Problems

A long history of scholarly interest has been focused on identifying the important factors that influence children's ability to solve word problems. Many factors have been hypothesized, including the size of the numbers, grammar or vocabulary in the problems, the age or development level of the child, and psychological aspects such as working memory, fact-recall ability, and specific cognitive structures (Baroody, 1987; Caldwell & Goldin, 1979; Carpenter et al., 2015; Fuchs et al., 2006; Jerman & Mirman, 1974; Nesher et al., 1982).

Two factors form the basis of a widely acknowledged taxonomy consisting of 11 additive[1] word problem types: *semantic structure* (i.e., whether quantities are combined, changed, or compared) and *location of the unknown* (Nesher et al., 1982). For example, in the problem, "Tom had 12 candies. He ate 5 of the candies. How many candies does Tom have now?," an action of eating changes the original number of candies, and the unknown is the number of candies that result from this change. This problem type contrasts with problems in which the values of the quantities do not change, such as the following: "Tom has 12 candies. His friend Sally has 5 candies. How many more candies does Tom have than Sally?"

Several decades of basic research conducted through the middle and latter part of the 20th century resulted in the taxonomy of additive word problems presented in Figure 1. Researchers have reported consistent patterns in the relative difficulty of various types of word problems in the taxonomy. For example, problems involving change situations with the unknown as the resulting quantity tend to be easier for young children to solve than problems that have the unknown in a different location (Carpenter et al., 1981; Carpenter & Moser, 1983; Fuson, 1992; Verschaffel et al., 2007).

Figure 1

Additive Word Problem Types

Join	*Result Unknown* Sofia saw a group of eight manatees in the river. Five more manatees joined the group. How many manatees are in the group now?	*Change Unknown* Sofia saw a group of eight manatees in the river. Some more manatees joined the group. Now there are 13 manatees in the group. How many manatees joined the group?	*Start Unknown* Sofia saw a group of manatees in the river. Five more manatees joined them. Now there are 13 manatees in the river. How many manatees were in the group to start with?
Separate	*Result Unknown* Sofia had 13 oranges. She ate five oranges. How many oranges does she have left?	*Change Unknown* Sofia had 13 oranges. She ate some oranges. Now she has eight oranges. How many oranges did she eat?	*Start Unknown* Sofia had some oranges. She ate five of the oranges. Now she has eight oranges. How many oranges did she have to start with?
Part-Part-Whole	*Whole Unknown* Sofia sees eight alligators on a log and five alligators in the river. How many alligators does Sofia see?	*Part Unknown* Sofia sees 13 alligators. Eight of the alligators are on a log, and the rest are in the river. How many alligators does Sofia see in the river?	
Compare	*Difference Unknown* Sofia has 13 oysters. Luka has eight oysters. How many more oysters does Sofia have than Luka?	*Compare Quantity Unknown* Luka has eight oysters. Sofia has five more than Luka. How many oysters does Sofia have?	*Referent Unknown* Sofia has 13 oysters. She has five more than Luka. How many oysters does Luka have?

Note. Reprinted from Schoen et al. (2021, p. 112).

[1] We use the term *additive*, in contrast with *multiplicative*, to convey the idea that these problems may be solved by simple addition or subtraction of two quantities (Van Dooren et al., 2010; Verschaffel et al., 1999). In other words, the units for all the quantities involved are the same (e.g., all three quantities in the cited problems about Tom are numbers of candies).

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

One explanation for the observed patterns in relative problem difficulty is based on the cognitive structures that researchers hypothesize are necessary for children to solve problems of a particular type (Nesher et al., 1982). Related is research concerning learning trajectories of children's strategies for solving world problems, which illustrates the point that certain problems are more difficult to solve for students who are using less advanced strategies (Carpenter et al., 2015; Clements & Sarama, 2014). According to such accounts, relative problem difficulty may be understood as a natural consequence of children's cognitive development and an inherent property of the type of word problem.

Other research findings suggest that differences in exposure to problems of various types may explain much of the variability in relative problem difficulty across cultures. Research involving students in Cyprus, Turkey, the U.S., and China indicated a positive association between the prevalence of certain types of word problems in textbooks and student performance on those types of word problems (Christou & Philippou, 1998; Olkun & Toluk, 2002; Stigler et al., 1986; Xin, 2007). Thus, relative word problem difficulty is potentially influenced more by the amount of children's exposure to or experience solving problems of certain types than it is by the semantic structure and position of the unknown in the various types of problems. A major limitation of the evidence cited here, however, is that it depends on cross-cultural comparisons, and different results may be influenced by many cultural factors (Wang & Lin, 2005).

Therefore, two competing hypotheses attempt to explain the underlying reasons for the relative difficulty among types of problems in the additive word problem taxonomy. The *inherence hypothesis* maintains that a combination of semantic structure and the location of the unknown quantity makes some types of problems inherently more difficult than others. The *exposure hypothesis*, by contrast, suggests that differential exposure to various types of problems explains their relative difficulty. However, differences in language and other experiences of children in countries represented in international comparisons have limited the ability of previous studies to isolate exposure to various problem types as the explanatory factor for differences in children's performance.

Recent policy shifts in the U.S. created an opportunity to compare relative problem difficulty within the same country in two different eras. The word problems included in U.S. first-grade textbooks of the 1980s were almost exclusively of the three traditional problem types, whereas CCE textbooks include a much more balanced distribution of problem types (Schoen et al., 2021). Thus, we see clear evidence of a case of policy influencing curriculum and have the opportunity to investigate the relationship between curriculum and student achievement.

The purpose of the present article is to perform a synthesis of historical data related to the performance of U.S. first graders on the 11 types of additive word problems in the taxonomy and to examine available empirical evidence to evaluate whether the estimates of relative difficulty of traditional and nontraditional types of problems differ significantly between the historical period and the CCE. In particular, the following research questions (RQs) guided our study.

RQ1. How do the frequency and distribution of additive word problems in mathematics textbooks used by a large sample of U.S. first-grade students during the 2010s compare with those used by U.S. first-grade students in the 1980s?

RQ2. What was the approximate difficulty[2] of each of the 11 types of additive word problems for U.S. first-grade students during the period 1966–1991?
   a. What were the discernable patterns in relative problem difficulty?
   b. How did the difficulty of the traditional problem types compare with that of the nontraditional problem types?

RQ3. What was the approximate difficulty of each of the 11 types of additive word problems for U.S. first-grade students (who used the same textbook analyzed in RQ1) during the 2010s?
   a. What were the discernible patterns in relative problem difficulty?
   b. How did the difficulty of the traditional problem types compare with that of the nontraditional problem types for these students?

RQ4. How does the relative difficulty of traditional versus nontraditional problem types for a sample of U.S. first-grade students in the 2010s (RQ3b) compare with that during the period 1966–1991 (RQ2b)?

[2] In this study, *problem difficulty* is empirically determined by the proportion of students who solve a problem correctly, which is consistent with the use of the term in psychometric practice that is based on classical test theory. Higher percent correct is interpreted as lower difficulty; lower percent correct is interpreted as higher difficulty.

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

Research Related to Relative Word Problem Difficulty

In this section, we review extant research related to word problem difficulty. First, we summarize the research that culminated in a widely used taxonomy of 11 additive problem types. Second, we review findings concerning relative difficulty by problem type. Third, we summarize literature indicating cross-cultural differences in patterns of word problem difficulty. In this review, our purpose is not to argue for one hypothesis over the other; it is to provide readers with a basic introduction and summary of the ideas and evidence presented in the corpus of literature concerned with categories and relative difficulty of additive word problems.

Problem Type Categorizations

Researchers have studied various aspects of word problems and categorized them in different ways. Features of interest have included action versus nonaction, abstract versus concrete, factual versus hypothetical, consistent versus inconsistent language, additive versus multiplicative, number of steps involved in solving the problem, location of the unknown quantity, mode of presentation, number of words, parts of speech, verbal cues or keywords, classes of numbers involved, and presence of distractors (e.g., Caldwell & Goldin, 1979; Carpenter et al., 1988, 2015; Gibb, 1956; Jerman & Mirman, 1974; Lewis & Mayer, 1987; Stigler et al., 1986). Whereas adults may simply categorize additive word problems by whether they can be solved by addition or subtraction of the given numbers, young children tend to think differently depending on specific problem characteristics (Carpenter & Moser, 1983; Carpenter et al., 1982). Thus, this line of research has sought to identify distinctions in problem type that affect how students approach the problems and which types tend to be more or less difficult to solve correctly.

These efforts culminated in the widely acknowledged taxonomy presented in Figure 1, which organizes 11 types of one-step, additive word problems on the basis of semantic structure and location of the unknown (Carpenter et al., 2015). In particular, this taxonomy consists of two types of change problems—Join and Separate—and two types of problems involving static relationships among sets that do not change—Part-Part-Whole and Compare—together with all possible locations for one unknown. This scheme emerged in the 1980s and 1990s on the basis of decades of research into children's mathematical thinking conducted worldwide. The taxonomy[3] of problem types in Figure 1 has influenced professional development programs (Carpenter et al., 1996, 2015; Fennema et al., 1996), further research on word problem difficulty (e.g., García et al., 2006; Olkun & Toluk, 2002; Stigler et al., 1986), and policy documents (e.g., NGA Center & CCSSO, 2010).

Throughout this manuscript, we refer to the following three problem types as *traditional*: Join Result Unknown (JRU), Separate Result Unknown (SRU), and Part-Part-Whole Whole Unknown (PWU). We describe the other eight problem types as *nontraditional*. We have three main reasons for this classification. First, JRU, SRU, and PWU problems are consistent with the meanings most commonly associated with addition and subtraction. Second, they are presented straightforwardly (with the result or whole being the unknown). Compare Difference Unknown (CDU) problems are also straightforward; however, SRU problems are consistent with the prominent take-away meaning for the subtraction operation (Whitacre et al., 2016), whereas CDU problems are inconsistent (or less consistent) with that meaning. Third and most relevant to the present study, our classification of problem types as traditional and nontraditional is supported by their prevalence in textbooks. JRU, SRU, and PWU problems accounted for more than 90% of the word problems presented in U.S. first-grade textbooks of the 1980s (Schoen et al., 2021).

Relative Difficulty by Word Problem Type

The relative difficulty of the 11 problem types in Figure 1 has had a major influence on the conceptualization of the categories and the multidimensional structure of the taxonomy. Researchers have generalized findings to conjecture that Change Unknown problems tend to be more difficult than Result Unknown problems, and Start Unknown problems are even more difficult (e.g., Berglund-Gray & Young, 1940; Gibb, 1956; Nesher et al., 1982). Some researchers have also found that problems involving actions—specifically, events in the story told in the problem that change the value of a quantity—to be easier for students to solve than problems involving static situations (Caldwell & Goldin, 1979; Carpenter et al., 1982, 2015; Nesher et al., 1982; Riley & Greeno, 1988). In particular, JRU problems are believed to be easier than PWU problems, and SRU problems are believed to be easier than CDU problems (Nesher et al., 1982). These results from basic research are summarized in popular methods texts used by mathematics teacher educators and prospective teachers (e.g., Van de Walle et al., 2018). Using the layout of the taxonomy in Figure 1, the patterns of difficulty can be described loosely as problem difficulty increasing from the upper-left corner to the lower-right corner.

[3] Although other taxonomies exist, including those with 14 categories (e.g., NGA Center & CCSSO, 2010; Nesher et al., 1982), the 11-item taxonomy is an appropriate choice for our purposes. Taxonomies with 14 problem types disaggregate more subtypes of Compare problems.

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

Nesher et al. (1982) offered a theoretical framework that explains relative problem difficulty on the basis of the cognitive structures required to solve additive problems of particular types. Those structures are hierarchical and develop in order from simplest to most complex. In particular, their Piagetian analysis posited four developmental levels and described the schemas associated with each level. On the basis of these schemas, the authors were able to explain children's performance. For example, children at Level 1 are able to "represent and operate on single sets" (p. 385). As a result, the authors explained that children at this level can solve Result Unknown problems but may have difficulty with Change Unknown problems (let alone Start Unknown problems).

Similarly, researchers have documented learning trajectories in the strategies that children use to solve word problems of various types (Carpenter et al., 2015). Most notably, Carpenter et al. described children as moving from the use of Direct Modeling to Counting to Number Facts strategies. *Direct Modeling* strategies are the most basic and involve acting out a word problem in a manner that follows the sequence of events in the story told by the word problem in order, including representing all the quantities involved, and typically counting by ones or tens. Thus, if a student is using Direct Modeling, solving Result Unknown problems would logically be easier than Change Unknown problems (and Start Unknown problems), because following the temporal sequence of the story in Result Unknown problems will produce the answer. Change or Start Unknown problems have information missing from the temporal sequence, which requires a more complex or abstract thought process for a child who may be thinking more literally or concretely about the story in the problem. Problems involving static relationships among the quantities in the sets in the story may also be more difficult, because a physical representation of the sets in the problem will include a set and two subsets, where one of those three pieces is not known, and the problem solver must realize this and determine the unknown quantity.

Such trajectories are typical of the populations studied. Thus, we can reasonably expect that Result Unknown problems would be easiest for first graders (many of whom use Direct Modeling strategies). Likewise, we can expect that problems in which the values of quantities change would be easier for students than would problems involving static situations. However, whether the documented patterns and trends in relative problem difficulty are universal (i.e., the inherence hypothesis) is unclear. Could differences in culture or curriculum alter relative problem difficulty, or at least reduce the disparity in difficulty between problem types (i.e., the exposure hypothesis)?

The inherence hypothesis has been challenged by more recent findings. For example, Christou and Philippou (1998) conducted a study in Cyprus with students in Grades 2–4. They used one-step word problems, including both additive and multiplicative structures, classified similarly to the taxonomy in Figure 1. Using statistical analyses of their data, the authors ranked groups of problems into four hierarchical levels such that "pupils were unable to solve a higher level problem unless they could solve problems of the preceding level" (p. 438). This hierarchy is intended to describe levels of cognitive development, which manifest in the ability to solve different types of problems in particular ways. Thus, this study was similar to that of Nesher et al. (1982) in theoretical orientation and purpose. However, Christou and Philippou (1998) reported a noteworthy difference in findings across studies:

> In contrast to students in similar studies in the United States (Carpenter, 1985) and Canada (Bergeron & Herscovics, 1990), Cypriot pupils found the separate problems easier than the join problems. (p. 439)

Such a finding seems to challenge the inherence hypothesis. We elaborate on this issue in the following section.

Cross-Cultural Investigations of Differences in Curricula and in Relative Problem Difficulty

Cross-cultural studies reveal associations between relative or absolute word problem difficulty and curriculum characteristics (Olkun & Toluk, 2002; Stigler et al., 1986; Xin, 2007). In particular, these findings raise the possibility that the relative difficulty of word problems in the taxonomy is not an inherent property of the type of problem. Rather, relative difficulty may be a product of children's experiences—especially those induced by interaction with their mathematics curricula.[4]

Stigler et al. (1986) compared the distribution of word problem types in four U.S. textbook series (dated 1983–1985) with the distribution in the Soviet national textbook series at the time. They found that the distributions of problem types in the U.S. textbooks were very similar to one another but contrasted with the Soviet texts. The U.S. textbooks focused almost exclusively on three of the 11 additive problem types—the two Result Unknown types and the Whole Unknown type, which we have labeled traditional—whereas the Soviet texts included a wider variety and a balance of problem types. Drawing on data from other studies, the authors pointed out that the most common problem types found in U.S. textbooks were also those that were easiest for U.S. children to solve. More recently, similar findings were reported in Turkey, where

[4] Another plausible explanation is that linguistic differences influence the ways in which the semantic structure of the story is processed in the child's mind (Wang & Lin, 2005). A comparative linguistic analysis is beyond the scope of the present study.

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

Olkun and Toluk (2002) found that Result Unknown problems were also the most prevalent in textbooks and the easiest for Turkish students to solve. Clearly, such evidence is insufficient to support causal inference, but it does raise questions about the relationship between exposure to problem types and relative problem difficulty.

Schoen et al. (2021) followed up on the study by Stigler et al. (1986) by comparing the distributions of additive word problem types in U.S. first-grade textbooks of the 1980s with those in four CCE U.S. first-grade textbooks. In particular, they analyzed the frequency and distribution of word problems in the first-grade student editions of *Investigations in Number, Data, and Space* (Akers et al., 2017), *Math Expressions* (Fuson, 2013), *enVisionmath 2.0* (Charles et al., 2016), and *Saxon Math* (Larson & Matthews, 2012). They found that the CCE textbooks contained many more additive word problems than did the historical texts. The CCE texts also featured much more variation with respect to the types of problems in the additive word problem taxonomy in Figure 1. Whereas the three traditional problem types—the three easiest types—accounted for 95% of the additive word problems in the 1980s texts, those types accounted for only 51% of the additive word problems in the CCE texts.

Xin (2007) compared the performance of U.S. and Chinese students on various types of multiplicative word problems, finding that Chinese students outperformed U.S. students overall. She also compared the distributions of word problem types in a U.S. and a Chinese textbook series and found different distributions, with the U.S. series being less balanced in its emphasis on the various problem types. Furthermore, she found that U.S. students performed worse than their Chinese peers on the types of problems that appeared less frequently in U.S. textbooks. She stated:

> Different performance profiles across U.S. and Chinese students may be caused by cultural (including student self-concept and expectations), language, and teaching-related factors (Wang & Lin [2005]). However, the relation between different patterns in student performance and the corresponding differences in word problem presentations in adopted textbooks seems to indicate that school curricula may have a role in shaping students' preference for certain problem types. . . . Having difficulty solving certain word problem types or activating a specific problem schema to represent and solve a problem (e.g., Riley et al., 1983) may be related to the failure of textbooks to provide sufficient opportunities for students to solve a range of problems to facilitate generalizable problem-solving skills. (p. 357)

The findings of Xin and others suggest that relative problem difficulty may be related to cultural and curricular influences, thus supporting the exposure hypothesis.

The perspective provided by these international comparisons of problem types and student performance offers insight into factors that influence relative difficulty of the types of problems in the additive word problem taxonomy. These authors' findings do not necessitate the conclusion that students perform better on certain types of problems *because* those problems appear more frequently in their textbooks. The findings do, however, raise the question of how the prevalence of certain types of word problems in textbooks may relate to students' problem-solving performance. More broadly, they raise important questions about whether students' problem-solving performance is influenced as much or more by their exposure to problems of particular types as by differences in the characteristics of the problems themselves.

Several important events occurred in the 1990s and 2000s that brought the problem-type taxonomy to the attention of researchers, teachers, and policymakers. Fuson (1992) published a chapter summarizing research on types of word problems in a widely disseminated book published by the National Council of Teachers of Mathematics. Carpenter et al. (1999) summarized the research on additive problem types and included the related taxonomy in their seminal cognitively guided instruction (CGI) book, one of the most widely distributed and influential books in mathematics education. Their book has sold more than 160,000 copies, and tens of thousands of teachers have participated in CGI-based professional learning opportunities since its publication (Schoen et al., 2022; Secada & Brendefur, 2000). State curriculum standards for mathematics began incorporating specific reference to the types of problems in the taxonomy by the mid-2000s (Florida Department of Education, 2007). CCSSM (NGA Center & CCSSO, 2010) included the full taxonomy of problem types, and related textbooks were written, adopted, and implemented in most U.S. schools by 2014 (Blazar et al., 2020; LaVenia et al., 2015). The apparent changes in the number and variety of word problems in the first-grade mathematics curriculum creates an opportunity to compare estimates of the difficulty of additive word problems for U.S. first graders exposed to different curricula in different time periods.

Methods

We drew on three different data sources in pursuit of answers to our RQs: (a) a CCE student textbook for Grade 1; (b) empirical reports of additive word problem difficulty for U.S. first-grade students before 1991; and (c) recent data concerning additive word problem difficulty for CCE first-grade students from schools using the textbook in the first source. Next, we describe the methods that we employed to answer each RQ.

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

RQ1: Analysis of the Frequency and Distribution of Additive Word Problems in a CCE Textbook

We coded the types of word problems presented in the CCE student textbook for Grade 1 in the district-adopted mathematics curriculum used in the schools in which we gathered student data (Dixon et al., 2013). For the purposes of coding, our operational definition of *word problem* was informed by and consistent with the definitions offered by Fuchs et al. (2006) and Stigler et al. (1986). In particular, we focused our investigation on classifying standard school mathematics word problems that involved a situation or context and described two quantities with known values, one quantity with unknown value, and a question about the unknown value. The textbook series we analyzed sometimes conveyed information through pictures. We included problems that communicated information through a combination of words and pictures, as well as problems that exclusively used words, but not those that exclusively used pictures. We included only one-step word problems (Fuchs et al., 2006); multistep problems were rarely encountered, and they were not coded or included in the count. For the coding of the CCE book, we included problems involving sums or minuends between 5 and 20, inclusive. Additional details regarding our definitions and coding procedures are located in Schoen et al. (2021).

Two of the authors identified and coded the word problems in the textbook according to the taxonomy of 11 problem types from Figure 1. Discrepancies were settled after a principled discussion of the coding criteria and our understanding of the way that Stigler et al. (1986) coded their data. In the end, the coders reached full agreement on all counts of occurrences of each problem type that appeared in the textbook.

We then conducted a secondary analysis of results reported by Stigler et al. (1986), who analyzed the frequency and variation of word problems in four U.S. textbook series of the 1980s. We used the first-grade textbook data from Stigler et al. to generate frequency data for each of the four first-grade books from the 1980s, calculating the total number of word problems and the number and proportion of each of the 11 types of additive word problems.[5]

RQ2: Estimating Additive Problem Difficulty for U.S. First-Grade Students: 1966–1991

To obtain empirical estimates of the difficulty of the various problem types before the curricular changes are thought to have occurred, we searched for all available research literature reporting the proportions of U.S. first-grade students who correctly solved additive word problems of any of the 11 types. Our search process began with articles that were already known to us and followed their reference lists. It also made use of ERIC, Education Full Text, JSTOR, PsycINFO, and Academic Search Complete. We used various combinations of search terms including "word problem," "story problem," "arithmetic problem," "addition," "subtraction," "problem difficulty," "performance," "Grade 1," "1st grade," and "first grade."[6]

We searched for original research conducted in the U.S.[7] and published before 1992, thus excluding works published after Fuson's (1992) chapter with its taxonomy of word problems. Using the same definition of word problem, we focused our attention on one-step additive word problems involving sums or minuends between 5 and 20, inclusive. Inclusion criteria required the reporting of the student sample size, the grade level(s) of students in each sample, the period of time in the school year when data were collected, the problem types and numbers used, and the proportion of Grade 1 students who solved the problems correctly. We included empirical estimates provided through dissertations, research reports, and articles published in refereed journals (see Table 1 for details).

We chose to focus on Grade 1 because that is the grade level when students can be expected to solve all types of problems in the additive word problem taxonomy, per CCSSM. Also, we suspect that the distinctions among the types of problems begin to disappear by the end of Grade 2. In our experience, we find that a large proportion of second graders will solve more of the types of problems in the taxonomy correctly when they involve sums or minuends of 20 or less, thereby creating a ceiling effect for many types of problems that obscures the distinctions among types of problems when they are examined through the lens of proportion of students who solve them correctly. We suspect that a larger proportion of students has developed a more abstract understanding of number, addition, and subtraction by the end of Grade 2, so the semantic

[5] Three of the publications used in the current study (i.e., Carpenter et al., 1981; Secada, 1991; Stigler et al., 1986) used Equalize-type problems in their coding schema. Whereas Equalize-type problems could also be recharacterized as Compare-type problems, our own analysis of the description of those types of problems in the original publications led to the conclusion that these were best recharacterized as Join or Separate problems with the Change Unknown, depending on whether they were reported to be Equalize (Add to) or Equalize (Take from) problems, respectively (Fuson, 1992). We sought feedback on this interpretation from one of the original authors of the 11-problem taxonomy, and he unequivocally agreed with this interpretation and coding decision (T. Carpenter, personal communication, October 1, 2014). We acknowledge that opinions vary about how to classify Equalize (Add to) or Equalize (Take from) problems. Classifying the former as JCU and the latter as SCU is appropriate for our purposes, because that transformation places both of them in our category of nontraditional problems.

[6] This process was not a simple keyword search. In addition to database searches and reference lists from other articles, we went to great lengths to obtain relevant data or information about such data that was cited in other articles, including contacting university libraries and authors' collaborators in search of details regarding students' dissertation or thesis work.

[7] We excluded studies involving data collected outside of the U.S. because of our interest in comparing the performance of U.S. first graders during different eras. An article by De Corte and Verschaffel (1987) represents a noteworthy contribution to the literature that was excluded for this reason.

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

Table 1

Historical Reports of Word Problem Difficulty for U.S. First Graders

Author(s)	Publication year	Number of students in sample	Number of problem type(s) assessed	Number of problem-difficulty estimates	Student demographic information
Carpenter et al.	1981	43	7	10	"A parochial school that draws students from a predominantly middle-class area of Madison, Wisconsin" (p. 31)
Carpenter and Moser	1979	150	6	12	"Predominantly white middle to upper-middle class neighborhoods" (p. 25)
Cummins	1991	24	4	4	"New Haven Public Schools . . . middle class and racially mixed" (p. 270)
Hiebert	1982	47	6	6	"A Lexington, Kentucky public school" (p. 342)
Riley and Greeno	1988	18	11	18	"A public elementary school in a predominantly White middle-class suburb" (p. 69)
Secada	1991	45	6	10	"Hispanic . . . lower-middle to middle [socioeconomic status] . . . working-class backgrounds . . . small urban school district in the metropolitan Chicago area" (p. 219)
Steffe	1966	341	2	16	"Unified School District of Racine, Wisconsin . . . randomly selected from this population" (p. 19)
Steffe and Johnson	1971	108	6	8	"Four schools from among the elementary schools in Walton County, Georgia" (p. 54)

structure and position of the unknown value are less prominent in the strategies students use and the relative difficulty among the problems for the students in the Grade 2 population.[8]

We applied exclusion criteria so that the number combinations from the historical studies would be comparable to those used in the CCE data. In particular, we decided a priori to exclude items that involved doubles facts (e.g., $4 + 4 = 8$), because such items were not included in the CCE data. This criterion resulted in the exclusion of two items from the historical data set. We also decided a priori to exclude problems involving a number plus or minus one (i.e., number triples in which one of the numbers is 1), but no such items were found in the historical data. We included reports of data generated through one-on-one mathematics interviews as well as through group-administered, written tests. We excluded data gathered before January of the first-grade year, because we wanted data that would substantially reflect the influence of instruction that took place during first grade. Table 1 lists the publications we found that met these criteria.

Drawing on the publications listed in Table 1, we produced (unweighted) sample mean difficulty estimates for each of the 11 types of additive word problems in Figure 1. We considered each published result to constitute its own sample. In other words, the proportion of students in each sample who correctly solved word problems meeting our inclusion criteria composed a single datum for a given type of problem. This resulted in 84 individual estimates of problem difficulty. In addition to the average difficulty for each type of problem, we were interested in the variability observed within each problem type. For the purpose of comparing the distributions, we generated maxima and minima (by sample) proportions of correct responses, as well as the first, second, and third quartiles and 95% confidence intervals (CIs) for each problem type. CIs and box plots were generated using JMP Pro 14 for Mac (v. 14.0.0). We used these descriptive statistics to make informal comparisons of relative difficulty for all 11 types of problems.

We also quantified the absolute and relative difficulty of traditional (i.e., JRU, PWU, and SRU) and nontraditional (i.e., Join Change Unknown [JCU], Join Start Unknown [JSU], Separate Change Unknown [SCU], Separate Start Unknown

[8] We acknowledge that the research literature does not offer clear support for our conjectures about Grade 2, and we think that they raise important empirical questions to be tested.

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

[SSU], Part-Whole Part Unknown [PPU], Compare Difference Unknown [CDU], Compare Compared Quantity Unknown [CQU], and Compare Referent Unknown [CRU] problem types). We grouped the 84 historical data points into these two categories, resulting in 37 data points for traditional problem types and 47 data points for nontraditional problem types. We calculated means and 95% CIs for these two categories of problems using JMP Pro. We also computed a difficulty ratio of the calculated means (i.e., traditional ÷ nontraditional).

RQ3: Estimating Relative Problem Difficulty in the CCE

Data to answer RQ3 came from research projects conducted in Florida during 2014–2017. These projects were randomized controlled trials involving large numbers of treatment and control teachers and their respective students. In some cases, the treatment-condition teachers experienced professional development in CGI, which provides teachers with opportunities to learn about these problem types and children's associated strategies. For the purposes of the present study, we exclusively used data obtained from the students in non-CGI classrooms (i.e., the classrooms of teachers who had not participated in CGI-based professional learning at the time of student data collection). We made this choice to avoid introducing confounding variables.

A total of 4,604 students are represented in the analytic sample for the present study of problem difficulty for U.S. first graders in 2014–2017. These students are nested in 58 schools in nine school districts, all of which were using the textbook series of Dixon et al. (2013) as their primary instructional resource for mathematics at the time of data collection. Gender identification was obtained for 3,857 of the students in the sample; 1,935 (50.2%) were identified as girls, and 1,922 (49.8%) as boys. Race and ethnicity data were given for 3,718 students. Of those students, 176 (4.7%) were identified as Asian or Pacific Islander (non-Hispanic), 614 (16.5%) were identified as Black or African American (non-Hispanic), 1,047 (28.2%) were identified as Hispanic or Latino, 162 (4.4%) were identified as two or more races or Native American (non-Hispanic), and 1,719 (46.2%) were identified as White (non-Hispanic). Free or reduced-priced lunch eligibility was indicated for 3,272 students, of whom 2,068 (65.5%) qualified. English language learner (ELL) status was obtained for 3,279 students, with 531 (16.2%) identified as ELLs at the time of data collection.[9] Data describing exceptionalities were given for 3,718 students. Of those students, 328 (8.8%) were identified as students with disabilities and 36 (2.4%) as gifted.

As with the historical data, we focused our analysis on 2014–2017 assessment data that were collected during spring semester of the first-grade year. The data were gathered using one-on-one mathematics interviews as well as group-administered tests using a paper-and-pencil format (Schoen et al., 2016, 2017). The interviews were conducted during the last three months of the school year and consisted of constructed-response type items.[10] Written tests were administered at the end of the school year and used constructed-response or selected-response formats. All items involved sums or minuends between 5 and 20, inclusive, and none of the items involved doubles facts. Individual student responses were scored and recorded as a dichotomous variable with a zero indicating an incorrect response and a one indicating a correct response. For each item, we calculated the proportion of students in each sample school who provided a correct response in a given year, using SPSS (version IBM SPSS Statistics 23). This process produced 36 school-level means (i.e., difficulty estimates) for interview items, 40 for constructed-response items on paper-and-pencil tests, and 272 for selected-response written items.

In some cases, schools were represented by a small number of students, especially for interview items. To avoid having school-level means using small numbers of students, we set a post hoc inclusion criterion of at least 12 students responding to a given item at a given school. As a result, 53 school-level means (using 1–11 students) were excluded. The remaining 348 school means (each mean representing the proportion of correct answers on a given item at a given school in a single school year) constitute our analytic sample for the new data.

Subsequent analyses were performed similarly to those done for RQ2. We analyzed the distributions of means for each problem type, computing minima, maxima, and quartile scores, as well as 95% CIs, using JMP Pro. As described next, we compared these estimates with the difficulty estimates from the 1966–1991 publications to answer RQ4.

As with RQ2, we also estimated the difficulty of traditional and nontraditional problems for the students in the sample. The 348 data points consisted of 91 traditional problems and 257 nontraditional problems. We calculated means and 95% CIs for the data in these two categories, using JMP Pro. We then quantified their relative difficulty by computing the difficulty ratio (i.e., traditional ÷ nontraditional).

[9] The sample included multilingual (English and Spanish) students, and instruction took place in both English and Spanish in at least one of the schools in the sample. Similarly, in the historical data, Secada (1991) assessed first graders who were bilingual in English and Spanish.

[10] To test for comparability of the different item formats and response types, we conducted a form of sensitivity analysis using a one-at-a-time approach. A one-way ANOVA revealed no significant effect of item format on student performance. A one-way ANOVA did reveal a statistically significant effect of response type on student performance, with a lower mean score on constructed response (.61) versus selected response items (.66).

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

RQ4: Comparing Relative Problem Difficulty Then and Now

We focused on comparing performance on traditional versus nontraditional problem types within each period (1966–1991 and 2014–2017). We compared the 95% CIs with respect to overlap and separation as well as the difficulty ratios from the two periods. In addition, we conducted a hypothesis test using a two-sample *t*-test for the effect of period (1966–1991 vs. 2014–2017) on performance for the two categories of problems. These analyses were performed using JMP Pro.

Results

We present our findings in order of the four RQs. We first present evidence that the textbook used by first graders in our recent sample was different from textbooks of the 1980s with respect to distribution of additive word problems. We then present difficulty estimates for each of the 11 problem types for the data from 1966–1991 and 2014–2017, respectively. Finally, we present the results of our analysis of the gap in performance in solving traditional and nontraditional problem types during the two periods.

Comparing the 1980s and CCE Textbooks

We found that the CCE textbook featured a greater total number of additive word problems and a much more balanced distribution of additive word problems compared with the 1980s U.S. texts analyzed by Stigler et al. (1986). The mean number of additive word problems in the 1980s textbooks was 82. The fewest number of additive word problems in any individual book in their sample was 64, whereas the most was 95. In the 2013 book, we counted 164 additive word problems—twice the average number of problems found in the 1980s books. Table 2 presents the proportion of each type of additive word problem in the 1983–1985 U.S. textbooks, as reported by Stigler et al. (1986). It also shows the analogous figures for the CCE textbook (Dixon et al., 2013) that was used in the schools in which our student-performance data were collected.

The CCE textbook had more variety in problem types than the 1980s books. As Table 2 shows, the 1980s U.S. textbooks included only a narrow range of problem types. The three most straightforward and traditional problem types—JRU, SRU, and PWU—together accounted for approximately 95% of the word problems that appeared in textbooks published in the 1980s, but they were only 45% of the word problems in the 2013 book. This implies that the CCE students in our study had considerably more exposure to the nontraditional problem types than U.S. students in the 1980s.

Estimates of Additive Word Problem Difficulty for First Graders in 1966–1991

Figure 2 presents a visualization of the results of the synthesis of historical data concerning the (unweighted) mean proportions of correct responses to each of the 11 types of word problems in Figure 1. The boxplots in Figure 2 provide a snapshot of the distributions of these sample-level results for proportions of correct responses. We note substantial variability within each type of problem. Table 3 shows 95% CIs for mean difficulty of each type of problem in the taxonomy (listed in the same order as in Figure 2). Using the historical data offered through the publications listed in Table 1, these results provide difficulty estimates for U.S. first graders in the 1960s–1990s on each of the 11 problem types.

The three traditional problem types were indeed easier than the nontraditional types for Grade 1 students in 1966–1991. In general, using the layout of Figure 1, our analysis and compilation of the historical data from individual studies confirms that problem difficulty increases from left to right and from top to bottom in Figure 1. As reported by many authors of the individual studies, Result Unknown problems tended to be easier for students than Change Unknown and Start Unknown problems. Likewise, PWU problems tended to be easier than Part Unknown problems, and CDU problems were easier, on average, than CQU and CRU problems.

Within the three traditional types, we found the median proportion correct to be highest for JRU, slightly lower for PWU, and lower yet for SRU. The sampling distributions for the JRU and PWU data had substantial overlap, and the difference between the two distributions is not statistically significant. Some separation is evident between those two and the SRU distribution, suggesting that the subtraction situations were more difficult for first graders than the addition situations.

The gap in difficulty between traditional and nontraditional problem types was large. The ratio of the means (i.e., traditional ÷ nontraditional) was .79 ÷ .42 = 1.88. On average, traditional problem types were answered correctly nearly twice as often as their nontraditional counterparts.

Estimates of Additive Word Problem Difficulty for First Graders in 2014–2017

Addressing RQ3, the boxplots in Figure 3 provide a snapshot of the distributions of our school-level results for proportions of correct responses. Table 3 displays the (unweighted) mean proportions of correct responses to each of the 11 additive word problem types, including standard errors and 95% CIs. Using the data gathered in the spring semesters

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

Table 2

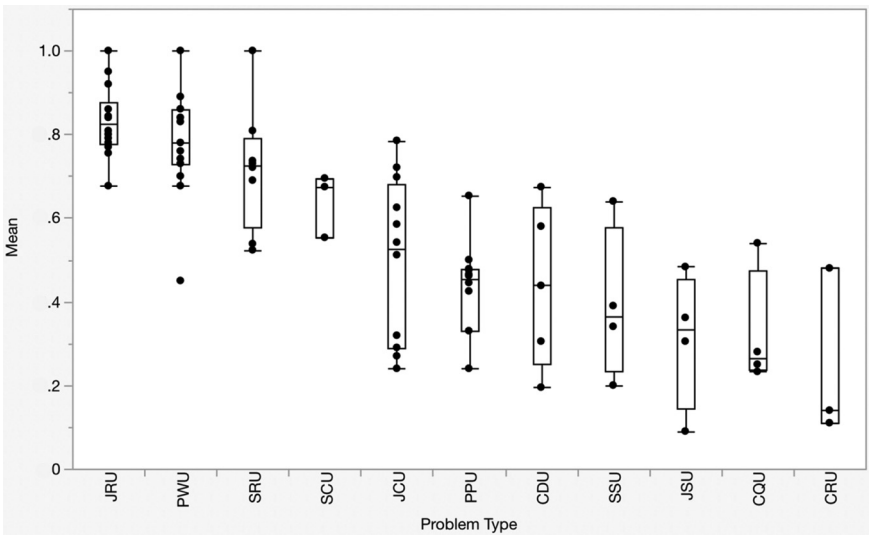
Proportion of Each Type of Word Problem in First-Grade Textbooks

Semantic structure	Position of the unknown[a]	1980s books[b] $M = 82.5$ problems			2013 book[c] $N = 182$ problems
		Min.	Mean	Max.	
Join	Result Unknown[T]	.12	.14	.16	.15
	Change Unknown	.00	.01	.05	.05
	Start Unknown	.00	.00	.00	.03^{+}
Separate	Result Unknown[T]	.38	.44	.53	.16^{-}
	Change Unknown	.00	.00	.00	.08^{+}
	Start Unknown	.00	.00	.00	.03^{+}
Part-Part-Whole	Whole Unknown[T]	.23	.37	.50	.14^{-}
	Part Unknown	.00	.03	.08	.16^{+}
Compare	Difference Unknown	.00	.00	.03	.10^{+}
	Compare Quantity Unknown	.00	.00	.01	.08^{+}
	Referent Unknown	.00	.00	.00	.02^{+}

[a] (T) indicates traditional problem types (all other problem types listed are classified as nontraditional).
[b] The values in this column report the mean of the proportion of word problems within each book, with the minimum and maximum proportions in parentheses.
[c] (+) indicates that the proportion of that type of word problem is higher than the maximum proportion for that problem type in the sample of 1980s textbooks. Likewise, (−) indicates that the proportion is lower than the minimum proportion in the 1980s textbooks.

Figure 2

Box Plot Representing Additive Problem Difficulty for U.S. First Graders in Spring, 1966–1991



Note. Problem types are ordered by median, from greatest (leftmost) to least (rightmost).

of 2014, 2015, 2016, and 2017, these results provide the best available difficulty estimates for U.S. first graders in the CCE on each of the 11 additive problem types.

Examination of the mean and median difficulty estimates for PWU and JRU problems reveals that students in the 2014–2017 sample scored higher on PWU problems than on JRU problems. As before, the difference is not statistically significant, so whether one type is more difficult than the other for U.S. first graders remains unclear.

Using the layout of problem types in Figure 1 for reference, these results are consistent with the general rule that problem difficulty increases from left to right and from top to bottom for first graders in the CCE, but we encountered some

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

Table 3

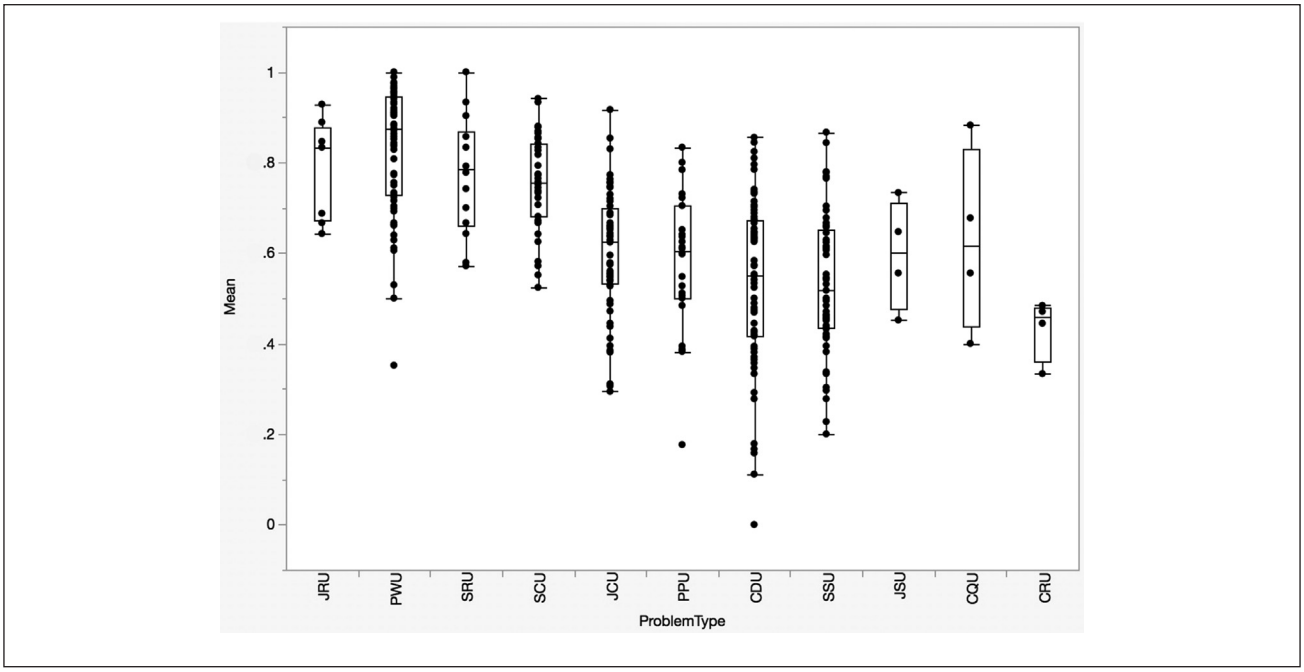
Difficulty Estimates for Additive Word Problems for U.S. First Graders in Spring, 1966–1991 cf. 2014–2017

	1966–1991				2014–2017			
Problem type	*n*	*M*	*SE*	95% CI	*n*	*M*	*SE*	95% CI
Traditional	37	.79	.03	[.74, .84]	91	.82	.02	[.79, .85]
JRU	14	.83	.04	[.75, .91]	8	.79	.05	[.69, .90]
PWU	15	.79	.04	[.71, .86]	69	.83	.02	[.80, .87]
SRU	8	.72	.05	[.62, .82]	14	.77	.04	[.69, .85]
Nontraditional	47	.42	.03	[.38, .47]	257	.59	.01	[.57, .61]
SCU	3	.64	.08	[.47, .81]	35	.76	.03	[.71, .81]
JCU	12	.49	.04	[.41, .57]	60	.60	.02	[.56, .64]
PPU	12	.43	.04	[.34, .51]	28	.58	.03	[.53, .64]
CDU	5	.44	.07	[.31, .57]	69	.53	.02	[.50, .57]
SSU	4	.39	.07	[.25, .54]	53	.54	.02	[.49, .58]
JSU	4	.31	.07	[.17, .46]	4	.60	.08	[.45, .74]
CQU	4	.33	.07	[.18, .47]	4	.63	.08	[.48, .78]
CRU	3	.24	.08	[.08, .41]	4	.43	.08	[.29, .58]

Note. M and *SE* are based on sample- or school-level data for unique items, not individual students.

Figure 3

Box Plot Representing Additive Problem Difficulty for U.S. First Graders in Spring 2014–2017



Note. To aid in visual comparison of distributions, this graph maintains the ordering of problem types from Figure 2.

exceptions to that rule. Indeed, Result Unknown problems tended to be easier for students than Change Unknown and Start Unknown problems. Likewise, Whole Unknown problems tended to be easier than Part Unknown problems. Mean performance on CQU items was higher than CDU items, but the standard error for CQU and CRU was three or four times as large as the standard error for many of the other problem types, possibly influenced by relatively small samples for those two types of problems. As discussed previously, Whole Unknown problems present another exception to the rule,

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

because the point estimate for the median difficulty suggests that this is the easiest type of problem for these first graders to solve correctly.

Student performance on traditional types of problems was substantially higher than it was on nontraditional types of problems. Because we see overlap in performance on SCU problems and the aggregate of the traditional problems, student performance on SCU problems are a noteworthy exception. The ratio of the means (i.e., traditional ÷ nontraditional) was .82 ÷ .59 = 1.39. Thus, a substantial difference in difficulty remains between traditional and nontraditional problem types.

First Graders in CCE Outperform Their 20th-Century Peers on Challenging Problems

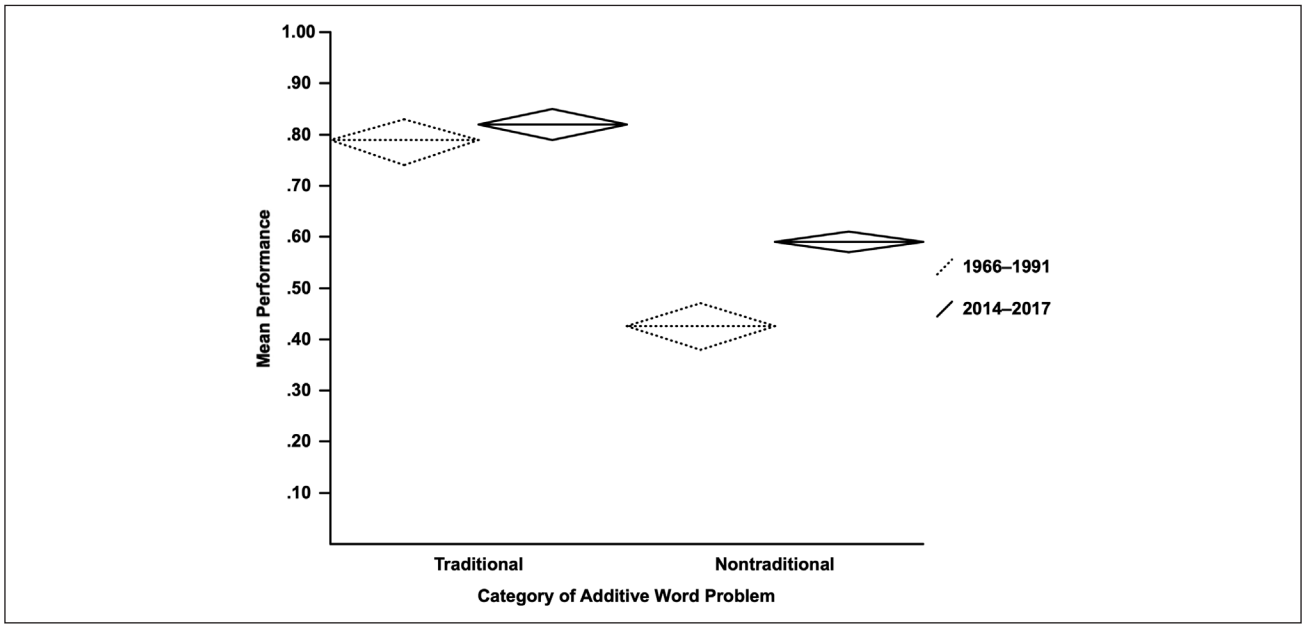
When we compare student performance on the traditional types of problems, we see a higher point estimate of mean performance for the CCE students but substantial overlap in the CIs. Figure 4 displays the means and CIs for nontraditional and traditional problems in the two time periods. The mean difficulty for traditional problems in our secondary analysis of historical data was .79, with a 95% CI of [.74, .84]. The mean difficulty for traditional problems in the data from students in the CCE was .82, with a 95% CI of [.79, .85]. The difference in means could be indicative of slightly better student performance on traditional problems during the CCE, but the overlapping CIs suggests that this may not be a meaningful difference.

However, a substantial difference in student performance on the nontraditional types of problems does appear when we compare performance both within and between the two time periods. The mean for nontraditional problems in our secondary analysis of the historical data was .43, with a 95% CI of [.38, .47]. By contrast, the mean for nontraditional problems in the data from students in the CCE was .59, with a 95% CI of [.57, .61]. For the period 1966–1991, we estimated a mean difficulty ratio of 1.88, whereas the analogous ratio for the period 2014–2017 was 1.39. Thus, we estimate a 26% reduction in the difficulty ratio (i.e., traditional ÷ nontraditional) for U.S. first graders from 1966–1991 to 2014–2017.

A two-sample *t*-test to compare the performance of students in the 1966–1991 ($M = .79$, $SD = .03$) and 2014–2017 samples ($M = .82$, $SD = .02$) revealed no significant difference in student performance on the traditional types of items ($t(126) = 1.187$, $p = .237$). A two-sample *t*-test did reveal a significant difference in performance of the students in 1966–1991 sample ($M = .42$, $SD = .03$) and 2014–2017 ($M = .59$, $SD = .01$) on the nontraditional types of items ($t(302) = 5.962$, $p < .001$). Thus, the performance of a sample of U.S. students near the end of first grade in the CCE was significantly better than that of their historical counterparts on nontraditional problem types but not on traditional ones.

Figure 4

Difficulty Estimates for U.S. First Graders on Traditional and Nontraditional Problem Types Historically and in the CCE



Note. Diamond shapes represent means and 95% CIs.

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

Discussion

Problem Difficulty Estimates in Historical Data

Although many studies have been concerned with the relative difficulty of the various types of word problems, we are not aware of a previous publication that surveyed all available data and investigated overall trends across samples. The empirical estimates of problem difficulty from historical data (1966–1991) represent a valuable and overdue contribution to the corpus of research findings related to word problem types and their relative difficulty. This synthesis of data largely confirms previously reported patterns of relative problem difficulty, but it may be the first one published using results from different samples.

Opportunities to Learn in the CCE Textbook

The results of our textbook analysis indicate that the students in our sample (along with large numbers of other U.S. first graders) were likely to have been exposed to a much wider variety of types of word problems than was typical in U.S. classrooms in the past. We found stark differences in the prevalence of various word problem types between the 1980s textbooks and the CCE textbook. Whereas the popular U.S. Grade 1 textbooks of the 1980s focused almost exclusively on three of the 11 additive problem types, the CCE textbook included a more balanced distribution, including all 11 types of additive word problems—ostensibly driven by their explicit inclusion in CCSSM (NGA Center & CCSSO, 2010). This fact, together with similar findings reported by Schoen et al. (2021) for four other CCE first-grade texts, is evidence of a rather dramatic curricular shift.[11]

According to Van de Walle et al. (2018), "In most curricula in the United States, the overwhelming emphasis is on the easier join and separate problems with the result unknown" (p. 153). We think that conclusion is fair for curricula being used before the CCE. In the early 1980s, the vast majority of additive word problems were of the traditional types (with JRU less prevalent than SRU or PWU; Stigler et al., 1986). On the basis of the CCE textbook that we analyzed—and four other popular U.S. Grade 1 texts, as reported by Schoen et al. (2021)—we find that the emphasis on the traditional problem types in curricula in the U.S. has decreased relative to the nontraditional types. We also find a stark contrast in opportunities to solve nontraditional types of word problems when we compare the prevalence of these problems in the 1980s books with that of the CCE books. With respect to the amount and variation in the types of additive word problems presented to students, U.S. textbooks have changed.

Updating the Body of Research on Relative Difficulty of Word Problems

The relative problem difficulty of the various types of problems within our recently gathered data roughly parallels those observed historically, but the observed differences in difficulty between the traditional and nontraditional types of problems appears to be considerably smaller now than in the past. These findings provide some evidence consistent with the exposure hypothesis. At the same time, the semantic structure and position of the unknown appear to continue to influence problem difficulty, because students in the CCE era still perform better on the traditional problem types. This finding is consistent with the inherence hypothesis.

Past research related to these competing hypotheses has involved cross-cultural comparisons. As such, whether we could reasonably expect to see differences in relative problem difficulty within the same country was unclear. With respect to ranking problem types by difficulty, we do not claim that such changes have taken place. We do see compelling evidence of change, however, with respect to how much more difficult the nontraditional problem types are versus the traditional ones. The sample of first graders in the CCE does appear to be better at solving nontraditional word problems compared with samples of first graders in studies in 1966–1991. They also appear to be at least equally able to solve the traditional problems correctly.

Many mathematics educators—including us—have long believed and taught that problems involving static situations are more difficult for children to solve than those that involve change or action. For example, one of the most popular elementary mathematics methods textbooks reads, "The structure of some problem types is more difficult than others. Problems in which a physical action is taking place, as in join and separate problems, are easier because children can model or act out the situation" (Van de Walle et al., 2018, p. 152). However, our analysis found that JRU problems were not easier, on average, than PWU problems for end-of-year U.S. first graders. PWU problems—which involve static situations—were the easiest type of problem for first graders in our CCE data set when comparing the point estimates for median difficulty. Further, we did not find a statistically significant difference in difficulty between PWU and JRU problems in our CCE data set or in the historical data.[12] Indeed, we are not the first to arrive at this conclusion. Reflecting the conventional

[11] We note that although the intended curriculum is clearly different from historical textbooks with respect to the number and types of word problems, a detailed assessment of the enacted curriculum is necessary to confirm that students were exposed to these problems.

[12] Although we invite readers to examine the descriptive statistics and consider whether the differences in problem difficulty within or between time periods might be meaningful, we caution readers against drawing inferences on the basis of such observed differences in these data. Instead, we encourage the curious reader to design and conduct tightly controlled studies to explore those relations in targeted ways and with adequate statistical power to perform hypothesis tests.

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

wisdom at the time, Baroody and Ginsburg (1986) wrote that empirical findings available "do not consistently show that change problems were easier than combine problems for kindergarten and first-grade children" (p. 80). Elementary mathematics educators should take heed of this apparent inconsistency between the widespread notions about relative problem difficulty and the empirical findings (see the Limitations and Future Directions section for more discussion on this topic).

Policy and Curriculum Implications

These findings concerning relative problem difficulty are not merely of academic interest. They also show a clear instance of policy and curriculum being informed by research. As noted in the Introduction, every problem type in Figure 1 appears explicitly in CCSSM. We also found that these problem types are prevalent in contemporary textbooks (Schoen et al., 2021). Thus, large numbers of students are apparently being exposed to these problem types, and teachers are expected to support the development of students' problem-solving abilities. The present examination of a focused—but central—component of the mathematics curriculum implies that increased exposure is paying dividends in the form of increased student performance on the nontraditional types of problems.

Implicit in the shift to including all types of additive word problems in mathematics standards is the assumption that doing so should make a difference. In this study, we found compelling evidence that the gap in difficulty between traditional and nontraditional additive word problem types has narrowed considerably for U.S. first graders, but many questions remain. Whereas continuing to expose students to a variety of word problems seems sensible, the optimal extent to which this should be done is unclear. As Tran (2016) pointed out:

> The challenge is that curriculum standards do suggest important topics to include in teaching and learning, but do not specify the level of treatment for the topics. Should a balance of task types be present . . . or are particular problem types . . . more worthy of textbook terrain? (pp. 294–295)

We believe that this question is important and worthy of rigorous investigation.

We wish to be clear that we do not recommend using the results of this study to create curricula designed around a strict adherence to an easy-to-difficult sequence of types of problems. We consider some of the types of problems to be more useful than others for the sake of instruction and assessment as well as support for specific learning goals. We believe that the type of problem(s) to be used in curriculum, instruction, and assessment should be selected purposefully and in consideration of the learner and a specific learning goal.

Limitations and Future Directions

We acknowledge numerous potential threats to internal validity and our ability to draw causal inference regarding the effect of exposure or inherent properties of different types of problems on problem difficulty using the results of this study. We attempted to estimate problem difficulty of comparable items by limiting the data sources (both historical and recent) to items administered to first graders in the latter part of the school year and to items involving number triples within 20 (as well as excluding items involving doubles facts). We were unable to control for all possible influences on student performance, such as wording or other factors, especially with regard to the secondary analysis of previously published data. All the schools in the CCE sample used the Dixon et al. (2013) text, but we do not have information about the extent to which students were expected to solve each type of problem in the book. Teachers could have deemphasized the nontraditional types of problems, for example. Other potential confounding factors that we have not already listed include changes in school accountability systems, changing demographics of students represented in the samples, conditions under which the problems were posed to students, and much more. Factors within the books that do not have anything to do with the number of word problems, such as the approaches taken to introduce different problem-solving strategies or the organization of content, also could have contributed to differences in performance. As suggested by Xin (2007), noncurricular cultural factors may influence relative problem difficulty, including differences in beliefs and values, language, and pedagogy. Such influences on relative problem difficulty continue to merit further research.

The generalizability of the results of this study is also limited. We conducted this study within the U.S., so the results may not generalize to students in other countries. For that matter, the current study does not claim to use a representative sample of U.S. students or curricula in either time period. We also do not know and cannot estimate the extent to which the books in the Stigler et al. (1986) study are representative of those used by students in the studies published in 1966–1991. We have no information about students whose parents or guardians did not consent to their participation in the studies, so we cannot speak to the potential influence of selection bias associated with the consent process. Generalizability of these results are also limited to additive problem types, and the current study cannot answer questions about whether the results would generalize to different topics in the mathematics curriculum, such as multiplication and division with whole numbers or

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

problems involving rational numbers. We hope that the current study will inspire similar studies in other aspects of the mathematics curriculum, such as multiplicative word problems, rational number, probability, or functions.

Children learn a tremendous amount about mathematics in a short period of time during early elementary school (Hill et al., 2008). Patterns of relative problem difficulty may differ for students in different grade levels or at different points in the school year within a given grade level. Relative difficulty among different types of problems may have been different if the students represented in the sample had been 6 months, 1 year, or 2 years younger. We think additional research to better understand the conditions under which that hypothesis is, and is not, supported by empirical results remains to be an important and unresolved topic.

Although we acknowledge the limitations of this observational study, we maintain that these results are interesting and that the methods, size, and diversity provide a unique insight into the validity of widely accepted conclusions about relative problem difficulty as well as into the ongoing discussion about the effects of exposure or inherency on problem difficulty. More tightly controlled studies that use the same number combinations, have more information about exposure to various types of problems, and compare student performance during specific windows of time during prekindergarten, kindergarten, first grade, and second grade may provide further elucidation. For example, dense, longitudinal assessment of children before first grade may reveal competence with JRU problems before PWU problems, because the former are more consistent with the informal Change (Add to) view of addition (Baroody & Ginsburg, 1986; Fischbein et al., 1985). If those studies manage to use large, representative samples, they may better support generalizations.

Conclusion

We end this article with two main conclusions. First, our synthesis of previously published data and our new data confirm that the general patterns of relative problem difficulty in the taxonomy were true in the latter part of the 20th century—when the traditional types of problems were the overwhelming majority of word problems represented in the curriculum—and remained true in the CCE—when students had more opportunities to be exposed to the nontraditional types of problems. Second, we found a significant and substantial difference in performance on nontraditional problem types for the two time periods. Increased exposure to nontraditional problem types seems to be the simplest—therefore, most plausible—explanation for these results. These results provide further support for the exposure hypothesis, although they also still support the overall patterns in relative difficulty of various types of problems in the taxonomy, consistent with the inherence hypothesis. In other words, relative problem difficulty appears to be affected both by inherent traits of the various types of problems and by varying levels of exposure to them.

This article made use of a substantial sample of student data to revisit relative problem difficulty. This investigation is valuable, but it is by no means the last word regarding the Inherence and Exposure Hypotheses. Given the decades-long tradition that led to important insights into student thinking and the interaction between children and mathematics, we hope that the findings we report in this article will inspire additional research, possibly through more tightly controlled and designed experiments to test the effects of assignment of students to instruction that exposes them to various types of problems.

Recent changes in standards and curricula present mathematics education researchers with an opportunity to explore important questions about the extent to which the relative difficulty of various mathematical tasks is influenced by (a) students' psychological processes and their cognitive development, and (b) students' opportunities to learn through exposure to particular types of tasks. We hope to rekindle a conversation that can be continued and expanded to offer fundamental insight into the ways that children learn mathematics and the influence of opportunities to learn, problem types, and other important factors in mathematical learning.

References

Agodini, R., & Harris, B. (2010). An experimental evaluation of four elementary school math curricula. *Journal of Research on Educational Effectiveness*, *3*(3), 199–253. https://doi.org/10.1080/19345741003770693

Akers, J., Bastable, V., Battista, M. T., Hickey, K., Clements, D., Cochran, K., Earnest, D., Economopoulos, K., Hollister, A., Horowitz, N., Kliman, M., Leidl, E., Mokros, J., Murray, M., Nemirovsky, R., Oh, Y., Perry, B. W., Rubin, A., Russell, S. J., . . . Tierney, C. (2017). *Investigations in number, data, and space: Grade 1* (3rd ed.). Pearson.

Baroody, A. J. (1987). *Children's mathematical thinking: A developmental framework for preschool, primary, and special education teachers*. Teachers College Press.

Baroody, A. J., & Ginsburg, H. P. (1986). The relationship between initial meaningful and mechanical knowledge of arithmetic. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 75–112). Erlbaum.

Berglund-Gray, G., & Young, R. V. (1940). The effect of process sequence on the interpretation of two-step problems in arithmetic. *The Journal of Educational Research*, *34*(1), 21–29. https://doi.org/10.1080/00220671.1940.10880969

Blazar, D., Heller, B., Kane, T. J., Polikoff, M., Staiger, D. O., Carrell, S., Goldhaber, D., Harris, D. N., Hitch, R., Holden, K. L., & Kurlaender, M. (2020). Curriculum reform in the Common Core era: Evaluating elementary math textbooks across six U.S. states. *Journal of Policy Analysis and Management*, *39*(4), 966–1019. https://doi.org/10.1002/pam.22257

Cai, J., Lo, J. J., & Watanabe, T. (2002). Intended treatments of arithmetic average in U.S. and Asian school mathematics textbooks. *School Science and Mathematics*, *102*(8), 391–404. https://doi.org/10.1111/j.1949-8594.2002.tb17891.x

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

Cai, J., Nie, B., & Moyer, J. C. (2010). The teaching of equation solving: Approaches in *Standards*-based and traditional curricula in the United States. *Pedagogies: An International Journal*, *5*(3), 170–186. https://doi.org/10.1080/1554480X.2010.485724

Caldwell, J. H., & Goldin, G. A. (1979). Variables affecting word problem difficulty in elementary school mathematics. *Journal for Research in Mathematics Education*, *10*(5), 323–336. https://doi.org/10.2307/748444

Carpenter, T. P., Fennema, E., & Franke, M. L. (1996). Cognitively guided instruction: A knowledge base for reform in primary mathematics instruction. *The Elementary School Journal*, *97*(1), 3–20. https://doi.org/10.1086/461846

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (1999). *Children's mathematics: Cognitively guided instruction*. Heinemann.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (2015). *Children's mathematics: Cognitively guided instruction* (2nd ed.). Heinemann.

Carpenter, T. P., Fennema, E., Peterson, P. L., & Carey, D. A. (1988). Teachers' pedagogical content knowledge of students' problem solving in elementary arithmetic. *Journal for Research in Mathematics Education*, *19*(5), 385–401. https://doi.org/10.5951/jresematheduc.19.5.0385

Carpenter, T. P., Hiebert, J., & Moser, J. M. (1981). Problem structure and first-grade children's initial solution processes for simple addition and subtraction problems. *Journal for Research in Mathematics Education*, *12*(1), 27–39. https://doi.org/10.2307/748656

Carpenter, T. P., & Moser, J. M. (1979). *An investigation of the learning of addition and subtraction* (Theoretical Paper No. 79). Wisconsin Research and Development Center for Individualized Schooling. https://files.eric.ed.gov/fulltext/ED188892.pdf

Carpenter, T. P., & Moser, J. M. (1983). The acquisition of addition and subtraction concepts. In R. A. Lesh & M. Landau (Eds.), *Acquisition of mathematics concepts and processes* (pp. 7–44). Academic Press.

Carpenter, T. P., Moser, J. M., & Romberg, T. A. (1982). *Addition and subtraction: A cognitive perspective*. Erlbaum.

Charles, R. I., Bay-Williams, J., Berry, R. Q., III, Caldwell, J. H., Champagne, Z. C., Copley, J., Crown, W., Fennell, F., Karp, K., Murphy, S. J., Schielack, J. F., & Suh, J. M. (2016). *enVisionmath 2.0: Grade 1*. Pearson.

Christou, C., & Philippou, G. (1998). The developmental nature of ability to solve one-step word problems. *Journal for Research in Mathematics Education*, *29*(4), 436–442. https://doi.org/10.2307/749860

Clements, D. H., & Sarama, J. (2014). *Learning and teaching early math: The learning trajectories approach* (2nd ed.). Routledge. https://doi.org/10.4324/9780203520574

Cummins, D. D. (1991). Children's interpretations of arithmetic word problems. *Cognition and Instruction*, *8*(3), 261–289. https://doi.org/10.1207/s1532690xci0803_2

De Corte, E., & Verschaffel, L. (1987). The effect of semantic structure on first graders' strategies for solving addition and subtraction word problems. *Journal for Research in Mathematics Education*, *18*(5), 363–381. https://doi.org/10.2307/749085

Dixon, J. K., Larson, M., Leiva, M. A., & Adams, T. L. (2013). *Go math! Florida: Grade 1*. Houghton Mifflin Harcourt.

Fan, L., & Zhu, Y. (2007). Representation of problem-solving procedures: A comparative look at China, Singapore, and U.S. mathematics textbooks. *Educational Studies in Mathematics*, *66*(1), 61–75. https://doi.org/10.1007/s10649-006-9069-6

Fennema, E., Carpenter, T. P., Franke, M. L., Levi, L., Jacobs, V. R., & Empson, S. B. (1996). A longitudinal study of learning to use children's thinking in mathematics instruction. *Journal for Research in Mathematics Education*, *27*(4), 403–434. https://doi.org/10.5951/jresematheduc.27.4.0403

Fischbein, E., Deri, M., Nello, M. S., & Marino, M. S. (1985). The role of implicit models in solving verbal problems in multiplication and division. *Journal for Research in Mathematics Education*, *16*(1), 3–17. https://doi.org/10.2307/748969

Florida Department of Education. (2007). Next generation sunshine state standards.

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., Schatschneider, C., & Fletcher, J. M. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, *98*(1), 29–43. https://doi.org/10.1037/0022-0663.98.1.29

Fuson, K. C. (1992). Research on whole number addition and subtraction. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 243–275). National Council of Teachers of Mathematics.

Fuson, K. C. (2013). *Math expressions: Grade 1 Common Core*. Houghton Mifflin Harcourt.

Fuson, K. C., Stigler, J. W., & Bartsch, K. (1988). Grade placement of addition and subtraction topics in Japan, Mainland China, the Soviet Union, Taiwan, and the United States. *Journal for Research in Mathematics Education*, *19*(5), 449–456. https://doi.org/10.5951/jresematheduc.19.5.0449

García, A. I., Jiménez, J. E., & Hess, S. (2006). Solving arithmetic word problems: An analysis of classification as a function of difficulty in children with and without arithmetic LD. *Journal of Learning Disabilities*, *39*(3), 270–281. https://doi.org/10.1177/00222194060390030601

Gibb, E. G. (1956). Children's thinking in the process of subtraction. *The Journal of Experimental Education*, *25*(1), 71–80. https://doi.org/10.1080/00220973.1956.11010564

Grouws, D. A., Tarr, J. E., Chávez, Ó., Sears, R., Soria, V. M., & Taylan, R. D. (2013). Curriculum and implementation effects on high school students' mathematics learning from curricula representing subject-specific and integrated content organizations. *Journal for Research in Mathematics Education*, *44*(2), 416–463. https://doi.org/10.5951/jresematheduc.44.2.0416

Herbel-Eisenmann, B. A. (2007). From intended curriculum to written curriculum: Examining the voice of a mathematics textbook. *Journal for Research in Mathematics Education*, *38*(4), 344–369.

Hiebert, J. (1982). The position of the unknown set and children's solutions of verbal arithmetic problems. *Journal for Research in Mathematics Education*, *13*(5), 341–349. https://doi.org/10.2307/749008

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hong, D. S., & Choi, K. M. (2014). A comparison of Korean and American secondary school textbooks: The case of quadratic equations. *Educational Studies in Mathematics*, *85*(2), 241–263. https://doi.org/10.1007/s10649-013-9512-4

Hsu, H.-Y., & Silver, E. A. (2014). Cognitive complexity of mathematics instructional tasks in a Taiwanese classroom: An examination of task sources. *Journal for Research in Mathematics Education*, *45*(4), 460–496. https://doi.org/10.5951/jresematheduc.45.4.0460

Jerman, M. E., & Mirman, S. (1974). Linguistic and computational variables in problem solving in elementary mathematics. *Educational Studies in Mathematics*, *5*(3), 317–362. https://doi.org/10.1007/BF01424553

Larson, N., & Matthews, L. (2012). *Saxon math 1*. Houghton Mifflin Harcourt.

LaVenia, M., Cohen-Vogel, L., & Lang, L. B. (2015). The Common Core State Standards initiative: An event history analysis of state adoption. *American Journal of Education*, *121*(2), 145–182. https://doi.org/10.1086/679389

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

Lewis, A. B., & Mayer, R. E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology*, *79*(4), 363–371. https://doi.org/10.1037/0022-0663.79.4.363

Mayer, R. E., Sims, V., & Tajika, H. (1995). A comparison of how textbooks teach mathematical problem solving in Japan and the United States. *American Educational Research Journal*, *32*(2), 443–460. https://doi.org/10.2307/1163438

Mesa, V. (2004). Characterizing practices associated with functions in middle school textbooks: An empirical approach. *Educational Studies in Mathematics*, *56*(2–3), 255–286. https://doi.org/10.1023/B:EDUC.0000040409.63571.56

National Council of Teachers of Mathematics. (2006). *Curriculum focal points for prekindergarten through Grade 8 mathematics: A quest for coherence*.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common core state standards for mathematics*. http://www.corestandards.org

Nesher, P., Greeno, J. G., & Riley, M. S. (1982). The development of semantic categories for addition and subtraction. *Educational Studies in Mathematics*, *13*(4), 373–394. https://doi.org/10.1007/BF00366618

Olkun, S., & Toluk, Z. (2002). Textbooks, word problems, and student success on addition and subtraction. *International Journal for Mathematics Teaching and Learning*. https://www.cimt.org.uk/journal/olkuntoluk.pdf

Otten, S., & de Araujo, Z. (2015). Viral criticisms of Common Core mathematics. *Teaching Children Mathematics*, *21*(9), 517–520. https://doi.org/10.5951/teacchilmath.21.9.0517

Remillard, J. T., Harris, B., & Agodini, R. (2014). The influence of curriculum material design on opportunities for student learning. *ZDM*, *46*(5), 735–749. https://doi.org/10.1007/s11858-014-0585-z

Remillard, J. T., & Heck, D. J. (2014). Conceptualizing the curriculum enactment process in mathematics education. *ZDM*, *46*(5), 705–718. https://doi.org/10.1007/s11858-014-0600-4

Riley, M. S., & Greeno, J. G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction*, *5*(1), 49–101. https://doi.org/10.1207/s1532690xci0501_2

Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In H. P. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153–196). Academic Press.

Sarama, J., Clements, D. H., Starkey, P., Klein, A., & Wakeley, A. (2008). Scaling up the implementation of a pre-kindergarten mathematics curriculum: Teaching for understanding with trajectories and technologies. *Journal of Research on Educational Effectiveness*, *1*(2), 89–119. https://doi.org/10.1080/19345740801941332

Schmidt, W. H., & Houang, R. T. (2012). Curricular coherence and the Common Core State Standards for Mathematics. *Educational Researcher*, *41*(8), 294–308. https://doi.org/10.3102/0013189X12464517

Schmidt, W., Houang, R., & Cogan, L. (2002). A coherent curriculum: The case of mathematics. *American Educator*, *26*(2), 1–18. https://www.aft.org/sites/default/files/periodicals/curriculum.pdf

Schmidt, W. H., McKnight, C. C., & Raizen, S. (1997). *A splintered vision: An investigation of U.S. science and mathematics education*. Springer.

Schoen, R. C., Anderson, D., & Bauduin, C. (2017). *Elementary mathematics student assessment: Measuring the performance of Grade K, 1, and 2 students in number, operations, and equality in spring 2016* (Research report No. 2017-22). Florida State University. http://doi.org/10.17125/fsu.1534964774

Schoen, R. C., Bray, W. S., Tazaz, A. M., & Buntin, C. K. (2022). *Description of the cognitively guided instruction professional development program in Florida: 2013–2020*. Florida State University. https://doi.org/10.33009/fsu.1643828800

Schoen, R. C., Champagne, Z., Whitacre, I., & McCrackin, S. (2021). Comparing the frequency and variation of additive word problems in United States first-grade textbooks in the 1980s and the Common Core era. *School Science and Mathematics*, *121*(2), 110–121. https://doi.org/10.1111/ssm.12447

Schoen, R. C., Erbilgin, E., & Haciomeroglu, E. S. (2011). Analyzing the Next Generation Sunshine State Standards for mathematics: Is the state curriculum still a mile wide and an inch deep? *Dimensions in Mathematics*, *31*(1), 30–39.

Schoen, R. C., & Koon, S. (2021). *Effects of an inquiry-oriented curriculum and professional development program on Grade 7 students' understanding of statistics and on statistics instruction* (Publication No. REL 2021–055). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=REL2021055

Schoen, R. C., LaVenia, M., Champagne, Z. M., Farina, K., & Tazaz, A. M. (2016). *Mathematics Performance and Cognition (MPAC) interview: Measuring first- and second-grade student achievement in number, operations, and equality in spring 2015* (Research report No. 2016-02). Florida State University. https://doi.org/10.17125/fsu.1493238666

Secada, W. G. (1991). Degree of bilingualism and arithmetic problem solving in Hispanic first graders. *The Elementary School Journal*, *92*(2), 213–231. https://doi.org/10.1086/461689

Secada, W. G., & Brandefur, J. L. (2000). CGI student achievement in region VI evaluation findings [Insert]. *The Newsletter of the Comprehensive Center–Region VI*, *5*(2). http://archive.wceruw.org/ccvi/zz-pubs/Newsletters/Fall2000_CGInSystemicReform/cgi_insert_only.pdf

Smith, J. P., III, Males, L. M., & Gonulates, F. (2016). Conceptual limitations in curricular presentations of area measurement: One nation's challenges. *Mathematical Thinking and Learning*, *18*(4), 239–270. https://doi.org/10.1080/10986065.2016.1219930

Son, J.-W. (2012). A cross-national comparison of reform curricula in Korea and the U.S. in terms of cognitive complexity: The case of fraction addition and subtraction. *ZDM*, *44*(2), 161–174. https://doi.org/10.1007/s11858-012-0386-1

Steffe, L. P. (1966). *The performance of first grade children in four levels of conservation of numerousness and three IQ groups when solving arithmetic addition problems* (Technical Report No. 14). Research and Development Center for Learning and Re-Education, University of Wisconsin. https://files.eric.ed.gov/fulltext/ED016535.pdf

Steffe, L. P., & Johnson, D. C. (1971). Problem-solving performances of first-grade children. *Journal for Research in Mathematics Education*, *2*(1), 50–64. https://doi.org/10.2307/748477

Stigler, J. W., Fuson, K. C., Ham, M., & Kim, M. S. (1986). An analysis of addition and subtraction word problems in American and Soviet elementary mathematics textbooks. *Cognition and Instruction*, *3*(3), 153–171. https://doi.org/10.1207/s1532690xci0303_1

Tarr, J. E., Grouws, D. A., Chávez, Ó., & Soria, V. M. (2013). The effects of content organization and curriculum implementation on students' mathematics learning in second-year high school courses. *Journal for Research in Mathematics Education*, *44*(4), 683–729. https://doi.org/10.5951/jresematheduc.44.4.0683

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC

Tarr, J. E., Reys, R. E., Reys, B. J., Chávez, Ó., Shih, J., & Osterlind, S. J. (2008). The impact of middles-grades mathematics curricula and the classroom learning environment on student achievement. *Journal for Research in Mathematics Education*, *39*(3), 247–280.

Tran, D. (2016). Statistical association: Alignment of current U.S. high school textbooks with the Common Core State Standards for Mathematics. *School Science and Mathematics*, *116*(5), 286–296. https://doi.org/10.1111/ssm.12179

Van de Walle, J. A., Lovin, L. H., Karp, K. S., & Bay-Williams, J. M. (2018). *Teaching student-centered mathematics: Developmentally appropriate instruction for grade pre-K–2* (3rd ed.). Pearson.

Van Dooren, W., De Bock, D., & Verschaffel, L. (2010). From addition to multiplication . . . and back: The development of students' additive and multiplicative reasoning skills. *Cognition and Instruction*, *28*(3), 360–381. https://doi.org/10.1080/07370008.2010.488306

Verschaffel, L., De Corte, E., & Vierstraete, H. (1999). Upper elementary school pupils' difficulties in modeling and solving nonstandard additive word problems involving ordinal numbers. *Journal for Research in Mathematics Education*, *30*(3), 265–285. https://doi.org/10.2307/749836

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

Wang, J., & Lin, E. (2005). Comparative studies on U.S. and Chinese mathematics learning and the implications for standards-based mathematics teaching reform. *Educational Researcher*, *34*(5), 3–13. https://doi.org/10.3102/0013189X034005003

Whitacre, I., Schoen, R. C., Champagne, Z., & Goddard, A. (2016). Relational thinking: What's the difference? *Teaching Children Mathematics*, *23*(5), 302–308. https://doi.org/10.5951/teacchilmath.23.5.0302

White, N., & Mesa, V. (2014). Describing cognitive orientation of calculus I tasks across different types of coursework. *ZDM*, *46*(4), 675–690. https://doi.org/10.1007/s11858-014-0588-9

Xin, Y. P. (2007). Word problem solving tasks in textbooks and their relation to student performance. *The Journal of Educational Research*, *100*(6), 347–360. https://doi.org/10.3200/JOER.100.6.347-360

Authors

Robert Schoen, Learning Systems Institute and School of Teacher Education, Florida State University, Tallahassee, FL 32306; rschoen@fsu.edu

Ian Whitacre, School of Teacher Education, Florida State University, Tallahassee, FL 32306; iwhitacre@fsu.edu

Zachary Champagne, The Discovery School, Jacksonville, FL 32217; zacharychampagne@gmail.com

Submitted May 8, 2021
Accepted August 13, 2021
doi:10.5951/jresematheduc-2020-0201

Unauthenticated | Downloaded 01/10/23 02:08 PM UTC