Universiteit Leiden
The Netherlands

Explanatory latent variable modeling of mathematical ability in primary school : crossing the border between psychometrics and psychology
Hickendorff, M.

The effects of presenting multidigit mathematics problems in a realistic context on sixth graders' problem solving

ABSTRACT

Mathematics education and mathematics assessments increasingly incorporate arithmetic problems in a context: a realistic situation that requires mathematical modeling. The aim of the present study was to assess the effects of presenting arithmetic problems in such a context on two aspects of problem solving: performance and strategy use. To that end, 685 sixth graders from the Netherlands solved a set of multidigit arithmetic problems on addition, subtraction, multiplication, and division. The total set consisted of eight pairs of problems; within each problem pair one problem was presented in a realistic context, and the parallel problem was in numerical format. Regarding performance, item response theory (IRT) models showed first that the same latent ability was involved in solving both types of problems, and second that the presence of a context affected the difficulty level only of the division problems, but not of the remaining operations. Regarding strategy use, results showed that strategy choice and strategy accuracy were not affected by the presence of a context in the problem. Importantly, the absence of context effects on performance and on strategy use was found to be independent of the student's gender, home language, and language achievement level. In sum, the present findings suggest that at the end of primary school the presence of a context in a mathematics problem had no marked effects on students' multidigit arithmetic problem solving behavior, contrary to expectations and common beliefs.

7.1 INTRODUCTION

Mathematics education has experienced a large international reform (e.g., Kilpatrick et al., 2001). A general characteristic of this reform is that mathematics education should no longer focus predominantly on decontextualized traditional mathematics skills, but that instead the process of mathematics problem solving and doing mathematics are important educational goals (e.g., National Council of Teachers of Mathematics, 1989, 2000). Word problems or the broader category of contextual problems[1] – typically a mathematics structure in a realistic problem situation – serve a central role for several reasons (e.g., Verschaffel et al., 2000): they may have motivational potential, mathematical concepts and skills may be developed in a meaningful way,

---

1 We defined *word problems* as problems containing only text, while the category *contextual* problems encompasses word problems, but also contains problems that include an illustration that may hold essential information for problem solving. In this study, we focus on the more general contextual problems.

and children may develop knowledge of when and how to use mathematics in everyday-life situations. Moreover, solving problems in context may ideally serve as a tool for mathematical modeling or mathematizing (e.g., Greer, 1997). As a consequence of this shift in educational goals, mathematics assessments include more and more contextual problems in their tests. For example, the PISA-2009 study (*Programme for International Student Assessment*; OECD, 2010) into mathematics included mainly problems presented in a real-world situation.

In the Netherlands, the reform is characterized by the principles of Realistic Mathematics Education (RME; Freudenthal, 1973, 1991; Treffers, 1993). In RME, contextual problems (defined as a problem that is experientially real to students) are central: they are the starting point for instruction, which is based on the principle of progressive schematization or mathematization by guided reinvention (Gravemeijer & Doorman, 1999). That is, contextual problems are postulated to elicit informal or naive solution strategies which are progressively abbreviated and schematized, a process guided by the teacher. In the last decades, RME has become the dominant instructional approach in mathematics curricula for Dutch primary education; in 2004, almost all elementary schools used a mathematics textbook based on RME principles (J. Janssen et al., 2005; Kraemer et al., 2005), although a return to more traditionally oriented mathematics textbooks has been observed recently (KNAW, 2009). These RME-based textbooks contain many problems in context, although there are substantial differences in this respect between the different textbooks. To link up with these developments, Dutch mathematics assessments (J. Janssen et al.; Kraemer et al.) and commonly used student monitoring tests also contain predominantly contextual problems. Therefore, today's Dutch primary school students' mathematics education and assessment consist for a large part of problems in realistic contexts.

This international shift toward the dominance of contextual mathematics problems gives rise to the main question asked in the current study: What is the effect on problem solving of presenting an arithmetic problem in a realistic context, as compared to the numerical problem format? Two aspects of problem solving are addressed: *performance* (i.e., accuracy) and *solution strategy use.* To our knowledge, there are no previous studies systematically investigating this issue. However, the growing importance of contextual problems in mathematics education as well as in mathematics assessments necessitates that we increase our understanding of the impact of contexts, both on a theoretical level (what aspects of mathematical cognition are involved?) as well as from a practical

educational perspective (what are the implications regarding testing and instruction practices?). Moreover, because the contexts in mathematics problems are usually verbal, special attention for students with low language level is called for.

7.1.1 Earlier studies into the effects of contexts on performance

Word problems can be considered a subcategory of the broader class of mathematics problems with a realistic context. Many studies have been carried out in the field of word problems, particularly in the domain of addition and subtraction. These studies focused mainly on the differences between different types of word problems (for an overview, see Verschaffel et al., 2007), thereby only allowing for comparisons *within* the class of word problems. Word problems contain only linguistic information, while the more general class of contextual problems can also contain other sources of information such as illustrations. A recent study investigated contextual mathematics problems (Berends & Van Lieshout, 2009), focusing on the effect of illustrations. This study, therefore, also allowed only for comparisons *within* different contextual problems. By contrast, research in which contextual problem solving is compared directly to solving bare numerical problems without a context is rare. Therefore, the current study aims to extend the existing literature on this issue.

Solving numerical and contextual problems (sometimes also called 'computations' and 'applications') is likely to involve different aspects of mathematical cognition. Solving contextual problems involves a complex process consisting of several cognitive processes or phases. Only after steps in which a situational and mathematical model of the problem situation have been formed accurately (mathematization), computational skill – and carefulness therein – comes into play. Therefore, other factors than 'pure' computational skills are likely to contribute to success in solving contextual problems (Fuchs et al., 2006, 2008; Wu & Adams, 2006). This also yields the expectation that contextual problems are more difficult to solve than numerical problems, as supported by early research findings of Cummins et al. (1988) in simple addition and subtraction word problems. So, the effects of contexts on performance can be of two kinds: different abilities may be involved in solving problems with and without a realistic context, and/or the context may affect the difficulty level of a problem.

Recently, some studies empirically investigated to what extent different abilities were involved in solving the two types of problems in American third graders (Fuchs et al.,

2006, 2008) and in Dutch first to third graders (Hickendorff, 2010b). These studies showed that solving numerical mathematics problems and solving contextual problems involved two highly related but distinct abilities, as evidenced by a less than perfect correlation between performance measures, as well as different cognitive correlates for the two measures. However, these studies did not allow a direct investigation of the effects of a realistic context on difficulty level of a problem, because problems with different numerical characteristics were used. A study in which such a design was employed was the study of Vermeer et al. (2000) into sixth graders' problem solving on parallel problems on computation and on application. Regrettably, a direct test comparing performance on the two types of problems was not reported. However, the proportion correct was slightly higher on applications (i.e., contextual problems) than on computations (i.e., numerical problems), thereby contradicting the expectation that contextual problems are more difficult to solve. Since theoretical hypotheses and empirical results are is inconclusive, systematic study is needed.

The present study extends the previous studies in two ways. Most importantly, it directly investigates the effect of problem format (with or without a context) on problem solving in a systematic test design, consisting of problem pairs in which one problem was presented with a realistic context and the parallel problem without without such a context. Second, a more complete account of problem solving was taken: besides addressing *performance* also *strategy use* was studied.

7.1.2 Solution strategies

Performance or accuracy is probably the most salient aspect of problem solving, and many studies into solving problems with and without a realistic context focused only on that aspect (Fuchs et al., 2006, 2008; Hickendorff, 2010b; Vermeer et al., 2000). However, another important aspect of problem solving is *strategic competence*. From cognitive psychology, it is well-established that adults and children know and use multiple solution strategies to solve mathematics problems (e.g., Lemaire & Siegler, 1995; Siegler, 1988a). Furthermore, solution strategies are important from the perspective of mathematics education as well, in at least two ways. First, the didactics for solving complex arithmetic problems have changed, from instructing standard written algorithms to building on children's informal or naive strategies (Freudenthal, 1973; Treffers, 1987, 1993), and mental arithmetic has become very important (Blöte et al., 2001). Second, mathematics

education reform aims at attaining adaptive expertise instead of routine expertise: instruction should foster the ability to solve mathematics problems efficiently, creatively, and flexibly, with a diversity of strategies (Baroody & Dowker, 2003; Torbeyns, De Smedt, et al., 2009b).

Lemaire and Siegler (1995) distinguished four aspects of strategic competence: strategy repertoire, strategy choice, strategy performance (such as accuracy), and strategy adaptivity. The current study focuses on the first three of these aspects, on the domain of multidigit arithmetic. In the domain of elementary or simple arithmetic, strategy use has been studied extensively: in elementary addition and subtraction (e.g., Carr & Jessup, 1997; Carr & Davis, 2001; Torbeyns et al., 2004b, 2005), in elementary multiplication (e.g., Anghileri, 1989; Imbo & Vandierendonck, 2007; Lemaire & Siegler, 1995; Mabbott & Bisanz, 2003; Mulligan & Mitchelmore, 1997; Sherin & Fuson, 2005; Siegler, 1988b), and in elementary division (e.g., Robinson et al., 2006). By contrast, research on solution strategies in complex or multidigit arithmetic problems is less extensive, but there is a growing body of studies in multidigit addition and subtraction (e.g., Beishuizen, 1993; Beishuizen et al., 1997; Blöte et al., 2001; Torbeyns et al., 2006) and in multidigit multiplication and division (e.g., Ambrose et al., 2003; Buijs, 2008; Hickendorff et al., 2009b; Hickendorff & Van Putten, 2010; Hickendorff et al., 2010; Van Putten et al., 2005).

The current study addressed multidigit arithmetic involving the four basic operations. Based on the solution strategies reported in the aforementioned studies in multidigit addition, subtraction, multiplication, and division (see also a recent review by Verschaffel et al., 2007), a classification scheme of written solution strategies was developed (i.e., the strategy repertoire). For each of the four operations, a basic distinction can be made between the *traditional standard algorithm* that proceeds digit-wise, *non-traditional procedures* that work with whole numbers, and *answers without written working*. A subcategory of the non-traditional strategies are the RME approaches (labeled 'columnwise arithmetic' by the developers, see Treffers, 1987, and Van den Heuvel-Panhuizen, Buys, & Treffers, 2001). These can be considered transitory between informal approaches and the traditional algorithm: they work with whole numbers instead of single-digits (like informal strategies), but they proceed in a more or less standard way (like the traditional algorithm). More details are given in the Method-section.

Based on the literature, we had the following expectations regarding the effects of problem format (contextual or numerical) on strategy use. Studies on elementary word problem solving with young children showed that different semantic structures of word

problems elicited different strategies (for a review, see Verschaffel et al., 2007). Although comparisons with bare numerical problems were not made explicitly in these studies, extending these findings would still lead to the expectation that contextual and numerical problems on multidigit arithmetic would also elicit different strategies. In particular, the theory behind the RME didactical approach yields the expectation that problems in a realistic context would be more likely to elicit more informal, less structured strategies (i.e., non-traditional strategies), while numerical problems would elicit more use of traditional algorithms (Van den Heuvel-Panhuizen et al., 2009). However, Van Putten et al. (2005) investigated Dutch fourth graders strategy use on multidigit division problems that did or did not include a context, and found no differences in strategy choice between the two types of problems. Given these inconsistent findings, the effects of contexts on strategy use requires further systematic study.

7.1.3   The role of language and gender

In the current study, the role of three student characteristics on problem solving was investigated: language ability level, language spoken at home, and gender. The first two characteristics were of interest because differential effects on solving numerical problems versus solving contextual problems were expected. Because the problem situation in a contextual problem is usually verbal, and a necessary condition for obtaining the correct answer is that this problem situation is accurately understood, it is likely that the student's language ability plays an important role. Support for the importance of language in word problem solving comes from the finding that language ability had smaller effects on computational skills (numerical problems) than on applied problem solving (contextual problem solving) (Fuchs et al., 2006, 2008; Hickendorff, 2010b). Additional support comes from the finding that a common source of errors in word problem solving appears to be misunderstanding of the problem situation (Cummins et al., 1988; Wu & Adams, 2006), and that conceptual rewording of word problems facilitated performance (e.g., Vicente et al., 2007). Therefore, we expect that the effect of language ability level is larger on performance in solving contextual problems than in solving numerical problems.

Ethnic minority pupils score lower on language ability tests than native pupils. In addition, they have been consistently found to lag behind in mathematics achievement too, as has been found in international assessments such as TIMSS-2007 (*Trends in International Mathematics and Science Study*; Mullis et al., 2008) as well as in Dutch

national assessments (J. Janssen et al., 2005; Kraemer et al., 2005). An obvious question is whether language level plays a role in the performance lag of ethnic minorities on mathematics problems that involve a verbal context. Several research findings with students in secondary education showed that difficulty of the problem text hampers particularly the pupils for whom the language in the test is not their native language (Abedi & Hejri, 2004; Abedi & Lord, 2001; Prenger, 2005; Van den Boer, 2003), due to text aspects like the use of unfamiliar vocabulary, passive voice construction, and linguistic complexity. Therefore, we expect differences with respect to the language spoken at home to be larger on the performance in solving contextual problems than in solving numerical problems.

The final student characteristic considered in the present study was gender. Gender differences in general mathematics performance have been reported frequently. Large-scale international assessments TIMSS-2007 (Mullis et al., 2008) and PISA-2009 (OECD, 2010) showed that boys tend to outperform girls in most of the participating countries, including the Netherlands. This pattern is supported by Dutch national assessments findings: on most mathematical domains boys outperformed girls in third and in sixth grade (J. Janssen et al., 2005; Kraemer et al., 2005). However, in grade 6, the multidigit operations were the exception: girls slightly outperformed boys. Moreover, Vermeer et al. (2000) found that in Dutch sixth graders, there were no gender differences in performance on computations, while boys outperformed girls on applications, possibly mediated by the finding that on these problems, girls had lower levels of subjective competence than boys and attributed bad results to lack of capacity and difficulty of the task. Based on these results, we expect that gender differences in performance to be larger on contextual problems than on numerical problems.

Regarding strategy choice, girls have been found to be more inclined to (quite consistently) rely on rules and procedures and use well-structured strategies, whereas boys have a larger tendency to use more intuitive strategies (Carr & Davis, 2001; Carr & Jessup, 1997; Gallagher et al., 2000; Hickendorff et al., 2009b, 2010; Hickendorff & Van Putten, 2010; Timmermans et al., 2007, Vermeer et al., 2000). There are no empirical findings on whether this pattern is the same for numerical problems as for contextual problems.

7.1.4 *The current study*

The current study's main objective was to systematically study the effect of the presence of a context in mathematics problems on two aspects of problem solving: performance and strategy use (strategy choice and strategy accuracy). To that end, a sample of Dutch students from Grade 6 (12-year-olds) were asked to solve a set of multidigit arithmetic problems on addition, subtraction, multiplication, and division. The set consisted of pairs of problems, and within each problem pair one problem was presented in a context, and the parallel problem was not. Based on the previous discussion of existing theoretical literature and empirical findings, we had the following expectations. Regarding performance, we expected that two highly related but distinct abilities would be involved in solving the two types of problems, and that contextual problems are more difficult, in particular for students with low language level as well as for girls. Regarding strategy use, we expected that contextual problems would elicit more use of informal, less structured strategies than numerical problems.

7.2 METHOD

7.2.1 *Participants*

Participants were 685 students from Grade 6 with mean age 12 years 0 months ($SD = 5$ months), originating from 24 different primary schools, with 3 to 82 students participating per school (on average 27.4 students per school). These schools were spread over the entire country of the Netherlands. There were 312 boys, 337 girls, and 36 students with missing gender information in the sample. In order to assess language level effects with sufficient power, the schools that were selected had relatively many ethnic minority pupils. As a consequence, the current sample of schools and pupils was not entirely representative for the population of Dutch primary schools.

Information on the language spoken at the students' home was gathered (missing data for 42 students). Students with observed home language data were classified into home language *Dutch* (either only Dutch (517 students, 80%) or Dutch as well as another language (46 students, 7%) or home language *non-Dutch* (80 students, 12%). The most prevalent non-Dutch language was Arabic (45% of student with home language other than Dutch), followed by Turkish (26%).

7.2.2   Material

Experimental task

The experimental task consisted of 16 multidigit arithmetic problems, built up with 8 pairs of one contextual and one numerical problem each. There were 2 pairs on multidigit addition, 2 pairs on multidigit subtraction, 2 pairs on multidigit multiplication, and 2 pairs on multidigit division, see also Appendix 7.A.

At the basis of each problem pair lied a contextual problem that was selected from the most recent Dutch national assessment (J. Janssen et al., 2005). For each operation, two problems were selected: one at the lower end of the ability scale (i.e., a relatively easy problem with small numbers), and one at the upper end of the scale (i.e., a relatively hard problem with large numbers). We used problems from the assessments to ensure that they were representative for the type of contextual problems that are used in current educational practices. For the numerical problems, these contextual problems were disposed of their contexts to yield the bare numerical operation required. In order to avoid testing effects that may have occurred if students had to solve exactly the same numerical operation twice (once with and once without a context), a parallel version of each problem was constructed with numbers and solution steps as similar as possible.

Two different test forms were created, so that item parallel version was counterbalanced over test form. That is, in form A item versions a were presented as contextual problems and item versions b as numerical problems, and in form B this pattern was reversed. For example, the first item pair on Addition in Appendix 7.A presents the problems as presented in form A. In form B, the numbers were switched: i.e., the text of the contextual problem said that 677.50 euro was sold on postcards and 975 euro on stamps, while the numerical problem was $466.50 + 985 = ?$. Figure 7.1 present the specific position of each problem in both task forms. Within each form, paired problems (e.g. A_{1a} and A_{1b}) were presented with 7 other problems in between, to prevent recency effects. The order of the 16 different problems was the same in both task forms, to rule out potentially confounding order effects in combining the data from the two forms.

In the task booklets that students received, there were at most 3 problems printed on the left side of a page (A4 size). The right side of each page was left blank, so that students could use that space as scrap paper in solving the problems.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
form A	A_{1b}	S_{2a}	A_{2a}	S_{1b}	M_{1b}	D_{2a}	M_{2a}	D_{1b}	A_{1a}	S_{2b}	A_{2b}	S_{1a}	M_{1a}	D_{2b}	M_{2a}	D_{1b}
form B	A_{1b}	S_{2a}	A_{2a}	S_{1b}	M_{1b}	D_{2a}	M_{2a}	D_{1b}	A_{1a}	S_{2b}	A_{2b}	S_{1a}	M_{1a}	D_{2b}	M_{2a}	D_{1b}

FIGURE 7.1 *Design of experimental task forms. A = Addition, S = Subtraction, M = Multiplication, and D = Division. Problem indices 1 (small numbers) and 2 (large numbers) denote the specific pair within each operation, indices a and b denote the two parallel versions within each problem pair. Problems in unshaded cells present numerical problems, problems in cells shaded gray are the contextual problems.*

Standardized tests

The students participating in the current study took part in the 2009 administration of CITO's End of Primary School Test (CITO, 2009) in February. This test is widely used in the Netherlands at the end of primary school, and its purpose is to give advice on the most suitable track of secondary education for each student. To that end, the instrument assesses achievement level on mathematics, language, and study skills. In 2009, over 150,000 Dutch sixth graders participated. The 100-item subtest on language skills consisted of items on writing, spelling, reading comprehension, and vocabulary, and had high internal reliability (KR20 = .89; CITO, 2009). In the current sample, the average number of language items correct was 73.0 (SD = 11.7; missing data for 29 students). This mean score was slightly lower than for the entire population of students participating in the End of Primary School Test, who scored on average 75.2 items correct (SD = 12.0). On the 60-item subtest on mathematics (KR20 = .91), the current sample scored on average 41.7 items correct (SD = 10.6), which was also slightly lower than the average of 42.8 correct (SD = 10.5) for all participants nationwide.

7.2.3 Procedure

Experimental task

The experimental task was administered as part of a pretest study for the CITO End of Primary School Test. A test booklet consisted of 6 tasks, divided over the different subjects mathematics, language, and study skills. Students completed each of these tasks

on a separate day in January 2009. One of the mathematics tasks included the current experimental task of 16 problems, and an additional 12 problems that were not part of the current study. One of the two experimental task forms (A or B) was assigned to each class.

The task was administered in the classroom, and each student worked individually. Teachers instructed their students that they were free to choose their solution strategy. Moreover, students were told that they could use the blank space next to each problem in the test booklet to make computations, and that they did not need separate scrap paper apart from their test booklet. Students could take as much time as they needed, so there was no time pressure.

Standardized tests

The students completed the 2009 End of Primary School Test (CITO, 2009) as part of their final year's standardized assessment in February 2009, at most one month after the students participated in the current study.

7.2.4 Solution strategies

The solution strategy used on each trial (student-by-item combination) was categorized based on the notes or solution procedures that students had written in the test booklet. Strategy data were available for 650 students. Three experts (the first author and two trained research assistants) each coded a separate part of the material. Table 7.1 shows the 9 (addition) or 10 (subtraction, multiplication, and division) different categories of solution strategies that were distinguished, and Appendix 7.B shows examples of categories 1 to 5 for each operation. Below, first the operation-specific categories 1 to 6 are discussed, and after that the operation-general categories 7 to 10 are explained.

Addition

Category 1 (traditional algorithm) coded strategies in which the standard algorithm for adding two or more multidigit numbers was applied. The addends have to be aligned vertically so that digits on the same position represent the same value, and addition proceeds digit-wise from right to left starting with the ones-digits (assuming there are no decimals), next the tens-digits, then the hundreds-digits, and so forth. If the outcome of any particular sub-addition is larger than 10, digits have to be carried to the next column.

TABLE 7.1 *Categories solution strategies.*

	addition	subtraction	multiplication	division
1	traditional algorithm	traditional algorithm	traditional algorithm	traditional algorithm
2	RME approach	RME approach	RME approach	repeated subtraction (HL)
3	partitioning 1 operand	partitioning 1 operand	partitioning 1 operand	repeated subtraction (LL)
4	partitioning $\geq$ 2 operands	partitioning 2 operands	partitioning 2 operands	repeated addition (HL)
5	-	indirect addition	repeated addition	repeated addition (LL)
6	other written strategy	other written strategy	other written strategy	other written strategy
7	no written working	no written working	no written working	no written working
8	wrong procedure	wrong procedure	wrong procedure	wrong procedure
9	unclear strategy	unclear strategy	unclear strategy	unclear strategy
10	skipped problem	skipped problem	skipped problem	skipped problem

Category 2 is the RME approach to addition. It contrasts with the traditional algorithm because it proceeds from left to right and it works with numbers instead of single-digits (e.g., $600 + 900 = 1500$ instead of $6 + 9 = 15$). Category 3 and 4 are partitioning strategies, in which either one or more than one of the operands is partitioned or split according to its place value (e.g., 975 is split into 900, 70, and 5). Partitioning of only the second operand is also called the jump or sequential strategy, while partitioning of two or more operands is called the split or decomposition strategy (e.g., Beishuizen, 1993). The final category 6 (note that we left out category number 5 for addition to be consistent with the other operations) included all kinds of strategies in which some calculations or intermediate solutions were written down from which could be inferred how the answer was obtained, but that did not fit in categories 1 to 4.

Subtraction

Category 1 (traditional algorithm) involved application of the standard algorithm for subtraction of two multidigit numbers. Similar to the addition algorithms, the two numbers have to be aligned vertically so that digits on the same position represent the same value, and subtraction proceeds digit-wise from right to left starting with the ones-digits (assuming there are no decimals), next the tens-digits, then the hundreds-digits, and so forth. In case a larger digit has to be subtracted from a smaller one (e.g., 0 - 9), borrowing from the column to the left is necessary. Category 2 is the RME approach to multidigit subtraction. It contrasts with the traditional algorithm because it proceeds from left to right and it works with numbers instead of single-digits. Moreover, there is

no need of borrowing: it works with negative numbers instead (e.g., $10 - 80 = -70$). Category 3 and 4 are partitioning strategies, in which either only the subtrahend or both operands are partitioned according to their place value (e.g., 689 is split into 600, 80, and 9). Similar to addition, partitioning of only the subtrahend is also called the jump or sequential strategy, while partitioning of both operands is called the split or decomposition strategy (Beishuizen, 1993). Category 5 involved indirect addition strategies. In these approaches, one starts from the subtrahend and adds on until the minuend is reached (see for example Torbeyns, Ghesquière, & Verschaffel, 2009). The final category 6 (other written strategy) included all kinds of strategies in which some calculations were written down, but that did not fit in categories 1 to 5.

Multiplication

The traditional standard algorithm for multiplication (category 1) involves writing the two operands below each other, and multiplying the upper number by each digit of the lower number separately, working from right to left. Then, these partial outcomes are added to obtain the solution. The RME approach (category 2) again contrasts with this algorithm because it proceeds from left to right. Furthermore, both numbers are partitioned and all sub-products are obtained and added. This strategy resembles the one in category 4 (partitioning of both operands), but the difference lies in the schematic notation that is applied in category 2. In category 3, only one of the operands is partitioned, while the other one is left intact (e.g., $36 \times 27 = 36 \times 20 + 36 \times 7$). Category 5 involved repeated addition, in which there is made use of the fact that multiplying by a factor n is equivalent to adding the multiplicand n times. This category included strategies in which either the multiplicand was added n times, or when doubling strategies were used (e.g., $2 \times 27 = 54$; $4 \times 27 = 108$; $8 \times 27 = 216$, …). Again, the final category 6 included all kinds of strategies in which some calculations were written down, but that did not fit in categories 1 to 5.

Division

The first category of division was the long division algorithm (note that notation may differ between countries). The algorithm is characterized by starting on the left side of the dividend, and trying to divide the first digit by the divisor (e.g., $7 \div 32$). If that yields a number smaller than 1, the first two digits are considered together (73) and the maximum number of times the divisor fits in ($2 \times 32 = 64$) is noted. Then, the difference

is determined ($73 - 64 = 9$) and the digit from the column to the right is pulled down (making 96). This procedure continues until the remainder is zero. Categories 2 and 3 involve repeated subtraction strategies (in the Netherlands this is the RME alternative for long division in the mathematics textbooks, Treffers, 1987). Multiples of the divisor are repeatedly subtracted from the dividend. This can be done efficiently with relatively few steps (high-level, HL) or less efficiently with many steps (low-level, LL). In the present study, we defined strategies as high-level when at most 3 steps (the minimum number of steps + 1) were taken. It is worth noting that the most efficient repeated subtraction strategy resembles the traditional algorithm, with the main difference that in the algorithm one proceeds digit-wise, while in repeated subtraction one works with whole numbers (e.g., 640 instead of 64). Categories 4 and 5 resemble categories 2 and 3, respectively, but they differ in the approach: repeatedly adding multiples of the divisor until the dividend is reached, as opposed to repeatedly subtracting from the dividend until zero is reached. The same distinction between high-level (maximum 3 steps) and low-level approaches was made. Like on the other operations, category 6 (other written strategy) involved strategies in which some calculations were written down, but that could not be classified in categories 1 to 5.

Remainder categories

The remaining strategy categories 7 to 10 were the same for the four operations. Category 7 (no written working) includes all trials in which an answers was written down, but nothing else (i.e., no calculations or intermediate solutions), so it is very likely that the answer was computed mentally (supported by findings of Hickendorff et al., 2010). Category 8 (wrong procedure) includes trials in which the wrong procedure was applied, such as adding the two numbers in a division problem. In trials classified in category 9 (unclear strategy) it was unclear how the student arrived at the answer (s)he had given, in some cases because the written solution steps were erased. The final category 10 (skipped problem) included trials in which the problem was skipped entirely, i.e., no answer was given and no solution steps were written down.

Reliability

The solution strategies of 45 students (720 trials; 180 trials per operation) were double-coded by two independent raters to assess the agreement in categorization. Cohen's

TABLE 7.2 *Descriptive statistics of performance (proportion correct) on numerical and contextual problems, by operation, gender, and home language.*

	numerical		contextual		total		
proportion correct	*M*	SD	*M*	SD	*M*	SD	*N*
item operation							
addition	.70	.36	.72	.33	.71	.28	685
subtraction	.72	.35	.73	.35	.72	.30	685
multiplication	.69	.36	.68	.36	.69	.31	685
division	.75	.36	.70	.37	.73	.31	685
gender							
boy	.69	.25	.70	.24	.69	.23	312
girl	.75	.22	.72	.24	.73	.22	337
home language							
Dutch	.72	.24	.71	.25	.71	.23	563
non-Dutch	.72	.21	.71	.22	.71	.19	80
total	.71	.24	.71	.24	.71	.22	685

kappa (Cohen, 1960) on the cross-tabulation of the categorization of the two raters was computed as a measure of inter-rater reliability. Kappa values were sufficiently high with .82, .79, .89, and .92 for addition, subtraction, multiplication, and division, respectively, indicating substantial and satisfactory agreement.

7.3   DATA ANALYSIS AND RESULTS

In all data analyses, the data were collapsed over the two test forms A and B, thereby counterbalancing potential differences between parallel item versions within problem pairs. The results are presented in two parts: performance and strategy use.

7.3.1   *Performance*

Table 7.2 presents descriptive statistics of performance (proportion of problems correct) on numerical and contextual problems, by the operation required, by gender, and by student's home language. It shows that the proportions correct on contextual problems and numerical problems were very close on each of the four operations (with the largest difference on division problems), as well as for boys and for girls and for students with home language Dutch or another home language.

Item response theory (IRT) modeling was used to statistically test the main question: What is the effect of presence of a context on performance? As discussed before, we studied this effect in two ways: first, we established whether different (latent) abilities are involved in solving items with and without a context (the multidimensionality hypothesis), and second, the effect of problem format on the difficulty level of an item was tested (the sources-of-difficulty hypothesis). Both hypotheses were explored with item response theory (IRT) models (e.g., Embretson & Reise, 2000; Van der Linden & Hambleton, 1997). In the most simple IRT model, the Rasch model, the probability P_{is} that person s solves item i correctly is modeled as a logistic function of the difference between the person's latent ability level θ_s and the item's difficulty level β_i: $P_{is} = \frac{\exp(\theta_s - \beta_i)}{1 + \exp(\theta_s - \beta_i)}$. We used the `lmer` function from the `lme4` package (Bates & Maechler, 2010) available in the statistical computing program R (R Development Core Team, 2009) to estimate the model parameters. For further details on how to fit IRT models with `lmer`, see De Boeck et al. (2011).

Multidimensionality

To explore whether different latent abilities were involved in solving items with and without a context, we used multidimensional IRT (MIRT) modeling (see Reckase, 2009). Specifically, a confirmatory two-dimensional IRT model was used, in which each item was assigned a priori to one of the two dimensions *solving numerical problems* and *solving contextual problems*. Figure 7.2 shows a graphical display of this model. Such a model belongs to the class of between-item or simple structure Rasch models, in which it is assumed that multiple related subscales or ability dimensions underlie test performance, and that each item in the test is only related to one of these subscales (Adams et al., 1997).

Our main interest lied in the estimate of the latent correlations between the two ability dimensions θ_1 and θ_2. A latent correlation estimate in a MIRT model is not attenuated by measurement error: it is an unbiased estimate of the true correlation between the latent variables (Adams & Wu, 2000; Wu & Adams, 2006). Therefore, it is a better alternative than computing the correlation on ability estimates of consecutive unidimensional models, or on classical test theory approaches that are based on the proportion of items solved correctly (as was done in the studies by Fuchs et al., 2006, 2008).

The results of fitting a 2-dimensional between-item Rasch model showed that the
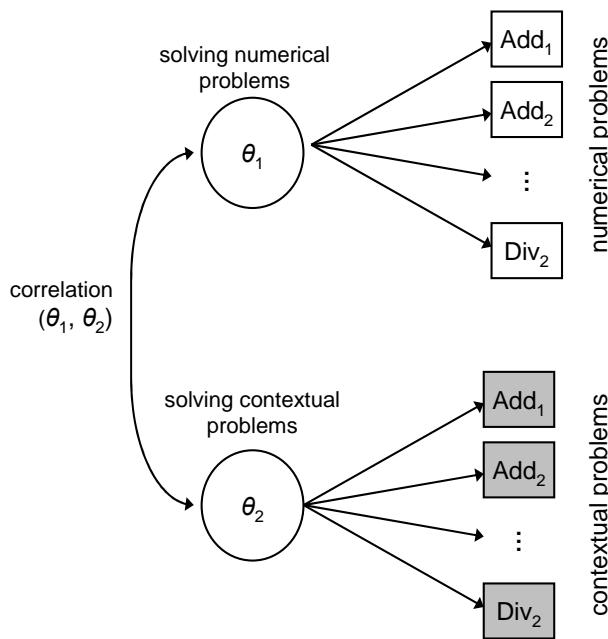
FIGURE 7.2 *Graphical representation of between-item two-dimensional IRT model.*

latent correlation between the ability to solve numerical problems and the ability to solve contextual problems was estimated at 1.000. [2] Therefore, we conclude that solving numerical problems and solving mathematics problems in a context involves one (latent) ability factor.

[2] We cross-checked this latent correlation estimate by using other estimation methods and software. In a first approach, item parameter were estimated for each dimension separately using conditional maximum likelihood (Verhelst & Glas, 1995) and the latent correlation between the dimensions was estimated in a separate step. Second, we used a Bayesian framework: the MIRT models were formulated as normal-ogive instead of logistic models, and parameters were estimated using an MCMC-procedure (see also Albert, 1992 for unidimensional IRT models and Béguin & Glas, 2001 for MIRT models), that was programmed into R(R Development Core Team, 2009)). Finally, we used the NLMIXED-procedure from SAS (see De Boeck & Wilson, 2004 and Hickendorff, 2010b) to fit the 2PL extension of the MIRT model, allowing for the nonzero discrimination parameters to be different from each other. In all three alternative programs, the latent correlation was estimated at 1.000, so we conclude that it is a robust result.

Sources of item difficulty

The next step was to assess the effect of item format on the difficulty level of an item. For that end, we used an extended Linear Logistic Test Model (LLTM; Fischer, 1987). The LLTM is an example of a broad class of explanatory IRT models, in which predictors on the item level, the person level, and on the item-by-person level can be incorporated in the model (De Boeck & Wilson, 2004). The LLTM allows for decomposition of item difficulty β_i into the effects of K different item features, in a multiple regression-like way: $\beta_i = \sum_{k=1}^{K} \tau_k q_{ik} + \tau_0$. The q_{ik} entries of the so-called Q-matrix specify the involvement of item feature k in item i, and have to be assigned a priori. The LLTM has the drawback that it assumes that the K item features predict item difficulty without error. To relax this assumption, the LLTM can be extended by incorporating error that is randomly distributed over items in the model, yielding the LLTM + e model (De Boeck, 2008; R. Janssen, Schepers, & Peres, 2004).

There were three item features: operation required (nominal variable with 4 categories: addition, subtraction, multiplication, and division; recoded into 3 dummy variables), number size of the problem (dichotomous variable with 2 categories: small or large), and item format (dichotomous variable with 2 categories: numerical or contextual). Our main interest was in the effect of item format, statistically correcting for the covariates operation and number size. In addition, we corrected for possible differences between students who were administered test form A or test form B, by including 'test form' in the model as well. Because this was a variable on the student level, the final IRT models that we used could be characterized as *latent regression* LLTM + e models.

Results showed that the main effect of item format was not significant ($\tau_{\text{context}} = .05$, $p = .30$). Testing for interactions between item format and the other two item features, it was found that the interaction between item format and number size was not significant (Likelihood Ratio $\chi^2(1) = .0$, $p = 1.00$), while the interaction between item format and operation was significant ($\chi^2(3) = 16.4$, $p < .001$). Further inspection of this latter interaction showed that only on division problems the effect of context format was significant, $\tau_{\text{context, division}} = .35$, $p < .001$, while for the other three operations the effect of context format was not significant ($\tau_{\text{context, addition}} = -.15$, $p = .10$; $\tau_{\text{context, subtraction}} = -.08$, $p = .36$, and $\tau_{\text{context, multiplication}} = .08$, $p = .35$). So, only for division problems the context made the item more difficult compared to the bare

numerical problem, and on the other three operations the contextual problems and the numerical problems were just as difficult.

Final analyses were carried out to assess whether the effect of item format (numerical vs. contextual) on item difficulty depended on either gender, language achievement level, or home language. First, the main effects of the student characteristics on performance are reported. Gender had a significant effect on performance: girls had significantly more correct answers than boys ($\beta_{\text{girl}} = .23$, $p = .03$). In contrast, the effect of home language on performance was not significant ($\beta_{\text{other than Dutch}} = -.04$, $p = .82$). Finally, the effect of language achievement level was highly significant and positive: $\beta_{\text{language}} = .58$, $p < .001$. Next, we focus on the interactions between student characteristics and item format. The interaction between item format and gender ($\chi^2(4) = 4.4$, $p = .35$), home language ($\chi^2(4) = 1.1$, $p = .89$), and language achievement level ($\chi^2(4) = 4.1$, $p = .39$) all turned out to be nonsignificant. So, the effects of presenting an item in a context did not depend on either gender, home language or language achievement level of the student.

7.3.2 Strategy use

Strategy choice

Table 7.3 shows the distribution of the strategy categories for addition, subtraction, multiplication, and division, split by problem format (numerical versus contextual). There were only small problem format differences in the distribution of strategies: The percentage of trials solved by each strategy were very similar for the numerical problems as for the contextual problems, with the largest difference being 3 percent points.

In contrast, operation required seemed to have large effects on strategy choice distribution. First, differences in the use of the traditional algorithm (category 1) emerged. It was the dominant strategy for addition and subtraction (used on around 70% of the trials), and also for multiplication it was the most prevalent strategy although its dominance was less pronounced, being used on over 50% of the multiplication trials. For division, however, the traditional algorithm was only used on 15% of all division trials. Second, the frequency of answering without written working (category 7) was about the same for each operation, occurring on between 14% and 17% of all trials. Moreover, wrong procedures (category 8), unclear strategies (category 9), and skipping the entire problem (category 10) occurred not very often, and with about the same frequency on each operation but slightly more often on division than on the other three operations.

TABLE 7.3 *Distribution in proportions of solution strategy categories of numerical (num) problems and contextual (con) problems, per operation. Strategy categories refer to Table 7.1.*

	addition		subtraction		multiplication		division	
strategy	num	con	num	con	num	con	num	con
1 (traditional)	.68	.70	.74	.73	.52	.51	.15	.14
2 *	.09	.09	.01	.01	.03	.02	.53	.50
3 *	.01	.01	.01	.00	.18	.17	.05	.04
4 *	.04	.03	.01	.01	.07	.07	.03	.03
5 *	n.a.	n.a.	.02	.04	.01	.02	.01	.01
6 *	.02	.01	.01	.01	.01	.01	.02	.03
7 (no written)	.14	.14	.17	.17	.16	.16	.14	.17
8 (wrong)	.00	.00	.00	.00	.00	.01	.01	.02
9 (unclear)	.01	.01	.01	.01	.01	.01	.02	.02
10 (skipped)	.01	.00	.01	.01	.01	.02	.04	.04
N observations	1300	1300	1300	1300	1300	1300	1300	1300

* Operation-specific strategy categories, see Table 7.1.

Third, some operation-specific patterns emerged. Most notably, on division the high-level repeated subtraction strategy was the most prevalent strategy, used on over 50% of all division trials. On multiplication, partitioning of 1 operand (category 3) was used quite often (17% of all trials). Finally, on addition, the RME approach was used relatively often (9% of all trials).

In order to statistically test the effects of operation and problem format on strategy choice distribution, we first collapsed some categories in order to make strategy categories comparable across operations, and to obtain categories filled with a substantial number of observations. We recoded the 9 or 10 operation-specific strategies into 4 operation-general categories: the traditional algorithm (former category 1), non-traditional strategies (former categories 2 to 6), no written working (former category 7), and other trials (former categories 8 to 10). Figure 7.3 presents the proportion of choice of each of these four strategy types on numerical and contextual problems, per operation.

Next, we estimated a multinomial logistic model for correlated responses using a random effects model (Hartzel, Agresti, & Caffo, 2001) in the SAS procedure NLMIXED, as described by Kuss and McLerran (2007). As predictor variables, we first included
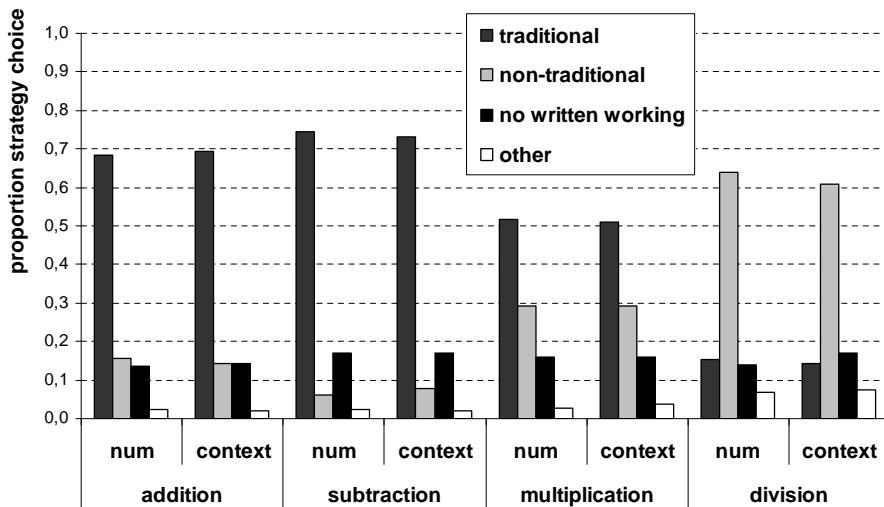
FIGURE 7.3 *Strategy choice proportion of recoded solution strategies on numerical (num) and contextual (context) problems, per operation.*

only variables on the item level: operation required (nominal variable with 4 categories: addition, subtraction, multiplication, or division; recoded into 3 dummy variables) and problem format (numerical or contextual), and their interaction.

Operation had a highly significant effect (Likelihood Ratio $\chi^2(9) = 3995.4$, $p < .001$) on the distribution of strategies, as is also obvious from Figure 7.3. However, the main effect of problem format was not significant ($\chi^2(3) = 4.9$, $p = .18$), and neither was the interaction between problem format and operation ($\chi^2(9) = 15.8$, $p = .07$). Therefore, we may conclude that the presence of a context did not influence the distribution of solution strategies on any of the four operations.

Final analyses were carried out to assess whether the effect of item format (numerical vs. contextual) depended on either gender, language achievement level, or home language. We first report the main effects of the student characteristics on strategy choice distribution. The effect of home language was not significant ($\chi^2(3) = .4$, $p = .95$), so students who spoke Dutch at home had the same strategy distribution as students who spoke another language at home. In contrast, gender ($\chi^2(3) = 36.4$, $p < .001$) as well as

TABLE 7.4 *Strategy choice distribution (in proportions), by gender and language achievement level.*

	gender		language achievement		
strategy	boy	girl	low	medium	high
traditional	.45	.59	.47	.55	.57
non-traditional	.29	.29	.29	.29	.29
no written working	.22	.09	.19	.14	.12
other	.04	.03	.06	.02	.02
N observations	4768	5216	3776	3712	2608

language achievement level ($\chi^2(3) = 187.7$, $p < .001$) had a significant effect on strategy distribution. To visualize the effect of language achievement level, we recoded it into three categories based on the population percentile rank (based on all participants of the End of Primary School Test 2009): low (up to percentile 33), medium (percentiles 34 to 66), and high (percentile 67 and higher). Table 7.4 shows that girls were more likely than boys to choose the traditional algorithm, and less likely to answer without written working. With respect to language achievement level, the probability to choose the traditional algorithm increased with higher language level, while the probability of answering without written working as well as choosing one of the other strategies decreased with higher language level.

Finally, the interaction effects between problem format (numerical vs. contextual) on the one hand, and the student characteristics on the other hand, were not significant regarding gender ($\chi^2(3) = .4$, $p = .95$), home language ($\chi^2(3) = .9$, $p = .82$), or language achievement level ($\chi^2(3) = 2.7$, $p = .44$). Thus, we may conclude that the finding that the presence of a context did not affect strategy choice distribution holds for all subgroups of students.

Strategy accuracy

Finally, we investigated to what extent the problem format (numerical vs. contextual) affected the accuracy of each of the strategies, i.e., the proportion correct per strategy. To that end, we again used explanatory IRT models (De Boeck & Wilson, 2004). This time, not only predictors on the item level were included, but also the strategy used was included as a person-by-item predictor (see also Hickendorff et al., 2009b, 2010). All trials

in which the solution strategy was classified in the Other category were excluded from the analyses, because this was a small heterogeneous group of trials with many skipped problems. As a result, we analyzed the accuracy differences between the traditional algorithm, non-traditional strategies, and no written working.

Results showed that these three strategies differed significantly in accuracy ($\chi^2(2) =$ 184.6, $p < .001$), and that the accuracy differences depended on the operation required ($\chi^2(6) = 55.0$, $p < .001$). However, the operation-specific accuracy differences between the strategies did not depend on the item format (numerical vs. contextual), $\chi^2(8) = 9.0$, $p = .34$. Figure 7.4 shows the estimated proportion correct of each strategy for students at the mean of the latent ability scale, by operation. The general pattern is that the traditional algorithm was more accurate than the non-traditional strategies, which in turn were more accurate than no written working. Statistical testing of these differences showed the following: the regression parameters of the accuracy difference between the traditional algorithm and non-traditional strategies were $\beta = .32$ ($p = .02$), $\beta =$ .96 ($p < .001$), $\beta = .55$ ($p < .001$), and $\beta = .10$ ($p = .60$), for addition, subtraction, multiplication, and division, respectively. The regression parameters for the accuracy difference between non-traditional strategies and no written working were $\beta = .58$ ($p < .001$), $\beta = .09$ ($p = .67$), $\beta = .74$ ($p < .001$), and $\beta = 1.72$ ($p < .001$), for addition, subtraction, multiplication, and division, respectively. So, there were two exceptions to the general pattern. First, on subtraction, non-traditional strategies were not significantly more accurate than no written working. Second, on division, the traditional algorithm was not significantly more accurate than non-traditional strategies. Moreover, it is worth noticing that the estimated accuracy of no written working on division problems was much lower than on the other three operations.

7.4 DISCUSSION

The current study aimed to assess the effects of presenting multidigit arithmetic problems in a realistic context on two aspects of problem solving: performance and solution strategy use. First, regarding performance, multidimensional IRT models showed that the same latent ability was involved in solving numerical problems and solving contextual problems. Moreover, explanatory IRT modeling showed that presenting an arithmetic problem in a context increased the difficulty level of the division problems, but did not affect the difficulty levels of addition, subtraction, and multiplication problems.
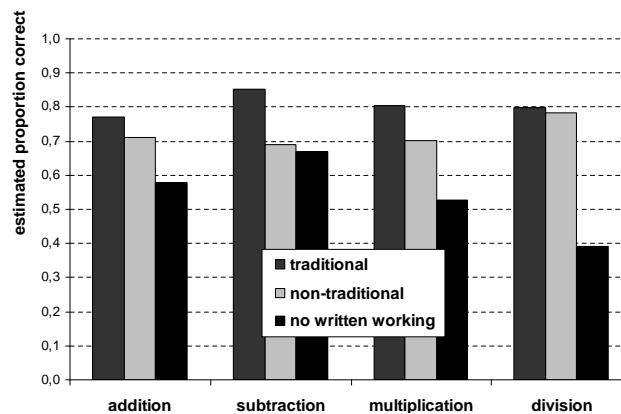
FIGURE 7.4 *Estimated mean accuracy of the three strategies, by operation.*

These performance effects were independent of student's gender, home language, and language achievement level. Second, the presence of a context did not affect the strategy choice distribution, nor the strategy accuracy, irrespective of student's gender, home language, and language achievement level. In summary, we conclude that, contrary to our expectations, the effects of presenting arithmetic problems in a realistic context were nonexistent on addition, subtraction, and multiplication, and that only a small difference in problem difficulty was found for division.

Regarding performance, based on earlier research findings (Fuchs et al., 2006, 2008; Hickendorff, 2010b) we expected that related but separate abilities would be involved in solving the two types of problems. A possible explanation for the difference between the current results and previous findings may lie in the differences between the age groups (first to third graders versus sixth graders). It has been argued that children in higher grades, who have had more years of formal schooling, have more developed cognitive schemata to solve word problems (De Corte et al., 1985; Vicente et al., 2007). Possibly, these cognitive schemata are so well-developed at grade 6 that students do not perceive differences in contextual problems and numerical problems anymore, which results both in indistinguishability of the latent ability dimensions involved, as well as in absence of an effect on problem difficulty (as was also found in another study in sixth grade by Vermeer et al., 2000). Importantly, this pattern held for girls, for students with low

language ability level, and for students from non-native origin. It thus seems that, at the end of primary school, these students are not hampered by the verbal nature of the contextual problems in mathematics. Furthermore, contextual problems did not elicit different solution strategies, contrary to expectations based on research on word problems with young children, as well as to expectations based on RME theory (e.g., Van den Heuvel-Panhuizen et al., 2009). However, the absence of effects on strategy choice is more congruent with the findings from Van Putten et al. (2005), who also found no difference in solution strategy choice on multidigit division problems in grade 4.

In the following, we will address the implications of these findings. In addition, we take a closer look into interesting patterns found in the present study, regarding multivariate solution strategies students used on the four operations, and regarding gender differences.

7.4.1 Practical implications

The results of the current study showed only minor effects of presenting multidigit arithmetic problems in a realistic context at the end of primary school. Because the problems used were taken from the Dutch national assessments at the end of primary school, we tentatively conclude that at least for multidigit addition, subtraction, multiplication, and division problems, the outcomes would have been the same if more or only numerical problems would have been included. This is an important observation, because the mathematics assessments have been criticized to appeal to language abilities too much because there are so many verbal problems. The results of the current study give no support for this criticism at the end of primary school, on the domain of multidigit arithmetic. Whether these results may be extended to other domains of mathematics (e.g., fractions) and/or other assessments such as TIMSS (grade 4 and grade 8) and PISA (15-year-olds) has to be studied in further research. Moreover, in the present study it was not possible to address the effects of specific characteristics of the context, such as linguistic complexity of the problem text (Abedi & Hejri, 2004; Abedi & Lord, 2001) and the effect of an illustration in the problem (Berends & Van Lieshout, 2009).

Although the presence of a context appeared not to affect arithmetic problem solving at the end of primary school, this does not mean that the shift towards dominance of contextual problems in mathematics tests and mathematics education is without consequences. Research findings in earlier grades (Fuchs et al., 2006, 2008; Hickendorff,

2010b) did show differences between mathematics problems with and without a context, and also showed that language ability had a larger effect on solving contextual problems. The present results implied that these differences have diminished and disappeared at the end of primary school, however, this does not preclude that students with low verbal abilities had more difficulties to obtain the same performance level on contextual problem solving as on numerical problem solving. Therefore, we still plead for a more balanced approach to mathematics education and testing, involving both bare numerical problems for computational fluency, as well as problems in a realistic context to apply these computational skills in real-life settings.

7.4.2 *Solution strategies for multidigit arithmetic*

Although there were no marked effects of the presence of a context on arithmetic problem solving, the present study gives unique empirical data on strategic competence of sixth graders from a reform-based educational environment. That is, strategy use (choice and accuracy) was studied across the four basic operations with multidigit numbers in one common framework. In particular, strategic competence in addition/subtraction on the one hand, and multiplication/division on the other hand, have not been studied simultaneously in one study before to our knowledge, and several interesting patterns emerged.

First, the dominance of the traditional algorithm decreased from addition (69%) and subtraction (74%) to multiplication (51%) and then again to division (15%). The Dutch national assessments showed a very similar trend in instructional practices: 69%, 72%, 57%, and 17% of the 118 participating grade 6 teachers reported instructing only the traditional algorithm for addition, subtraction, multiplication, and division, respectively (J. Janssen et al., 2005, p. 44). Therefore, it may be that students' choice for the traditional algorithm is determined to a large extent by the teacher's instruction. The traditional algorithm turned out to yield the highest probability of a correct answer on addition, subtraction, and multiplication, but on division it did not differ in accuracy from the non-traditional strategies (similar to Hickendorff et al., 2009b). However, because students were free to choose their solution strategy, strategy accuracy figures may be biased by *selection effects* (cf. Siegler & Lemaire, 1997). That is, different students choose different strategies on different items, thereby affecting the accuracy rates. A possible way to assess strategy accuracy unbiasedly would be to implement the *choice/no choice* method

(Siegler and Lemaire).

Second, division stood out compared to the other operations in the frequency of use of non-traditional strategies: 56% of all division trials were solved by repeated subtraction. Again, this is in line with the instructional approach in mathematics textbooks and teacher reports, in which repeated subtraction is the dominant strategy. This may also have reflected in the strategy accuracy: on each operation except from division, non-traditional strategies were less accurate than the traditional algorithm.

Third, on each operation, students gave an answer without any written work on a minority but still substantial number of trials (14-17%). This frequency fell in the range of frequencies reported in previous studies on multidigit multiplication and division (Hickendorff et al., 2009b, 2010; Hickendorff & Van Putten, 2010; Van Putten et al., 2005), and a previous study showed that it predominantly involved mental computation (i.e., students calculating in their head; Hickendorff et al., 2010). Importantly, it was the least accurate strategy (although the difference with non-traditional strategies was not significant on subtraction), and this is a consistent research finding. Therefore, we plead for more systematic research into this phenomenon, that may yield educational recommendations for instructing students when and when not to use a written strategy.

7.4.3 Gender differences

Interestingly, girls outperformed boys on the multidigit arithmetic problems, in contrast to international (PISA, TIMSS) educational assessments findings. Dutch national assessments, however, reported the trend that multidigit operations were the only domain of mathematics on which boys *did not* outperform girls, and in the 2004 assessment, girls even showed a small advantage on this domain (J. Janssen et al., 2005). A possible explanatory mechanism is strategy use: girls were found to use the more accurate traditional algorithm more often than boys, who in turn were more inclined to answer without written working. This pattern has been reported consistently (Hickendorff et al., 2009b, 2010; Hickendorff & Van Putten, 2010) and is also in line with more general research findings concerning gender differences in strategy choice (Carr & Davis, 2001; Carr & Jessup, 1997; Gallagher et al., 2000; Timmermans et al., 2007; Vermeer et al., 2000).

Furthermore, we expected differential effects of the presence of a context for boys and girls, based on findings of Vermeer et al. (2000). This expectation was not confirmed

in the current study. A tentative explanatory mechanism may be that girls may better understand the problem text of a contextual problem than boys, because they have been consistently found to have a higher reading ability level than boys in all countries participating in PISA-2009 (OECD, 2010) as well as in PIRLS-2006 (*Progress in International Reading Literacy Study*; Mullis, Martin, Kennedy, & Foy, 2007). This linguistic advantage for girls compared to boys in solving contextual problems may have cancelled out the pattern that girls prefer numerical problems over contextual ones (e.g., Vermeer et al., 2000).

APPENDIX 7.A   THE 8 PROBLEM PAIRS IN TEST FORM A, TEXTS TRANSLATED FROM DUTCH

addition problems

A class of students sells postcards and stamps for charity. They sold for € 466.50 on postcards and for € 985 on stamps. How much did they earn?	677.50 + 975 =

scarf: € 18,90 coat: € 298 shirt: € 119,50 hat: € 9,95 Here, you see the price tags of the clothes that Francien has bought. How much did she pay in total?	19.95 + 198.50 + 129 + 8.80 =

subtraction problems

Mrs. De Vries has € 3010 on her bank account. She withdraws € 689 to buy a new bike. Hoe much is left on her account?	4020 − 787 =

How much more was there on the account on 1-1-2004 compared to 1-1-2003?	3618.88 − 2923.95 =

multiplication problems

Charles has to photocopy 36 pages. He needs 27 copies of each page. How many copies are that in total?	37 x 24 =

Mother buys 17 meters of this curtain material. How much does she have to pay?	26 x 20.1 =

division problems

Invoice for "The Sunflower"	864 : 36 =

	number	price per book	total price
history textbooks	**32**		**€ 736**

The Sunflower has bought 32 new history textbooks. How much is the price per book?
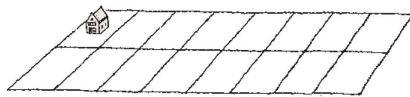
From a 5120 m² piece of land, 16 building parcels are made. Each building parcel will have the same area. What area will each building parcel have?	5880 : 14 =

APPENDIX 7.B   EXAMPLES OF SOLUTION STRATEGY CATEGORIES OF TABLE 7.1

addition

①
```
    11
677,50
 975 +
1652,50
```

②
```
677,50
 975 +
1500
 140
  12
  0,50
1652,50
```

③
```
677,50 + 900 = 1577,50
1577,50 + 70 = 1647,50
1647,50 + 5 = 1652,50
```

④
```
600 + 900 = 1500
70 + 70 = 140
7,50 + 5 = 12,50
1500 + 140 + 12,50 =
               1652,50
```

subtraction

①
```
    9 10
  2 10 10
 3010
 689 -
 2321
```

②
```
3010
 689 -
3000
-600
 -70
  -9
2321
```

③
```
3010 - 600 = 2410
2410 - 80 = 2330
2330 - 9 = 2321
```

④
```
3000 - 600 = 2400
2400 - 80 = 2320
2320 + 10 - 9 = 2321
```

⑤
```
689 + 11 = 700
700 + 300 = 1000
1000 + 2010 = 3010
11 + 300 + 2010 = 2321
```

multiplication

①
```
  36
  27x
 252
 720+
 972
```

②
```
  36
  27x
  42
 210
 120
 600+
 972
```

③
```
36 x 20 = 720
36 x 7 = 252
720 + 252 = 972
```

④
```
30 x 20 = 600
30 x 7 = 210
6 x 20 = 120
6 x 7 = 42+
        972
```

⑤
```
27
27
...
27+
972
```

division

①
```
32/736\23
    64
    96
    96
     0
```

②
```
736
640-   20x
 96
 96-   3x
  0    23x
```

③
```
736
320-   10x
416
320-   10x
 96
 64-   2x
 32
 32-   1x
  0    23x
```

④
```
10 x 32 = 320
20 x 32 = 640
 3 x 32 = 96
23 x 32 = 736
```

⑤
```
 2 x 32 = 64
 4 x 32 = 128
 8 x 32 = 256
16 x 32 = 512
20 x 32 = 640
22 x 32 = 704
23 x 32 = 736
```