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

Hickendorff, M.

Explanatory latent variable modeling of mathematical ability in primary school

Crossing the border between psychometrics and psychology

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

Cover design by Moon grafisch ontwerp
Printed by Proefschriftmaken.nl, Oisterwijk

ISBN 978-90-8891-326-6

Explanatory latent variable modeling of mathematical ability in primary school

Crossing the border between psychometrics and psychology

PROEFSCHRIFT

ter verkrijging van de graad van Doctor aan de Universiteit Leiden,
op gezag van de Rector Magnificus prof. mr. P. F. van der Heijden,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 25 oktober 2011
klokke 16.15 uur

door Marian Hickendorff
geboren te Leiden in 1981

PROMOTIECOMMISSIE

Promotor	prof. dr. W. J. Heiser	
Copromotores	dr. C. M. van Putten	(Universiteit Leiden)
	prof. dr. N. D. Verhelst	(Cito Instituut voor Toetsontwikkeling)
Overige leden	dr. A. A. Béguin	(Cito Instituut voor Toetsontwikkeling)
	prof. dr. P. A. L. de Boeck	(Universiteit van Amsterdam)
	dr. E. H. Kroesbergen	(Universiteit Utrecht)
	prof. dr. L. Verschaffel	(K.U. Leuven, België)

Contents

List of Figures

List of Tables

Introduction

Children's mathematical ability is a hotly debated topic in many countries, including the Netherlands. One point of discussion is mathematics *education*. A reform movement of international scope has taken place, which can roughly be described as a shift away from teachers directly instructing arithmetic skills that children have to drill, towards an approach that considers children's existing pre-knowledge as the basis on which to build mathematical knowledge, attempting to attain not only procedural expertise but also mathematical insight, flexibility, and creativity. Another point of discussion is the the mathematics *performance* level of students in primary school and in secondary school. Results from large-scale national and international assessments of students' mathematical ability, reporting trends over time, international comparisons, and deviations from educational standards that hold within a country, usually form the starting point of this discussion.

In this thesis – the result of a collaborative research project of the Institute of Psychology of Leiden University and CITO, the Dutch National Institute for Educational Measurement – the focus is on primary school students' mathematical ability in the Netherlands. The findings of the most recent national mathematics assessment at the end of primary school (sixth grade; 12-year-olds) carried out by CITO in 2004 (called PPON [*Periodieke Peiling van het OnderwijsNiveau*]; J. Janssen, Van der Schoot, & Hemker, 2005; see also Van der Schoot, 2008) were the starting point. The 2004-assessment was the fourth cycle, with earlier assessments carried out in 1987, 1992, and 1997; the fifth cycle is planned in 2011. Trends over the time period from 1987 to 2004 showed diverse patterns: in some mathematics domains students' performance increased, while in other domains it decreased. Moreover, in general students' performance lagged behind the educational standards, in some domains more than in others. In the newspapers and other platforms

of the public debate people expressed their opinions on these developments. One returning element is the didactical theory of Realistic Mathematics Education (RME; e.g., Freudenthal, 1973, 1991; Treffers, 1993) that has become the dominant theory in primary school mathematics education in the 80s and 90s of the previous century, which evokes strong feelings. In the public debate, however, commonsense beliefs and personal sentiments with anecdotal foundations usually prevail over robust insights based on empirical study of what students know and can do in mathematics and of what are the performance outcomes of different mathematics programs. The purpose of the current thesis is to provide these empirically-based insights.

First, to give an overview of what is known empirically – and what is *not* known – about performance outcomes of different mathematics programs or curricula, a research synthesis of empirical studies that address this question for primary school students in the Netherlands is presented. Next, with respect to what primary school students know and can do in mathematics, CITO's mathematics assessments are a rich source of information on students' performance level compared to the educational standards, as well as on differences between students (e.g., boys and girls) and on trends over time. However, these assessments are surveys and therefore limited to descriptive analyses. Explanations for apparent differences or trends require further study. That is exactly what has been done in the current research project and what is reported on in six empirical studies in this thesis.

These empirical studies, addressing determinants of students' ability in the domain of arithmetic (addition, subtraction, multiplication, and division), cross the border between the academic fields of substantive educational and cognitive psychology on the one hand and psychometrics on the other hand. Substantively, *solution strategies* are a key element of all but one of the empirical studies. Strategies were deemed relevant both from an educational psychology perspective, because they are a spearhead in mathematics education reform, as well as from a cognitive psychology perspective where mechanisms of strategy choice and concepts as strategic competence are important research topics. In the current studies, solution strategies were considered both as outcome measures in analyses of determinants of strategy choice, and as explanatory variables in analyses of determinants of mathematics performance. Related recurring elements are individual differences in strategy choice and in performance, and differences between groups of students such as boys and girls.

Psychometrically, the data in the empirical studies reported are complex, requiring

advanced statistical modeling. In the substantive fields of educational and cognitive psychology, such techniques are not very common. That is why the current thesis can be said to be an attempt to integrate psychometrics and psychology, such as advocated by Borsboom (2006). One salient complicating aspect of the data in the current studies is that they involve repeated observations within subjects (i.e., each student responds to several mathematics problems), leading to a correlated or dependent data structure. To take these dependencies into account, it is argued that *latent variable models* are appropriate: one or more latent (unobserved) variables reflect individual differences between students, and the dependent responses within each student are mapped onto these variables. Latent variables can be either categorical, modeling qualitative individual differences between students, or continuous, modeling quantitative individual differences. Furthermore, the influence of explanatory variables such as assessment cycle or students' gender can be addressed by analyzing the effect on the latent variable.

In the studies reported, the responses on each trial (when a student is confronted with an item) are of categorical measurement level. That is, two types of responses are dealt with: the strategy used to solve the problem (several unordered categories) and the accuracy of the answer given to the problem (dichotomous: correct/incorrect). To analyze individual differences in strategy choice, latent class analysis (LCA; e.g. Goodman, 1974; Lazarsfeld & Henry, 1968) was used. In latent class models, it is assumed that there are unobserved subgroups (classes or clusters) of students that are characterized by a specific response profile, in this case, a specific strategy choice profile. In order to address the influence of student characteristics on latent class membership, latent class models with covariates (Vermunt & Magidson, 2002) were used. To analyze individual differences in performance, item response theory (IRT; e.g., Embretson & Reise, 2000; Van der Linden & Hambleton, 1997) models were used, in which the probability of giving the correct answer is determined by one or more continuous latent (ability) dimensions. In particular, measurement IRT models were extended with an explanatory part, in which predictors at the person level, at the item level, or at the person-by-item level can be incorporated (De Boeck & Wilson, 2004; Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). One innovating application of these explanatory IRT models was to use the strategy used on an item as a person-by-item predictor, thereby modeling strategy accuracy (the probability to obtain a correct answer with a certain strategy) while statistically accounting for individual differences in overall ability and for differences in difficulty level between problems, something that had not been accomplished before in psychological

research into solution strategies.

OUTLINE

The thesis starts with Chapter 1 reporting a research synthesis of empirical studies that were carried out in the Netherlands into the relation between mathematics education and mathematics proficiency. This chapter is based on work that was done for the KNAW (*Royal Dutch Academy of Arts and Sciences*) Committee on Primary School Mathematics Teaching[1], whose report came out in 2009. Starting with an overview of results of Dutch national assessments and the position of Dutch students in international assessments, the main body of the chapter is devoted to a systematic review of studies in which the relationship between instructional approach and students' performance outcomes was investigated. The main conclusion that could be drawn was that much is unknown about the relation between mathematics programs and performance outcomes, and that methodologically sound empirical studies comparing different instructional approaches are rare, which may be because they are very difficult to implement. In the remainder of this thesis, the focus is shifted to other determinants of students' mathematics ability related to contemporary mathematics education, such as the strategies students used to solve the problems and characteristics of the mathematics problems.

First, two studies are reported in which secondary analyses on the raw student material (test booklets) of the two most recent national mathematics assessments of 1997 and 2004 were carried out. They both focus on complex or multidigit arithmetic: a mathematics domain on which performance decreased most severely over time, as well as stayed furthest behind the educational standards. Furthermore, the RME approach has changed the instructional approach as to how to solve these problems, paying less attention to the traditional algorithms and instead focusing more on informal whole-number approaches (Van den Heuvel-Panhuizen, 2008). Therefore, both studies focus on *solution strategies* as explanatory variables of performance, a recurring issue in this thesis. Specifically, in Chapter 2, solution strategies that students used to solve complex or multidigit *division* problems were studied, aiming to give more insight in the performance decrease between 1997 and 2004. The complex nature of the data necessitated advanced psychometric modeling, and latent variable models – latent class

1 I worked as an associate researcher supporting the Committee. In particular, the Committee requested me to carry out the systematic literature review that formed the basis of chapter 4 in the report. Chapter 1 in the current thesis is based on this work.

analysis (LCA) and item response theory (IRT) – with explanatory variables are introduced in this chapter. Subsequently, in Chapter 3 the domain of division is broadened to include complex or multidigit *multiplication* problems as well. Furthermore, the influence of teachers' instructional approach to solving multiplication and division problems on students' strategy choice is addressed.

The subsequent part of this thesis reports on two studies in which new data were collected to answer specific research questions that were raised based on the findings of the secondary analyses on the division problems data in Chapter 2. Specifically, one important conclusion was that students increasingly answered without any written working, and that that this shift was unfortunate with respect to performance, since it was the least accurate strategy. In Chapter 4 individual differences in strategy use in complex division problems were studied in a systematic research design: a partial *choice/no choice* design (Siegler & Lemaire, 1997). Sixth graders solved division problems in two different conditions: in the choice condition, they were free to choose how they solved the problem (with a written or a mental strategy), while in the subsequent no-choice condition, they were forced to write down how they solved the problem. In addition, individual interviews with students using a non-written strategy in the choice condition were carried out to investigate how they had solved the problem without using paper and pencil. Next, Chapter 5 reports on a study in which a complete choice/no-choice design was implemented, in which there was an additional no-choice condition in which students were forced to use a mental strategy. In addition, solution times were recorded, so that two aspects of strategy performance – accuracy and speed – could be taken into account simultaneously. In this study, it was possible to address the issue of *strategy adaptivity* at the student level: the extent to which a student chooses the best strategy for him or her on a particular division problem.

The final part of the current thesis addresses another aspect of contemporary mathematics education: an increased focus on mathematics problems in a *realistic context* – including word problems – in instruction as well as in tests. These contexts usually consist of a verbal description of a mathematical problem situation, which may be accompanied by an illustration. Such problems serve a central role for several reasons (e.g., Verschaffel, Greer, & De Corte, 2000): they may have motivational potential, mathematical concepts and skills may be developed in a meaningful way, and children may develop knowledge of when and how to use mathematics in everyday-life situations. Little is known, however, on the the differences between solving computational problems

(bare numerical problems) and solving such contextual problems. Therefore, this question is addressed in two studies in the domain of the four basic arithmetical operations: one focusing on children in the lower – first, second, and third – grades of primary school, reported in Chapter 6; the other focusing on students in grade six, reported in Chapter 7. In both studies, special attention is paid to the influence of students' language level, because students need to understand the problem text in order to be able to successfully solve the problem. Chapter 6 focuses on performance only, modeling students responses (correct/incorrect) to mathematics problems of both types in a multidimensional IRT framework. Chapter 7 extends this focus by investigating strategy use as well.

Chapter 8 concludes this thesis with a general discussion. Besides reflecting on the substantive findings regarding mathematical ability in Dutch primary school students, attention is also paid to the psychometric modeling techniques that are used.

Finally, note that because the seven main chapters of this thesis are separate research papers, a certain amount of overlap is inevitable.

CHAPTER 1

Performance outcomes of primary school mathematics programs in the Netherlands: A research synthesis

This chapter is based on research I have done for the KNAW Committee on Primary School Mathematics Teaching, reported in KNAW (2009). Note that this report is written in Dutch, and the reproduction of ideas in English is on my account.

ABSTRACT

The results of a systematic quantitative research synthesis of empirical studies addressing the relation between mathematics education and students' mathematics performance outcomes is presented. Only studies with primary school students carried out in the Netherlands were included. In total, 25 different studies were included: 18 intervention studies in which the effects of different mathematics interventions (instructional programs) were compared, and 7 curriculum studies in which differential performance outcomes with different mathematics curricula (usually textbooks) were assessed. In general, the review did not allow drawing a firm univocal conclusion on the relation between mathematics education and performance outcomes. Some more specific patterns emerged, however. First, performance differences were larger within a type of instructional approach than between different instructional approaches. Second, more time spent on mathematics education resulted in better performance. Third, experimental programs implemented in small groups of students outside the classroom had positive effects compared to the regular educational practice. Fourth, low mathematics performers seemed to have a larger need for a more directing role of their teacher in their learning process.

1.1 INTRODUCTION

1.1.1 Background

Recently, there has been a lot of criticism on mathematics education in primary school in the Netherlands, originating in growing concern on children's mathematical proficiency. This public debate – both in professional publications as well as in more mainstream media – is characterized by its heated tone and its polarizing effect. That caused the *Royal Netherlands Academy of Arts and Sciences* (KNAW) to set up a Committee on Primary School Mathematics Teaching in 2009. When the State Secretary, Ms. Sharon Dijksma, announced a study on mathematics education, these two initiatives were combined. The Committee's mission was *"To survey what is known about the relationship between mathematics education and mathematical proficiency based on existing insights and empirical facts. Indicate how to give teachers and parents leeway to make informed choices, based on our knowledge of the relationship between approaches to mathematics teaching and mathematical achievement."* (KNAW, 2009, p. 10).

The current chapter is based on the systematic quantitative review of empirical studies addressing the relation between mathematics education or instruction and children's mathematical proficiency in the Netherlands, one of the core parts of the committee's report (KNAW, 2009, ch. 4[1]). In the remainder of the Introduction, first a short overview of the state of primary school students' mathematical proficiency level is presented, based on findings of national and international large-scale educational assessments. Then a brief discussion of existing international reviews and meta-analyses of research on the effects of mathematics instruction follows. In the main part of this chapter, the methodology and results of the current systematic quantitative review are presented. This review is largely along the lines of what Slavin (2008) proposed as a *best-evidence synthesis*: a procedure for performing syntheses of research on educational programs that resembles meta-analysis, but requires more extensive discussion of key studies instead of primarily aiming to pool results across many studies (Slavin & Lake, 2008). In the current review into the effect of primary school mathematics programs in the Netherlands, a distinction is made between *intervention studies* in which the researchers intervened in the educational practice, and *curriculum studies* in which no intervention took place, the mathematics programs compared were self-selected by schools. This chapter ends with a summary of the research synthesis, conclusions, and implications.

1.1.2 The state of affairs of Dutch students' mathematical performance

To describe the state of Dutch primary school students' mathematical performance level, empirical quantitative results of national and international assessments were used. Such large-scale educational assessments aim to report on the outcomes of the educational system in various content domains such as reading, writing, science, and mathematics. At least two aspects are important (Hickendorff, Heiser, Van Putten, & Verhelst, 2009a). The first aspect is a description of students' learning outcomes: what do students know, what problems can they solve, to what extent are educational standards reached, and to what extent are there differences between subgroups (such as different countries in international assessments, or boys and girls within a country)? The second aspect concerns trends: to what extent are there changes in achievement level over time?

1 I carried out this research review at request of the KNAW Committee, for which I worked as an associate researcher.

At the national level, CITO carried out educational assessments – PPON [*Periodieke Peiling van het Onderwijsniveau*] – of mathematics education in grade 3 (9-year-olds) and in grade 6 (12-year-olds) in cycles of five to seven years since 1987. In the current overview only the results for grade 6 are discussed, because these concern students' proficiency at the end of primary school. At the international level, there is TIMSS (*Trends in International Mathematics and Science Study*): an international comparative study in the domains of science and mathematics, carried out in grade 4 (10-year-olds) and in grade 8 (14-year-olds, second grade of secondary education in the Netherlands), with assessments in 1995, 2003, and 2007. Only the grade 4 results concern primary school, so we focus on those.

Dutch national assessments: PPON in grade 6

Van der Schoot (2008) presented an overview of the grade 6 mathematics assessment results. Thus far, there have been four cycles: 1987, 1992, 1997, and 2004 (the next assessment is planned in 2011). The domain of mathematics is structured in three general domains: (a) numbers and operations, (b) ratios/fractions/percentages, and (c) measures and geometry. In each general domain, several subdomains are distinguished. In total, there were 22 different subdomains in the most recent assessment of 2004 (J. Janssen et al., 2005).

Students' results were evaluated in two ways: the trend over time since 1987, and the extent to which the educational standards were reached. For the latter evaluation, the standards set by Dutch Ministry of Education, Culture, and Sciences (1998) were operationalized by a panel of approximately 25 experts, ideally consisting of 15 primary school teachers, 5 teacher instructors, and 5 educational advisors. In a standardized procedure, these panels agreed upon two performance levels: a *minimum* level that 90-95% of the students at the end of primary school should reach, and a *sufficient* level, that should be reached by 70-75% of all students. Table 1.1 presents the relevant results. First, it shows the effect size (ES, standardized mean difference) of the performance difference between the baseline measurement (usually 1987), interpreted as $.00 \leq |\text{ES}| < .20$ negligible to small effect, $.20 \leq |\text{ES}| < .50$ small to medium effect, $.50 \leq |\text{ES}| < .80$ medium to large effect, and $|\text{ES}| \geq 0.80$ large effect. Second, it shows the percentage of students reaching the educational standards of minimum and sufficient level.

The trends over time show varying patterns, with the most striking developments

TABLE 1.1 *Dutch mathematics assessments results, from Van der Schoot (2008, p. 20-22).*

	trend in ES (baseline 1987 = 0)			reaching standard in 2004	
	1992	1997	2004	min.	suff.
numbers and operations					
numbers and number relations	+.28	+.46	+.94	96%	42%
simple addition/subtraction	*	−.11	+.24	92%	76%
simple multiplication/division	*	−.30	−.20	90%	66%
mental addition/subtraction	n.a.	+.49	+.53	92%	50%
mental multiplication/division	n.a.	−.12	−.11	92%	66%
numerical estimation	n.a.	+.94	+1.04	84%	42%
complex addition/subtraction	−.12	−.17	−.53	62%	27%
complex multiplication/division	−.17	−.43	−1.16	50%	12%
combined complex operations	−.40	−.44	−.78	50%	16%
calculator	*	+.29	+.26	73%	34%
ratios/fractions/percentages					
ratios	+.11	+.26	+.14	92%	66%
fractions	+.09	+.23	+.15	95%	60%
percentages	+.12	+.28	+.51	88%	58%
tables and graphs	n.a.	*	+.10	84%	50%
measures and geometry					
measures: length	+.00	−.03	−.13	79%	38%
measures: area	−.32	−.04	+.05	67%	21%
measures: volume	+.10	.00	−.03	67%	21%
measures: weight	+.02	+.20	+.33	88%	58%
measures: applications	−.05	−.21	−.25	92%	50%
geometry	.00	+.12	−.08	95%	62%
time	+.17	+.23	.00	92%	50%
money	−.21	−.31	n.a.	84%	42%

* Earlier results not available, alternative baseline.

in the domain of numbers and operations. Differences were negligible to medium-sized ($|ES| < .50$) on 14 of the 21 subdomains for which trends could be assessed. Positive developments of at least medium size ($ES \geq .50$) were found in percentages, mental addition/subtraction, numbers and number relations, and numerical estimation. Negative trends of at least medium size ($ES \leq -.50$), however, were found for complex addition and subtraction, combined complex operations, and complex multiplication and division.

Regarding attainment of the educational standards, Table 1.1 shows that on only one subdomain (simple addition/subtraction), the desired percentage of 70% or more students attaining the sufficient level was reached. On eleven domains, this percentage was between 50% and 70%, and on five domains it was between 30% and 50%. Finally, on five domains the percentage of students attaining sufficient level did not exceed 30%. So, in particular performance in the *complex operations* (addition/subtraction, multiplication/division, and combined operations; all concern multidigit problems on which the use of pen and paper to solve them is allowed) and in the *measures* subdomains *weight* and *applications* is worrisome according to the expert panels.

International assessments: TIMSS in grade 4

The Netherlands participated in the grade 4 international mathematics assessments in 1995, 2003, and 2007 (Meelissen & Drent, 2008; Mullis, Martin, & Foy, 2008). Worldwide, 43 countries participated in TIMSS-2007. In this TIMSS cycle there were mathematics items from three mathematical content domains – number, geometric shapes and measures, and data display – crossed with three cognitive domains – knowing, applying, and reasoning. Curriculum experts judged 81% of the mathematics items suited for the intended grade 4 curriculum in the Netherlands. Conversely, only 65% of the Dutch intended curriculum was covered in the TIMSS-tests.

Dutch fourth graders' mathematics performance level was in the top ten of the participating countries; only in Asian countries performance was significantly higher. Interestingly, the spread of students' ability level was relatively low, meaning that students' scores were close together. Another way to look at this is to compare performance to the TIMSS International Benchmarks: the advanced level was attained by 7% of the Dutch students, high level by 42%, intermediate level by 84%, and low level by 98% of the students. Although these percentages were all above the international median, compared to other countries that had such a high overall performance as the Netherlands, there were relatively many students attaining the low performance level, but relatively few students reaching the advanced level. Furthermore, developments over time showed a small but significant negative trend in total mathematics performance since 1995 (average score 549), via 2003 (average score 540), towards 2007 (average score 535). Internationally, more countries showed improvements in fourth grade performance than declines, so the Netherlands stand out in this respect.

Students' attitudes toward mathematics were investigated with a student questionnaire with questions on positive affect toward mathematics and self-confidence in own mathematical abilities (Mullis et al., 2008). Students reported a slightly positive affect toward mathematics, although it showed a minor decrease compared to 2003. Moreover, in the Netherlands there were proportionally many students (27%; international average 14%) at the low level of positive affect, and proportionally few students (50%; international average 72%) at the high level. Dutch students had quite high levels of self-confidence, and the distribution was comparable to the international average distribution.

Finally, we discuss some relevant results on the teacher and the classroom characteristics and instruction. Dutch fourth grader teachers were at the bottom of the international list in participating in professional development in mathematics. Still, they reported to feel well prepared to teach mathematics for 73% of all mathematics topics (international average 72%). Furthermore, Dutch fourth grade teachers reported experiencing much fewer limitations due to student factors than the international averages. A last relevant pattern was that Dutch students reported relatively frequently to work on mathematics problems on their own, while they reported explaining their answer relatively infrequently.

Summary national and international assessments

The national assessments (PPONs) were tailor-made to report on the outcomes of Dutch primary school mathematics education. Results showed that in many subdomains there were only minor changes in sixth graders' performance level between 1987 and 2004, and opposed to subdomains where performance declined there were subdomains in which performance improved. International assessments (TIMSS) showed that Dutch fourth graders still performed at a top level from an international perspective.

However, these results do not justify complacency (KNAW, 2009). In TIMSS, too few students reached the high and advanced levels, there was a small performance decrease over time causing other countries to come alongside or even overtake the position of the Netherlands, and students too often reported low positive affect toward mathematics. Moreover, it seems unwise to cancel out the positive and negative developments that were found in PPON. In addition, students' performance level lagged (far) behind the educational standards for primary school mathematics in most subdomains, also in the

subdomains showing improvement over time.

1.1.3 International reviews, research syntheses, and meta-analyses

We briefly review some patterns that emerge from international reviews and meta-analyses into the effects of mathematics instruction on achievement outcomes[2]. Note that this discussion is by no means exhaustive. Moreover, the findings are to a large extent based on studies carried out in the US. A first important observation is that the authors of most of the reviews stated that there are few studies that meet methodological standards that permit sound, well-justified conclusions about the comparison of the outcomes of different mathematics programs. The number of well-conducted (quasi-)experimental studies is low, and in particular studies meeting the 'golden standard' of randomized controlled trials are rarely encountered. For example, the US *National Mathematics Advisory Panel*, that had a similar assignment as the Dutch KNAW Committee, reviewed 16,000 research reports and concluded that only a very small portion of those studies met the rigorous methodological standards that allowed conclusions on the effect of instructional variables on mathematics learning outcomes (National Mathematics Advisory Panel, 2008). This review, however, has been heavily criticized for its stringent inclusion criteria that resulted in exclusion of relevant research findings, as well from its narrow cognitive perspective on mathematics education (see Verschaffel, 2009, for an overview of reactions in the US).

We primarily focus on two recent research syntheses: one by Slavin and Lake (2008) of research on achievement outcomes of different approaches to improving mathematics in regular primary education, and the other by Kroesbergen and Van Luit (2003) of research on the effects of mathematics instruction for primary school students with special educational needs.

Slavin and Lake (2008) conducted a 'best-evidence synthesis' of research on the achievement outcomes of three types of approaches to improving elementary mathematics: mathematics curricula, computer-assisted instruction (CAI), and instructional process programs. In total, 87 studies were reviewed, meeting rather stringent methodological criteria based on the extent to which they contribute to an unbiased, well-justified quantitative estimate of the strength of the evidence supporting each program.

2 This section is partly based on contributions of prof. dr. Lieven Verschaffel to chapter 3 of the KNAW (2009) report.

Regarding mathematics curricula, the results of the synthesis showed that there was little empirical evidence for differential effects. A noteworthy shortcoming of these studies was that they mainly used standardized tests that focused more on traditional skills than on concepts and problem solving that are addressed in reform-based mathematics curricula. However, in the cases when outcomes on these 'higher-order' mathematics objectives were considered, they do not suggest a differential positive effect of reform-based curricula. This observation contrasts with that of Stein, Remillard, and Smith (2007), who reviewed US-studies comparing 35 different mathematics textbooks (written curricula), of which approximately half could be characterized as reform-based or constructivistic, and the other half as traditional or mechanistic. They concluded that students trained with reform-based textbooks performed at about an equal level on traditional skills, but did better on higher-order goals such as mathematical reasoning and conceptual understanding, compared to students trained with traditional textbooks. An important remark, however, is that Stein et al. found that variation in teacher implementation of traditional curricula was smaller than in teacher implementation of reform-based curricula, hampering sound conclusions on differential effects of mathematics curricula.

CAI-supplementary approaches had moderate positive effects on students learning outcomes, especially on measures of computational skills (Slavin & Lake, 2008). Although the effects reported were very variable, the fact that in no study effects favoring the control group were found, and that the CAI-programs usually supplement the classroom instruction by only about 30 minutes a week, Slavin and Lake claimed that the effects were meaningful for educational practice. CAI primarily adds the possibility to tailor the instruction to individual students' specific strengths and weaknesses. In a meta-analysis of intervention research of word-problem solving in students with learning problems, Xin and Jitendra (1999) also found that CAI was a very effective intervention, but Kroesbergen and Van Luit (2003) found negative effects of CAI compared to other interventions in their meta-analysis of mathematics intervention studies in students with special educational needs.

Finally, Slavin and Lake (2008) found the largest effects for instructional process programs, that primarily focus on what teachers do with the curriculum they have, not changing the curriculum. The programs reviewed were highly diverse. Programs with positive effects either used various forms of cooperative learning, focused on classroom management strategies, used direct instruction models, or supplemented

traditional classroom instruction (including small group tutoring). These are quite general characteristics of how teachers use instructional process strategies. In line with these findings are results from a recent investigation of the Dutch Inspectorate of Education (2008) into school factors that are related to students' mathematics performance in primary school. They found that the educational process (quality control, subject matter, didactical practice, students' special care) was of lower quality in mathematically weak schools than in mathematically strong schools. In particular, there were nine school factors in which mathematically weak schools lagged behind: (a) yearly systematic evaluation of students' results; (b) quality control of learning and instruction; (c) the number of students for whom the subject matter is offered up to grade 6 level; (d) realization of a task-focused atmosphere; (e) clear explaining; (f) instructing strategies for learning and thinking; (g) active participation of students; (h) systematic implementation of special care; and (i) evaluation of the effects of special care.

Slavin and Lake (2008, p. 475) concluded their research synthesis with stating that *"the key to improving math achievement outcomes is changing the way teachers and students interact in the classroom."* The central and crucial role of the teacher in improving mathematics education is also subscribed to by others, such as Kroesbergen and Van Luit (2003) and Verschaffel, Greer, and De Corte (2007). An important concept is teachers' *Pedagogical Content Knowledge* (PCK), a blend of content knowledge and pedagogical knowledge of students' thinking, learning, and teaching. Fennema and Franke (1992) and Hill, Sleep, Lewis, and Ball (2007) pointed at the potential of pre-service and in-service training programs to improve teachers' mathematical PCK, but at the same time they acknowledge that there is little empirical evidence about the causal relation between teachers' PCK and students' achievement outcomes.

A lot of research attention is devoted to interventions for students with special educational needs, sometimes distinguished in students with learning disabilities (LD) and students with (mild) mental retardation (MR). Kroesbergen and Van Luit (2003) carried out a meta-analysis into the effects of mathematics interventions for these students, reviewing 58 studies addressing three mathematical domains: preparatory arithmetic, basic skills, and problem solving. The meta-analysis showed that intervention effects were largest in the domain of basic skills, implying that it may be easier to teach students with mathematical difficulties basic skills than problem-solving skills. Further relevant conclusions were that regarding treatment components of the interventions, self-instruction and direct instruction (more traditional instructional approaches) were more

effective than mediated/assisted instruction (more reform-based approach). The results favoring direct instruction were in in line with other meta-analyses of intervention studies with students with learning disabilities (e.g., Gersten et al., 2009; Swanson & Carson, 1996; Swanson & Hoskyn, 1998), stressing the importance of the role of the teacher to help students with special educational needs and to evaluate their progress. Similarly, the National Mathematics Advisory Panel (2008) also concluded that explicit instruction is effective for students struggling with mathematics. Apart from this instructional component, Kroesbergen and Van Luit's meta-analysis did not find effects of other characteristics of Realistic Mathematics Education. Kroesbergen and Van Luit therefore concluded that the mathematics education reform does not lead to better performance for students with special educational needs.

Another review worth mentioning is that of Hiebert and Grouws (2007) into the effects of classroom mathematics teaching on students' learning. Their first conclusion was that *opportunity to learn*, which is more nuanced and complex than mere exposure to subject matter, is the dominant factor influencing students learning. Secondly, they distinguish between teaching for skill efficiency and teaching for conceptual understanding. In teaching that facilitates skill efficiency, the teacher plays a central role in organizing, pacing, and presenting information or modeling to meet well-defined learning goals; in short: teacher-directing instruction. Teaching that facilitates conceptual understanding, however, is characterized by an active role of students and explicit attention of students and teachers to concepts in a public way.

1.2 Method of the current review

The basic approach of the current review was along the lines of Slavin's (2008) best evidence synthesis procedure. This technique *"seeks to apply consistent, well-justified standards to identify unbiased, meaningful, quantitative information from experimental studies"* (Slavin & Lake, 2008, p. 430). Slavin contended that the key focus in synthesizing (educational) program evaluations is minimizing the bias in reviews of each study, because there are usually only a small number of studies per program. The scarceness of studies also precludes pooling of results over studies and statistically testing for effects of study characteristics or procedures like in meta-analysis (Lipsey & Wilson, 2001). Instead, a more extensive discussion of the nature and quality of each study is incorporated. For each qualifying study not only effect sizes are computed, but also the context, design,

and findings of each are discussed (Slavin & Lake).

The objective of the current review was to *"investigate what is known scholarly about the relation between instructional approaches and mathematical proficiency"* (KNAW, 2009, p. 12). To that end, a quantitative synthesis of achievement outcomes of alternative mathematics programs was carried out. In this synthesis, quantitative results of other outcomes such as motivation or attitudes were not included, although relevant findings are discussed in the text. Two types of empirical studies addressing this objective are distinguished, similar to Slavin and Lake (2008): *intervention studies* and *curriculum studies*.

Intervention studies aim to assess the effect of one or more mathematics programs that are implemented with an intervention in the regular educational practice. These programs either replace or supplement (part of) the regular curriculum, and usually address a specific delimited content area such as addition and subtraction below 100. The programs are highly diverse. Furthermore, the implementation of the (experimental) programs is under researcher control, but the extent of control varies. It may be that external trainers implement the programs – yielding much control – or that the regular teacher was trained to implement the program. Combinations are also possible. Assignment to conditions (i.e., programs) may be either on individual student level or at the level of whole classrooms or schools. Furthermore, assignment may be random (experimental design) or non-random (quasi-experimental design). Finally, in most studies a pretest is administered before start of the program under study, in others not.

Curriculum studies aim to investigate differential achievement outcomes of different mathematics curricula, usually operationalized as mathematics textbook (series). The researchers have no control on assignment to curricula or on the implementation of the curriculum, and therefore these are observational studies. A disadvantage is that selection effects cannot be ruled out: factors that determine which mathematics textbook a school uses are likely to be related to achievement, biasing the results. Moreover, there is usually only one measurement occasion, so that correcting for differences between groups is also not possible.

1.2.1 Search and selection procedures

A number of inclusion criteria for a study to qualify for the review were set up, based on their potential to address the review's objective. The criteria were:

1. the study specifically addresses mathematics, or at least it should be possible to parcel out the mathematics results;
2. it should be possible to examine the results for children in the age range 4-12 years;
3. the study is executed less than 20 years ago[3];
4. the study is carried out in the Netherlands, with Dutch classes and students, or in case of an international study it should be possible to parcel out the effects for the Netherlands;
5. the study is empirical, meaning that conclusions are based on empirical data;
6. the study's results are published, preferably in (inter)national journals, books, and doctoral theses;
7. at least two different mathematics programs are compared,
8. there is enough statistical information in the publication to compute or approximate the effect size (see section 1.2.2)[4].

Compared to Slavin and Lake (2008) and Slavin's (2008) recommendations, we were less strict in excluding studies. Specifically, we were less stringent in excluding studies based on the research design (i.e., studies with non-random assignment and without matching were not excluded), based on pretest differences (i.e., studies with more than half a standard deviation difference at pretest were not excluded *per se*, but rather were marked as yielding unreliable effect sizes), based on study duration, and based on outcome measures. Our approach to including studies was this liberal because we argue that compromises on study quality are necessary, because there are so few studies in number. Moreover, by including studies liberally but clearly describing each study's limitations, readers have a comprehensive overview of the existing literature and can judge the studies' quality themselves.

To search for relevant studies, the KNAW Committee asked 50 experts in mathematics education research in the Netherlands to give input on studies to include. This resulted in 76 proposed publications, 17 of which met the inclusion criteria as set in the current chapter. Additional literature searches resulted in a total of 25 different studies (18 intervention studies and 7 curriculum studies) that met the inclusion criteria, reported in 29 different publications.

3 We were more strict on this criterion than in KNAW (2009), thereby excluding one study that was included in that report.

4 This was not one of the original inclusion criteria in KNAW (2009, p. 43-44), and thereby one more study was excluded.

1.2.2 Computation of effect sizes

To compare and synthesize quantitative results from many different studies they need to be brought to one common scale. To that end, results are reported in effect sizes (ES): the standardized mean difference between conditions (e.g., Lipsey & Wilson, 2001). The difference in mean posttest achievement scores in condition or program 1 ($\overline{X}_1$) and condition 2 ($\overline{X}_2$) is divided by the pooled standard deviation s_p, i.e.,

$$\text{ES} = \frac{\overline{X}_1 - \overline{X}_2}{s_p}, \tag{1.1}$$

with

$$s_p = \sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-1}}, \tag{1.2}$$

with n_1 and n_2 the number of students in program 1 and 2, respectively, and s_1 and s_2 the standard deviation in program 1 and 2. Guidelines for interpreting these effect sizes are commonly: $.00 \leq |\text{ES}| < .20$ negligible to small effect, $.20 \leq |\text{ES}| < .50$ small to medium effect, $.50 \leq |\text{ES}| < .80$ medium to large effect, and $|\text{ES}| \geq .80$ large effect, see for example Cohen (1988). Furthermore, Slavin (2008) qualified an ES of at least .20 as practically relevant in educational research. If there were multiple achievement outcomes, effect sizes were computed and reported for each measure separately. For studies that did not report means and standard deviations, other statistical information was used to compute and approximate the mean difference and the pooled standard deviation (e.g., Kroesbergen & Van Luit, 2003).

An important possible threat to the validity of comparisons of program outcomes is the influence of pre-existing group differences. These differences were accounted for in the following ways. If the study reported posttest means that were corrected for pretest measures or background variables (for example from an analysis of covariance or a multiple regression analysis), these adjusted means were used in computing the effect size. If such adjusted means were not reported, correction was approximated by subtracting the standardized mean difference in pretest scores from the standardized mean difference in posttest scores, as recommended by Slavin (2008). If no data from before the start of the program were reported, statistically correcting for pre-existing differences was not possible, and this should be held in mind in evaluating the reported effect sizes.

1.2.3 Study characteristics coded

For each study, several characteristics were coded, and they are described in the Summary Tables in Appendices 1.A and 1.B. The characteristics were:

1. *reference*: the publication reference(s) in which the study is reported;
2. *domain*: the mathematical content domain the study addressed;
3. *participants*: several characteristics of the students participating in the study: the sample size N, the number of classes or schools they originated from, the type of primary school they attended (regular or special education), and whether all students or only low math performers participated;
4. *intervention* or *curriculum*: the programs evaluated [intervention studies] or the mathematics curricula used [curriculum studies];
5. *duration and implementation*: the duration of the mathematics programs or curricula and who implemented it [intervention studies only];
6. *design and procedure* [intervention studies only]: the study design (measurement occasions and intervention) and the procedure of assigning students to conditions;
7. *corrected*: per outcome measure, for which pre-existing differences the comparison was statistically corrected for;
8. *(posttest) results*: per outcome measure, the results of the comparison of posttest scores between programs [intervention studies] or of performance measures with different curricula [curriculum studies], in which it is indicated whether the difference was significant (indicated with < and >) or not significant (n.s.);
9. ES: per outcome measure, the effect size computed (standardized mean difference on posttest), statistically corrected as indicated in column *corrected*.

If applicable, in the columns *(posttest) results* and ES the mean score in the least innovating program was subtracted from the mean score in the more innovating program. Furthermore, if the results were separated by subgroups of students in the original publication, this was also done in the *results* and ES.

1.3 Intervention studies

The didactical approach used can differ greatly between studies.Furthermore, in the programs studied it is very common that more than one didactical element is varied, such as the models used (e.g., the number line), the type of instruction and the role of

the teacher (varying from very directive to very open), the type of problems used (very open problem situations, contextual math problems, or bare number problems), and type of solution strategies instructed (standard algorithms or informal strategies). This mixing of program elements makes it impossible to investigate which of the elements caused the effect reported. The study characteristics of the intervention studies reviewed are displayed in the Summary Table in Appendix 1.A.

In discussing the relevant findings of the intervention studies, we distinguish the results according to the type of comparison that was made. The first type involved comparisons of outcomes of *two or more different experimental programs*, second, the second type comparisons of outcomes of an *experimental program with a control program* (the latter usually the self-selected curriculum), and the third type, comparisons of outcomes of a *supplementary experimental program with a control group* that did not receive any supplementary instruction or practice. In some studies, comparisons of more than one of these categories were made (for instance when there were two experimental programs and one control condition). The findings of these studies were split up accordingly.

1.3.1 Comparing the outcomes of different experimental programs

In this section, study findings regarding comparisons of achievement outcomes of at least two experimental mathematics instruction programs are discussed. For a comparison to qualify in this category, the programs had to be implemented similarly, i.e., by the same kind of instructor in the same kind of instructional setting with the same duration. Six studies compared two specific instructional interventions (guided versus direct instruction) in low mathematically achieving students, in regular education as well as in special education. In another study, two different remedial programs for low mathematics achievers in regular education were compared. Finally, two more studies addressed instructional programs for all students (not only the low achieving ones) in regular education.

Guided versus direct instruction in low mathematics achievers

Six studies focusing on low mathematics achievers, both in special education and in regular education, were quite comparable in their instructional interventions, and are therefore discussed together. Each of these studies compared *guided instruction* (GI)

versus *direct instruction* (DI)[5] in a particular content domain. Guided or constructivistic instruction involved either students bringing up possible solution strategies, or teachers explaining several alternative ways to solve a problem. Students choose a strategy to solve a problem themselves. By contrast, in direct (also called explicit or structured) instruction, students were trained in one standard solution strategy. In one study (Milo, Ruijssenaars, & Seegers, 2005), there were two direct instruction conditions: one (DI-j) instructing the 'jump' strategy (e.g., $63-27$ via $63-20=43$; $43-7=36$), and the other (DI-s) instructing the 'split' strategy (e.g., $63-27$ via $60-20=40$; $3-7=-4$; $40-4=36$, see also Beishuizen, 1993).

The intervention programs consisted of between 26 and 34 lessons. One study (Van de Rijt & Van Luit, 1998) addressed 'early mathematics' in preschoolers, the other studies addressed the domain of multiplication (Kroesbergen & Van Luit, 2002; Kroesbergen, Van Luit, & Maas, 2004) or addition and subtraction below 100 (Milo et al., 2005; Timmermans & Van Lieshout, 2003; Timmermans, Van Lieshout, & Verhoeven, 2007) with students between 9 and 10 years old. With respect to the outcomes, often a distinction was made in automaticity/speed tests, performance measures (achievement on the content domain addressed in the program), and transfer tests (performance on problems that students were not exposed to in the intervention programs). All six studies had a pretest - intervention - posttest design, thereby making statistical correction for pre-existing group differences possible. Either whole classes were randomly assigned to programs, or students within classes were matched and then assigned to programs (however, in Milo et al. (2005) the assignment procedure was unclear). Table 1.2 synthesizes the main findings of these six comparable studies.

In four studies, *automaticity* was an outcome measure. In two studies, a small to medium disadvantage of guided instruction was found, while in the other two studies, differences were negligible. Thus, guided instruction resulted in comparable or lower automaticity outcomes than direct instruction.

All six studies reported on *performance* in the domain of study. Two studies reported a small to medium advantage for guided instruction, two studies found negligible to small advantage of guided instruction, and two studies reported a small to medium advantage for direct instruction. Two additional patterns are worth mentioning. First, in Milo et al. (2005) there were two direct instruction conditions: one (DI-j) instructing the

5 If reported, the comparisons between outcomes of the GI and DI conditions on the one hand and a control condition on the other hand, are discussed in section 1.3.2.

TABLE 1.2 *Synthesis of results from six studies comparing guided instruction (GI) and direct instruction (DI) in low mathematics performers.*

		effect size GI - DI		
study	school type	automaticity	performance	transfer
Kroesbergen & Van Luit (2002)	reg. + spec.	[–.51]	+.43	+.52
	special	*[–2.42]*	*+.32*	*+.36*
	regular	*[+.61]*	*+.86*	*+.95*
Kroesbergen et al. (2004)	reg. + spec.	+.03	–.30	n.a.
Milo et al. (2005)	special	n.a.	–.73 (DI-j)	+.07* (DI-j)
		n.a.	–.21 (DI-s)	+.59* (DI-s)
Timmermans & Van Lieshout (2003)	special	–.23#	.00#	–.57*
Timmermans et al. (2007)	regular	+.05	+.13	n.a.
	girls	*+.07*	*+.84*	*n.a.*
	boys	*+.03*	*–.53*	*n.a.*
Van de Rijt & Van Luit (1998)	regular	n.a.	+.20	n.a.

Note. ES between [ ]: pretest difference > .5 SD, adequate statistical correction not possible.
* no statistical correction for pre-existing differences possible.
# mean difference approximated with available data, in which ES was set to 0 if the only information reported was that the difference was not significant.

'jump' strategy and the other (DI-s) instructing the 'split' strategy. Although in both DI-conditions outcomes were better than in the GI-condition, direct instruction in the jump strategy led to better performance than direct instruction in the split strategy (ES = .52). Second, in Timmermans et al. (2007) differential instruction effect for boys and girls were observed. For girls, guided instruction resulted in better performance, while for boys, direct instruction had better performance outcomes.

Finally, three studies reported results on *transfer*. Again, results were mixed: small to medium differences were found favoring guided instruction as well as favoring direct instruction.

Next to achievement outcomes, other outcomes investigated (not reported in the Summary Table) were strategy use and motivational/affective variables. With respect to strategy use (Kroesbergen & Van Luit, 2002, 2005; Milo & Ruijssenaars, 2005; Timmermans & Van Lieshout, 2003; Timmermans et al., 2007), findings showed that students who received direct instruction in a standard strategy more frequently used that strategy than students who received guided instruction. However, the latter students were not more flexible in their strategy use, meaning that they did not use their larger strategy repertoire adaptively to solve different problems. Finally, there were only minor instruction effects found on variables regarding motivation and affect (Kroesbergen et al., 2004; Milo, Seegers, Ruijssenaars, & Vermeer, 2004; Timmermans et al., 2007).

Remedial programs for low mathematics achievers in regular education

Willemsen (1994, study 2) compared two experimental remedial programs[6] for low mathematics achievers in regular education (grade 4) in the domain of written subtraction. These programs were the 'mapping' program aiming to remediate misconceptions that are at the basis of systematic computational errors, and the 'columnwise' program introducing an alternative strategy replacing the traditional subtraction algorithm. Students trained with the mapping program performed better than students trained with the columnwise program at posttest (ES = +.92) and at retention test (ES = +.64), medium to large differences. Furthermore, students in the mapping program made fewer systematic computational errors than students in the columnwise program (not in the Summary Table). In conclusion, the mapping program for remediating misconceptions that are at the basis of systematic computational errors had small to medium positive effects on written subtraction performance, compared to the columnwise program in which an alternative for the traditional algorithm was instructed.

Other instructional programs in regular education

Two studies compared the outcomes of two experimental programs in regular education students: Klein (1998) compared two instructional programs for addition and subtraction in grade 2, while Terwel, Van Oers, Van Dijk, and Van Eeden (2009; see also Van Dijk, Van Oers, Terwel, & Van Eeden, 2003) compared two instructional programs on 'mathematical modeling' in grade 5.

6 The comparisons with the control program are discussed in section 1.3.2.

First, Klein (1998; see also Blöte, Van der Burg, & Klein, 2001; Klein, Beishuizen, & Treffers, 1998) compared the *Realistic Program design* (RPD) with the *Gradual Program Design* (GPD) in instruction of 2-digit addition and subtraction. In the RPD, the focus was on letting students create and discuss their solution strategies. Realistic contexts for mathematics problems were used, and flexible strategy use was emphasized. Note that the authors contended that this program differed from the principles of realistic mathematics education, with instruction in the RPD being more directive and with students having more opportunity to practice. In the GPD, instruction was more traditional with knowledge being built up stepwise, starting from one basic addition and subtraction procedure: the jump strategy (see before).

No pretest was administered before the program started, so it was not possible to correct for pre-existing group differences. On the posttest, the performance differences (RPD - GPD) in speed tests (ES = +.19), strategy test (ES = +.15), paper-and-pencil addition and subtraction test (ES = +.10), standardized mathematics test LVS (CITO's Student Monitoring System - Mathematics; ES not estimable, difference was not significant), transfer test (ES = –.03), and retention test (ES = +.20) were all negligible to small favoring the RPD. On the speed tests, strategy test, and paper-and-pencil test, the program effects were assessed separately for low and high mathematics achievers. In the low achieving group, students in the RPD program performed better than those in the GPD, with a small to medium effect size (ES +.57, +.31, and +.36, respectively). In the high achieving group, students in the RPD performed better on the speed test (ES = +.47), almost the same on the strategy test (ES = +.02), and lower on the paper-and-pencil test (ES = –.15) than their counterparts in the GPD. However, before the start of the program, the high achievers in the RPD program performed better at the standardized mathematics test LVS (ES = +.50) than the high-achievers in the GPD, a pre-existing difference that could not be statistically accounted for. Furthermore, students in the RPD (low and high achievers) showed more flexible strategy use (not in the Summary Table) than students in the GPD. Finally, there were negligible to small differences in diverse affective and motivational outcomes, usually in the advantage of the RPD.

In summary, achievement outcomes differences were minor to small in favor of the Realistic Program Design over the Gradual Program Design. In addition, the RPD resulted in more flexible strategy use than the GPD, as well as in slightly better outcomes on affective and motivational measures.

Second, Terwel et al. (2009; see also Van Dijk et al., 2003) compared the outcomes of

two instructional programs on mathematical modeling in the domain of percentages and graphs. In the 'co-constructing/designing' program, students were instructed how to make models or representations of the open, complex problem situations that were offered, in co-operation with their classmates and under guidance of their teacher. In the 'providing' program, students were instructed to work with ready-made models that the teacher provided. Furthermore, students worked individually on the problems, followed by a classroom discussion. Note that the authors contended that this latter condition resembles common practice in Dutch education. Results showed that students in the co-constructing/designing program performed better than students in the providing program on problems on percentages and graphs (ES = +.32) and on transfer problems (ES = +.55). The co-constructing/designing program thus appeared to have a small to medium positive effect on achievement, compared to the providing program.

Summary

First, results of six studies on achievement outcomes of guided versus direct instruction in low mathematics performers (special and regular education) were mixed. Differences were found in both directions, and that even within a particular study on different outcome measures as well as between studies within one outcome measure. It seems that factors that were not measured or controlled for, such as the teacher, the composition of the class, and the program implementation, were more important than the instructional approach. The differential gender effect merits further research: in only one study, program effects were reported separately for boys and girls, and large differences in instruction effects were found. Finally, students receiving guided instruction showed a larger strategy repertoire than students receiving direct instruction, but did not use these strategies more adaptively or flexibly.

Second, for low mathematics achievers in regular education, a remedial program based on remediating misconceptions that are at the basis of systematic computational errors had medium to large positive effects on written subtraction performance, compared to a program in which an alternative (RME-based) solution strategy was instructed as replacement of the traditional algorithm. Finally, two studies in regular education showed that the more RME-based instructional programs (RPD in Klein, 1998, and co-constructing/designing program in Terwel et al., 2009) had negligibly small to medium positive effects on achievement, compared to the more traditional instructional

programs.

1.3.2 Experimental programs versus a control program

In this category of intervention studies, we discuss studies in which performance of students who followed an experimental program was compared to performance of students who followed a control program, commonly the regular mathematics curriculum. The majority of the programs addressed low mathematics achievers, both in special and in regular education. There were results of four studies in preschoolers (three with low math achievers), in three studies experimental remedial programs for low mathematics achievers were evaluated, and in the remainder four studies (three with low math achievers) experimental programs for 9 to 10 year-olds were compared to a control program. It is worth noting that besides the instructional program, usually also the instructor (external person in experimental program versus regular teacher in control group) and the instructional setting (small groups of students outside the classroom in experimental program versus whole class in the control group) differed between conditions. Therefore, it is not possible to assign found differences to any of these elements separately.

Preschoolers

In four studies, outcomes of students trained in an experimental program addressing early mathematical skills for preschoolers were compared with outcomes of peers in the regular preschool mathematics curriculum, that in practice was or was not characterized by the use of a specific mathematics textbook. Two studies were carried out in regular education (Poland & Van Oers, 2007; Van de Rijt & Van Luit, 1998), and the other two in special education (Schopman & Van Luit, 1996; Van Luit & Schopman, 2000).

Poland and Van Oers (2007; see also Poland, 2007) developed an experimental program for preschoolers in which schematizing activities were taught in meaningful situations. Preschoolers (not selected on their mathematics achievement level) who followed the program performed at about equal level as their control group peers on a mathematics test halfway the intervention (ES = –.05) and at the end of the intervention (ES = +.02). Eight months after the intervention, they performed better than the controls (ES = +.57), a medium to large difference. At the end of first grade (twelve months after the intervention), this difference reduced (ES = +.18) to a small advantage of the

experimental group. Furthermore, preschoolers in the experimental program showed more schematizing activities during and after the intervention than the controls (not in the Summary Table). In conclusion, the experimental program for preschoolers in which schematizing activities were taught in meaningful situations had a negligibly small to medium sized positive effect on first grade mathematics performance, compared to the control group.

In Van de Rijt and Van Luit (1998; see also section 1.3.1), low achieving preschoolers trained with the Additional Early Mathematics (AEM) program (either in the guided instruction or in the direct instruction variant) outperformed their control group peers in early mathematics skills, with large differences (ES = +1.06 and ES = +1.26, respectively). Thus, the AEM-program had a large positive effect on low achieving preschoolers' early mathematics skills.

There were two intervention studies with programs for preschoolers with low mathematics achievement level in special education. Schopman and Van Luit (1996) investigated the effect of an intervention program addressing counting to 10 as preparation for formal mathematics education that starts in first grade in special education. Preschoolers with a low mathematics level who were trained with this experimental program[7] performed better on a test of preparatory arithmetic skills (ES = +1.07) than preschoolers in the control group, a large effect. In the second study, Van Luit and Schopman (2000) extended the intervention program to more sessions and to numbers up to 15. Again, preschoolers in the experimental program performed better than their peers in the control group on a test of early numeracy (ES = +.73), and also on a transfer test (ES = +.22). In conclusion, in both studies, preschoolers who followed a preparatory program on counting skills to 10 or 15 performed better on a test of early numeracy than preschoolers in the control group, with medium to large differences.

Remedial programs

In three studies (one in special education, and two in regular education) the effects of an experimental remedial program compared to the regular mathematics curriculum were addressed.

7 In Schopman and Van Luit (1996) there were actually two experimental conditions: one with guiding instruction, and one with directing instruction. However, these instructional variants appeared not to differ from each other in practical implementation. Therefore, the results of these two experimental conditions were combine in the current review.

Harskamp and Suhre (1995) developed a remedial program for instruction in addition and subtraction below 100 for low mathematics achievers (10-11 years old) in special education. The program aimed to build on students' individual solution strategies, and it replaced two regular mathematics lessons a week. The program turned out to have a large positive effect compared to the control group that followed just the regular lessons on posttest and retention test achievement in addition and subtraction (ES = +3.22, but adequate statistical correction not possible), also separately for students with learning disabilities (LD) (ES = +3.13, but adequate statistical correction not possible) and for students with learning difficulties (MR) (ES = +3.69). Furthermore, the program also had a large positive effect on application problems in LD students (ES = +3.58, but adequate statistical correction not possible) and MR students (ES = +3.58). In conclusion, the experimental remedial program had large positive effects on addition and subtraction performance in LD and MR students in special education, compared to the control group.

Willemsen (1994) compared one (study 1) or two (study 2) experimental remedial programs[8] for low mathematics achievers in regular education (grade 4) in the domain of written subtraction with a control program, in which the subject matter was systematically rehearsed and practiced. In study 1, students in the 'mapping' program performed better at posttest than students in the control program (ES = +.32), a small to medium difference. In study 2, students in the mapping program again performed better than students in the control program at posttest (ES = +.74) and at retention test (ES = +.84), medium to large differences. Students in the columnwise program, however, performed somewhat less well than students in the control program at posttest (ES = –.17), but somewhat better at retention test (ES = +.20). Furthermore, students in the mapping program made fewer systematic computational errors than students in the control program (study 1 and 2, not presented in the Summary Table). In conclusion, the mapping program for remediating misconceptions that are at the basis of systematic computational errors had small to medium positive effects on written subtraction performance compared to the control program (systematic rehearsal and training). By contrast, the outcomes differences of the other experimental remedial program 'columnwise' versus the control program were only small and in both directions.

8 See section 1.3.1 for the comparison of the outcomes of the two experimental remedial programs.

Other studies

The results of four studies in which the outcomes of an experimental program were compared with the outcomes of a control group who followed the regular curriculum remain.

Keijzer and Terwel (2003; see also Keijzer, 2003) developed a program for instruction in fractions in fourth grade. This program was innovating compared to the RME-based textbook *Wereld in Getallen* (WIG) used in the control group on two aspects: the fractions model (number line versus circles or bars in WIG) and the instructional approach ('negotiation of meaning' in whole class discussions versus students working individually in WIG). On standardized LVS mathematics tests, differences between the groups were negligible in the domain of numbers and operations (ES = –.01), but students in the experimental group performed better than the controls in the domain of measures and geometry (ES = +.35), a small to medium difference. On fraction problems that were administered in interviews with standardized support, students in the experimental program performed better than the controls (uncorrected ES = +.52). In conclusion, the fractions program had no effects to medium sized positive effects on fourth graders' mathematics performance, compared to the control group.

Van Luit and Naglieri (1999) developed the MASTER program for students (age 10-12 years) in special education, focused on the development of solution strategies for multiplication and division up to 100. The program used principles of self-instruction, discussion, and reflection. Students who followed this program performed much better than students from the control group (ES = +2.16), which also held separately for LD students (ES = +2.50) and for MR students (ES = +3.08). Furthermore, there were also positive effects on a follow-up test (LD and MR students) and far transfer (only LD students; not in the Summary Table). In conclusion, the MASTER-training, aimed at development of strategies for multiplication and division below 100 making use of self-instruction, discussion, and reflection, had very large positive performance effects compared to the control group.

Finally, in both studies of Kroesbergen (Kroesbergen & Van Luit, 2002; Kroesbergen et al., 2004) from section 1.3.1 a modified version of the MASTER program was used. The comparisons between the experimental conditions (GI and DI) on the one hand and the control conditions on the other hand fit in the current section. In Kroesbergen and Van Luit (2002), posttest differences between students in the GI-condition and control

students were zero to large, with ES .00, +.89, and +.96 in automaticity, multiplication ability, and transfer, respectively. Comparisons between students in the DI-condition and control students should be evaluated with caution because pretest differences were too large to adequately statistically account for, but nevertheless all results favored the experimental program with ES +.51, +.46, and +.44 in automaticity, multiplication ability, and transfer, respectively. Similarly, in Kroesbergen et al. (2004) students in the experimental programs variant performed better than control students in automaticity (ES +.35 for GI and +.32 for DI) and in multiplication ability (ES = +.23 for GI and +.53 for DI). In conclusion, there were small, medium, and large positive effects of the program found compared to the regular curriculum, both in special education students and in regular education students.

Summary

The experimental programs investigated had negligibly small to large positive effects on mathematics performance, compared to the control group in which students usually followed the regular curriculum implemented by the regular teacher. These experimental programs each incorporated aspects of RME: development of solution strategies by self-instruction, discussion, and reflection; schematizing in meaningful situations; the number line as model; and whole-class discussion aiming at 'negotiation of meaning'. However, it is impossible to disentangle the effects of these elements from the general implementation differences between experimental and control conditions, such as instructor and instructional setting.

1.3.3 Supplemental programs for low mathematics achievers

There were two studies in which the effects of supplemental remedial or training programs for low mathematics achievers in regular education were investigated.

Harskamp, Suhre, and Willemsen (1993) compared performance of regular education students (grade 2 and 3) in six different combinations of a mathematics textbook based on RME principles on the one hand (*Wereld in Getallen, Operatoir Rekenen*, or *Rekenen & Wiskunde*), and a remedial program that was either structuralistic (more traditional: *Rekenspoor* or *Gouds Rekenpakket*) or RME-based (*Remelka*) on the other hand, with performance of students in the control group who did not receive this supplemental remedial training. Because the practical implementation of the six different combination

appeared not to differ from each other, we will not differentiate between them here. Supporting this equivalence was the result that performance of students in the six combinations of RME-textbook and RME-based or structuralistic remedial program did not differ from each other on either the number problems or the application problems. Compared to the control group, however, posttest performance in bare number problems was higher in the six remedial conditions in grade 2 (ES = +1.18, but pretest differences to large for adequate statistical correction) and in grade 3 (ES = +.39). On application problems, small positive effects of the remedial conditions compared to the control condition were found in grade 2 (ES = +.17) and in grade 3 (ES = +.24). In conclusion, the remedial programs seemed mainly to improve low mathematics achievers' abilities in number problems, irrespective of the didactical characteristics of the remedial program and the combination with didactical characteristics of the regular mathematics textbook.

Finally, Menne (2001) developed a supplemental 'productive practice' (in contrast to 'reproductive practice') program. This program addressed basic counting with units and tens, aiming to make students jump fluently and flexibly on the (empty) number line with varying step lengths. She implemented this program in grade 2 of regular education, and compared it to a control group of students who only followed their regular lessons. Students following the supplemental training program performed better than their control group peers: on LVS tests the ES was approximately +.44, and the performance difference between students who did and who did not follow the training program was larger for ethnic minority students (approximated ES = +.59) than for native Dutch students (approximated ES = +.41). In conclusion, the supplemental productive practice program had a small to medium positive effect on mathematics performance compared to the control group, in particular for ethnic minority students.

Summary

In these two studies, a positive effect of supplemental programs on students' achievement was found, compared to the control students who followed their regular mathematics lessons and did not receive extra training.

1.4 Curriculum studies

As said, curriculum studies are observational studies aiming to investigate differential achievement outcomes of different mathematics curricula, usually different mathematics

textbooks. They are discussed in three sections: domain-specific studies that address one specific delimited content domain of mathematic, large-scale curriculum studies carried out in 1980s that addressed general mathematics achievement, covering a range of mathematical domains, and differential outcomes by mathematics textbook in the Dutch national assessments. All study characteristics are in the Summary Table in Appendix 1.B.

1.4.1 Domain-specific curriculum studies

Two studies analyzed performance difference between students with different mathematics curricula on a specific content domain: one on addition and subtraction in special education (Van Luit, 1994) and the other on division in regular education (Van Putten, Van den Brom-Snijders, & Beishuizen, 2005).

Van Luit (1994) compared special education students' (age 9-11 years) addition and subtraction performance who followed a structuralistic or an RME-based curriculum. On the posttest[9] involving addition and subtraction without crossing tens, MR-students in the RME-based curriculum performed somewhat worse (ES = –.22; a small difference) than MR-students in the structuralistic curriculum, while in addition and subtraction with crossing tens there was only a negligible difference (ES = +.04). In LD-students, performance differences were in disadvantage of the RME-based curricula, with respectively ES = –.62 and ES = –1.00. On problems involving a realistic context, performance differences between LD-students in structuralistic or RME-based curricula were minor (ES = –.08). In conclusion, addition and subtraction performance of special education students (MR and LD) in RME-based curricula was equal to or lower than in structuralistic curricula.

Van Putten et al. (2005) compared fourth graders' division performance with two different textbooks, *Rekenen & Wiskunde* (R & W) and *Wereld in Getallen* (WIG) in regular education. Both textbooks are based on RME-principles, but WIG has a more (pre-)structured learning trajectory for division than R & W. Halfway fourth grade, R & W students had lower performance than WIG students (ES = –.43), while at the end of grade four the performance difference was reversed (ES = +.35). Furthermore, strategy use (not in the Summary Table) developed positively over time on the aspects schematizing (R & W more increase than WIG) and number relations (R & W and WIG same increase,

9 Although a pretest was administered, differences were not corrected for, because at the time the pretest was administered the students already had six months instruction in addition and subtraction according to a structuralistic or RME-based curriculum.

but WIG higher overall score). These results from Dutch students were also compared with UK students from the same age (Anghileri, Beishuizen, & Van Putten, 2002; not in the Summary Table). In the UK, the learning trajectory for division is characterized by a rather abrupt transition of informal solution strategies to the traditional long division algorithm. By contrast, in the Dutch RME-based textbooks R & W and WIG, informal strategies are progressively schematized toward more structured and efficient strategies (not the traditional algorithm). At the end of fourth grade, Dutch students outperformed the UK students, an indication that the progressive schematization of informal solution strategies was effective. In summary, Dutch students with mathematics textbook *Rekenen & Wiskunde* had a lower division performance than students with the textbook *Wereld in Getallen* halfway fourth grade, but reversed this to an advantage at the end of grade four. Both groups of Dutch students outperformed UK counterparts.

1.4.2 Large-scale curriculum studies from the 1980s

In the 1980s, two large-scale curriculum studies were carried out, in which modern (at that time) mathematics textbooks were compared to traditional textbooks.

First, the MORE-project was carried out at the Freudenthal Institute, a study in which students were longitudinally followed from first to third grade (Gravemeijer et al., 1993). The two mathematics curricula compared were the traditional textbook series *Naar Zelfstandig Rekenen* (NZR) and the modern RME-based textbook series *Wereld in Getallen - edition 1* (WIG-1). Results were corrected for students' mathematics level in first grade, socio-economical background, and intelligence scores. On general mathematics, WIG-1 students performed at approximately equal level as NZR students in grade 1 (ES = –.02), but were outperformed in grade 2 (ES = –.10; negligible to small difference) and in grade 3 (ES = –.32; small to medium difference). On automatization WIG-1 students were outperformed by NZR students with medium to large differences, in grade 2 (ES = –.60) and in grade 3 (ES = –.58). Furthermore, investigation of the implementation of the two textbooks (not in the Summary Table) showed that the instruction in NZR-teachers was reasonably mechanistic (traditional), while the instruction by WIG-teachers was RME-based only to a limited extent. Thus, the implemented curriculum in NZR-teachers was more in accordance with the didactical theory in the textbook series than in WIG-teachers. In conclusion, in grade 1 to 3, students in the RME-based *Wereld in Getallen - edition 1* curriculum performed negligibly to substantially lower than students in the

traditional *Naar Zelfstandig Rekenen* curriculum, in particular on automatization.

Second, Harskamp (1988) compared sixth graders' achievement outcomes with 8 different mathematics curricula (textbook series), 3 of which he classified as 'modern' (NB. not *Wereld in Getallen*) and 5 as 'traditional' (among which *Naar Zelfstandig Rekenen*). Corrected for intelligence scores, performance differences on the CITO End of Primary School Test (ES = +.09) and at mathematics tests developed at RION (ES = +.06) were negligible to small in the advantage of modern textbooks. Furthermore, there were implementation factors associated with mathematics performance (not in the Summary Table): two general factors with a positive relation with performance were the number of mathematics lessons per week and the percentage of students for whom the basic subject matter from the textbook was covered, and two content-specific factors were variation in subject matter (positive relation) and differentiation in subject matter (negative relation). In summary, students with 'modern' mathematics textbook series performed somewhat better than students with 'traditional' mathematics textbook series.

1.4.3 Differential mathematics outcomes by mathematics textbook in national assessments

In CITO's national assessments of mathematics education, usually performance differences between students who were taught with different textbook series are reported.

The most recent assessment halfway primary school (grade 3) carried out in 2003 analyzed performance differences between students with seven different mathematics textbooks, corrected for several background variables (Kraemer, Janssen, Van der Schoot, & Hemker, 2005). Students who were instructed with *Talrijk* and *Rekenrijk* performed best, students with *Wereld in Getallen - edition 1* and *Rekenen & Wiskunde* had the lowest performance level. The difference between highest and lowest average performance by textbook was medium to large (ES = +.64). Because all textbooks used were based on RME-principles, it was not possible to compare performance between RME-based and traditional curricula. What was reported, though, is that the shift in mathematics textbooks shares between the cycles of 1997 and 2004 had in general a small positive effect on students' mathematics performance (ES = +.18). It is hard to characterize this shift in terms of RME-based versus traditional curricula, because in the 1997 cycle less that 5% of the textbooks used was still traditional.

The most recent assessment at the end of primary school (grade 6) carried out in 2004

did not report performance differences between different mathematics textbooks used, because 80% of all schools started using a new textbook in reaction to the introduction of the new currency (the euro) in 2002, and therefore sixth graders had experienced a change in textbook in their primary school trajectory (J. Janssen et al., 2005). Only the summative effects of shifts in mathematics textbooks shares were reported. In the period 1997 to 2004 this summative effect was positive but very small (ES = +.12), in the period 1992 to 2004 this effect was also positive but small (ES = +.18). Thus, in total, the shift in market share of mathematics textbooks between 1992 and 2004, from 37% to 100% RME-based textbooks, had a negligible to small positive effect on sixth graders' mathematics performance.

In the third assessment cycle (1997) at the end of primary school, performance differences by mathematics textbook used were still reported (J. Janssen, Van der Schoot, Hemker, & Verhelst, 1999). Students instructed with *Wereld in Getallen - edition 2* performed best, students with *Niveau Cursus Rekenen* and *Naar Zelfstandig Rekenen* performed at the lowest level. The difference between highest and lowest average performance by textbook was medium (ES = +.53). Importantly, differences *within* a curriculum type (RME-based or traditional) were larger than *between* the two curriculum types.

In summary, in third grade, mathematics performance of students instructed with one of seven different RME-based mathematics textbooks differed from each other. Similarly, in sixth grade, the 1997 cycle showed performance differences by mathematics textbook, in which differences within a curriculum type (RME-based or traditional) were larger than between curriculum types. Both in third grade and in sixth grade, the shift in market shares of mathematics textbooks over time had a very small to small positive effect on mathematics performance. In third grade, this shift in textbooks used was almost entirely within the spectrum of RME-based curricula. In sixth grade, this entailed both shifts from traditional and hybrid textbooks toward RME-based textbooks, as well as shifts within the different RME-based textbooks.

Summary

The domain-specific curriculum studies showed that in special education, students in a structuralistic curriculum outperformed students in an RME-based curriculum on addition and subtraction. In regular education, fourth graders' division performance

with two RME-based textbooks showed a varying pattern over the school year, but both curricula seemed more effective than the UK approach that was more traditionally oriented.

Within the large-scale curriculum studies covering many domains of mathematics, one of the studies from the 1980s showed that students with the RME-based textbook *Wereld in Getallen - edition 1* (WIG-1) were outperformed by students with the traditional textbook *Naar Zelfstandig Rekenen* (NZR). By contrast, the other curriculum study from the 1980s showed that students with 'modern' textbooks slightly outperformed students with 'traditional' textbooks.

The Dutch national assessments of mathematics education showed first and foremost that there are no univocal results in comparing students' performance with RME-based and traditional mathematics curricula. There were substantial performance differences within both curricula types. Furthermore, the shift from older to newer RME-based textbooks resulted in somewhat better performance in grade 3. In grade 6, the shift from traditional, hybrid, and (older) RME-based textbooks toward (newer) RME-based textbooks also had a small positive effect on mathematics performance.

1.5 SUMMARY, CONCLUSIONS, AND IMPLICATIONS

The results of all empirical studies reviewed taken together do not give an unequivocal picture on the relation between mathematics instruction and mathematics performance in the Netherlands. There is remarkably little research that allows for well-grounded and univocal conclusions, a similar conclusion as was reached in other research syntheses (e.g., Kroesbergen & Van Luit, 2003; National Mathematics Advisory Panel, 2008; Slavin & Lake, 2008). *Intervention studies* compare the effects of different mathematics interventions, i.e., instructional programs. The intervention studies reviewed do not yield firm conclusions, because they are limited in content domain, sample size, duration or magnitude of the intervention, and range of outcome variables. In addition, usually several didactical and instructional aspects were varied simultaneously, making it impossible to disentangle their effects. *Curriculum studies* make a large-scale comparison between performance of students who were instructed with different mathematics curricula or textbook series. However, these studies are limited in the amount of control on the practical implementation of the curricula and in correction for confounding variables: they are carried out in everyday educational practice.

Furthermore, the results of the curriculum studies did not point univocally in one direction either. More generally speaking, the idea of extending the concept of 'evidence-based medicine' towards the field of education (e.g., National Mathematics Advisory Panel, 2008) is hampered by many practical limitations.

Regarding the public debate on RME-based mathematics education versus traditional education, it is important to note that – even if it would be possible to speak of 'the' RME approach and 'the' traditional approach – the studies reviewed rarely compared these instructional approaches directly. Rather, specific elements that can be characterized as reform-based/constructivistic or more traditional/mechanistic were varied. The effects found were, generally speaking, small. Synthesizing all results, one could say that a very small advantage of RME-based programs was found, in particular regarding the higher-order goals such as flexibility. This advantage, however, was too small to draw firm conclusions. In addition, there are studies that favored more traditional programs instead.

The lack of a firm general conclusion, however, does not preclude conclusions on some more specific patterns. First, it is striking that *within* a type of instructional approach performance differences were larger than *between* instructional approaches. Apparently, didactical principles play a less important role than the practical implementation by the teacher and the teacher-student interaction, implications that agree with the findings of for example Slavin and Lake (2008). Second, more time spent on mathematics education leads to better performance, as comes forward from the positive results of remedial and training programs that are supplemental to the regular curriculum, which is congruent with the notion of 'opportunity to learn' being the most important predictor of mathematics achievement outcomes (Hiebert & Grouws, 2007). Third, with educational time held equal, experimental programs implemented in small groups of students outside the classroom had positive effects compared to the regular educational practice. Similarly, Slavin and Lake (2008) also reported a positive effect of small-group tutoring. Fourth, in the studies reviewed there was a lot of attention for low mathematics performers. These students seemed to benefit less from a free form of instruction and to have a larger need for a more directing role of their teacher in their learning process, similar to findings of international reviews (Gersten et al., 2009; Kroesbergen & Van Luit, 2003; Swanson & Carson, 1996; Swanson & Hoskyn, 1998). Much less research has been done into the relation between mathematics instruction and mathematics performance in medium and high performers. Likewise, little is known

about differential instruction effects for boys and girls.

The main implication of the current research synthesis may be that the key to improving mathematics education seems to be in the teacher quality (KNAW, 2009). The crucial role of the teacher in mathematics education also comes forward from international reviews (Hiebert & Grouws, 2007; Hill et al., 2007; Kroesbergen & Van Luit, 2003; Slavin & Lake, 2008; Verschaffel et al., 2007). It is widely accepted that teachers differ in their effectiveness, although empirical evidence is weak (Nye, Konstantopoulos, & Hedges, 2004). Furthermore, the teacher is the pivot in the learning/teaching process. We put high demands on our teachers, in particular in instructional approaches that focus on the input of and interaction with the students, such as in mathematics education reform (see also Stein et al., 2007). These demands necessitate an investment in teacher training to consolidate and improve teachers' (pedagogical) content knowledge. Noteworthy in this light is the finding from TIMSS-2007 that the Dutch teachers participated the least in professional development in mathematics from all participating countries (Mullis et al., 2008). Furthermore, Dutch pre-service teacher training programs has received a lot of criticism, but promising developments (e.g., future teachers need to show a sufficient level of mathematical content knowledge and skills in a mandatory test, and a national framework has been developed of knowledge that future mathematics teachers need to have) have taken place recently. A final conclusion is that more research into effective instructional elements and ways to improve mathematics instruction and teachers' quality is indispensable for further educational recommendations.

APPENDIX 1.A STUDY CHARACTERISTICS OF INTERVENTION STUDIES

references	domain	participants	intervention	duration and implementation	design / procedure	corrected	posttest results	ES
Harskamp & Suhre (1995)	addition and subtraction < 100	*N* = 24 4 schools M = 10.2 years (LD) M = 11.4 years (MR) special education low math performers	1. remedial program 2. control (regular math curriculum; all structuralistic or traditional textbooks)	remedial program: 13 weeks, two 45-minutes lessons per week replacing two regular math lessons implementation remedial program by remedial assistant with 2 students outside the classroom implementation control curriculum by regular teacher in whole class	pretest - intervention - posttest selection and assignment procedure unclear, but assignment to conditions within schools	pretest add. and subtr. < 100	addition and subtraction < 100 (posttest and retention test) remedial > control [+3.22] *LD students [+3.13]* *MR students +3.69*	
						pretest add. and subtr. < 100	application problems remedial > control ?? *LD students [+3.58]* *MR students +3.58*	
Harskamp, Suhre & Willemsen (1993)	grade 2: addition & subtraction < 20 grade 3: addition & subtraction < 100, and multiplication tables	grade 2: *N* = 357 89 schools grade 3: *N* = 242 81 schools regular education low math performers	1. remedial: 6 combinations of 1 out of 3 remedial programs and 1 out of 3 math textbooks (self-selected) 2. control (no remedial program)	duration remedial program: 12 blocks of lessons (in total 16 hours) in 4-5 months implementation remedial program by trained remedial assistants	pretest - intervention - posttest assignment to remedial programs at school level, but precise procedure unclear	pretest number problems	numerical problems add. and subtr. (and mult. in grade 3) gr. 2: remedial > control [+1.18] gr. 3: remedial > control +.39	
						pretest application problems	performance application problems gr. 2: remedial – con. n.s. +.17 gr. 3: remedial – con. n.s. +.24	

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Keijzer (2003) Keijzer & Terwel (2003)	fractions	*N* = 20 2 classes (1 school) grade 4 (9-10 years old) regular education	1. experimental fractions program (exp) 2. control (regular math textbook *Wereld in Getallen*)	one school year experimental program implemented by researcher control curriculum by regular teacher	pretest - intervention - posttest assignment to conditions on class level, not random post-hoc matching on student level	pretest LVS – numbers and operations	LVS – numbers and operations exp − control n.s.	−.02
						pretest LVS – measuring and geometry	LVS - measuring and geometry exp − control n.s.	+.35
						no correction	skills in fractions (interviews) exp > control	+.52
Klein (1998) Klein, Beishuizen & Treffers (1998) Blöte, Van der Burg & Klein (2001)	two-digit addition and subtraction < 100	*N* = 275 10 classes (9 schools) grade 2 regular education	1. Realistic Program Design (RPD) 2. Gradual Program Design (GPD)	one school year, program replacing 75% of regular math curriculum implementation by regular teacher in whole class	intervention - intermediate test - posttest random assignment to programs on class level matched pairs of classes (based on LVS general math scores)	no correction	speed-tests add. and subtr. RPD − GPD n.s. *low math performers* *high math performers*	+.19 *+.57* *+.47*
						no correction	strategy-test add. and subtr. RPD − GPD n.s. *low math performers* *high math performers*	+.15 *+.31* *+.02*
						no correction	p & p test add. and subtr. RPD − GPD n.s *low math performers* *high math performers*	+.10 *+.36* *−.15*
						no correction	LVS-test End Grade 2 RPD − GPD n.s.	??
						no correction	transfer (add. / subtr. > 100): RPD − GPD n.s.	−.03
						no correction	retention (add. / subtr < 100) RPD > GPD	+.20

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Kroesbergen & Van Luit (2002)	multiplication up to 10 x 10	*N* = 75 7 schools M = 9 years regular education + special education low math performers	1. guided instruction (GI) 2. structured instruction (SI) 3. control (regular math curriculum, mixture of different instructional approaches)	4 months, GI and SI involved 30 sessions of 30 minutes, replacing 2 regular math lessons a week GI and SI implemented by researcher in small groups of students control condition by regular teacher in whole class	pretest - intervention - posttest random assignment to experimental condition (GI or SI) vs. control condition on student level random assignment within experimental conditions to either GI or SI on class level	pretest multiplication automaticity	multiplication automaticity GI − control: n.s. SI > control GI < SI *regular ed.: GI > SI* *special ed.: GI < SI*	 .00 [+.51] [−.51] *[+.64]* *[−2.42]*
						pretest multiplication ability	multiplication ability GI > control SI > control GI > SI *regular ed.: GI > SI* *special ed.: GI > SI*	 +.89 [+.46] +.43 *[+.85]* *+.32*
						pretest transfer	transfer (up to 10 x 20) GI > control SI > control GI > SI *regular ed: GI > SI* *special ed: GI > SI*	 +.96 [+.44] +.52 *[+1.03]* *+.36*
Kroesbergen, Van Luit & Maas (2004)	multiplication	*N* = 265 23 schools M = 9.7 years regular education + special education, only low math performers	1. constructivist instruction (CI) 2. explicit instruction (EI) 3. control (regular math curriculum)	4 months, GI and SI involved 30 sessions of 30 minutes, replacing 2 regular math lessons a week GI and SI implemented by externals in small groups of students control implemented by regular teacher in whole class of students	pretest - intervention - posttest random assignment to experimental condition vs. control condition on student level random assignment within experimental conditions to either GI or SI on class level	pretests, months of prior multiplication instruction, gender, general math level, IQ, school type (regular / special)	multiplication automaticity CI > control EI > control CI − EI: n.s.	 +.35 +.32 +.03
							multiplication ability CI > control EI > control CI < EI	 +.23 +.53 −.30

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Menne (2001)	mental calculation < 100	*N* = 225 12 schools (with sufficient number of low math achievers) grade 2 regular education	1. productive training program 2. control (regular math curriculum, no extra training sessions)	whole school year, at least 3 times a week 15 minutes training program implemented by regular teacher	pretest - intervention - posttest assignment to program at school level, but precise procedure unclear	pretest LVS (general math)	LVS-tests (general math) [NB. means and SDs approximated] training > control *ethnic minorities* *native Dutch*	 +.44 *+.59* *+.41*
Milo, Ruijssenaars & Seegers (2005)	addition and subtraction < 100	*N* = 70 3 schools M = 9.8 years special education	1. directing instruction - jump (DI-j) 2. directing instruction - split (DI-s) 3. guiding instruction (GI)	6 months, 30 lessons, replacing 2 lessons a week implemented by trained university students in groups of 3-5 students outside the class	pretest -intervention - posttest assignment of groups of students (uniform MR or LD) to programs procedure selection and assignment unclear	pretest add. and subtr. < 100	performance add. and subtr. GI < DI-j GI − DI-s: n.s. DI-j − DI-s: n.s.	 −.73 −.21 +.52
						no correction	transfer (add. / subtr. > 100) GI − DI-j: n.s. GI − DI-s: n.s. DI-j − DI-s: n.s.	 +.07 +.59 +.52
Poland (2007) Poland & Van Oers (2007)	general mathematics	*N* = 133 6 schools (with 'develop-mental education' approach) preschool through grade 1 regular education	1. experimental program 'schematizing' 2. control (standard preschool curriculum)	one school year (2nd preschool year) experimental program implemented by regular teacher supported by researcher control by regular teacher	pretest - intervention - posttest matching schools in pairs, assignment procedure within pairs unclear	pretest number sense	general math: halfway exp − control n.s.	 −.05
						pretest number sense	general math: end exp − control n.s.	 +.02
						pretest number sense	general math: 8 months after exp > control n.s.	 +.57
						pretest number sense	general math: 12 months after exp − control n.s.	 +.18

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Schopman & Van Luit (1996)	preparatory arithmetic skills	*N* = 60 unknown number of preschools preschoolers (5-7 years old) special education low math performers	1. experimental program (guiding or directing instruction*) 2. control (standard curriculum)	experimental program: 3 months, 13 lessons, ½ hour sessions twice a week implementation experimental program: unclear by whom, in groups of 4 students control curriculum by regular teacher	pretest - intervention - posttest assignment to program at student level with matching of students, but precise assignment procedure unclear	pretest preparatory arithmetic skills	preparatory arithmetic skills exp > control	+1.07
***** In the practical implementation, the instructional approaches did not differ from each other. Therefore, these groups were combined here.								
Timmermans & Van Lieshout (2003)	subtraction < 100	*N* = 16 2 schools M = 10.5 years special education students low performing in subtraction	1. guided instruction (GI) 2. direct instruction (DI)	34 lessons (of which 24 specific GI vs. DI) all lessons implemented by same trainer in groups of 4 students	pretest - intervention - posttest students within classes assigned to GI vs. DI students in DI and GI matched on pretest and age	pretest speed-test	speed-test addition (without regrouping) GI < DI	−.99
						pretest speed-tests	remaining speed-tests add. and subtr. GI − DI n.s.	??
						pretest performance	performance test GI − DI n.s.	??
						no correction	transfer (add. / subtr. > 100 without regrouping) GI < DI	−.96
						no correction	transfer (add. / subtr. > 100 with regrouping) GI − DI n.s.	−.18

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Timmermans, Van Lieshout & Verhoeven (2007)	subtraction < 100	*N* = 40 5 schools M = 9.3 years regular education students low performing in subtraction	1. guiding instruction (GI) 2. directing instruction (DI)	34 lessons (of which 24 specific GI vs. DI), 2 lessons a week within a school, all GI / DI lessons implemented by same trainer in groups of 4 students	pretest -intervention - posttest assignment: matched pairs of students within classes, within each pair random assignment to GI vs. DI	pretest speed-tests	speed-tests add. and subtr. GI − DI n.s. *girls: GI − DI n.s.* *boys: GI − DI n.s.*	 +.05 *+.07* *+.03*
						pretest power test	power test add. and subtr. GI − DI n.s. *girls: GI > DI* *boys: GI < DI*	 +.13 *+.84* *−.51*
						no correction	strategy- test subtraction GI − DI n.s. *girls: GI > DI* *boys: GI − DI n.s.*	 +.17 *+.60* *−.25*
Van de Rijt & Van Luit (1998)	early mathematics	*N* = 136 20 schools M = 5.9 years regular education low math performers	1. AEM program - guiding instruction (AEM-GI) 2. AEM program - structured instruction (AEM-SI) 3. control (regular math curriculum)	13 weeks, AEM-GI and AEM-SI twice a week 30 minutes lesson replacing the regular math curriculum AEM-GI and AEM-SI: instruction in groups of 4-5 students, probably (?) by regular teacher	pretest -intervention - posttest matching of students on pretest, age, and gender assignment to conditions on student level	pretest early mathematics	early mathematics AEM-GI > control AEM-SI > control AEM-GI − AEM-SI: n.s.	 +1.06 +1.26 +.20
Van Dijk, Van Oers, Terwel & Van Eeden (2003); Terwel, Van Oers, Van Dijk & Van Eeden (2009)	mathe-matical modeling: percentages and graphs	*N* = 238 10 classes (8 schools) grade 5 10-11 years regular education	1. program 'co-constructing / designing' (CCD) 2. program 'providing' (P)	3 weeks, 13 daily lessons of 1 hour implementation by regular teacher	pretest - intervention - posttest random assignment to programs on class level	standardized pretest general math	percentages and graphs CCD > P	 +.33
						standardized pretest general math	transfer: CCD > P	 +.55

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Van Luit & Naglieri (1996)	multiplication and division < 100	*N* = 84 unknown number of classes special education M = 10.8 years (LD) M = 12.7 years (MR) low math performers	1. experimental MASTER program 2. control (standard curriculum)	MASTER program: 17 weeks, 3 times a week 45-minutes lesson, replacing all regular math lessons implementation MASTER program by remedial teacher in groups of 5-6 students outside the class control by regular teacher	pretest - intervention - posttest random assignment to program at student level, within type of student (LD or MR)	pretest multiplication and division	multiplication and division < 100 exp > control *LD students* *MR students*	 +2.16 *+2.50* *+3.08*
Van Luit & Schopman (2000)	early numeracy	*N* = 124 9 schools preschoolers (5-7 years old) special education low math performers	1. experimental intervention program 2. control (standard curriculum)	exp program: 6 months, 20 lessons, ½ hour sessions twice a week, replacing all regular math lessons implementation experimental program by trained assistant, in groups of 3 students control curriculum by regular teacher	pretest - intervention - posttest assignment to program at student level with matching of students, but precise assignment procedure unclear	pretest early numeracy	early numeracy exp > control	 +.75
						no correction	transfer exp − control n.s.	 +.22

references	domain	participants	intervention	duration and implementation	design and procedure	corrected	posttest results	ES
Willemsen (1994) - study 1	written subtraction	*N* = 40 5 classes (5 schools) grade 4 regular education low math performers	1. remedial program 'mapping' 2. control remedial program 'systematic practice'	10 weeks, 10 50-minutes lessons mapping program implemented by remedial teacher in groups of 3-5 students control program implemented by regular teacher	pretest - intervention - posttest random assignment to conditions on class level	pretest written subtraction	written subtraction mapping > control	 +.32
Willemsen (1994) - study 2	written subtraction	*N* = 34 3 schools grade 4 regular education low math performers	1. remedial program 'mapping' 2. remedial program 'columnwise' 3. control remedial program 'systematic practice'	10 weeks, 10 50-minutes lessons remedial programs (1+2) implemented by remedial teacher in groups of 3-5 students control program implemented by regular teacher	pretest -intervention - posttest random assignment to conditions on student level	pretests written subtraction & mental computation	written subtraction mapping > control column − control: n.s. mapping > column	 +.75 –.17 +.92
						pretests written subtraction & mental computation	retention written subtraction mapping > control column − control: n.s. mapping > column	 +.84 +.20 +.64

APPENDIX 1.B STUDY CHARACTERISTICS OF CURRICULUM STUDIES

references	domain	participants	curriculum	duration	corrected	results	ES
Gravemeijer et al. (1993) [MORE-study]	general mathematics	*N* = 430 18 schools longitudinal setup grade 1 – grade 3 regular education	1. RME-based curriculum *Wereld in Getallen – edition 1* (WIG-1) 2. traditional curriculum *Naar Zelfstandig Rekenen* (NZR)	one to three school years	grade 1 math level, SES, intelligence	general mathematics grade 1 grade 2 grade 3	-.02 -.10 -.32
					grade 1 math level, SES, intelligence	automatizing grade 2 grade 3	-.60 -.58
Harskamp (1988)	general mathematics	*N* = 2579 120 schools grade 6 regular education	math textbook used 1. modern (3 textbook series) 2. traditional (5 textbook series)	school years up to measurement	intelligence	CITO test (general math) modern vs. trad. n.s.	+.09
					intelligence	RION-tests (general math) modern vs. trad. n.s.	+.06
Janssen, Van der Schoot, Hemker & Verhelst (1999) = PPON grade 6 - 1997	general mathematics	*N* = 7890 (321 schools) in 1987 *N* = 4335 (241 schools) in 1992 *N* = 5314 (253 schools) in 1997 grade 6 regular education	math textbook used (8 different)	school years up to measurement	students' SES, gender, nr. of school years; school's SES; assessment cycle	general mathematics performance compared to *NZR* *Wereld in Getallen - ed. 2* *Nieuw Rekenen* *Pluspunt* *Rekenen & Wiskunde* *Operatoir Rekenen - ed. 1* *Wereld in Getallen - ed. 1* *Niveau Cursus Rekenen* *Naar Zelfstandig Rekenen*	+.53 +.31 +.29 +.25 +.23 +.22 +.02 ref = 0
Janssen, Van der Schoot, & Hemker (2005) = PPON grade 6 - 2004	general math	*N* = 4335 (241 schools) in 1992 *N* = 5314 (253 schools) in 1997 *N* = 3078 (122 schools) in 2004 grade 6, regular education	math textbook used (NB. 80% new textbooks in 2004)	school years up to measurement	students' SES, gender, nr. of school years; school's SES;	general math performance: summative effect of shift in textbooks 2004 vs. 1997 2004 vs. 1992 1997 vs. 1992	+.12 +.18 +.06

references	domain	participants	curriculum	duration	corrected	results	ES
Kraemer, Janssen, Van der Schoot, & Hemker (2005) = PPON grade 3 - 2003	general math	*N* = 3350 (164 schools) in 1992 *N* = 5972 (130 schools) in 1997 *N* = 2032 (77 schools) in 2004 grade 3 regular education	math textbook used (7 different)	school years up to measurement	students' origin, gender, nr. of school years; assessment cycle	general math performance compared to *R&W*	
						Talrijk	+.64
						Rekenrijk	+.62
						Wereld in Getallen - eds. 2+3	+.46
						Pluspunt - eds. 1+2	+.44
						Wereld in Getallen - ed. 1	+.18
						Rekenen & Wiskunde	ref = 0
						summative effect of shift in textbooks	
						2003 vs. 1997	+.18
Van Putten, Van den Brom-Snijders & Beishuizen (2005)	multidigit division	*N* = 256 10 schools grade 4 regular education	1. math textbook *Rekenen & Wiskunde* (R&W) 2. math textbook *Wereld in Getallen* (WiG)	school years up to measurement	speed-test performance	halfway grade 4	
						R&W < WiG	-.43
					speed-test performance	end grade 4	
						R&W > WiG	+.35

CHAPTER 2

Solution strategies and achievement in Dutch complex arithmetic: Latent variable modeling of change

This chapter has been published as Hickendorff, M., Heiser, W. J., Van Putten, C. M., & Verhelst, N. D. (2009). Solution strategies and achievement in Dutch complex arithmetic: Latent variable modeling of change. *Psychometrika, 74*, 331-350.

The research was supported by CITO, National Institute for Educational Measurement. For their efforts in coding the strategy use, we would like to thank Meindert Beishuizen, Gabriëlle Rademakers, and the Bachelor students from Educational and Child studies who participated in the research project into strategy use.

ABSTRACT

In the Netherlands, national assessments at the end of primary school (Grade 6) show a decline of achievement on problems of complex or written arithmetic over the last two decades. The present study aims at contributing to an explanation of the large achievement decrease on complex division, by investigating the strategies students used in solving the division problems in the two most recent assessments carried out in 1997 and in 2004. The students' strategies were classified into four categories. A data set resulted with two types of repeated observations within students: the nominal strategies and the dichotomous achievement scores (correct/incorrect) on the items administered.

It is argued that latent variable modeling methodology is appropriate to analyze these data. First, latent class analyses with year of assessment as a covariate were carried out on the multivariate nominal strategy variables. Results show a shift from application of the traditional long division algorithm in 1997, to stating an answer without writing down any notes or calculations in 2004, especially for boys. Second, explanatory IRT analyses showed that the three main strategies were significantly less accurate in 2004 then they were in 1997.

2.1 INTRODUCTION

2.1.1 National assessments of mathematics achievement

In the Netherlands, the level of mathematics achievement has changed over the last two decades. Large scale national assessments of mathematics education at the end of primary school by the National Institute for Educational Measurement (CITO) on four consecutive occasions (1987, 1992, 1997 and 2004) showed diverse trends (J. Janssen et al., 2005). On the one hand, achievement has increased strongly on numerical estimation and general number concepts, and has increased to a lesser extent on calculations with percentages and mental addition and subtraction. However, results show a steady and large decline of performance on complex (written) arithmetic. Specifically, students at the end of Grade 6 in 2004 performed less well than students at the end of Grade 6 did in 1987 on complex addition and subtraction, and especially on complex multiplication and division. In the period from 1987 to 2004, achievement in complex multiplication and division has declined with more than one standard deviation on the ability scale, with an accelerating trend (J. Janssen et al., 2005).

2.1.2 Mathematics education

Mathematics education has experienced a reform process of international scope over the last couple of decades (Kilpatrick, Swafford, & Findell, 2001). Although several countries differ in their implementation, there are common trends. These are globally described by a shift away from transmission of knowledge, toward investigation, construction, and discourse by students (Gravemeijer, 1997b).

In the Netherlands this reform movement is in effect by the name of Realistic Mathematics Education (RME) (Freudenthal, 1973; Gravemeijer, 1997b). The content of mathematics education has shifted from the product of mathematics to the process of doing mathematics (Gravemeijer, 1997b). Instruction is based on the key principle of guided reinvention (Freudenthal, 1973). This principle entails that teachers should give students the opportunity to reinvent the mathematics they have to learn for themselves, according to a mapped out learning route. The informal strategies of students are a possible starting point. Mathematics problems are often embedded in experientally real situations.

At present, Dutch primary schools have almost uniformly adopted mathematics textbooks based on the principles of RME (J. Janssen et al., 2005), although these books differ in their emphasis on prestructuring of students' solutions (Van Putten et al., 2005).

2.1.3 Complex division

In this paper, the focus is on complex or written division, for two reasons. First, the largest decline in performance is observed in this domain. This development is worrisome , since it is a core educational objective set by the Dutch government that students at the end of primary education *"can perform the operations addition, subtraction, multiplication, and division with standard procedures or variants thereof, and can apply these in simple situations"* (Dutch Ministry of Education, Culture, and Sciences, 1998, p. 26). This objective has not changed since its first publication in 1993, and it was still valid in the most recent publication of the educational objectives in 2005. A panel of several experts on mathematics education (such as experienced teachers and teachers' instructors) set up norm levels, to offer a frame of reference for evaluating to what extent these core objectives are reached by the educational system (Van der Schoot, 2008). If a majority (70-75%) of the students attains these norm levels, the core objectives are sufficiently reached, according to the expert panel. In 1997, only half of the students reached this level on

complex multiplication and division (J. Janssen et al., 1999), and in 2004 this dropped even further to only 12% of the students (J. Janssen et al., 2005). So, the objectives of primary education on complex division seem not to be reached by far, particularly not in 2004.

Second, with the introduction of RME in the Netherlands, complex division has served as a prototype of the alternative informal approach (Van Putten et al., 2005). So, that makes a further study into changes in this domain of mathematics education particularly interesting. This is especially true if the solution strategies that students applied are incorporated in the analysis. By including this information on the cognitive processes involved in solving these problems, we aim to give more insight in the decrease in achievement level.

Several studies investigated the informal strategies young children develop for division (Ambrose, Baek, & Carpenter, 2003; Mulligan & Mitchelmore, 1997; Neuman, 1999). Main strategies observed in these studies are counting, repeatedly adding or subtracting the divisor, making multiples of the divisor (so called *chunking*), decomposing or partitioning the dividend, and (reversed) multiplication.

In RME, the didactical approach to complex division starts from these informal strategies. Treffers (1987) introduced column arithmetic according to progressive schematization, resulting in a division procedure of repeated subtraction of multiples (chunks) of the divisor from the dividend, as shown in the right hand panel of Figure 2.1. This learning trajectory starts with dividing concretely (piece-by-piece or by larger groups), and is then increasingly schematized and abbreviated. In the final phase, the maximum number of tens and ones (and hundreds, thousands, and so forth, depending on the number size of the problem) is subtracted in each step. However, not all students need to reach this optimal level of abbreviation.

In contrast, in the traditional algorithm for long division (see left panel of Figure 2.1) it is necessary that each subtraction of a multiple of the divisor is optimal. Furthermore, the number values of the digits in the dividend are not important for applying the algorithm in a correct way.

Van Putten et al. (2005) studied this kind of written calculation methods for complex division at Grade 4, and designed a classification system to categorize the solution strategies. Different levels of abbreviation or efficiency of chunking of the divisor were distinguished. In addition, partitioning of dividend or divisor was observed. Chunking and partitioning strategies are based on informal strategies, and were therefore labeled

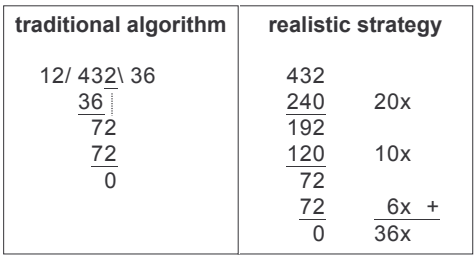


FIGURE 2.1 *Examples of the traditional long division algorithm and a realistic strategy of schematized repeated subtraction for the problem 432 ÷ 12.*

realistic strategies. Another strategy was the traditional long division algorithm. The final category involved students who did not write down any solution steps, and mental calculation was inferred as the strategy used for obtaining an answer to the problem.

2.1.4 Goals of present study

The present study has a substantive and a methodological aim. Substantive aim is to gain more insight in the worrisome large decrease of achievement in complex division. The analysis is extended beyond achievement, by including information on the strategies students used to solve the division problems of the national assessments. The first substantive research question is whether and how strategy use has changed over the two most recent assessments. The second research question is how strategy use can predict the probability of solving an item correctly and how these strategy accuracies relate to the observed decrease in achievement.

Methodological aim is to discuss analysis techniques that are appropriate for these kind of substantive research questions. One important characteristic of the data set is that it contains multivariate strategy and score information. Furthermore, to explain observed changes in a cross-sectional design one needs to establish a common frame of reference, for strategy use as well as for achievement. Together with some other properties of the data, these characteristics call for advanced psychometric modeling. Aim is to provide future research into strategy use and achievement with suitable modeling methodology, that can be implemented within flexible general software platforms.

2.2 METHOD

2.2.1 Sample

In the present study, parts of the material of the two most recent national assessments of CITO are analyzed in depth. These studies were carried out in May/June 1997 (J. Janssen et al., 1999) and in May/June 2004 (J. Janssen et al., 2005). For each assessment, a national sample was obtained of students at the end of their primary school (in the Netherlands Group 8, equivalent to Grade 6 in the US). These samples were representative for the total population in terms of social-economical status, and schools were spread representatively over the entire country. Each sample consisted of approximately as many girls as boys. Various mathematics textbooks were used, although the large majority (over 90% of the schools in 1997, and almost 100% of the schools in 2004) used textbooks based on RME principles.

A subset of the total sample was used in the present analysis: we included only students to whom items on complex division were administered. In 1997, that subset consisted of 574 students from 219 different primary schools. It consisted of 1044 students from 127 schools in 2004. So, the sample used in the present study contains 1618 students.

2.2.2 Design of the tests

Figure 2.2 displays the design of the tests of these two assessments. In the 1997 assessment, 10 different division problems were administered in a complete design (J. Janssen et al., 1999). In 2004, there were 13 division problems, but these were administered in an incomplete design: each student was presented a subset of 3 to 8 of these problems (J. Janssen et al., 2005). In total, there were 8 different subsets of item combinations, each administered to around 130 students. Four problems were included in both assessments (items 7 to 10). Consequently, linking of the of 1997 to the 2004 results was possible through these common items. The total number of items was 19: 6 items unique in 1997, 4 common items, and 9 items unique in 2004.

These items were constructed such that their difficulty levels had an even spread, from quite easy to quite hard. Most of these 19 items presented the division problem in a realistic situation. On most items, students had to deal with a remainder (i.e. the outcome was not a whole number). On those items, the answer either had to be calculated with

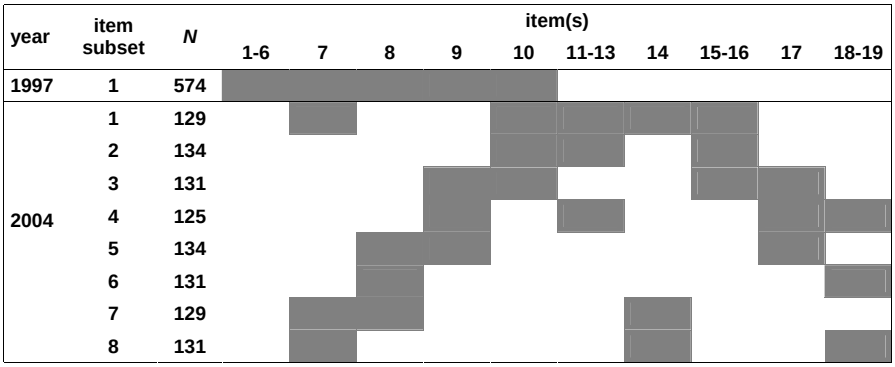


FIGURE 2.2 *Design of the assessments.*

the precision of two decimals, or (on 4 items) the answer had to be rounded to a whole number in a way that was appropriate given the situation presented in the problem.

Table 2.1 displays several specifications of the items: the numbers involved in the division problem, whether the problem was presented in a realistic context, what the correct answer was given the context, and the percentage correct answers in either 1997, 2004, or both. However, because CITO will use several of these items in upcoming assessments, not all items are released for publication. Therefore, Table 2.1 displays (in italics) parallel forms (with respect to size of dividend, divisor, and outcome) of the original items.

In the 2004 assessment, students were instructed as follows: *"In this arithmetic task you can use the space next to each item for calculating the answer. You won't be needing scrap paper apart from this space."* In addition, the experimenter from CITO explicitly instructed students once more that they could use the blank space in their booklets for making written calculations. In the 1997 assessment these instructions were not as explicit as in 2004. In 1997 as well as in 2004, on a single page several items were printed. For all items there was enough space left blank where the students could write down their calculations.

TABLE 2.1 *Specifications of the items.*

				% correct	
item nr.	division problem	context	answer*	1997	2004
1	*19 ÷ 25*	yes	0.76	18.3	-
2	*64800 ÷ 16*	yes	4050	55.2	-
3	*7040 ÷ 32*	no	220	60.3	-
4	*73 ÷ 9*	no	8.11	44.1	-
5	*936 ÷ 12*	yes	78	44.8	-
6	*22.8 ÷ 1.2*	no	19	42.2	-
7	872 ÷ 4	yes	218	75.4	54.8
8	1536 ÷ 16	yes	96	53.1	36.3
9	736 ÷ 32	yes	23	71.3	51.5
10	*9157 ÷ 14*	yes	654	44.3	29.2
11	*40.25 ÷ 7*	yes	5.75	-	43.0
12	139 ÷ 8	yes	17 R 3	-	59.3
13	*668 ÷ 25*	yes	27	-	52.8
14	6.40 ÷ 15	yes	0.43	-	12.6
15	*448 ÷ 32*	yes	14	-	51.3
16	157.50 ÷ 7.50	yes	21	-	60.4
17	*13592 ÷ 16*	yes	849.5	-	21.3
18	80 ÷ 2.75	yes	29	-	22.1
19	*18600 ÷ 320*	yes	59	-	24.4

*Answer that was scored correct, given the item context

2.2.3 Responses

Two types of responses were obtained for each division problem in these two tests. First, the answers given to the items were scored correct or incorrect. Skipped items were scored as incorrect. Second, by looking into the students' written work, the strategy used to solve each item was classified. We used a similar classification scheme as the one applied by Van Putten et al. (2005). Four main categories were distinguished. First, students solved division problems with a traditional long division algorithm. Second, realistic strategies (chunking and partitioning) were observed. Third, it occurred quite often that students did state an answer, but did not write down any calculations or notes (No Written Working). Finally, a category remained including unclear or erased strategies, wrong procedures such as multiplication instead of division, and skipped problems (Other strategies).

For parts of the material, the strategies were coded by two different raters, and Cohen's κ (Cohen, 1960) was computed to assess the interrater reliability. For the 1997 data, solution strategies of 100 students were coded by two raters, resulting in a value of Cohen's κ of .89. In 2004, solution strategies of 65 students were coded by two raters, resulting in a Cohen's κ of .83. So, in both assessments a satisfactory level of interrater reliability was attained.

In addition to the response variables, three student characteristics were available. First, gender of the student was recorded. Second, an index of parental background and educational level (PBE) was available, with 3 categories: students with at least one foreign (non Dutch) parent with a low level of education and/or occupation, students with Dutch parents who both have a low level of education and/or occupation, and all other students. Third, a rough indication of general mathematics level (GML) of the students was computed, based on performance of the students on all mathematics items (other than complex division) presented to them. In each assessment sample, the students were divided into three equally sized groups, labeled as weak, medium, and strong general mathematics level.

2.2.4 Properties of the data set

In discussing what psychometric modeling techniques are appropriate to obtain answers to the research questions, we have to take a further look into the specific properties of the present data set. Two aspects deserve attention. They are also illustrated in Table 2.2, presenting part of the data set.

First, because each student had several items administered, the different responses within each student are correlated. Analysis techniques should take this correlated data structure into account. In addition, each of these repeatedly observed responses is bivariate: the item was solved correct or incorrect (dichotomous score variable) and a specific strategy was used (nominal variable).

Second, both research questions involve a comparison of the results from 1997 and 2004. The incomplete design of the data set impedes these comparisons, because different students completed different subsets of items. Analysis on the item level would be justified, but does not take the multivariate aspect of the responses into account. In addition, univariate statistics would be based on different samples of students. Furthermore, analyses involving changes in performance would be limited to

TABLE 2.2 *Part of the data set.*

					item 7		item 8		item 19		...
student	year	gender	PBE	GML	Str	Sc	Str	Sc	Str	Sc	...
1	1997	b	1	weak	R	1	N	0	-	-	...
2	1997	b	2	strong	T	1	T	1	-	-	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
574	1997	g	1	medium	T	0	N	0	-	-	...
575	2004	g	3	weak	-	-	R	0	R	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
705	2004	b	3	medium	O	0	-	-	R	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
1618	2004	b	1	strong	-	-	R	1	-	-	...

Note 1. Str = strategy: T = Traditional, R = Realistic, N = No Written Working, O = Other
Note 2. Sc = score (1 = correct, 0 = incorrect)
Note 3. - = item not administered

the four common items and would therefore not take all information into account.

Therefore, we need analysis techniques that can take into account the multivariate aspect of the data, and are not hampered by the incomplete design. This aim can elegantly be attained by introducing a latent variable. Individual differences are modeled by mapping the correlated responses on the latent variable, while the student remains the unit of analysis.

Finally, it should be possible to include at least one predictor variable: year of assessment. For both research questions, we discuss appropriate techniques next.

2.2.5 Latent Class Analysis

The first research question is directed at changes in strategy use between the two assessments. So, the nominal strategy responses are the dependent variables. We argue that a categorical latent variable is best to model this multivariate strategy use, because differences between students are qualitative in this respect. Latent class analysis (LCA) accomplishes this goal, by introducing a latent class variable that accounts for the covariation between the observed strategy use variables (e.g. Goodman, 1974; Lazarsfeld

& Henry, 1968). The basic latent class model is:

$$f(\mathbf{y}|D) = \sum_{k=1}^{K} P(k) \prod_{i \in D} P(y_i|k). \tag{2.1}$$

Classes run from $k = 1, \ldots, K$, and $\mathbf{y}$ is a vector containing the nominal strategy codes on all items i that are part of the item set D presented to the student. Resulting parameters are the class probabilities or sizes $P(k)$ and the conditional probabilities $P(y_i|k)$. The latter reflect the probability of solving item i with each particular strategy, for each latent class. So, we search for subgroups (latent classes) of students that are characterized by a specific pattern of strategy use over the items presented.

Predictor effects

To assess differences in strategy use between the assessments of 1997 and 2004, year of assessment was introduced as a covariate in the LCAs. This entails that classes are formed conditional upon the level of the covariate, so that year of assessment predicts class membership (Vermunt & Magidson, 2002). The LC-model with one observed covariate z can be expressed as:

$$f(\mathbf{y}|D, z) = \sum_{k=1}^{K} P(k|z) \prod_{i \in D} P(y_i|k). \tag{2.2}$$

Class probabilities sum to 1, conditional on the level of the covariate, i.e. $\sum_{k=1}^{K} P(k|z) = 1$. Parameters estimated are the class probabilities conditional on year of assessment, and for each class, the probability of using each particular strategy on each item (the conditional probabilities).

To study how the other background variables were associated with strategy use, we carried out some further analyses. Inserting all these variables and their interactions as covariates in the latent class analysis would yield an overparameterized model. Therefore, all students were assigned to the latent class for which they had the highest posterior probability (modal assignment). Next, this latent class variable was analyzed as the response variable in a multinomial logit model (e.g., Vermunt, 1997). The associations of each of the explanatory variables with latent class are modeled conditional on the joint distribution of all explanatory variables. Cell entries f_{kz} of the 5-way frequency table, with k the value on the response variable latent class, and z the joint distribution of the

explanatory variables year of assessment, gender, parental background/education, and general math level, are modeled as

$$\log f_{kz} = \alpha_k + \sum_j \beta_j x_{jkz}. \tag{2.3}$$

The design matrix x_{jkz} specifies the j associations or effects in the model.

Software

Analyses were carried out in the program LEM (Vermunt, 1997), a general and versatile program for the analysis of categorical data. Input data for the latent class analyses consisted of the strategy used on each of the 19 items, and the level of the covariate year of assessment. The incompleteness of the design (Figure 2.2) yielded 9 different patterns of missing values (for the items that were not administered). Input data for the multinomial logit models were the values on each of the 4 explanatory variables and the latent class each student was assigned to.

2.2.6 Explanatory IRT

Research question 2 asks how strategy use can predict the probability of solving an item correctly, and how these strategy accuracies relate to the observed decrease in achievement. So, the repeatedly observed correct/incorrect scores are the dependent variables, and the nominal strategies take on the role of predictors. We argue that in these analyses, a continuous latent variable is appropriate. This latent variable models the individual differences in proficiency in complex division by explaining the correlations between the observed responses. Item Response Theory (IRT) modeling accomplishes this goal. Through the four common items, it was possible to fit one common scale for 1997 and 2004 of proficiency in complex division, based on all 19 items.

In the most simple IRT measurement model, the probability of a correct response of subject p on item i can be expressed as follows:

$$P(y_{pi} = 1 | \theta_p) = \frac{\exp(\theta_p + \beta_i)}{1 + \exp(\theta_p + \beta_i)}. \tag{2.4}$$

Latent variable θ expresses ability or proficiency, measured on a continuous scale. The item parameters β_i represent the easiness of each item.

Such descriptive or measurement IRT models can be extended with an explanatory part (Rijmen et al., 2003; Wilson & De Boeck, 2004). This implies that covariates or predictor variables are included, of which the effects on the latent scale are determined. These can be (a) item covariates, that vary across items but not across persons, (b) person covariates, that vary across persons but not across items, and (c) person-by-item or dynamic covariates, that vary across both persons and items. The latent regression model SAUL (Verhelst & Verstralen, 2002) is an example of a explanatory IRT model with person covariates.

The present data set includes person predictors and person-by-item predictors (the strategy used on each item). Person predictors are denoted Z_{pj} ($j = 1,\ldots,J$), and have regression parameters ζ_j. Person-by-item predictors are denoted W_{pih} ($i = 1,\ldots,I$ and $h = 1,\ldots,H$), and have regression parameters δ_{ih}. These explanatory parts enter the model in (2.4) as follows, with indices i for items, p for persons, h for strategy, and j for the person covariate used as predictor variable:

$$P(y_{pi} = 1 | Z_{p1} \ldots Z_{pJ}, W_{pi1} \ldots W_{piH}) =$$

$$\int \frac{\exp\left(\sum_{j=1}^{J} \zeta_j Z_{pj} + \sum_{h=1}^{H} \delta_{ih} W_{pih} + \beta_i + \epsilon_p\right)}{1 + \exp\left(\sum_{j=1}^{J} \zeta_j Z_{pj} + \sum_{h=1}^{H} \delta_{ih} W_{pih} + \beta_i + \epsilon_p\right)} g(\epsilon) d\epsilon. \tag{2.5}$$

It is assumed that all person specific error parameters ϵ_p come from the common density $g(\epsilon)$. Usually, it is assumed that $g(\epsilon)$ is a normal distribution, with mean fixed to 0 to get the scale identified, i.e. $\epsilon_p \sim N(0, \sigma_\epsilon^2)$.

Fitting the models

In the present data set, there are 2 binary person predictors (year of assessment and gender). Furthermore, there are 2 categorical person predictors with each 3 categories (parental background/education and general mathematics level). These can both be dummy-coded in 2 binary predictors, respectively. The strategy used on each item yields 19 categorical person-by-item predictors, each with 4 categories. However, the Other strategies are not of interest in the present analysis into strategy accuracies. These Other strategies are a small heterogeneous category of remainder solution strategies, consisting mainly of skipped items, which of course, result in incorrect answers. Therefore, we excluded item-student combinations solved with an Other strategy from the explanatory

TABLE 2.3 *Part of the data set in long matrix format.*

student	year	gender	PBE	GML	d7	d8	d19	...	Str	Sc
1	1997	b	1	weak	1	0	0	...	R	1
1	1997	b	1	weak	0	1	0	...	N	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮
2	1997	b	2	strong	1	0	0	...	T	1
2	1997	b	2	strong	0	1	0	...	T	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮	⋮
1618	2004	b	1	strong	0	1	0	...	R	1

IRT analyses. Dummy coding the remaining 3 strategies, taking the No Written Working strategy as reference category on each item, yielded a total of $19 \times (3-1) = 38$ binary strategy predictors. For each of these 38 strategy predictors a regression parameter is estimated. So, this model with strategy predictors specified for each item separately yields many parameters, which is an unpleasant property of the model, as discussed later.

Software

Model (2.5) is equivalent to a general linear mixed model, a GLMM (McCulloch & Searle, 2001). Advantage of formulating the model in the GLMM framework is that existing and newly formulated models can be estimated in general purpose statistical software. All explanatory IRT models in this study were estimated using Marginal Maximum Likelihood (MML) estimation procedures within the NLMIXED procedure from SAS (SAS Institute, 2002; see also De Boeck & Wilson, 2004; Rijmen et al., 2003; Sheu, Chen, Su, & Wang, 2005). We chose nonadaptive Gaussian quadrature for the numerical integration of the marginal likelihood, with 90 quadrature points, and Newton Raphson as the optimization method.

To use the NLMIXED procedure, the data have to be transposed into a long matrix, in which each row represents the response of one student to one item. Separate dummy variables (d1, d2, ..., d19) indicate which item is at stake. So, in the long data matrix, each student is replicated as many times as the number of items he or she was administered. Table 2.3 shows this transformation of part of Table 2.2.

TABLE 2.4 *Strategy use in proportions.*

	common items										all items	
	item 7		item 8		item 9		item 10		total		total	
	'97	'04	'97	'04	'97	'04	'97	'04	'97	'04	'97	'04
Traditional	.31	.08	.34	.11	.42	.19	.41	.19	.37	.14	.35	.13
Realistic	.22	.15	.21	.16	.24	.33	.22	.25	.22	.22	.21	.24
No Written Working	.41	.61	.26	.54	.22	.30	.17	.35	.26	.45	.26	.44
Other	.06	.16	.19	.19	.12	.19	.20	.21	.14	.19	.18	.19
# observations	574	386	574	392	574	388	574	392	2296	1558	5740	5704

2.3 RESULTS

2.3.1 Research Question 1

Table 2.4 displays proportions of use of the four main strategies, separately for the 1997 and the 2004 assessment. In the first 8 columns, strategy proportions are presented for the four common items. Next, these are totaled over these four items. The final two columns contain the strategy use totaled over all items presented in each assessment, so these proportions for 1997 and 2004 are based on different item collections. From Table 2.4, we see that the four common items were solved less often by the Traditional algorithm in 2004 than in 1997, but that the proportion of Realistic strategies did not change. Instead, it appears that stating an answer without writing down any calculations has increased in relative frequency. A similar pattern of strategy shifts is observed when all items are included.

Latent class models with year of assessment as covariate were fitted with 1 to 6 latent classes. Table 2.5 gives the log-likelihood (LL), Bayesian Information Criterion (BIC), and number of parameters ($\#p$) for each of these models. The BIC is a criterion that penalizes the fit (LL) of a model with the loss in parsimony. It is computed as -2LL+$\#p \cdot \ln(N)$, with N the sample size. Lower BIC-values indicate better models in terms of parsimony. From Table 2.5, the 4-class model has the best fit, according to the BIC. So, we choose to interpret the model with 4 classes.[1]

1 As Table 2.5 shows, the number of parameters increases rapidly when the number of latent classes increases. When estimating models with more than 150 parameters, LEM does not report standard errors of parameters. Moreover, for the 5 and 6-class models, several locally optimal solutions were found. Therefore, we have also estimated models with 1 to 6 classes, based only on the strategies used on the four common items. On this less complex problem, again the 4-class model has the best fit according to the BIC. The interpretation of this 4-class model is very similar to the one reported here.

TABLE 2.5 *Latent class models.*

classes	LL	BIC	# *p*
1	-15373.9	31279.8	72
2	-12798.8	26565.6	131
3	-11790.2	24984.3	190
4	-11385.7	24611.5	249
5	-11219.2	24714.2	308
6	-11106.3	24924.4	367

Figure 2.3 displays the probabilities of using each strategy on the 19 items, for each particular class. First note that each class-specific strategy profile is more or less dominated by one strategy type used all items. So, apparently students are quite consistent in their strategy use on a set of items. From these strategy profiles reflected in the conditional probabilities, we interpret the classes as follows. The first class is dominated by the Traditional algorithm, although this dominance is not uniform. Especially item 16 and to a lesser extent item 18 are exceptions, because these items are as likely or more likely to be answered without written working. However, we think the best way to summarize this latent class is to label it the Traditional class. The second class is characterized by a very high probability on all items to state the answer without writing down any calculations or solution steps (No Written Working class). The third class (Realistic class) is dominated by Realistic strategies, but again items 16 and 18 also have a substantial probability of No Written Working. Finally, the fourth class mainly consists of high probabilities of Other strategies, supplemented with answering without written working. In this Other class, Traditional and Realistic strategies have a very low usage probability on most of the items.

Effects of predictors on class membership

To qualify the effect of year of assessment, Table 2.6 shows the sizes of the classes, conditional on year of assessment. The Traditional class has become much smaller in 2004 than it was in 1997. In 1997, 43% of the students were using mainly the Traditional algorithm, but this percentage decreased to only 17% in 2004. The Realistic class did not increase accordingly. In 1997 as well as in 2004, little more than one quarter of the students could be characterized as a Realistic strategy user. Instead of an increase in

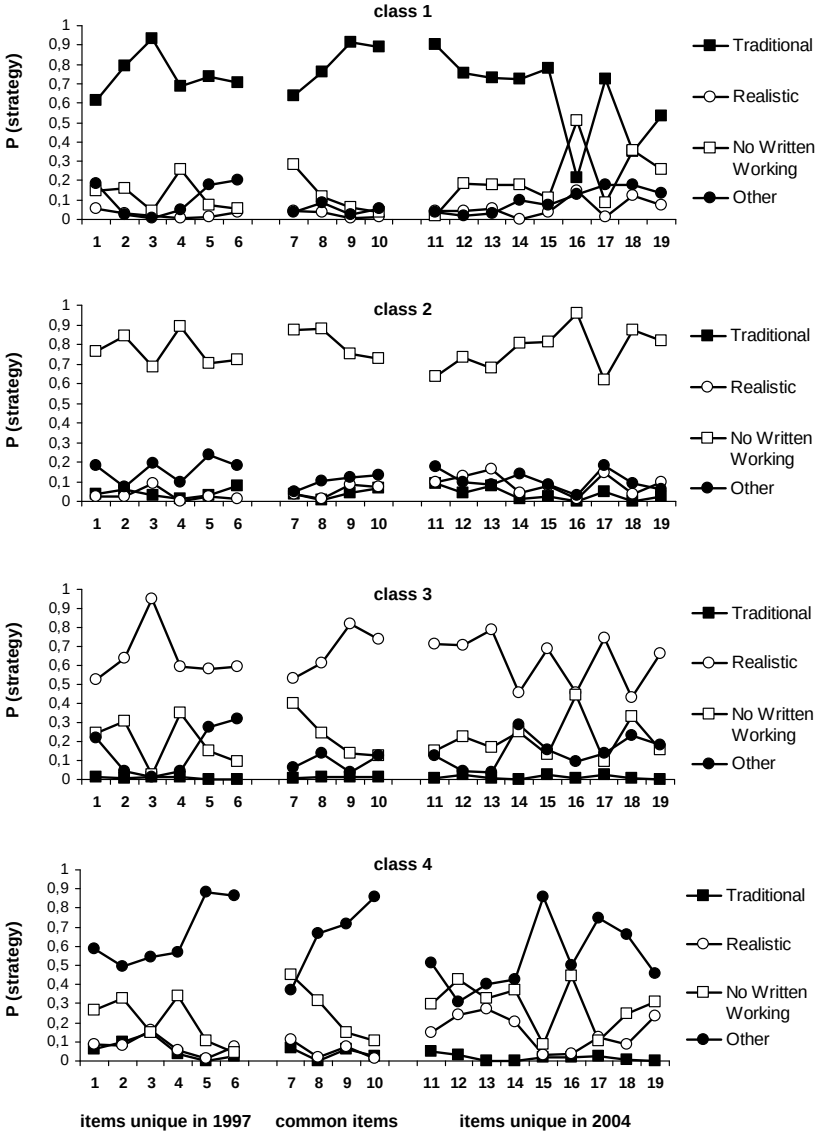


FIGURE 2.3 *Conditional probabilities of the 4-class LC-model.*

TABLE 2.6 *Class sizes in 1997 and 2004.*

	class			
year	1 (T)	2 (N)	3 (R)	4 (O)
1997	.43	.16	.27	.14
2004	.17	.36	.31	.16

the Realistic class, the No Written Working class has become larger in 2004 compared to 1997. In 1997, only 16% of the students could be classified as quite consistent in not writing down any calculations, while in 2004 this percentage increased to 36%. Finally, the remainder class of Other strategies did not change much between 1997 (14%) and 2004 (16%).

Further associations of the other background variables with latent class membership were studied by multinomial logit models. From these analyses, 59 students were excluded because they had one or more missing values on the background variables.

The model with effects of year of assessment (Year), gender, general math level (GML), and parental background/education (PBE) on class membership had a χ^2 value of 111.2, $df = 87$, $p = .04$. Removing any of these four predictor effects yielded a significant decrease in fit statistic, according to Likelihood Ratio tests (the difference between the deviances (-2LL) of two nested models is asymptotically χ^2-distributed, with df the difference between the number of parameters between the two models). So, each of the background variables had a significant relation with class membership. Next, we included interaction effects between the predictors, and LR-tests showed that only the interaction between Year and Gender had a significant effect on class membership (LR-test statistic = 8.3, $df = 3$, $p = .03$). This model had a χ^2 value of 100.7, $df = 84$, $p = .10$, indicating that this model adequately fitted the observed frequency table. Adding other interaction effects between predictors did not result in a significantly better model fit.

So, the final multinomial logit model indicated that GML and PBE each had an effect on class membership, and that Year and Gender interacted in their effect on class membership. Therefore, we present the relevant cross-tabulations in Table 2.7. [2]

2 Note that the marginal class proportions of 1997 and 2004 in Table 2.7 are slightly different from the conditional class probability parameters in Table 2.6. This difference is due to the modal assignment of students to latent classes prior to fitting the multinomial logit model, a procedure in which the uncertainty of this classification is not taken into account. In contrast, classification uncertainty does not play a role if the predictor variable Year is inserted as a covariate in the LCA.

TABLE 2.7 *Relevant proportions of Year, Gender, GML and PBE crossed with class membership.*

		class				
		1 (T)	2 (N)	3 (R)	4 (O)	*N*
1997	boy	.43	.20	.27	.10	261
	girl	.47	.13	.30	.11	290
2004	boy	.14	.49	.25	.12	499
	girl	.20	.26	.39	.15	509
weak		.15	.43	.18	.24	509
medium		.28	.25	.36	.11	529
strong		.37	.23	.37	.03	521
PBE 1		.28	.27	.33	.13	1077
PBE 2		.29	.31	.30	.10	287
PBE 3		.19	.46	.20	.16	195

The three-way cross-tabulation of Year, Gender and class membership shows that, apart for the effect of year of assessment described earlier, in 1997 the distribution over the four classes was about equal for boys and girls. However, in 2004, boys were more often than girls classified in the No Written Working class, and less often in the Realistic class. So, although boys and girls both shifted away from applying mainly the Traditional algorithm, for boys this was replaced by answering without writing anything down, while for girls this was replaced mostly with Realistic strategies.

The cross-tabulation of GML with class membership shows that students with a weak mathematics level were classified much more often in the No Written Working class, and less often in the Realistic class, than students with either a medium or a strong level of mathematics. Furthermore, class sizes for the Traditional class are positively related with mathematics level, and class sizes for the remainder class of Other strategies decreased with increasing mathematics level.

Finally, the cross-tabulation of PBE with class membership shows that compared to students with Dutch parents, either with low education/occupation (PBE 2) or not (PBE 1), students from the third group (PBE 3) who have at least one foreign parent with low education/occupation are classified more in the No Written Working class and less in the Realistic and Traditional classes.

TABLE 2.8 *Explanatory IRT models.*

model	predictor effects	LL	BIC	# *p*	LR-test stat	*df*
M0	-	-5003.0	10152.8	20		
M1	Year	-4963.4	10081.0	21		
M2	(M1) + Strat (item-specific)	-4592.5	9618.1	59		
M3	(M1) + Strat (restricted)	-4640.5	9449.8	23	96.0**	36
M4	(M3) + Strat x Year	-4636.2	9455.9	25	8.6*	2
M5	(M4) + Gender + PBE + GML	-4307.6	8835.4	30		
M6	(M5) + Strat x GML	-4294.9	8839.4	34	25.4**	4
M7	(M6) + Year x GML	-4294.4	8853.1	36	1.0	2
M8	(M6) + Year x Gender	-4293.8	8844.5	35	2.2	1
M9	(M6) + Strat x Gender	-4292.4	8849.1	36	5.0	2
M10	(M6) + GML x Gender	-4294.7	8853.7	36	.4	2

Note. LR-tests involve comparison to Models between brackets in column *predictor effects.*
* LR-test significant with $.01 \leq p < .05$; ** LR-test significant with $p < .01$.

2.3.2 Research Question 2

Starting from the measurement model without explanatory variables, we fitted a series of models by successively adding predictor variables. From all these analyses, 59 students were excluded because they had one or more missing values on the background variables. Furthermore, as discussed earlier, all 1778 observations (student by item combinations) involving Other strategies were excluded. In total, 1542 students yielding 8868 observations were included in the analyses. Model fit statistics are presented in Table 2.8.

First, the null model without any predictor effects (as in equation (2.4)) was fitted (M0), assuming that the θ_p come from one normal distribution. Therefore, 20 parameters are estimated: 19 item parameters β_i and the variance of θ_p. The mean of the distribution of θ_p was fixed at 0 for identification purposes. Next, the effect of year assessment was estimated (M1), which resulted in a substantial decrease in BIC.

Strategy effects

Next, type of strategy used on an item was inserted as a predictor of the probability of solving an item correct. First, in model M2, effects of dummy coded strategies were estimated for each item separately. The large decrease in BIC-value from model M2 compared to model M1 indicated that strategy use is an important explanatory variable. Figure 2.4 shows these strategy effects. In the upper panel we can see the direction of the effects of the different strategies within each item. However, the different easiness levels of the items make it hard to compare these strategy effects over the items. Therefore, the lower panel displays the strategy effects relative to the item-specific effects β_i of model M2.

On all items, the Traditional algorithm as well as the Realistic strategies had a consistent positive effect on success probability, compared to answering the item without writing down any calculations. The effect of using the Traditional algorithm compared to using a Realistic strategy was not unidirectional. On the 1997-items, applying the Traditional algorithm was more successful than using Realistic strategies. However, on the 2004-items, this differed per item, and on some items the Realistic strategies were more successful than the Traditional algorithm. This difference suggests an interaction effect of year of assessment and strategy use.

Estimating the effects of strategy use on success probability for each item separately results in many parameters, making standard errors large and interpretation cumbersome. Furthermore, if we want to estimate interaction effects of background variables such as year of assessment with strategy use, the number of parameters proliferates fast and interpretation gets even more difficult. Therefore, in model M3 the effects of the strategy used were restricted to be equal for all items ($\delta_{ih} = \delta_h$ for all items $i = 1, \ldots, 19$). Because most item-specific strategy effects were in the same direction for all items, we argue it is also a substantively sensible procedure.

These restrictions yielded a much more parsimonious model with only 23 parameters instead of 59. Model M3 is nested within model M2, so a Likelihood Ratio (LR) testing procedure could be applied. Relevant LR-test statistics are presented in Table 2.8. Although the result of the LR-test between model M3 and M2 indicated a significant decrease in model fit, the lower BIC-value of model M3 compared to model M2 (Table 2.8) indicated a much better trade-off between model fit and parsimony. Therefore, the model with restricted strategy effects was taken as the base model to which other effects were

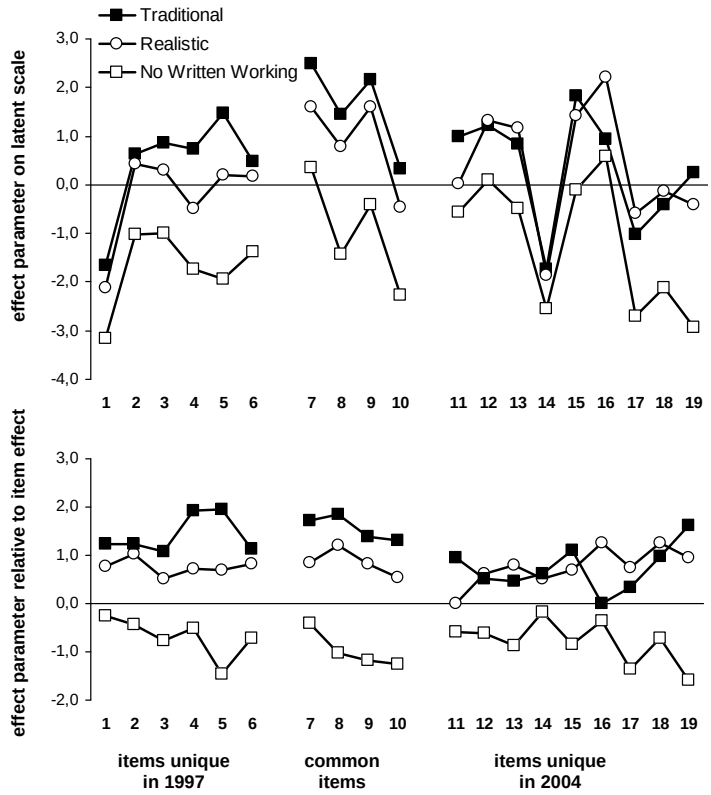


FIGURE 2.4 *Item-specific effect parameters of each strategy, from model M2.*

added.

First, we expected a different effect of the strategy used for the 1997 assessment and for the 2004 assessment, as already suggested by the item-specific strategy effects. Therefore, we estimated the interaction effect of (restricted) strategy use and year of assessment in model M4. The LR-test comparing model M4 and M3 was significant, so the strategy effects changed differently between 1997 and 2004.

Background variables

Next, in model M5 the background variables gender, parental background/education (PBE) and general mathematics level of the student (GML) were included. This again resulted in a large drop in BIC-value. The effects of mathematics level were very large: the effect of medium compared to weak students was 1.20 (SE = .10) and the effect of strong compared to weak students was 2.51 (SE = .10). The effects of the levels of PBE were also significant. Compared to students with Dutch parents with a certain level of education/occupation, students with Dutch parents from low education/occupation performed less well (effect is –.20, SE = .10), and also having at least one foreign parent with low education/occupation had a negative effect on performance (–.29, SE = .12). The effect of gender was not significant (girls compared to boys .12 (SE = .08)). This finding is important, because on most domains of mathematics boys outperform girls at the end of primary school in the Netherlands (Janssen et al., 2005).

In the final model building steps, several interaction effects were added. First, the interaction between strategy use and general mathematics level in model M6 yielded a significant improvement of model fit compared to model M5. Adding other two-way interaction effects of year of assessment with general mathematics level (model M7) or with gender of the student (model M8), did not improve model fit significantly. Other interaction effects of gender with strategy use (model M9) or with general mathematics level (model M10) also could not improve the fit significantly. So, according to the Likelihood Ratio tests, model M6 has the best fit. However, the BIC-value for model M6 is not the lowest of all models, but we argue that the slight difference in BIC-values does not countervail against the significant LR-tests.

Interpretation of the selected model

Figure 2.5 graphically displays the interaction effects of strategy with year of assessment, and of strategy with general mathematics level. In the following, all effects reported are significant at the .05-level, as assessed with a Wald test.

The left-hand panel reveals that in both assessments, Realistic strategies and the Traditional algorithm were significantly more accurate than stating an answer without written working. The effects of Realistic strategies and the Traditional algorithm did not differ significantly from each other in either 1997 or in 2004. Furthermore, changes in the strategy accuracies from 1997 to 2004 are present. The three main strategies were

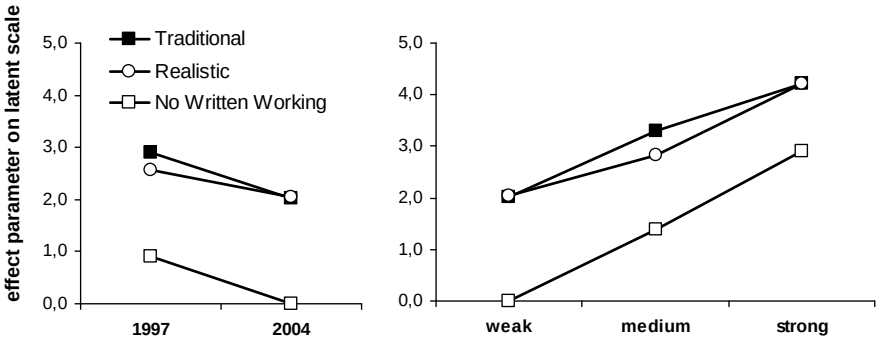


FIGURE 2.5 *Interaction effects of strategy use with year of assessment (left panel) and with general mathematics level (right panel) from model M3b.*

less accurate in 2004 than they were in 1997: the Traditional Algorithm (difference = -.88, SE = .17), stating an answer without written working (difference = –.90, SE = .11), and applying Realistic strategies (difference = –.53, SE = .13). Moreover, a differential effect is present. The decline in accuracy from 1997 to 2004 is significantly less for the Realistic strategies (–.53) compared to the decline of No Written Working (–.90).

The right-hand panel in Figure 2.5 shows that the strong students are more accurate in using all different strategies than the medium students. These medium students are in turn more accurate than the weak students in using all strategies. The interaction effect comprises first that there is a larger variation in the accuracy of the strategies for the weak and medium students, than for the strong students. Second, weak and strong students have as much success with the Traditional algorithm as with the Realistic strategies. In contrast, medium students perform better with the Traditional algorithm than with Realistic strategies (difference = .48, SE = .18).

Conclusions Research Question 2

All three main strategies have become less accurate in 2004 than in 1997, but Realistic strategies show the least decline. Realistic strategies have reached the same level of accuracy as the Traditional algorithm in 2004, but that level is still lower than it was in 1997. Both Realistic strategies and the Traditional algorithm are much more accurate than stating an answer without writing down any notes or calculations. Furthermore,

the general mathematics level of the students also plays an important role. Weak and medium students benefit more from writing down their solution strategy than strong students. Strong students do quite well without writing down their workings; they are even more accurate when they do not write down calculations than the weak students are when they apply either a Realistic or a Traditional strategy. Students with a medium mathematics level perform less well using a Realistic strategy than when applying the Traditional algorithm.

2.4 Discussion

Our study started from the observation that achievement on complex arithmetic (especially on complex multiplication and division) decreased considerably between 1987 and 2004 in the Netherlands. We believe the extent of this development is worrisome, because it is an educational objective that students at the end of primary school are able to solve these complex mathematics problems. This objective was far from reached on complex division: not in 1992 or 1997, but even less so in 2004. Therefore, our goal was to get more insight into the achievement drop on complex division. We searched for changes in strategy use and strategy accuracy between the two most recent national assessments.

First, strategy use has changed. With latent class analyses, multivariate strategy use was characterized. Changes in strategy use between the two assessments could be quantified by including a covariate in the analysis. As could be expected from the implementation of RME in Dutch classrooms and mathematics textbooks, the percentage of students that mostly apply the Traditional algorithm for long division has dropped considerably. However, the amount of students applying mostly Realistic strategies did not increase accordingly. Instead, more and more students did not write down any calculations or solution steps in solving the problems. Furthermore, a multinomial logit model showed that this shift toward No Written Working could mainly be contributed to the boys, and much less so to the girls.

Second, the accuracy of each particular strategy changed, as was assessed in explanatory IRT-analyses. The strategy used to solve an item fitted well in this flexible framework for including predictor effects. Equality restrictions of the strategy effects over the items made the model much more parsimonious and easy to interpret, and interaction effects with strategy use could be assessed without the need for many more

parameters. Results showed that stating an answer without showing any written working was much less accurate than either using the Traditional algorithm, or using some form of a Realistic strategy. So, the observed strategy shift seems rather unfortunate. Moreover, students in 2004 were less proficient in using all three main strategies (Traditional algorithm, Realistic strategies, and No Written Working) than they were in 1997.

So, not only did strategy use shift to less accurate strategies, also each of the three main strategies turned out to be less accurate. These two changes together seem to have contributed to the considerable decrease in achievement.

2.4.1 Limitations

This study comprised additional analyses on material that was collected for national assessment purposes. Therefore, the data were not collected with the present research questions in mind, resulting in several methodological limitations.

First, a large drawback of the present analysis of strategy use is that we do not know how students who did not write down anything in solving these problems, reached their answer. Did they solve the problem in their head by mental calculation, did they give an estimation, or did they perhaps just guess?

A second limitation is that the characteristics of the different strategies, such as the accuracies, may be biased by selection effects: selection by students, and selection by items (Siegler & Lemaire, 1997). For example, we found that mainly weak students answered without notations, which could have affected the accuracy of answering without written working negatively. Furthermore, it may seem that performance of those weak students who answer without written working would increase if they applied either the Traditional algorithm or Realistic strategies, since these are more accurate strategies. However, these strategy accuracies are based on different students who selected them, and it is unknown what these accuracies would be for students who did not select these strategies. A way to to obtain unbiased strategy characteristics would be to use the *Choice/No-Choice* methodology, proposed by Siegler and Lemaire (1997). Students then would have to answer a set of items in two different types of conditions. In the first condition type, students are free to choose what strategy they use (such as in the assessments under consideration). In the second condition type, students are obliged to use a particular strategy.

Third, it was not possible to take item characteristics into account as predictors of

strategy use or item difficulty. In large scale assessment programs such as the one currently studied, it is not common to systematically vary item characteristics. In the present item set, characteristics such as size of the numbers involved, whether the problem was presented in a context or not, and whether the problem involved a remainder or not, were confounded. So post-hoc analyses would involve contaminated effects.

A final limitation is that there were only four items common in the 1997 and the 2004 assessment. So, linking of the results of the two different assessments was only based on those four items. However, we believe that those items are representative problems for the domain of complex division, so that they are suitable link items.

2.4.2 Methodological considerations

Methodologically, we started with a complex data set, containing correlated nominal strategy variables, accompanied by correlated dichotomous score variables. We were interested in comparisons between two different samples of students, that were administered a partly overlapping item set. We argue that latent variable models are very appropriate for these kind of research questions about changes in strategy use and achievement. Specifically, latent class analyses and explanatory IRT model building both resulted in interpretable results and clear conclusions. Furthermore, we have shown that these models can be implemented in flexible software platforms, giving future researchers the possibility to build latent variable models according to their specific needs.

With respect to the explanatory IRT models fitted, several decisions were made. First, the measurement part of the IRT model used assumed a common slope for all items (the Rasch model). As an alternative, we also used a less restrictive IRT model in which for each item also a discrimination parameter was estimated. This analysis yielded very similar estimates of the effects of interest.

Second, the measurement part and the explanatory part of the IRT models were fitted simultaneously. An advantage of such a simultaneous approach is that measurement error of the estimated item parameters is taken into account when predictor effects are estimated. A potential disadvantage of this approach is that item parameter estimates may be affected by the inclusion of predictors. Moreover, it is not possible to establish the fit of the measurement model and assess the importance of the predictors separately.

For a more detailed discussion of disadvantages of the simultaneous approach, see Verhelst and Verstralen (2002). Therefore, as an alternative we also applied a sequential approach. In the first step, the measurement model was estimated. In the second step, this measurement scale was fixed, and effects of explanatory variables were estimated with the item parameters inserted as known constants. Again, very similar parameter estimates were found as in the analyses presented.

Finally, in fitting the item parameters of the measurement model, we used Marginal Maximum Likelihood (MML) estimation. In MML formulation, it is assumed that person parameters θ_p or ϵ_p arised from a normal distribution. MML estimation is therefore population-specific. As an alternative estimation procedure we also used Conditional Maximum Likelihood (CML) estimation, in which the model is fitted without making assumptions on the distribution of the latent scale in the population (Verhelst & Glas, 1995). Again, very similar results were obtained. A disadvantage of CML estimation is that it is not possible to estimate the easiness parameters and discrimination parameters jointly with CML, if one is interested in a 2-parameter IRT model. It is also not possible to estimate the effects of the explanatory variables with CML, so one needs to do this in a second step.

In conclusion, several alternative approaches to the presented explanatory IRT analyses were tested: incorporating item discrimination parameters, using a sequential approach for fitting the measurement part and explanatory part of the model, and using CML estimation for the measurement part of the model. All alternative approaches resulted in the currently presented model (M6) as the best fitting model, and the interpretation of the parameter estimates was very similar. Therefore, we presented the results of the most simple model, and we believe that these results are robust against potential model misspecifications.

2.4.3 Educational implications

The present findings of changes in strategy use and strategy accuracy may have several educational implications. A first issue is the relative accuracies of Realistic strategies and the Traditional algorithm, since the latter strategy is disappearing. Realistic strategies were as accurate as the Traditional algorithm, and also decreased the least in that accuracy. So, from these figures it seems that replacing the Traditional algorithm with Realistic strategies is not a bad development with respect to accuracy, but it only holds if

students apply those strategies in a structured way, by writing down their solution steps.

A second educational issue is also related to the the gradual disappearance of the Traditional algorithm for long division. The decrease in the use of the Traditional algorithm did not occur parallel with the introduction of mathematics textbooks adhering to the RME principles. In 1997 as well as in 2004, almost all schools used textbooks that do not cover the Traditional algorithm for division. However, we see that a substantial number of students still used that algorithm in 1997, and even in 2004 (albeit much less students). So, this may call the implementation of RME into question: it seems that teachers do not always follow the instructional design from their textbooks. This possibility is supported by results from a questionnaire for teachers in the assessment of 2004 (J. Janssen et al., 2005), in which 41% of the teachers reported that they still instructed the Traditional algorithm, either as the preferred strategy, or in combination with Realistic strategies.

Finally, there seems to be a trend that students (especially boys and students with a weak mathematics level) do not find it necessary to write down solution steps or calculations, or that these students are less able to do so. However, based on our current findings we believe the decreasing use of pen and paper in solving problems on complex arithmetic is unfortunate, because answering without written working turns out to be the least accurate strategy, especially for the weak and medium students. We find it worrisome that students do not seem to recognize that writing down solution steps helps them in recording key items and in schematizing information (Ruthven, 1998). It remains an open question what brought about this trend, and whether the value of writing down notes or calculation should obtain more emphasis in primary education.

CHAPTER 3

Complex multiplication and division in Dutch educational assessments: What can solution strategies tell us?

This chapter has been submitted for publication as Hickendorff, M. & Van Putten, C. M. *Complex multiplication and division in Dutch educational assessments: What can solution strategies tell us?*

The research was supported by CITO, National Institute for Educational Measurement. We would like to thank all Psychology undergraduate students who participated in the coding of strategy use.

ABSTRACT

The aim of the current study was to get more insight in sixth graders' performance level in multidigit multiplication and division that was found to be decreasing over time and lagging behind educational standards in large-scale national assessments in the Netherlands, where primary school mathematics education is characterized by reform-based learning/teaching trajectories. In secondary analyses of these assessment data, we extended the focus from achievement to aspects of strategic competence, by taking solution strategies that students used into account. In the first part of this paper, the negative performance trend between the 1997 and 2004 assessment cycles in multiplication problem solving was examined, by analyzing changes in strategy choice, overall differences in accuracy between strategies, and changes in these strategy accuracies. Findings showed that two changes contributed to the performance decline: a shift in students' typical strategy choice from a more accurate strategy (the traditional algorithm) to less accurate ones (non-traditional partitioning strategies and answering without written work, the increase in the latter strategy mainly observed in boys), as well as a general decline of accuracy rate within each strategy. In the second part, the influence of instruction on students' strategy choice in multiplication and division problems was analyzed. Findings showed that the teacher's instructional approach affected students' strategy choice, most profoundly in division problem solving.

3.1 INTRODUCTION

National and international large-scale educational assessments aim to report on the outcomes of the educational system in various content domains such as reading, writing, science, and mathematics. To evaluate the learning outcomes, a reference framework is needed. This can be either a comparison between countries as is done in the international comparative assessments (e.g., TIMSS, PIRLS, PISA), a comparison to the educational standards or attainment targets that are set within a country, or a comparison to performance levels from previous assessment cycles to find a trend over time.

The reports of educational assessments are usually limited to descriptive and correlational data on students' achievement, and therefore explanations for found differences or trends require further study. In such further studies, insights from educational psychology are essential to give direction to the exploration of potential explanatory mechanisms. In the current study, the focus is on one candidate mechanism: solution strategy use. The main research question is to what extent (change in) students'

strategy choice explains (change in) their achievement, and in turn, to what extent instructional approach influences students' strategy choice, in the domain of complex or multidigit multiplication and division. We tried to answer this question by carrying out secondary analyses on data of the two most recent Dutch national assessments at the end of primary school (1997 and 2004 cycles), extending the focus on achievement to aspects of strategic competence (e.g. Lemaire & Siegler, 1995) by studying solution strategies that students used. The aim of the current study was to get more insight in the performance level of Dutch sixth graders in complex multiplication and division, that was found to be decreasing over time and lagging behind educational standards.

3.1.1 Dutch results of educational assessments of mathematics achievement

Recent national and international assessments showed a varying pattern of results regarding mathematics performance in primary schools in the Netherlands. On the positive side, national results of the most recent cycle of PPON (Dutch assessment of mathematics education at the end of primary school, i.e., 12-year-olds) in 2004 showed improvements over time on some mathematics competencies, in particular on numerical estimation and on number sense (J. Janssen et al., 2005; Van der Schoot, 2008; see also Figure 3.1). Moreover, TIMSS-2007 (Meelissen & Drent, 2008; Mullis et al., 2008) results showed that Dutch fourth graders performed at the top level internationally, and also in PISA-2009 (OECD, 2010) Dutch 15-year-olds' mathematics performance took in an international top position. On the downside, however, there are also some results that are less positive. Both TIMSS and PISA reported a negative ability trend over time in the Netherlands. In addition, national assessments showed that on some mathematics domains performance decreased substantially since the first assessment in 1987 (see Figure 3.1). Furthermore, in many mathematics domains the educational standards were not reached (Van der Schoot, 2008).

Particularly, performance in *complex operations* – i.e., addition, subtraction, multiplication, division, and combined operations with multidigit numbers on which paper and pencil may be used – is worrisome. Not only did performance decrease most severely on these domains, with an accelerating trend (Figure 3.1), but also the percentage of students who reached the educational standards was lowest. A group of experts operationalized the educational standards and defined a 'sufficient' level of performance per domain that had to be reached by 70-75% of all students. In PPON 2004, this level was reached by

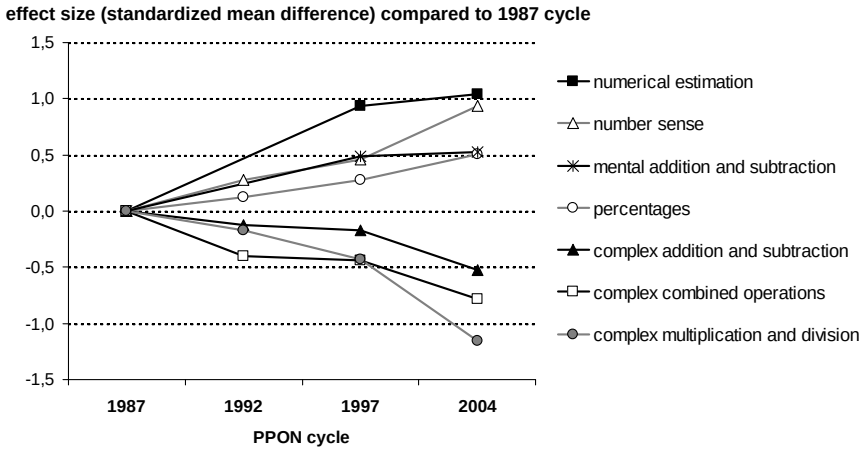


FIGURE 3.1 *Largest trends over time from Dutch national assessments (PPONs) of mathematics education at the end of primary school (Van der Schoot, 2008, p. 22), in effect sizes (standardized mean difference) with 1987 as baseline level. Effects statistically corrected for students' gender, number of school years, and socio-economical background, socio-economical composition of school, and mathematics textbook used.*

27% of the students in addition and subtraction, by 12% in multiplication and division, and by 16% in problems involving combining operations. On these domains, learning outcomes thus lagged far behind the goals.

The aim of the current study was therefore to gain more insight in students' lagging and decreasing performance level in the domain of complex multiplication and division. Our main approach was to extend the focus on achievement by including aspects of *strategic competence* (Lemaire & Siegler, 1995). We focused on complex multiplication and division for several reasons. First, as discussed above, performance decreased most severely on these operations, and it stayed furthest away from the educational standards. Second, compared to addition and subtraction, multiplication and division have received far less research attention, and especially multidigit multiplication and division are understudied research topics. Finally, instruction in how to solve multidigit operations has changed under influence of mathematics education reform, in particular on complex division, where the traditional algorithm for long division has completely

disappeared from mathematics textbooks and the learning/teaching trajectory (Van den Heuvel-Panhuizen, 2008).

3.1.2 Solution strategies

It has been well-established that children know and use multiple strategies in mathematics, and these strategies have different characteristics such as accuracy and speed (e.g., Beishuizen, 1993; Blöte et al., 2001; Lemaire & Siegler, 1995; Torbeyns, Verschaffel, & Ghesquière, 2004b, 2006; Van Putten et al., 2005). Therefore, solution strategy use may be an important predictor of achievement, and thereby also a potential mediator between (change in) instruction and (change in) achievement.

Mathematics education and instruction in primary school have undergone a large reform of international scope (e.g., Kilpatrick et al., 2001). In the Netherlands, the reform movement goes by the name of realistic mathematics education (RME), and it has become the dominant didactical theory in mathematics education practice. In the 1997 assessment, over 90% of the schools used a mathematics textbook that was based on the RME principles (J. Janssen et al., 1999); in the 2004 assessment this increased to nearly 100% (J. Janssen et al., 2005).

Solution strategies play an important role in this reform in at least two ways. First, the learning/teaching trajectory for solving complex arithmetic problems has changed, from top-down instruction of standard written algorithms to building on children's informal or naive strategies that are progressively formalized (Freudenthal, 1973; Treffers, 1987, 1993; Van den Heuvel-Panhuizen, 2008), a process in which mental arithmetic has become very important (Blöte et al., 2001). Second, the 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, Ghesquière, & Verschaffel, 2009b). The question is to what extent the instructional changes in complex arithmetic affected strategy use, and consequently, achievement.

Hickendorff, Heiser, Van Putten, and Verhelst (2009b) investigated the role of solution strategies in explaining the performance decrease in complex division problems observed in the Dutch national assessments. They carried out secondary analyses on the assessment material of 1997 and 2004 by coding the solution strategies that students used to solve the division problems (based on their written work). Findings showed shifts

between the two assessment cycles in strategy choice as well as in strategy accuracy, both contributing to the explanation of the performance decrease. The use of the accurate traditional long division algorithm decreased at the cost of an increase in problems that were answered without any written work (most likely mental calculation), a strategy that was much less accurate. Moreover, each of the main strategies led to fewer correct answers (i.e., was less accurate) in 2004 than it was in 1997.

In a follow-up study, Hickendorff, Van Putten, Verhelst, and Heiser (2010) analyzed the most relevant strategy split – mental versus written computation – more rigorously by collecting new data according to a partial *choice/no-choice* design (Siegler & Lemaire, 1997). Findings showed that for students who spontaneously chose a mental computation strategy to solve a complex division problem, the probability of a correct answer increased on average by 16 percent points on a parallel problem on which they were forced to write down their working. This suggested that the choice for a mental strategy on these problems was not optimal or adaptive with respect to accuracy, contrasting with the prediction in cognitive models of strategy choice that individuals choose their solution strategy adaptively (e.g., Shrager & Siegler, 1998; Siegler & Shipley, 1995). Moreover, the findings had clear implications for educational practice: encouraging students to write down their solution steps in solving complex division problems would probably improve performance.

These two studies illustrate the mutual value of bringing together the field of large-scale educational assessments and the field of educational and cognitive psychology. In the current study, Hickendorff et al.'s (2009b) analyses of strategy use on the complex division problems in the Dutch assessments are extended in two important ways. First, the domain of study is broadened to complex multiplication. Second, information on the instructional approach the teachers applied (that was, unfortunately, only available in the 2004 assessment) was used as a predictor of strategy use. Below, we discuss these two topics in more detail.

3.1.3 Complex or multidigit multiplication strategies and instruction

The majority of studies that focus on multiplication strategies in children and adults considered simple multiplication under 100, i.e., multiplying two single-digit numbers (Anghileri, 1989; Imbo & Vandierendonck, 2007; Lemaire & Siegler, 1995; Mabbott & Bisanz, 2003; Mulligan & Mitchelmore, 1997; Sherin & Fuson, 2005; Siegler, 1988b). The

following solution strategies were identified for solving simple multiplication problems like 3×4: counting procedures (unitary counting, $1,2,3,4,\ldots,5,6,7,8,\ldots,9,10,11,12$, as well as using a counting string, $4,8,12$), repeated addition (adding an operand the appropriate number of times, $4+4+4=12$), transforming the problem (referring to related operations or related facts, $2\times4=8$, $8+4=12$), and retrieval (knowing the answer by heart). With increasing age and experience, retrieval becomes the dominant strategy for simple multiplication.

In contrast, in multidigit or complex multiplication problems retrieval is not a feasible strategy, and computational strategies are required to derive the answer. Ambrose et al. (2003) analyzed the development of multidigit multiplication strategies and described three classes of strategies: concrete modeling strategies (which the authors note to be of limited use when two multidigit numbers have to be multiplied), adding and doubling strategies (including repeated addition), and partitioning strategies using tenfolds of one or both of the operands (see also Figure 3.2). Note that combinations of these classes of strategies are also possible (as was also described by Sherin and Fuson (2005), who called this hybrid strategies). For example, in Figure 3.2, the strategy in which one of the operands is decimally split also includes the additive strategy of doubling.

The RME learning-teaching trajectory in multidigit multiplication has its roots in the aforementioned developmental pattern, and can be characterized by progressive schematization and abbreviation of the informal solution strategies (Treffers, 1987; Van den Heuvel-Panhuizen, 2008). Buijs (2008) analyzed the recent RME-based textbooks, and found a common learning trajectory that starts from the repeated addition strategy, that is abbreviated by grouping, eventually with groups of ten times one of the operands. This leads to splitting or partitioning strategies in which one of the operands is decimally split. Partitioning strategies in which the solution steps are written down systematically in a more or less fixed order (which Buijs labeled 'stylized mental strategies', also called 'column multiplication' in the RME literature Van den Heuvel-Panhuizen, 2008) are suitable as transition phase toward the standard written algorithm for multiplication: it works with whole numbers instead of single-digits (like informal strategies), but it proceeds in a more or less standard way (like the traditional algorithm).

For multiplication, the end point of the RME-based learning trajectory still is the traditional algorithm in which calculation proceeds by multiplying single digits in a fixed order, from small to large (see Figure 3.2), although it does not have to be attained by all students; 'column multiplication' is considered a full alternative. In contrast, in the RME-

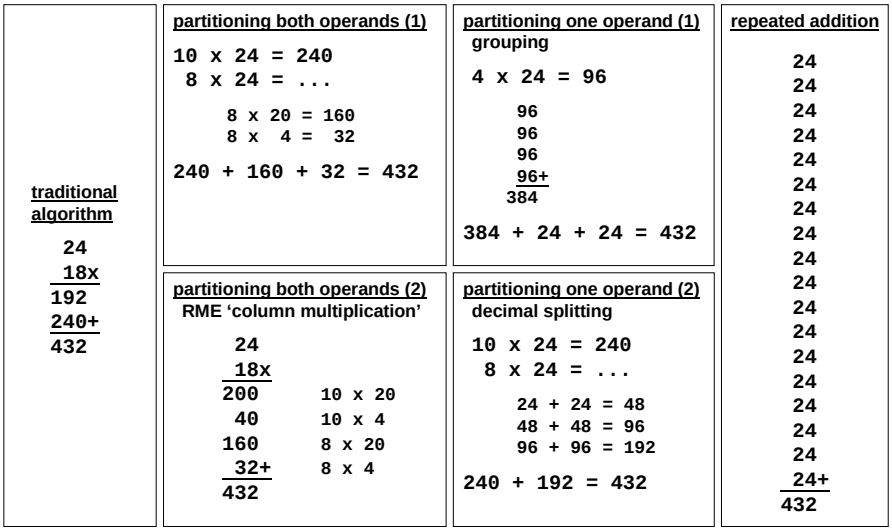


FIGURE 3.2 *Example strategies for multidigit multiplication for the problem* 18×24.

based learning trajectory for multidigit division, the traditional long division algorithm has completely disappeared (Van den Heuvel-Panhuizen, 2008). These instructional differences call into question to what extent they affect students' strategy choices in these operations. Therefore, the influence of teacher's instructional approach on students' strategy choice in complex multiplication and complex division problem solving is compared. The results may yield insights into the extent that teachers can influence students' strategic behavior, and by that, their achievement too. Furthermore, in contrast to the relation between simple multiplication and division (e.g., Campbell, Fuchs-Lacelle, & Phenix, 2006; De Brauwer & Fias, 2009; Mauro, LeFevre, & Morris, 2003), the relation between complex multiplication and division problem solving has not been studied before to our knowledge, so the current study extends the existing research body by studying these operations simultaneously.

3.1.4 Differences between students

Student level variables have been found to influence strategy choice and performance in mathematics. We focus in particular on the student characteristics gender, mathematics achievement level, and socio-economical background. Arguably, other variables such as students' motivation and attitudes (Vermeer, Boekaerts, & Seegers, 2000) and other home background and resources variables (Mullis et al., 2008; Vermeer et al., 2000) are found to be important determinants of mathematics achievement as well, but unfortunately, we have no information on that in the data.

Regarding mathematics achievement level, it has been frequently (but not uniformly, see Torbeyns, Verschaffel, & Ghesquière, 2005) reported that students of higher mathematical ability choose more adaptively or flexibly between strategies than students of low mathematical ability (Foxman & Beishuizen, 2003; Hickendorff et al., 2010; Torbeyns, De Smedt, et al., 2009b; Torbeyns, Verschaffel, & Ghesquière, 2002, 2004a; Torbeyns et al., 2006). In complex division, Hickendorff et al. (2009b, 2010) reported that sixth graders with a higher mathematics achievement level more often used written strategies (the traditional long division algorithm as well as repeated addition/subtraction strategies, see also Figure 1 in Hickendorff et al., 2010) than students with a lower mathematics level. Moreover, differences in accuracy between the strategies decreased with higher mathematics level. In other words, for high achievers it made less difference regarding accuracy which strategy they chose than it did for low achievers.

Gender differences in mathematics performance have often been reported. 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 grade (Kraemer et al., 2005) and in sixth grade (J. Janssen et al., 2005). Furthermore, boys and girls have been found to differ in the strategy choices they make on mathematics problems: girls have a higher tendency to (quite consistently) rely on rules and procedures, whereas boys are more inclined to use more intuitive strategies (Carr & Davis, 2001; Carr & Jessup, 1997; Gallagher et al., 2000; Hickendorff et al., 2010; Timmermans et al., 2007). Furthermore, Hickendorff et al. (2010) found that the shift in strategy use towards mental computation in solving complex division problems was mainly attributable to boys.

Finally, students socio-economical background has an effect on mathematics performance. TIMSS-2007 reported effects of parents' highest level of education (positively related to mathematics performance), the language spoken at home (lower performance if different than the test language) and whether parents were born in a different country (lower performance) (Mullis et al., 2008). Results from the Dutch national assessments on the effects of parents' origin and education were similar (J. Janssen et al., 2005). Moreover, in complex division, students with lower socio-economical background more often answered without written work and less often with one of the two written strategies (traditional algorithm and non-traditional strategies) (Hickendorff et al., 2009b).

3.1.5 The current study

The central aim of the current study was to get more insight in Dutch sixth graders' performance level in complex multiplication and division that was found to be decreasing over time and lagging behind educational standards, by using insights from educational psychology. In secondary analyses of national assessment data, we studied the role of (change in) solution strategy use in explaining (change in) achievement, and in turn, the influence of instructional approach on students' strategy choice. Because information on the instructional approach was only available in the 2004 cycle and not in the 1997 cycle, we set up this study in two separate parts. In the first part, we focused on the effect of solution strategy use on achievement in complex multiplication, thereby extending previous work of Hickendorff et al. (2009b) in complex division. Specifically, we aimed to get more insight in the performance decrease between 1997 and 2004 in multiplication, by analyzing changes in students' typical strategy choice, overall differences in accuracy between strategies, and changes in these strategy accuracies. Moreover, we also addressed the effects of the student characteristics gender, mathematics achievement level, and socio-economical background. Findings may yield educational implications and recommendations on how to turn the negative trend around.

In the second part, we focused on the – possibly different – influence of teacher's strategy instruction on students' strategy choice in multiplication and division. To that end, solution strategy data on multiplication and division problems from the 2004 assessment data were combined. Because instruction in how to solve multiplication problems (end point is traditional algorithm) differs from instruction in how to solve

division problems (traditional algorithm disappeared from the Dutch mathematics textbooks), the question is to what extent that reflects in students' strategy choice, potentially yielding implications for educational practice on the influence of the teacher's instruction on students behavior (strategy choice and performance).

3.2 Part I: Changes in strategy choice and strategy accuracy in multiplication

3.2.1 Method

Sample

In the present study, parts of the material of the two most recent national assessments carried out by CITO (Dutch National Institute for Educational Measurement) are analyzed in depth. These studies were carried out in May/June 1997 (J. Janssen et al., 1999) and in May/June 2004 (J. Janssen et al., 2005). For each assessment cycle, schools were sampled from the national population of primary schools, stratified with respect to three socio-economical status categories. In 1997, 253 primary schools with in total 5314 sixth graders (12-year-olds) participated. In the 2004 sample, there were 122 primary schools with in total 3078 students. Schools used various mathematics textbooks, although the large majority (over 90% of the schools in 1997, and almost 100% of the schools in 2004) used textbooks based on RME principles.

Subsets of the total samples of 1997 and 2004 were used in the present analysis: we included only students to whom items on complex multiplication were administered. In 1997, that subset consisted of 551 students with mean age 12 years 4 months (SD = 5 months; range = 11;2 - 14;0) from 218 different primary schools. It consisted of 995 students with mean age 12 years 4 months (SD = 4 months, range = 11;1 - 14;0) from 123 schools in 2004. So, the analyses in part I of this study are based on observations of 1,546 students in total.

In the 1997 sample, there were 45.9% boys and 49.9% girls (remaining 4.2% missing data); in the 2004 sample there were 49.6% boys and 48.8% girls (1.5% missing data). Information on the socio-economical background of the students was available too, based on the background and education of the parents: students with foreign parents

with low level of education/occupation (SES-2) and all other students (SES-1)[1]. In 1997, the distribution of students was 87.0% SES-1 and 9.1% SES-2 (4.0% missing data). In 2004, these percentages were 84.0% SES-1 and 14.5% SES-2 (1.5% missing data).

Material and Procedure

In the two assessment cycles together, there were 16 different complex multiplication problems administered, of which five problems were administered in both 1997 and 2004. These five problems were the anchor items, serving as a common basis for comparisons over time. Table 3.1 shows several characteristics of the multiplication problems: the actual multiplicative operation required, whether or not the problem was presented in a realistic context, and the proportion correct in 1997 and 2004 (if observed). On the five common problems (items 7-11), the proportion correct was lower in 2004 than in was in 1997 with differences ranging from .05 (item 10) to .16 (item 11), illustrating the achievement decrease between the two consecutive assessments.

The design of the assessment tests was different in 1997 than it was in 2004. In the 1997 assessment, there were in total 24 different mathematics content domains, and for each domain a subtest was assembled. Students were administered three to four of these subtests. One content domain was *complex multiplication and division*, and its subtest contained 12 problems on multiplication (of which one was eliminated from the scale in the test calibration phase) and 12 on division (also one item was eliminated). Therefore, from the 1997 cycle there were responses of 551 students to 11 different multiplication problems, see also Figure 3.3. In the 2004 cycle, each subtest contained items from different content domains instead of from only one domain as in 1997. Specifically, items were systematically distributed over test booklets in an incomplete test design. In total, there were 18 different test booklets, of which 8 booklets contained items on complex multiplication (and division). There were 10 different multiplication problems used in 2004. Figure 3.3 shows the distribution of these problems (7-16) over the test booklets.

1 In the Dutch educational system, funding of schools is based on an index of parental background and education of the students. There are three major categories: at least one foreign (non Dutch) parent with a low level of education and/or occupation, Dutch parents with a low level of education and/or occupation, and all other students. The definition of the second category has become more stringent between the 1997 and 2004 cycles: in 1997, students were in this category if only one of the parents had a low level of education/occupation, while in 2004 both parents had to have a low level of education/occupation (J. Janssen et al., 2005). As a consequence, the first two categories are incomparable between the two cycles. Therefore, these two categories were combined in the current study, in the category SES-1 (as was also done by J. Janssen et al., 2005).

TABLE 3.1 *Specifications of the multiplication problems*.*

			% correct	
item	problem	context	1997	2004
1	25 × 22	yes	.86	-
2	704 × 25	yes	.62	-
3	178 × 12	yes	.73	-
4	1.800 × 1.75	yes	.31	-
5	86 × 60	no	.77	-
6	109 × 87	no	.70	-
7	*24 × 57*	yes	.76	.62
8	*9.6 × 6.4*	no	.43	.30
9	0.18 × 750	no	.51	.41
10	*16 × 13.2*	yes	.48	.43
11	38 × 56	yes	.75	.59
12	*1.500× 1.60*	yes	-	.53
13	*28 × 27.50*	yes	-	.48
14	*4380 × 3.50*	yes	-	.31
15	99 × 99	no	-	.43
16	42 × 52	no	-	.61

*Italicized problems concern problems that are not released for publication by CITO, and therefore a parallel version (with respect to number characteristics of the operands and outcome) is presented here.

995 students in the 2004 cycle completed one of these eight booklets, and thus responded to three to five different multiplication problems per student.

The testing procedure was very similar in both assessment cycles. Test booklets were administered in classroom setting and each student worked through the problems individually, without time pressure. On each page of the test booklet, several items were printed. Next to each item there was blank space that students could use to write down calculations. In 2004, test instruction was as follows: "*In this arithmetic task, you can use the space next to each item for calculating the answer. You won't be needing scrap paper apart from this space.*" In addition, the experimenter from CITO explicitly stressed that students could use the blank space in their booklets for writing down calculations. Students were free to choose their own solution strategy, including choosing whether or not to make written calculations. In the 1997 assessment, instructions were somewhat less explicit in this respect.

For each student, a measure of general mathematics achievement level (GML) was

	test	item															
cycle	booklet	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	*N*
1997	–	x	x	x	x	x	x	x	x	x	x	x						551
2004	1							x			x		x	x		x		120
2004	2								x	x		x			x			131
2004	3							x								x	x	129
2004	4							x	x						x	x		122
2004	5								x						x		x	123
2004	6									x	x	x	x	x				127
2004	7										x		x	x			x	120
2004	8									x		x					x	123
***N* per item**		551	551	551	551	551	551	922	927	932	918	932	367	367	376	371	495	

FIGURE 3.3 *Distribution of multiplication items over test booklets, in the 1997 and in the 2004 assessment cyles. Symbol* × *indicates item was administered.*

computed, based on their performance on all mathematics problems presented to them in their test booklets. In the 1997 cycle, students completed – besides multiplication problems – also other problems from the domain of *numbers and operations*. Using Item Reponse Theory (IRT; e.g., Embretson & Reise, 2000; Van der Linden & Hambleton, 1997, see also below), a latent ability scale was fitted to the responses to all non-multiplication items. Consecutively, each students' position on the latent scale was estimated, and we standardized these estimates in the 1997-sample; range (–3.79, 3.15). For the 2004 cycle, a similar procedure was used, but because the item sampling design was different, students completed different sets of mathematics items from all mathematics domains (*numbers and operations*, *measurement*, and *percentages/fractions/ratios*). A general mathematical ability scale was fitted with IRT, and students' ability estimates were standardized, but now with respect to the 2004-sample; range (–4.52, 3.52). So, the general mathematics level (GML) measure used in the analyses indicates the relative standing of the student compared to the other students in his/her assessment cycle. Three students (one from 1997, two from 2004) with extreme scores (absolute standardized value larger than 3.50) were excluded from the analyses.

Responses

Two types of responses were obtained for each multiplication problem. First, the numerical answer given was scored as correct or incorrect. Skipped items were scored as incorrect. Second, by looking into the students' written work, the strategy used to solve

each item was classified. Seven categories were distinguished, see also Figure 3.2. The first strategy (*Traditional*) was the traditional algorithm for multiplication. The second category (*Partitioning both operands*) included strategies in which both the multiplier and the multiplicand were split. An example of this strategy is the RME approach of 'column multiplication' (Van den Heuvel-Panhuizen, 2008). In the third category of strategies (*Partitioning one operand*) only one of the operands was split. The fourth category contained all *Other written* strategies, including only repeated addition. The fifth category (*No Written Working*) consisted of trials (student-by-item combinations) in which nothing was written down except the answer. The sixth category (*Wrong/Unclear*) consisting of erased or unclear strategies, and wrong procedures such as adding the multiplicands. The final category (*Skipped*) contained skipped problems (no written working and no answer).

Solution strategies were coded by 8 independent trained research assistants who each coded a separate part of the data. To assess the reliability of this coding, the work of 256 students was recoded by 2 external independent trained research assistants, and the interrater-reliability coefficient Cohen's κ (Cohen, 1960) was computed. The average κ on categorizing solution strategies was .87, which was more than satisfactory.

Statistical analyses

Several properties of the data set necessitated advanced psychometric modeling. These properties were, first, that the responses *within* each student were not independent, because each student completed several items (i.e., there were repeated observations). Hence, this correlated data structure should be accounted for in the psychometric modeling approach. Second, each of these repeatedly observed responses was bivariate: the item was solved correct or incorrect (dichotomous score variable) and a specific strategy was used (nominal variable). Third, the incomplete design of the data set complicated the comparisons between 1997 and 2004, because different students completed different subsets of items. Analysis on the item level would be justified, but would not take the multivariate aspect of the responses into account, and univariate statistics would be based on different subsets of students. Furthermore, analyses involving changes in performance would be limited to the common items and would therefore not make use of all available information. A final consideration was that it should be possible to include student characteristics as predictor variables in the analysis.

In sum, analysis techniques were needed that can take into account the multivariate aspect of the data and are not hampered by the incomplete design. These demands can elegantly be met by introducing a latent variable. Individual differences are modeled by mapping the correlated responses on the latent variable, while the student remains the unit of analysis. The latent variable can be either categorical or continuous.

Recall that we aimed to analyze changes in students' typical strategy choice, overall differences in accuracy between strategies, and changes in these strategy accuracies. For the analysis of changes in strategy choice, we argue that a categorical latent variable is best suited to model multivariate strategy choice, because individual differences between students are qualitative in this respect. Latent class analysis (LCA) accomplishes this goal, by introducing a latent class variable that accounts for the covariation between the observed strategy choice variables (e.g. Lazarsfeld & Henry, 1968; Goodman, 1974).

The basic latent class model is $f(\mathbf{y}|D) = \sum_{k=1}^{K} P(k) \prod_{i \in D} P(y_i|k)$. Classes run from $k = 1, \ldots, K$, and $\mathbf{y}$ is a vector containing the nominal strategy codes on all items i that are part of the item set D presented to the student given the test design. Resulting parameters are the class probabilities or sizes $P(k)$ and the conditional probabilities $P(y_i|k)$. The latter reflect for each latent class the probability of solving item i with each particular strategy. So, we search for subgroups (latent classes) of students that are characterized by a specific pattern of strategy use over the items presented. To analyze changes between 1997 and 2004 in the relative frequency of the different strategy classes, year of assessment was inserted as a covariate, so that assessment cycle predicted class membership (Vermunt & Magidson, 2002). All latent class models were fitted with the `poLCA` package (Linzer & Lewis, 2010, 2011) available for the statistical computing program R (R Development Core Team, 2009). Because latent class models on variables with 7 different categories were very unstable, we recoded the solution strategies into four main categories: Traditional, Non-Traditional (partitioning both operands, partitioning one operand, other written strategies), No Written Working, and Other (wrong/unclear and skipped).

The second portion of the research question focused on strategy accuracy: how can the strategy used predict the probability of solving an item correctly? We argue that in these analyses a continuous latent variable is appropriate, to be interpreted as (latent) ability or proficiency. The repeatedly observed correct/incorrect scores are the dependent variables, and the nominal strategies take on the role of predictors. The latent variable accounts for the individual differences in proficiency in complex multiplication

by explaining the correlations between the observed responses. Item Response Theory (IRT) modeling (e.g., Embretson & Reise, 2000; Van der Linden & Hambleton, 1997) accomplishes this goal. Through the five common items, it was possible to fit one common scale for 1997 and 2004 of proficiency in complex multiplication, based on all 16 items.

In the most simple IRT measurement model (the Rasch model), the probability of a correct response of subject p on item i can be expressed as $P(y_{pi}=1|\theta_p)=\frac{\exp(\theta_p+\beta_i)}{1+\exp(\theta_p+\beta_i)}$ Latent variable θ_p expresses ability or proficiency, measured on a continuous scale. The item parameters β_i represent the *easiness* of each item. Such descriptive or measurement IRT models can be extended with an explanatory part (Wilson & De Boeck, 2004; Rijmen et al., 2003), meaning that covariates or predictor variables are included of which the effects on the latent scale are determined. These can be (a) item covariates, that vary across items but not across persons, (b) person covariates, that vary across persons but not across items, and (c) person-by-item covariates, that vary across both persons and items. In the present analyses, the strategy chosen on an item was dummy coded and included as person-by-item predictor variables (for further details, see Hickendorff et al., 2009b). Like in the LCA, we used the four main solution strategy categories. Moreover, the category of Other strategies was not of interest in analyzing strategy accuracies, since it was a small heterogeneous category of remainder solution strategies, consisting mainly of skipped items. Therefore, all student-by-item combinations (i.e., trials) solved with an Other strategy were excluded from the explanatory IRT analyses.

All IRT models were fitted using Marginal Maximum Likelihood (MML) estimation procedures within the NLMIXED procedure from SAS (SAS Institute, 2002, see also De Boeck & Wilson, 2004; Rijmen et al., 2003; Sheu et al., 2005). We chose nonadaptive Gaussian quadrature for the numerical integration of the marginal likelihood, with 20 quadrature points, and Newton Raphson as the optimization method.

3.2.2 Results

Strategy choice

Table 3.2 displays proportions of use of the seven strategies, separately for the 1997 and the 2004 assessment. In the first 2 columns, strategy proportions are totaled over the five common items. The traditional algorithm was the most prevalent strategy in both years, but its use decreased markedly between 1997 and 2004. The non-traditional

TABLE 3.2 *Strategy use on multiplication problems in proportions, based on 1997 and 2004 data.*

	common items		all items	
multiplication strategy	1997	2004	1997	2004
traditional	.65	.45	.59	.39
partitioning - both operands	.03	.08	.03	.07
partitioning - one operand	.03	.08	.05	.09
other written strategy	.01	.02	.01	.03
no written working	.17	.25	.23	.31
wrong/unclear	.02	.03	.02	.03
skipped	.08	.09	.07	.09
N observations	2755	1876	6061	3852

strategies (partitioning both operands, partitioning one operand, and other written strategies) each increased in relative frequency of choice: on the common items, from a total of 7% of the trials in 1997 to 18% of the trials in 2004. Furthermore, the frequency of answering without written working also increased between the two cycles. The final two strategy categories (wrong/unclear and skipped items) remained more or less stable in frequency. In the final 2 columns of Table 3.2, strategy proportions are totaled over all items presented in each assessment, so these proportions are based on different item collections for 1997 and 2004. Although these distributions seem slightly different from those based only on the common items, the pattern of shifts between 1997 and 2004 was very similar.

Latent class models on strategy choice, recoded in four main categories, with year of assessment as covariate were fitted with 1 to 6 latent classes. The Bayesian Information Criterion (BIC) was used to select the optimal number of classes. The BIC is a criterion that penalizes the fit (log-likelihood, LL) of a model with the model complexity (the number of parameters; P), and it is computed as -2LL $+P\log(N)$, with N the sample size. Lower BIC-values indicate better models in terms of parsimony. The model with 4 classes showed the lowest BIC-value, and was therefore selected as the best-fitting model. The relative entropy of this latent class model, a measure of classification uncertainty ranging between 0 (high uncertainty) and 1 (low uncertainty) (Dias & Vermunt, 2006), was .84, indicating that the latent classes were well separated.

Figure 3.4 graphically displays the estimated parameters of this 4-class model. These

are first the conditional probabilities: for each particular class, the probabilities of choosing each of the four strategies on each of the 16 items. The second parameters were the class sizes in the two assessment years, showing changes over time. First note that the class-specific strategy profiles of the first three classes are more or less dominated by one strategy type chosen on all items. So, apparently students were quite consistent in their strategy choice on this set of multiplication problems.

From these strategy profiles we interpret the latent classes as follows. The first class is dominated by the Traditional algorithm, although item 12 and to a lesser extent item 4 are clear exceptions with a large probability of being solved without written working. Nevertheless, we argue that the best way to summarize this latent class is to label it the Traditional class. In the 1997 assessment, the majority of the students (67%) belonged to this class, while this decreased to less than half (44%) of the students in 2004. The second class is characterized by a very high probability on all items to state the answer without writing down any calculations or solution steps (No Written Working class). This class nearly doubled in size, from 13% in 1997 to 22% in 2004. The third class (Non-Traditional class) is dominated by Non-Traditional strategies, but again items 12 and 4 are exceptions with the modal probability of No Written Working. This class tripled in size, from 7% in 1997 to 22% in 2004. Finally, the fourth class is a mishmash of Other strategies, No Written Working, and Traditional strategies. This Remainder class did hardly change in size between 1997 (13%) and 2004 (12%).

Next, we studied whether the effect of assessment cycle on latent strategy class depended on students' gender, general mathematics level, or socio-economical status. Because inserting these many variables as covariates in latent class analysis would render the model statistically unstable, an alternative approach was used consisting of two steps. First, all students were assigned to the latent class for which they had the highest posterior probability (modal assignment; mean classification error .08). Next, this 4-category class membership variable was used as dependent variable in a multinomial logistic regression model (see for example Agresti, 2002). Fifty-one students were excluded because they had missing or extreme values on at least one of the predictor variables.

The main effects of year (Likelihood Ratio (LR) test[2] $= 72.5$, $df = 3$, $p < .001$), gender (LR $= 59.7$, $df = 3$, $p < .001$), GML (LR $= 88.8$, $df = 6$, $p < .001$), and SES (LR $= 27.8$,

2 The Likelihood Ratio test can be used to statistically test the difference in fit of two nested models. The test statistic is computed as 2 times the difference between the LL of the general model and the LL of the specific model, and it is asymptotically χ^2-distributed with df the difference in number of estimated parameters between the 2 models.

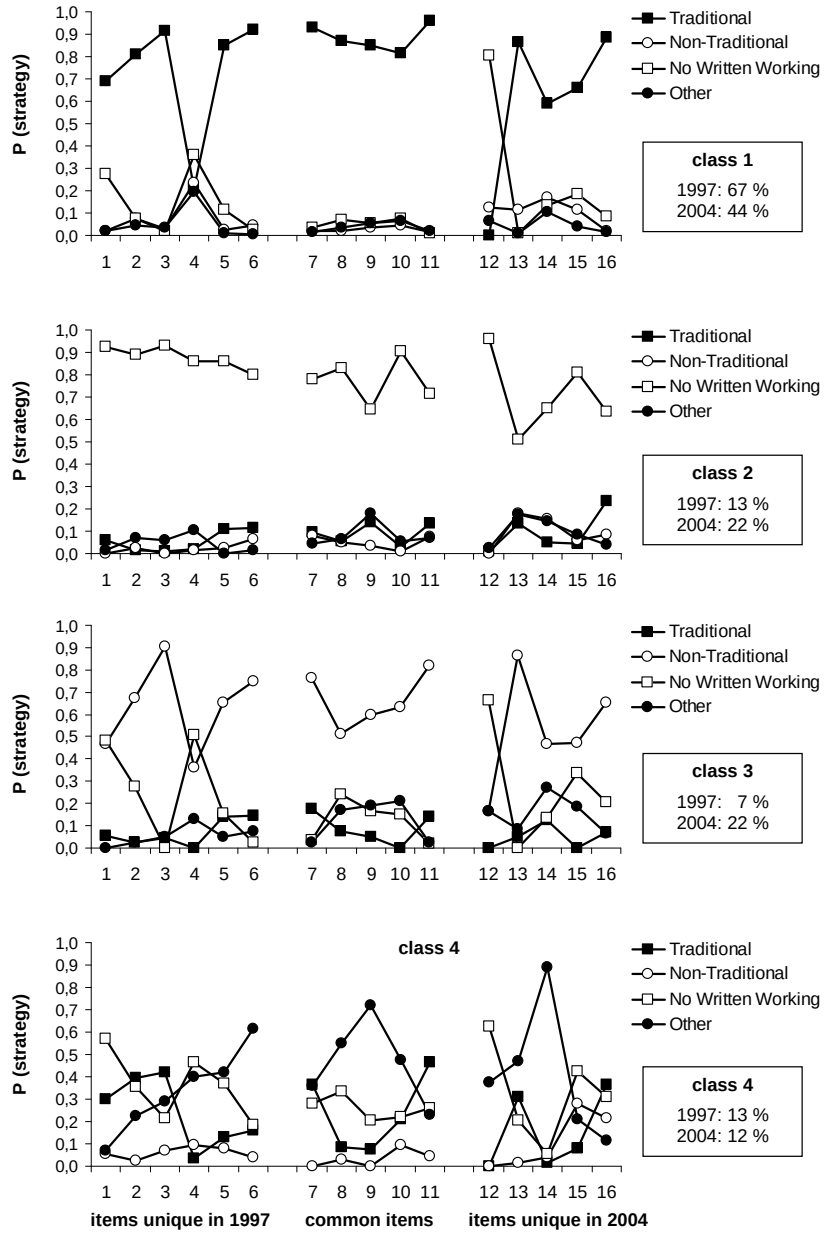


FIGURE 3.4 *Conditional probabilities of strategy choice on multiplication problems of the 4 latent classes model, 1997 and 2004 data.*

TABLE 3.3 *Cross-tabulations of the student background variables general mathematics level, gender, and SES with latent strategy class membership (in proportions); multiplication problems, 1997 and 2004 data.*

		Latent strategy class				
		1 (T)	2 (NWW)	3 (N-T)	4 (R)	N
boys	1997	.65	.17	.10	.08	252
	2004	.39	.26	.22	.13	488
girls	1997	.73	.11	.05	.11	274
	2004	.58	.12	.17	.13	481
low GML		.43	.23	.13	.20	480
medium GML		.58	.16	.15	.10	508
high GML		.65	.14	.16	.05	509
SES-1		.58	.14	.16	.12	1306
SES-2		.42	.34	.12	.12	191

Note. T = Traditional class; NWW = No Written Working class; N-T = Non-Traditional class; R = Remainder class.

$df = 3$, $p < .001$) on class membership were all significant. Moreover, the interaction between gender and assessment cycle was also significant (LR $= 8.6$, $df = 3$, $p = .035$), implying that the shift in relative frequency of the strategy choice classes was not the same for boys and girls. The other interaction effects, between GML or SES on the one hand and assessment cycle on the other hand, were not significant (ps $> .05$).

The top portion of Table 3.3 shows the interaction between gender and assessment cycle on latent class membership. Gender differences in overall strategy choice patterns clearly emerge: In both assessment cycles, girls more often typically used the Traditional algorithm than boys, while they were less often classified in the No Written Working or Non-Traditional classes. Interestingly, the shift in strategy choice between 1997 and 2004 was different for boys than for girls. Boys were increasingly classified in the No Written Working class, while the proportion of girls in this class was about stable. Furthermore, the decrease over time in the Traditional strategy class was larger for boys than for girls. Apparently, the shift away from the traditional algorithm towards answering without written working should be mainly attributed to boys. Table 3.3 also shows the main effects of GML (trichotomized based on percentile scores, to facilitate interpretation) and SES. The proportion of students being classified in the Traditional class increased

with increasing mathematics achievement level, while the proportion of students being classified in the Remainder class as well as in the No Written Working class decreased with increasing GML. The proportion of students classified in the Non-Traditional class was relatively unaffected by GML. Finally, SES-1 students were more often classified in the Traditional class than SES-2 students, and less often in the No Written Working class.

Strategy accuracy

To evaluate how the found shift in strategy choice should be evaluated with respect to achievement, we investigated whether the multiplication strategies differed in accuracy rate, with (explanatory) IRT models. Starting from the Rasch measurement model without explanatory variables, a model was built with a forward stepwise procedure by successively adding predictor variables and retaining those that had significant effects. All 1,027 trials (student-by-item combinations) involving Other strategies (wrong, unclear, of skipped) were excluded. In total, 1,542 students yielding 8,886 observations were included in the analyses.

First, the null model without any predictor effects was fitted, assuming that the θ_p come from one normal distribution. The 17 parameters were 16 item easiness parameters β_i with estimates ranging between –.86 and 2.19, and the variance of the ability scale θ estimated at 1.35 (the mean of θ was fixed at 0 for identification of the latent scale). Next, the effect of assessment cycle was estimated, which resulted in a substantial decrease in BIC as well as in a significant increase in model fit; LR $= 42.4$, $df = 1$, $p < .001$. The latent regression parameter of 2004 compared to 1997 was –.64 on the logit scale[3], which was highly significant ($z = -6.52$).

Next, type of strategy used on an item was inserted as a predictor of the probability of solving an item correct. In order to keep the number of parameters manageable and interpretation feasible (see also Hickendorff et al., 2009b), these strategy effects were restricted to be equal for all items. Adding strategy effects yielded a highly significant increase in model fit (LR $= 393.2$, $df = 2$, $p < .001$). Compared to No Written Working, both using a Traditional strategy (difference on logit scale $\delta_{\text{T vs. NWW}} = 1.53$, $z = 19.58$)

3 The effect of –.64 on the logit scale can be transformed to the odds ratio scale or the probability scale. The odds ratio is computed as $\exp(-.64) = .53$, and implies that the odds of a correct answer for 2004-students is about half the size of the odds for 1997-students. On the probability scale, we can compute that on an item on which 1997 students had a 50% probability to obtain a correct answer, this probability was $\frac{\exp(-.64)}{1+\exp(-.64)} \times 100\% = 35\%$ for students in the 2004 assessment.

and using a Non-Traditional strategy ($\delta_{\text{N-T vs. NWW}} = .93$, $z = 9.74$) yielded a significantly higher probability to obtain a correct answer. Moreover, the Traditional strategy was significantly more accurate than Non-Traditional strategies ($\delta_{\text{T vs. N-T}} = .59$, $z = 6.64$). Clearly, the three main strategy categories differed in accuracy. By accounting for strategy choice shifts between 1997 and 2004, the regression parameter of year decreased to –.43 ($z = -4.49$). Furthermore, the interaction effect of Year and Strategy was not significant (LR = 4.9, $df = 2$, $p = .09$), implying there was a general and equally-sized decrease in success rates from 1997 to 2004 for each of the three strategies.

Subsequently, we tested whether the achievement change over time or the effect of strategy used depended on either gender, general mathematics level (GML), or socio-economical status (SES). Excluding an additional 50 students (201 trials) from the analyses because they had missing or extreme values on one or more of the background variables, these three student characteristic variables were added to the explanatory IRT model, and we tested the interaction effects with year and strategy. None of the two-way interaction effects of the student characteristics with year were significant (year × gender: LR = .2, $df = 1$, $p = .63$; year × GML: LR = .6, $df = 1$, $p = .45$; year × SES: LR = 2.3, $df = 1$, $p = .13$). This implied that the accuracy decrease between assessment cycles was about the same size for boys and girls, for students with low or higher SES, and for students with different mathematics achievement level. By contrast, strategy significantly interacted with gender (LR = 7.2, $df = 2$, $p = .027$) and GML (LR = 37.8, $df = 2$, $p < .001$), but not with SES (LR = 2.3, $df = 1$, $p = .13$), the largest interaction effect being with GML. The strategy-by-gender interaction was no longer significant up and above the interaction between Strategy and GML (LR = 4.6, $df = 2$, $p = .10$); apparently, it was mediated by gender differences in general mathematics achievement level.

Figure 3.5 displays the interaction effects between GML and strategy used on the logit IRT ability scale. It shows that students' general mathematics level was positively related to performance on the multiplication problems, within each particular strategy used. Furthermore, the effect of GML was significantly stronger when the strategy No Written Working was used ($\zeta_{\text{GML in NWW}} = 1.18$, $z = 18.35$) than it was when either the Traditional algorithm ($\zeta_{\text{GML in T}} = .73$, $z = 15.56$) or one of the Non-Traditional strategies ($\zeta_{\text{GML in N-T}} = .84$, $z = 10.02$) was used. The difference between the latter two regression parameters was not significant. Interpreting these effects, it seems that with increasing general mathematics level it became less important which strategy students used on complex multiplication. In particular, for low performers, answering without written

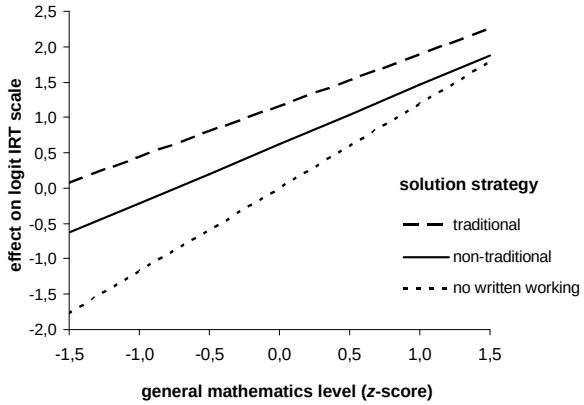


FIGURE 3.5 *Graphical display of interaction effect between strategy used and student's general mathematics level on IRT ability scale, based on multiplication problems in 1997 and 2004 cycles.*

work was much less accurate than using one of the two written strategies; for high performers this difference disappeared.

Importantly, even after accounting for all significant (interaction) effects of student characteristics and strategy used, the performance decrease between 2004 and 1997 remained substantial ($-.50$) and significant ($z = -5.96$), so shifts in strategy choice only partially accounted for the performance decrease.

3.2.3 Conclusions part I

In the first part of this study, we aimed to get more insight in the lagging and decreasing performance level in multiplication, by analyzing changes in strategy choice and in strategy accuracies between 1997 and 2004. Both descriptive statistics and latent class models showed that strategy choice has shifted from 1997 to 2004: The use of the traditional algorithm decreased, while the use of non-traditional strategies as well as no written working solutions increased, the latter two by approximately the same amount. Moreover, the shift away from typically using the traditional algorithm towards typically answering without written working was observed mainly in boys.

To evaluate how the found shift in strategy choice should be evaluated with respect to accuracy, we investigated whether the multiplication strategies differed in accuracy rate.

Results showed that the traditional algorithm was more accurate than non-traditional strategies, which in turn were more accurate than answering without written working (these differences were smaller for students with higher mathematics achievement level). Consequently, the observed shift in strategy choice – replacing traditional strategies by non-traditional and no written working strategies – can be characterized as unfortunate with respect to achievement, and is one contributor to the general performance decline. However, this did not explain the complete performance decrease: even after accounting for the shift in strategy choice between the two years, still a significant decrease in performance from 1997 to 2004 remained. So, each solution strategy on its own was carried out significantly less accurately in 2004 than it was in 1997.

In conclusion, two changes regarding strategy use appeared to have contributed to the general performance decline on complex multiplication problems: a shift in choice of more accurate to less accurate ones, and a general accuracy decline within each strategy on its own. A relevant next question is what influences students' strategy choice. The effect of student characteristics gender, general mathematics level, and SES were already addressed in the first part of the study. In the next part, we try to get more insight in the effect of teacher's instruction on strategy choice, focusing on differences between complex multiplication and division.

3.3 Part II: Effect of teachers' strategy instruction on students' strategy choice

3.3.1 Method

Sample

The sample used for the second part of this study consisted of the 995 students of the 2004 assessment, who were also part of the sample of part I of this study. These students not only completed the complex multiplication problems, but also problems on complex division.

Material and Procedure

In total, there were 10 complex multiplication problems (see part I of this study) and 13 problems on complex division (see Hickendorff et al., 2009b, Table 1) in the 2004

assessment[4], ranging from the easiest problem $157.50 \div 7.50$ (60.4% correct) to the most difficult one $6.40 \div 15$ (12.6% correct). These 23 problems were administered in an incomplete test design: there were 8 different test booklets containing between 6 and 13 problems. The testing procedure was the same as in part I of this study.

In the schools participating in the 2004-assessment, teachers in grade 4 ($N = 116$), 5 ($N = 115$), and 6 ($N = 118$) filled in a questionnaire about the mathematics curriculum and teaching practices. There were questions included on their approach in teaching multidigit operations (addition, subtraction, multiplication, and division). For each operation, they were asked to choose, from two worked-out examples, which approach best matched the practice in their classroom: (a) the traditional algorithm, or (b) so-called 'column calculation' (Van den Heuvel-Panhuizen, 2008), the RME-alternative to the standard algorithm. Column calculation in multiplication entailed strategies in which both operands were partitioned (see Figure 3.2); in division it entailed repeated subtraction of multiples of the divisor from the dividend (see below). If teachers taught the column calculation procedure first and the traditional algorithm later, they could mark both approaches.

Figure 3.6 shows the distribution of teachers' responses to these questions on multiplication and division. It shows that between grade 4 and grade 6 there is a gradual shift from the RME approach to the traditional approach, in both multiplication and division. However, sixth grade teachers instructed the traditional algorithm much less frequently for solving division problems than for solving multiplication problems. This difference between multiplication and division is in line with the learning/teaching trajectories differences and the mathematics textbooks, which do no longer cover the traditional algorithm for long division (Van den Heuvel-Panhuizen, 2008). In this part of the current study, sixth grade teachers' approach to multiplication problem solving (missing data for 39 students) and to division problem solving (missing data for 57 students) were used as variables predicting students' strategy choice.

4 For reasons of consistency, we used the same item numbers in part II of as in part I of this study for the multiplication problems (7 - 16) and as in Hickendorff et al. (2009b) for the division problems (7 - 19).

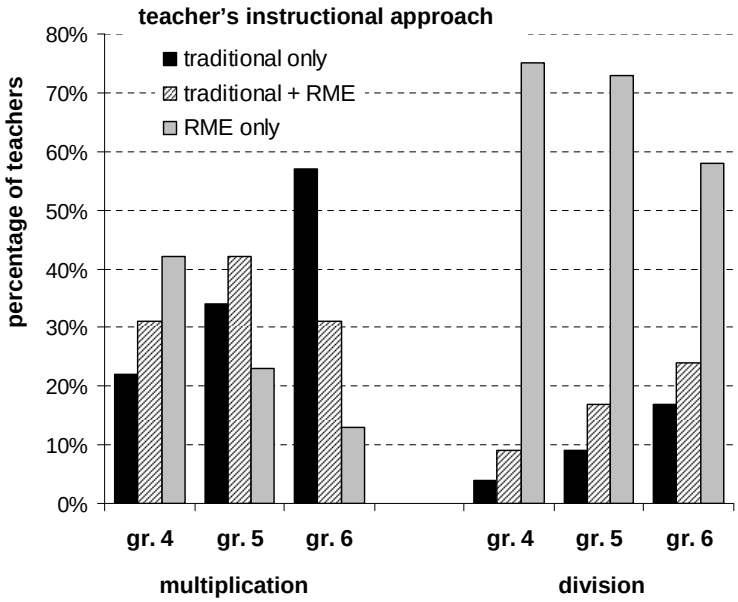


FIGURE 3.6 *Fourth grade, fifth grade, and sixth grade teachers' approach to complex multiplication and division problem solving, as reported in J. Janssen et al. (2005, p. 44).*

Responses

For the multiplication problems, the strategy categorizations from part I of this study were used. For the division problems, we distinguished seven main strategies[5] (see Hickendorff et al., 2009b, 2010 for examples): (a) the traditional algorithm of long division, (b) repeated subtraction of multiples of the divisor from the dividend (the RME alternative to the traditional algorithm; Van den Heuvel-Panhuizen, 2008) (c) repeated addition of multiples of the divisor towards the dividend (multiplying-on), (d) other written strategies, (e) answering without written working, (f) unclear strategies or wrong procedures, and (g) skipping the problem.

5 In Hickendorff et al. (2009b), these 7 categories were recoded into 4 main solution strategies, by combining first the repeated subtraction with the repeated addition strategy and other written strategies into a category labeled Realistic strategies, and second, by combing the last two categories of unclear strategies/wrong procedures and skipped problems into a category labeled Other strategies.

TABLE 3.4 *Strategy use on multiplication and division problems, split by teacher's instructional approach, based on 2004 data.*

	teacher's approach to multiplication			
multiplication strategy	trad. only	trad. + RME	RME only	total
traditional	.43	.34	.31	.39
partitioning - both operands	.08	.11	.09	.07
partitioning - one operand	.04	.10	.16	.09
other written strategy	.02	.05	.02	.03
no written working	.32	.28	.32	.31
wrong/unclear	.03	.03	.02	.03
skipped	.09	.09	.08	.09
number of trials	2242	1084	383	3709
	teacher's approach to division			
division strategy	trad. only	trad. + RME	RME only	total
traditional	.43	.18	.02	.14
repeated Subtraction	.04	.17	.24	.19
repeated Addition	.04	.05	.05	.05
other written strategy	.01	.01	.02	.01
no written working	.39	.41	.47	.44
wrong/unclear	.04	.04	.06	.05
skipped	.07	.14	.13	.12
number of trials	897	1225	2674	4796

3.3.2 Results

Table 3.4 presents the distribution of strategy choice trials on multiplication and division problems, split by the instructional approach of the student's sixth grade teacher. First, it shows that the overall distribution of strategy choice is different for multiplication than for division problem solving. That is, the traditional algorithm for multiplication was used much more frequently (39% of all trials) than the traditional algorithm for division (14% of the trials); while answering without written work was more common on division problems (44% of the trials) than on multiplication (31%).

Second, there was a clear influence of the teacher's approach to problem solving

on students' strategy choice, in particular in division.[6] For multiplication, choosing the traditional algorithm increased when the teacher instructed this approach, in particular if it was the only strategy. Moreover, the use of the partitioning-one-operand strategy increased when teachers instructed the RME approach in combination with the traditional algorithm or in particular when it was the only strategy instructed. For division, the use of the traditional algorithm clearly depended on whether the teacher instructed this approach or not. Furthermore, the use of the repeated subtraction strategy increased when teachers instructed it. Finally, both answering without written working and skipping problems appeared to be influenced by the teacher's approach: no written working was most prevalent with teachers instructing only the RME approach to division, while skipping a problem occurred least often when the teacher instructed only the traditional algorithm for division.

3.3.3 Conclusions Part II

In the 2004 assessment, information on the teacher's approach to instruction in multidigit multiplication and division problem solving was available. This variable appeared to affect students' strategy choice on both operations. In multiplication, choice for the traditional algorithm and for partitioning one operand was influenced by the teacher's instructional approach. The effect of instructional approach was particularly strong in division, however: nearly exclusively students whose teacher instructed the traditional algorithm for long division used that algorithm. Moreover, students whose teacher instructed the RME approach to division more frequently used this RME strategy (repeated subtraction), but also more often answered without written working or skipped the problem entirely.

6 Straightforward statistical testing of dependency of rows and columns in Table 3.4 was not possible, however, because observations within cells were not independent. To provide support for the statistical significance of the relation between teacher's approach and students' strategy choice, we tested it using students' latent strategy class membership as dependent variable. For multiplication, strategy choice latent class membership (4 classes) of part I of this study was used, and the effect of teacher's approach to multiplication problem solving was highly significant ($\chi^2(6, N = 956) = 43.2$, $p < .001$). For division, we used strategy choice latent class membership (also 4 classes: mainly Traditional, mainly Non-Traditional, mainly No Written Working, and mainly Other strategies) from p. 340-343 Hickendorff et al. (2009b). The effect of teacher's approach to division problem solving on strategy choice class was even more significant ($\chi^2(6, N = 938) = 251.9$, $p < .001$) than it was in multiplication.

3.4 GENERAL DISCUSSION

In the current study, we aimed to get more insight in Dutch sixth graders' performance level in complex or multidigit multiplication and division, that was found to be decreasing over time and lagging behind educational standards, in a reform-based mathematics learning/teaching trajectory. In secondary analyses of national assessment data we focused on the solution strategies students used as an explanatory mechanism between (change in) instruction and (change in) achievement. In the first part, the relation between solution strategy use and achievement in complex multiplication was investigated to analyze the negative performance trend between 1997 and 2004. Findings showed that two changes regarding solution strategies contributed to the performance decline: a shift in strategy use from a more accurate strategy (the traditional algorithm) to less accurate ones (non-traditional partitioning strategies and answering without written work, the latter shift attributable to boys), as well as a general decline in each strategy's accuracy rate. In the second part, students' strategy choice in multiplication and division in the 2004 assessment appeared to be influenced by the instructional approach held by their teachers, most profoundly in division problem solving. In the following, we discuss the conclusions and the implications in more detail.

3.4.1 Complex multiplication problem solving

Strategy use in complex multiplication shifted between 1997 and 2004. The use of the traditional algorithm decreased, while answering without written work (most likely mental calculation, see also Hickendorff et al., 2010) and the use of non-traditional partitioning strategies increased. An important subsequent question is: How do we evaluate this strategy shift in multiplication? The current findings showed clear differences in accuracy between the strategies, leading us to argue that, first, the increase in mental calculation is a worrisome development, because the success rate of non-written strategies was substantially lower than of the two written strategies. Second, because the non-traditional strategies were less accurate than the traditional algorithm for students of all mathematics levels, an increase in this strategy also does not seem desirable.

Another worrisome – and more difficult to grasp – development is the finding that each of the three main multiplication strategies dropped in accuracy rate between the two assessment cycles. That means that on the same problem with a particular strategy,

sixth graders in 1997 had a higher probability to derive the correct answer than sixth graders in 2004 had. It leaves us with another negative trend that needs explanation, that should probably be sought in the educational practice. Potential mechanisms include the amount of time and practice spent on these topics, i.e., opportunity-to-learn (OTL), which has been argued to be the single most important predictor of student achievement (Hiebert & Grouws, 2007). However, the educational assessments do not offer enough information to analyze this rigorously (see also Hickendorff et al., 2009a), and further research is needed. It would also be very informative to carry out an international comparative study between countries that differ in their opportunity-to-learn with respect to multidigit multiplication and division.

Furthermore, there were differences between students in multiplication problem solving. First, there were clear gender differences in strategy choice: girls had a larger tendency than boys to consistently use the traditional algorithm and were less inclined to consistently use non-written strategies. These findings are in congruence with previous findings on gender differences in strategy use, with girls showing a larger reliance on structured strategies and algorithms, and boys having a higher tendency to use more informal, less structured strategies (Carr & Davis, 2001; Carr & Jessup, 1997; Gallagher et al., 2000; Hickendorff et al., 2009b, 2010; Timmermans et al., 2007). These gender differences may be related to the consistent finding that girls have lower levels of confidence with mathematics (Mullis et al., 2008; Timmermans et al., 2007; Vermeer et al., 2000), so they may act more cautiously than boys and therefore choose the safety of using well-structured written strategies. Moreover, gender differences changed over time: between 1997 and 2004, boys and girls showed a strategy shift from the traditional algorithm towards non-traditional strategies, but boys additionally shifted from the traditional algorithm toward answering without written working. The shift towards mental calculation should thus be attributed mainly to boys.

Second, students' mathematics achievement level affected individual differences in strategy choice in multiplication too: the tendency to quite consistently use non-written strategies decreased with higher mathematics achievement level, while the tendency to use written strategies (traditional and non-traditional) increased with higher mathematics level. So, lower performers seem to choose their strategy less adaptively than high performers, congruent with findings of Foxman and Beishuizen (2003), Hickendorff et al. (2010), Torbeyns, De Smedt, et al. (2009b), and Torbeyns et al. (2002, 2004a, 2006). Moreover, for students with lower mathematics achievement level

the accuracy gap between written and non-written strategies (see Figure 3.5) was larger, so they seemed to be doubly disadvantaged by choosing a non-written strategy.

These results have theoretical implications for cognitive models of strategy choice, in which it is hypothesized that children choose their strategies adaptively, i.e., they choose the fastest strategy that yields the correct answer (e.g., Shrager & Siegler, 1998). The present findings seem to signal suboptimal strategy choices when students, in particular the lower performing ones, chose non-written strategies. However, cautiousness is called for: because students were free to select their strategies (the so-called choice method; Siegler & Lemaire, 1997) it is likely that selection effects biased strategy accuracy data. For example, the finding that the use of written strategies increased with students' mathematics achievement level may have biased the accuracy of those strategies upwardly, although we were able statistically correct for this. Further research addressing strategy efficiency in an unbiased manner, such as has been done by Hickendorff et al. (2010) in the domain of division, is needed to make firmer conclusions regarding the adaptivity of students strategy choices in the domain of complex multiplication.

3.4.2 Multiplication and division: similarities and differences

The present study shows some remarkable similarities and differences between the domains of multiplication and division problem solving (as reported in part II of the present study and in Hickendorff et al., 2009b).

First, regarding shifts in strategy choice between the two assessment cycles, a similarity was the decrease in use of the traditional algorithm. This is in accordance with a general shift away from algorithmic procedures in mathematics education reform (although it is worth noticing that already in 1997 the large majority of mathematics textbooks used were based on reform principles). A second similarity is an increase in the answering without written working (most likely mental calculation, see also Hickendorff et al., 2010), that was mainly attributable to boys. Although mental calculation plays an important role in mathematics education reform (Blöte et al., 2001; Buijs, 2008) it was not anticipated that this would also affect the way students solve complex arithmetic problems with multidigit numbers, on which the use of paper and pencil was allowed. One would expect that, rather than an increase in mental computation, predominantly the use of non-traditional strategies would increase, because these are part of the learning trajectories (Van den Heuvel-Panhuizen, 2008). Herein also lies a striking difference

between multiplication and division. In division instruction, the traditional algorithm has been replaced entirely by the RME approach, but surprisingly, students' behavior does not show an increase of the RME approach. In contrast, in multiplication, students increasingly used the RME strategies, while this was not the end point of the RME learning trajectory. One would expect to find the opposite pattern.

Second, the accuracy differences between the main solution strategy categories in multiplication and division were also characterized by similaries and differences. In both operations, answering without written working was the least accurate strategy, in particular for the lower performers. The accuracy difference between the traditional algorithm and non-traditional strategies, however, depended on the operation. In multiplication, non-traditional partitioning strategies were less accurate than the traditional algorithm, for students of all mathematics levels. In contrast, in division the non-traditional repeated addition/subtraction strategies were equally accurate as the traditional algorithm for most students (although for medium performers the traditional division algorithm was significantly more accurate). A possible explanation for this difference may be that for division, repeated addition/subtraction strategies (the RME approach) are actually the only approach being taught (at least in the mathematics textbooks) and hence serve as a full alternative to the traditional algorithm, while this is not the case for multiplication.

Finally, in the 2004 assessment, information on the teachers' instructional approach to solving multiplication and division problem solving was available from a teacher questionnaire. The teacher's instruction appeared to be quite different for multiplication where the traditional algorithm was still dominant, than it was for division where the RME approach was dominant. Moreover, it appears that in both domains a switch towards an increase in the traditional algorithm has taken place between grade 4 and grade 6. In multiplication, this is in line with the RME learning/teaching trajectory, while that is not the case in division (Van den Heuvel-Panhuizen, 2008). The observation that teachers apparently diverged from the mathematics textbook in division also illustrates the fact that the *enacted* curriculum (the actual instruction taking place in the classroom, e.g. Porter, 2006; Stein et al., 2007) can differ from the *intended* curriculum that is based on written documents such as textbooks and educational standards, and that it is therefore important to take both curricula into account.

The instructional approach of the teacher had substantial effects on the strategy choice of the students. Particularly the approach to solving complex division problems

was important: almost exclusively students whose teacher instructed the traditional algorithm for division (as the only strategy or in combination with RME strategies) actually used this strategy. So, only when teachers departed from their textbook and instructed the division algorithm students used it, which is not unexpected since it is probably not a strategy that is easy to self-invent. A further interesting finding is that when teachers instructed the RME approach to division problem solving, the frequency of answering without written working and of skipping problems entirely was higher than when teachers instructed the traditional algorithm. It thus seems that when students were instructed the RME approach to division problem solving, they were less inclined to apply their standard written procedure (the RME approach) and were also less able or confident to attempt solving the problem at all. A tentative explanation for this pattern may be that the RME approach to division is less well structured than the traditional algorithm, so that students know less well how to start and what to do. In addition, it may be that teachers who instruct the traditional algorithm for division, thereby diverging from the mathematics textbook, value standard solution procedures more, affecting their students' behavior.

For multiplication, the relation between teacher's instructional approach and students' strategy choice was less marked than for division. Still, the use of the traditional algorithm was higher in students whose teachers instructed it than it was in students whose teachers instructed only the RME approach. Moreover, students whose teacher instructed the RME approach to multiplication problem solving more often used partitioning strategies. Contrary to division, the frequency of answering without written working was rather unaffected by the teacher's approach.

3.4.3 Educational implications

Regarding multidigit multiplication, the present findings would lead to the educational recommendation that teachers encourage students to use the traditional algorithm. Moreover, for both multiplication and division it seems legitimate to encourage the use of written strategies over non-written strategies for problems with multidigit numbers, in particular for the lower performing students. The current findings give initial support for the idea that changing teacher's strategy instruction may be an effective way to influencing students' strategy choice, although further research is needed. Moreover, one could think of other mechanisms to affect students' problem solving behavior as well,

such as for example crediting written solution steps on top of crediting only the correct answer, or changing the appraisal of written compared to mental strategies. Furthermore, the entire domain of multiplicative reasoning performance (the tables, mental arithmetic, and complex arithmetic with the use of paper and pencil allowed) showed a negative trend (J. Janssen et al., 2005), so we plead for vigilance of the educational community regarding the position of this domain in the mathematics curriculum.

All results taken together – an unfortunate shift in strategy choice (mainly in boys), the traditional algorithm being the most accurate strategy (at least in multiplication), the relatively high proportion of lower achievers who answer without written work while for them this is a particularly unsuccessful strategy, and the general decrease in accuracy within each strategy – we argue that reconsideration of several elements of the implementation of the RME approach is called for. These elements are not unique to the Dutch mathematics education reform, so it is also important in an international perspective. For example, students' informal strategies are very important in the reform, which seems not to be without problems. In particular, we think that the transition from informal strategies to the traditional algorithm needs further consideration, such as for example also came forward from findings in the UK (Anghileri et al., 2002). In addition, the idea that students are free to choose how to solve problems may have a negative side-effect in boys, who are on average more inclined to intellectual risk-taking than girls (Byrnes, Miller, & Shafer, 1999) and may overestimate their ability to solve problems without writing down solution steps or intermediate answers. Moreover, like Geary (2003), Torbeyns et al. (2006), and Verschaffel, Luwel, Torbeyns, and Van Dooren (2009) we plead for more research-based evidence into the feasibility of striving for adaptive expertise in mathematics education, especially for the lower performing students who seem to be doubly disadvantaged by making suboptimal strategy choices.

3.4.4 Final considerations

The present study is limited in several ways, since it is based on large-scale educational assessments data (see also Hickendorff et al., 2009a; Van den Heuvel-Panhuizen, Robitzsch, Treffers, & Köller, 2009). Because assessments are surveys, they are descriptive by nature, which has its limitations such as allowing only correlational analyses with explanatory variables. Therefore, our present study is also limited in several ways. For example, it was not possible to study the effect of item characteristics (such as whether a

problem was presented in a context or not) on strategy use or accuracy, because these features were not varied in a systematic way. Furthermore, the classroom administration procedure – although making large sample sizes feasible – had the drawback that for gaining insight in solution strategy use we had to revert to students' written work. Therefore, we were left with no further information on the instances in which students did not write down any work but the answer. Presumably, they used mental computation on those trials (as was mostly confirmed by Hickendorff et al., 2010), but we cannot be certain about that. Finally, the results of the present study are limited to the situation in the Netherlands, and the question is to what extent it would generalize to other countries. We believe, however, that the Dutch situation is interesting to study, because of the influential theory of realistic mathematics education gaining international popularity, which contains several elements of the international reform movement. In addition, because there is, unlike the US, a nationally coherent curriculum, trends in mathematics achievement can be linked quite closely to shifts at a national level in instructional approach.

Acknowledging these limitations, we argue that studying solution strategies in data from large-scale assessments was a valuable enterprise, both from a practical educational viewpoint as well as from the perspective of educational psychology. The present findings can be a valuable starting point for evaluating learning outcomes more comprehensively and for raising research questions for further research, and they advanced our insight into performance trends, strategy choice, and strategy accuracy in multidigit multiplication and division in a reform-based educational environment significantly.

CHAPTER 4

Individual differences in strategy use on division problems: Mental versus written computation

This chapter has been published as Hickendorff, M., Van Putten, C. M., Verhelst, N. D., & Heiser, W. J. (2010). Individual differences in strategy use on division problems: Mental versus written computation. *Journal of Educational Psychology, 102*, 438-452.

The research was supported by CITO, National Institute for Educational Measurement.

ABSTRACT

Individual differences in strategy use (choice and accuracy) were analyzed. A sample of 362 Grade 6 students solved complex division problems under two different conditions. In the choice condition they were allowed to use either a mental or a written strategy. In the subsequent no-choice condition, they were required to use a written strategy. Latent class analysis showed that there were 3 subgroups of students with respect to their pattern of strategy choices: primarily using a written strategy (more girls than boys); primarily using a mental strategy (more boys than girls); and using a written strategy on more difficult items, but a mental strategy on the easier ones (almost no weak mathematical achievers). Strategy accuracies were analyzed with explanatory IRT modeling. A between-subjects comparison in the choice condition showed that written strategies were usually more accurate than mental strategies, especially for the weak achievers. A within-subject comparison showed that the performance of students who used mental calculation on a particular item in the choice condition, improved by requiring the use of a written strategy in the no-choice condition.

4.1 INTRODUCTION

In arithmetic, children know and use multiple strategies (e.g., Lemaire & Siegler, 1995; Siegler, 1988a). These strategies and their characteristics have been extensively studied in elementary addition and subtraction (e.g., Carr & Jessup, 1997; Carr & Davis, 2001; Torbeyns et al., 2004b), in elementary multiplication (e.g., Lemaire & Siegler, 1995; Siegler & Lemaire, 1997) and in mental multidigit addition and subtraction (e.g., Beishuizen, 1993; Beishuizen, Van Putten, & Van Mulken, 1997; Blöte et al., 2001; Fuson et al., 1997; Torbeyns et al., 2006). In these domains of mathematics, near consensus is reached on what strategies children use (Torbeyns et al., 2006). In contrast, strategies for division have received considerably less attention (Robinson et al., 2006). In particular, few studies are devoted to more complex division problems in higher grades of primary school (Van Putten et al., 2005). However, complex arithmetic in general (i.e., operations with multidigit numbers for which one may use a written procedure) and complex division in particular are interesting domains in the light of the reform in mathematics education. In addition, the most recent national assessment in the Netherlands reported a large performance decline on these domains since 1987 (J. Janssen et al., 2005). Therefore, it is important to systematically study solution strategies for solving complex division

problems, which was the purpose of the present study. Specifically, the aim was to analyze individual differences in strategy choice and strategy accuracy in solving complex division problems by students at the end of primary school (Grade 6).

4.1.1 Solution strategies

Strategy competence is a much-studied topic in cognitive and educational psychology. Lemaire and Siegler (1995) distinguished four dimensions of strategic competence on which individuals may differ: their strategy repertoire (which strategies are used), their strategy distribution (the frequency with which the strategies are used), their strategy efficiency (strategy speed and/or accuracy), and their strategy selection (how strategies are chosen, related to problem and individual strategy characteristics). These dimensions are central to the current study.

In cognitive models such as the Adaptive Strategy Choice Model (ASCM, Siegler & Shipley, 1995), the choice of a strategy on a particular problem is a function of individual strategy performance characteristics for that problem and strategy-choice criteria held by the individual. People tend to choose their strategy adaptively: they choose the fastest and most accurate strategy for a given problem out of their strategy repertoire. Strategy speed and accuracy may vary from individual to individual (Siegler & Lemaire, 1997). Furthermore, individuals differ in the stringency of the threshold for choosing a strategy, the confidence criterion (Siegler, 1988a, 1988b). For example, Siegler (1988a) has found three subgroups of first-graders with regard to their strategy choices, and labeled them as good students, not-so-good-students, and perfectionists. The latter group contrasted with the first two groups with respect to the required certainty that a particular strategy yields the correct answer for choosing that strategy: the perfectionists had a high confidence-criterion (Siegler, 1988a, 1988b). Hecht (2006)) found similar subgroups in adult's multiplication.

In addition to task and individual strategy performance characteristics, individual differences in strategy choices are also the result of the influence of other cognitive and socio-emotional or socio-cultural variables (Torbeyns, Ghesquière, & Verschaffel, 2009). In mathematics, gender differences and achievement level effects on strategy choices have been found. Previous studies reported that girls show a greater reliance on rules and procedures (i.e., they may set their confidence criterion higher), whereas boys seem to use more intuitive strategies (Carr & Davis, 2001; Carr & Jessup, 1997; Gallagher

et al., 2000; Timmermans et al., 2007). These findings may also be related to more general gender-related differences in mathematics, such as that girls have lower levels of confidence with mathematics (Mullis et al., 2008; Timmermans et al., 2007; Vermeer et al., 2000). With respect to achievement level, Torbeyns et al. (2006)) found that above-average achievers were more adaptive in their strategy choices than below-average achievers on addition and subtraction up to 100. Similar differences were found by Foxman and Beishuizen (2003) in mental calculation strategy choices of top, middle, and bottom attainment band students. Finally, on complex (multidigit) division problems, weak mathematical achievers more often used mental strategies and less often used written strategies, compared to medium and strong mathematical achievers (Hickendorff et al., 2009b).

Individual differences in strategy use can only be found by analyzing individual profiles of strategy choice and strategy performance over items; aggregate measures such as means or correlations obscure the profiles (Gilmore & Bryant, 2006; Mabbott & Bisanz, 2003). Until recently, strategy use was usually studied by the so-called *Choice* method: letting students solve several problems, subsequently coding their overt and/or covert strategy use, and relate these strategies to the correctness of the answers and time needed to execute the strategy. Such a procedure was also followed in the few studies devoted to solution strategies for complex division (Anghileri et al., 2002; Van Putten et al., 2005). Siegler and Lemaire (1997) have argued that such studies are flawed by selection effects, because strategy performance was assessed in a condition in which students could choose which strategy they used. In that case, estimates of strategy characteristics may be biased by selection effects in two ways. Selection of students could play a role: for example, it could be that the accuracy of a particular strategy is overestimated because it is applied relatively more often by better than by weaker students. Second, selection of items could play a role: it could for example be that the accuracy of a particular strategy is underestimated because it is applied more often on difficult problems than on easy ones.

To estimate strategy characteristics in an unbiased manner, Siegler and Lemaire (1997) have proposed the *Choice/No-Choice* method. Each participant is tested under two different types of conditions. In the *Choice* condition, participants are free to choose their strategy. In the *No-Choice* conditions, participants have to use a specific strategy to solve all items, so that accuracy and speed of that strategy are assessed unbiasedly. Several studies in mathematics have applied the Choice/No-Choice methodology to

study solution strategies (e.g., Lemaire & Lecacheur, 2002; Luwel, Verschaffel, Onghena, & De Corte, 2003; Torbeyns et al., 2004b, 2006). An important aspect of the Choice/No-Choice methodology is that the adaptiveness of strategy choices can be evaluated on an individual level, i.e., whether a subject chooses that strategy that *for him or her* is most efficient (Luwel et al., 2003).

4.1.2 Complex division

In the present paper, complex division is defined as division problems in which the quotient is a multidigit number (e.g., $872 \div 4 = 218$), and the divisor may be multidigit too (e.g., $736 \div 23 = 32$). This contrasts with simple division (division problems from the multiplication tables), in which the quotient is a single digit (e.g., $48 \div 8 = 6$; Robinson et al., 2006). Complex division is of special interest for three reasons. First, complex division is an understudied topic thus far. Second, with the mathematics education reform in the Netherlands, instruction in solving complex division problems has changed to the largest extent compared to other arithmetical domains. Third, a large decline in achievement has been observed on this domain in the Netherlands. These latter two points will be discussed below.

Mathematics education has experienced a reform process of international scope over the last couple of decades (Kilpatrick et al., 2001). A common international trend is that students should become active learners who construct their own mathematics (Blöte et al., 2001). In the Netherlands this reform movement is known by the name of Realistic Mathematics Education (RME) (Freudenthal, 1973; Gravemeijer, 1997b; Treffers, 1987). In RME, students' deep understanding of mathematics is pursued, instead of mastery of rules and procedures. Students should acquire insight and flexibility in their use of strategies (Kilpatrick et al., 2001). Instruction is based on the key principle of guided reinvention (Freudenthal, 1973), implying that instructors should give students the opportunity to reinvent the mathematics they have to learn for themselves, according to a mapped out learning route, starting at the informal or intuitive strategies students have. Another characteristic of the reform is that mental arithmetic plays a central role in mathematics education (Blöte et al., 2001). At present, nearly all Dutch primary schools use mathematics textbooks based on RME principles (J. Janssen et al., 2005).

The RME approach to solving complex division problems starts from the informal strategies that young children employ for division. These are direct counting, repeated

addition, use of a multiplicative operation (reversed multiplication), and repeated subtraction (Ambrose et al., 2003; Mulligan & Mitchelmore, 1997; Neuman, 1999; Robinson et al., 2006). In the Netherlands, Treffers (1987) introduced progressive schematization, building on the informal strategy of repeated subtraction of multiples (chunks) of the divisor. The resulting solution strategy is increasingly schematized and abbreviated. Therefore, two aspects of the solution strategy can vary: the level of abbreviation and the level of schematization. Figure 4.1 shows examples of strategies varying in abbreviation of chunking (low-level versus high-level, discussed in more detail later) and in whether or not the schematic notation of repeated subtraction is used. In general, there are two fundamental differences between these types of chunking strategies and the traditional algorithm for long division (also in Figure 4.1). First, in the traditional algorithm it is necessary that each subtraction of a multiple of the divisor is optimal, while this is not necessary in the chunking approaches. Second, understanding the place values of the digits in the dividend is not important for applying the traditional algorithm in a correct way, while the place values of the numbers are left intact in the progressive schematization approach. In RME-based textbooks, the traditional long division algorithm is replaced by the alternative approach as introduced by RME. Therefore, complex division can be said to be a prototype of the RME approach (Van Putten et al., 2005), which makes it an interesting domain to study.

Another reason why research on strategies for complex arithmetic, especially division, is needed, is that national assessment results from the Netherlands showed that achievement on these domains has decreased considerably since 1987 (J. Janssen et al., 2005). This decrease occurred between four consecutive large scale national assessments of mathematics achievement at the end of primary school, carried out in 1987, 1992, 1997, and 2004 by the Dutch National Institute for Educational Measurement (CITO). Results showed that achievement has increased on numerical estimation and general number concepts, and to a lesser extent on calculations with percentages and mental addition and subtraction. However, results showed a decline of performance on complex arithmetic. Students who were in their final year of primary school in 2004 performed less well than students who were at their final year in 1987 on complex addition and subtraction, and especially on complex multiplication and division. Between 1987 and 2004, achievement in complex multiplication and division has dropped with more than one standard deviation on the ability scale, with an accelerating trend (J. Janssen et al., 2005).

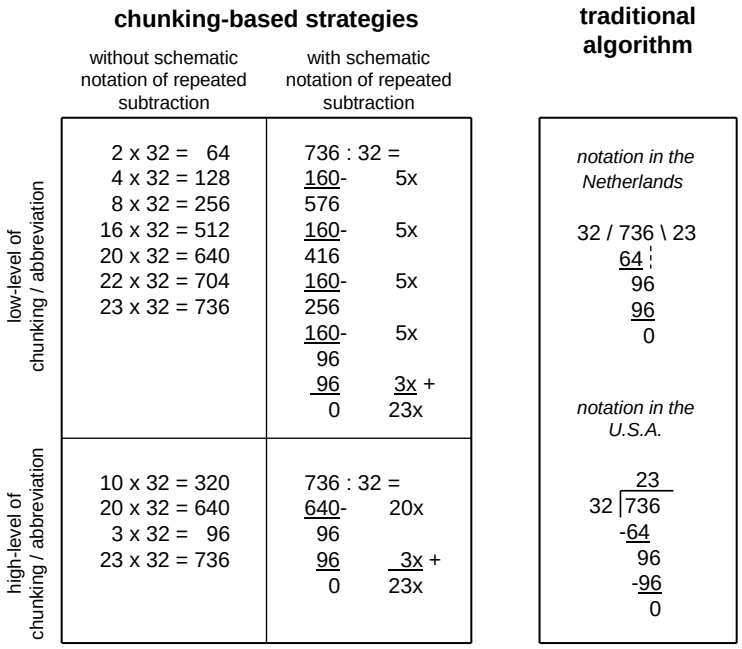


FIGURE 4.1 *Examples of solution strategies for the problem 736 ÷ 32.*

A recent study related this achievement decline on complex division on the two most recent national assessments of 1997 and 2004 to the solution strategies used (Hickendorff et al., 2009b). Results showed that two changes appeared to have contributed to the decline. First, strategy use had shifted: use of written procedures decreased (attributable to a decrease in the use of the traditional long division algorithm), while an increased percentage of items was answered without calculations written down on scrap paper. This shift could be attributed to boys much more than to girls. The strategy shift was unfortunate, since answering these problems with a nonwritten strategy yielded fewer correct answers than when students used a written strategy. Second, each of the solution strategies yielded less correct answers in 2004 than in 1997, with approximately the same amount of accuracy decrease per strategy. So, the performance decline on complex division seems to be related to a change in strategy use, in particular to a decrease in

the use of written strategies, and to a change in strategy accuracy as well (Hickendorff et al., 2009b). However, this study was descriptive and therefore limited in several aspects, among which the aforementioned selection effects.

4.1.3 Current study

The purpose of the current study was to study strategy use for solving complex division problems in a more systematic way than has been done in the descriptive studies devoted to this subject. Therefore, a partial Choice/No-Choice design was used: in the Choice condition students could choose whether they used a written or mental strategy in solving a set of complex division problems, and in the subsequent No-Choice condition they had to use a written strategy on a set of parallel problems. Students were interviewed on their nonwritten strategies in the Choice condition.

The main focus in this study was on the distinction between written and mental strategies, because national assessment findings showed that this was a very relevant distinction with respect to the observed performance decrease: use of mental strategies increased over time, but their success rates stayed far behind those of written strategies (Hickendorff et al., 2009b). In the present study, mental computation was defined as carrying out arithmetical operations without the use of any recording devices such as pen and paper, similar to the definitions of for example Reys (1984), Timmermans et al. (2007), and Varol and Farran (2007). Written strategies included all forms of calculations in which some part of the solution process was written down on paper, ranging from only recording intermediate solution steps to written algorithmic procedures.

The design of the current study was a partial implementation of the Choice/No-Choice design, because there was only one No-Choice condition (forced written strategy use). So, there was no No-Choice condition in which students had to answer these problems using mental computation. We believed it would be too large a burden for a great number of students if they had to use a mental strategy on this type of problems on which they would usually need scrap paper. Many students may have become frustrated and unmotivated to continue such a task. Another difference with many of the other studies using the Choice/No-Choice methodology was that response times were not recorded, because we applied a classroom administration procedure, comparable to the procedure in the national assessments.

The present study's first aim was to describe the repertoire and distribution of the

written strategy types, as well as of the mental calculation strategies. Two aspects were of particular interest. One aspect concerned the mental strategies. From previous studies, it was not entirely clear that students who did not write down their strategy or intermediate solution steps had used a mental calculation strategy to reach their answer. It could also be that they provided only an estimate of the precise answer, or that they just guessed because they did not know how to solve the problem. The other aspect concerned the repertoire and distribution of (forced) written strategies that were used by students who applied a mental strategy when they could choose. Specific points of focus were to investigate whether these students did have written procedures in their repertoire, and how their written strategy choices compared to those of students who already used a written procedure when they were free to choose.

The second aim was to analyze individual differences in the extent to which students chose mental or written procedures, and relate these individual differences to problem characteristics and to the student characteristics gender and general mathematics level. Regarding problem features, we distinguished problems with a large cognitive demand from problems requiring less cognitive effort. In the former category, the dividend, divisor, and outcome were either large or decimal numbers. On these problems, more written strategies were expected. Problems in the latter category were such that either the dividend could be easily split up, or that a compensation procedure (rounding off the dividend) could be used. We expected that solving these problems would require less cognitive effort, and therefore, less need to write down a solution strategy or intermediate steps. Regarding student characteristics, we first hypothesized that boys would be more inclined than girls to use a mental strategy, replicating the findings of the national assessments (Hickendorff et al., 2009b) and other studies on gender differences in strategy choice (e.g., Carr & Davis, 2001; Carr & Jessup, 1997; Timmermans et al., 2007). Second, we expected that students with lower levels of mathematics achievement would make less adaptive strategy choices than higher achieving students, such as has been found by Foxman and Beishuizen (2003) and Torbeyns et al. (2006).

The final aim of this study was to compare the relative accuracy of written and mental calculation strategies in a more systematic way than was possible in the national assessments (Hickendorff et al., 2009b) and previous studies into complex division (Anghileri et al., 2002; Van Putten et al., 2005). The national assessments showed that the accuracy of written procedures was higher than the accuracy of answering without writing anything down, suggesting that encouraging the use of of a written strategy would

improve performance. However, these results were based on comparing the performance results *between* students and items, and thus biased by selection effects. In the present study, these comparisons were made *within* students and items. We hypothesized that the accuracy of students who used mental calculation on a particular item would increase on a parallel item, by forcing them to use a written strategy, since writing down of the solution procedure helps both in recording key information and in schematizing the solution process (Ruthven, 1998). For students who already chose to write down their solution steps in the Choice condition, we expected that forcing them to do so in the No-Choice condition would not be harmful (but also not beneficial) since it would be their usual way of solving the problem.

4.2 METHOD

4.2.1 Participants

The present sample consisted of 362 students from Grade 6, with a mean age of 12.0 years (SD = .51), ranging from 10.4 to 14.5 years. There were 193 boys, 161 girls, and 8 students with missing gender information. The students originated from 12 different schools, located in different regions in the Netherlands with varying levels of urbanization. Each of these schools used a mathematics textbook based on the RME principles, although they did not use the same textbook.

4.2.2 Materials

Experimental task

The experimental task consisted of a total of 13 items: 4 pairs of parallel items yielding 8 items, and 5 additional unpaired items. [1] The complete item set is presented in Appendix 4.A. Parallel item versions were constructed such that the two items within each pair were as similar as possible with respect to item context and number characteristics of the divisor, dividend, and outcome, but would not result in testing effects that would occur when 2 sets of identical items were presented to each student. All 13 items presented a complex division problem that was embedded in a realistic situation of sharing or dividing.

1 The complete task administered to the students consisted of 22 items. The 9 items that were not part of the present study were administered for other purposes of CITO.

As discussed in our hypothesis on the effect of problem features, we distinguished items with large cognitive demands – the four items pairs – from items requiring less cognitive effort – the five unpaired items. In the four item pairs, several item characteristics were varied, taking pair 1 as the base. In pair 2, the dividend and divisor were decimal numbers. In pair 3, the outcome (quotient) was much larger than in the other items, so that students had to subtract more chunks and/or larger chunks. Finally, in pair 4 the outcome was a decimal number, so students had to deal with a remainder. The five unpaired items (items 5 to 9 in Appendix 4.A) were constructed so that they could well be solved mentally, or by only recording some key numbers. Specifically, the dividend and divisor of items 6, 8, and 9 were chosen so that they could be mapped directly (e.g., $3240 \div 4$ could easily be split up in $32(00) \div 4$ and $40 \div 4$). Items 5 and 7 were constructed so that a compensation procedure (rounding the dividend) would be an efficient approach (e.g., $2475 \div 25 = 2500 \div 25 - 1$).

The task was divided into two parts containing separate instructions. Lay-out was designed in such a way that when students had finished the first part (items in the Choice-condition), they could not see the next part.

Standardized mathematics test

The general mathematical ability of each student was measured by the standardized mathematics subtest from the 2007-version of CITO's End of Primary School Test (CITO, 2007). This instrument is widely used in the Netherlands, and assesses the level of achievement of Grade 6 students on mathematics, language, study skills, and world orientation. The 60-item subtest on mathematics has high internal reliability (KR20 = .92; CITO, 2007). In addition to the sum score (the number of items answered correctly), the percentile rank for each student, based on a total of more than 150,000 participants of the End of Primary School Test 2007, could be calculated.

4.2.3 Conditions

In the experimental task, items were administered in 2 different conditions: 9 items were administered in the Choice condition, and 4 items in the No-Choice condition. The 9 items in the Choice condition consisted of a particular version of each of the 4 pairs of parallel items, and of the 5 unpaired items. In this condition, students were free to choose whether they used the scrap paper or not in solving each of the problems presented. If

students wanted to write down notes or solution procedures, they had to use the space next to each item for doing this.

In the No-Choice condition, the parallel versions of the first 4 items in the Choice condition were administered. In this condition, students were instructed that they had to write down their solution procedure in a calculation box presented next to each of the items. To encourage the use of the calculation box even more, students were told that if they would not write anything down in the calculation box, the answer they would state would be scored as incorrect.

The Choice condition always preceded the No-Choice condition, to prevent carry-over effects of having to write down solution steps on the free strategy choice. The assignment of item version to condition was counterbalanced, yielding two different forms of the task. In Form A, item versions *a* of each of the 4 item pairs were presented in the Choice condition, and their counterparts (versions *b*) were presented in the No-Choice condition. In Form B, this assignment was reversed. Half of the sample completed Form A, the other half Form B.

4.2.4 Procedure

Classroom administration of the experimental task

The experimental task was administered in the classroom. One of the two specific task forms (A or B) was assigned to each class. The teacher instructed the students that this test consisted of two parts, and that they could start with the first part (the items in the Choice condition), but that they could not start the second part before all students in the classroom had finished the first part. In addition, students were instructed that they could use the space next to each item to write down notes or calculations, and that it was not allowed to use a separate piece of paper.

When the last student in the classroom had finished the first part of the task, students could turn the page to the next part of the task, and the teacher read the instructions about using the calculation box out loud. All students then started the second part at the same time. When all students had finished both parts of the task, they handed in their test booklet. They could take as much time as they needed, so there was no time pressure.

Individual interviews

In 8 of the 12 schools, students interviews took place. After all students had finished both parts of the paper-and-pencil task and had handed in their booklets, the experimenters selected those who failed to write out calculations on at least 1 of the first 4 items. Due to time limitations, only a sample of 89 students from this selection (stratified on a teacher-based judgment of their general mathematics level) were interviewed. They were asked about their solution procedure on problems in the Choice condition on which they had not written down anything but their answer.

Interviews took place individually in a room outside the class, approximately one to three hours after they had finished the experimental task. The experimenters emphasized that they were only interested in *how* the student had solved the problem, so that the students needed not to worry about making mistakes. The experimenters asked the students whether they could remember how they had calculated the answer they had given and whether they could demonstrate the solution steps thinking aloud, for each item solved without written working. These interviews were audiotaped, and the experimenter also made notes.

Standardized mathematics test

The students completed the 2007 End of Primary School Test (CITO, 2007), as part of their final year's standardized assessment. This assessment took place approximately one month after the students participated in the current study.

4.2.5 Solution strategies

The strategy use on the experimental task of all 362 students was categorized. First, strategies were crudely categorized into one of three categories, based on the notes or solution procedures that were written down in the booklets containing the items. These categories were *written* strategies, *mental* strategies (answers stated without written work), or *skipped* items. A more fine-grained classification specified the type of written strategies or the type of mental strategies, respectively.

Types of written strategies were classified based on the work students had written down. A first consideration was whether a traditional strategy or a chunking-based approach was used. The latter category was subdivided into 4 categories, based on the combination of two aspects (see also Figure 4.1). First, this was the level of abbreviation

or chunking: divided into *low-level chunking* or *high-level chunking*. For a strategy to be categorized as high-level chunking, the first chunk needed to be at least the largest possible power of 10 times the divisor that fits into the dividend (for example, when solving $782 \div 34$ the first chunk should be at least 10×34, and when solving $4080 \div 20$ it should be at least 100×20). Furthermore, the remaining chunks needed also to be sufficiently efficient (so, no long tail after a first efficient chunk). If these criteria did not hold, the strategy was categorized as low-level chunking. This category included also: trying out several solutions, no chunking at all (for instance, when solving $782 \div 34$ repeatedly subtracting single 34s from 782), or splitting of the dividend (e.g., when solving $33 \div 12$ doing $30 \div 12 + 3 \div 12$). The second aspect on which written strategies were classified was whether or not a schematic notation of repeated subtraction was used. A final category was included containing *wrong* procedures (either the wrong operation or splitting of the divisor, e.g., when solving $782 \div 32$ doing $782 \div 30 + 782 \div 2$) and *unclear* strategies. Interrater reliability of categorization was assessed by computing Cohen's κ (Cohen, 1960) on 200 randomly selected observations (student-by-item combinations) that were coded by two independent experts. For the crude classification (written, mental or skipped) κ was .95, indicating almost perfect agreement. For the fine-grained classification κ was .76, indicating substantial and satisfactory agreement.

Based on audiotaped interviews, the type of strategies that the sample of students interviewed used in solving the items they had answered without written work in the experimental task was inferred. In the large majority of these instances (92%), students reported that they calculated the answer mentally ("in their head"). Within these mental calculation strategies, the following four categories were distinguished. Similar to the types of written strategies, these were first *low-level chunking* and second *high-level chunking* (of course, always without the schematic notation of repeated subtraction). In the third category, students reported that they did not try to calculate the exact answer and that the answer given was a *guess* or a *numerical estimation*. The final category comprised *wrong* and *unclear* mental calculation procedures.

4.3 RESULTS

For 354 subjects, the CITO mathematics test score as well as gender information were available. On the mathematics test, an average score of 45.2 items correct (SD = 9.3) out of a total of 60 items was obtained. The present sample of students performed

TABLE 4.1 *Descriptive statistics of strategy use and strategy accuracy.*

	Choice									No-Choice			
strategy	it. 1	it. 2	it. 3	it. 4	it. 5	it. 6	it. 7	it. 8	it. 9	it. 1	it. 2	it. 3	it. 4
strategy choice: number of students using each strategy													
mental strategy	44	106	42	92	141	155	157	235	196	3	8	2	13
written strategy (total)	313	253	317	264	218	205	202	125	161	354	350	355	343
high-level schema	263	122	234	176	124	114	107	77	93	304	187	281	222
high-level no schema	27	58	39	40	59	57	67	24	55	26	80	30	62
low-level	18	63	37	26	25	26	22	19	10	19	64	37	37
wrong/unclear	5	10	7	22	10	8	6	5	3	5	19	7	22
total (non-skipped)	357	359	359	356	359	360	359	360	357	357	358	357	356
strategy accuracy: proportion correct per strategy													
mental strategy	.57	.57	.40	.53	.82	.81	.75	.95	.71	n.a.	n.a.	n.a.	n.a.
written strategy (total)	.79	.82	.75	.59	.76	.79	.85	.90	.83	.82	.77	.76	.62
high-level schema	.83	.87	.81	.66	.81	.89	.92	.99	.83	.88	.84	.81	.68
high-level no schema	.78	.90	.77	.60	.83	.77	.84	.83	.87	.73	.84	.73	.71
low-level	.39	.75	.49	.62	.52	.65	.73	.89	.90	.32	.73	.54	.46
total (non-skipped)	.76	.75	.72	.58	.79	.80	.81	.93	.77	.82	.76	.76	.60

slightly better than the national sample of test takers: the median (population) percentile rank was 57.0, with (population) quartiles of 38.3 and 78.0 (these values would have been 50, 25, and 75, respectively, if the distribution of mathematics achievement of this sample would have been completely representative of the total population of participants of the End of Primary School Test 2007). Furthermore, there were gender differences in standardized mathematics achievement: girls answered significantly fewer items correctly ($M = 42.8$, SD $= 9.3$) than boys did ($M = 47.2$, SD $= 8.9$), $t(352) = 4.43$, $p < .001$. In the total population, there were similar differences between boys and girls on this mathematics test, so the present sample is representative in that respect.

Table 4.1 shows descriptive statistics of strategy choice (upper part) and strategy accuracy (lower part) for the items in the Choice as well as in the No-Choice condition on the experimental task. The two low-level chunking categories (with and without schematic notation) were taken together because of small numbers of observations. Furthermore, in the present sample there turned out to be only 3 out of the 362 students who used the traditional algorithm. Therefore, the traditional algorithm was grouped with the strategies of high-level chunking with a schema of repeated subtraction. Several aspects of Table 4.1 will be discussed in the following sections in which each of the research objectives is addressed.

4.3.1 Repertoire and distribution of written and mental solution strategies (first aim)

Table 4.1 shows that in the Choice condition, items 1 to 4 were solved with a mental strategy by 12% to 29% of the students. Not surprisingly, these percentages were higher on items 5 to 9, that were devised such that they could be solved mentally more easily than the first 4 items, ranging from 39% to 65%. Items were skipped only by very small numbers of students (2 to 6 students). In the No-Choice condition, items 1 to 4 were answered without written working by 3, 8, 2, and 13 students, respectively. These observations will be left out of the analyses comparing the Choice and No-Choice conditions. Evidently, the experimental manipulation to force students to write their solution steps in the calculation box was successful for the large majority of observations.

Written strategies

Table 4.1 also shows that the vast majority of written procedures consisted of high-level chunking, ranging from 71% on item 2 to 93% on item 1. High-level chunking was applied usually with the schematic notation of repeated subtraction. However, on items 5 to 9 and also on item 2, this schematic notation was relatively less often used than on items 1, 3, and 4. Another interesting result (not presented in Table 4.1) is that on items 5 and 7, which were devised so that applying a compensation strategy would be an efficient approach, compensation was not used very often within the written solution strategies. On item 5, 13% of the 218 written solutions involved compensation of the dividend, and on item 7 this was only in 7% of the 202 written solutions.

Table 4.2 shows the relative distributions of written strategies on items 1 to 4 in the No-Choice condition, separately for students who used a written or a mental computation procedure on the parallel item in the Choice condition. Per item, students were excluded who skipped the item in Choice and/or in No-Choice. The Fisher's Exact test-statistic testing the association between using a mental or written procedure on an item in the Choice condition on the one hand, and the distribution of written strategies used on that item on the other hand, is reported.[2] For each item, the association was significant (see last row of Table 4.2).

2 Statistical testing of the Pearson χ^2-statistic for independence of rows and columns of a contingency table was not feasible. This was because the assumption of most of the cells having an expected cell count of at least 5 was violated in 2 of the contingency tables. Fisher's Exact test was used instead. This test uses exact distributions instead of large sample approximations, such as the Pearson χ^2 does. See for example Agresti (2002).

TABLE 4.2 *Distributions of written strategies in the No-Choice condition, separate for students who solved that item with a mental (m) or written (w) strategy in the Choice condition.*

	item 1		item 2		item 3		item 4	
strategy	m	w	m	w	m	w	m	w
low-level	.07	.05	.16	.19	.18	.10	.18	.08
high-level	.22	.05	.26	.22	.23	.06	.27	.16
high-level schema	.66	.89	.44	.57	.51	.83	.43	.72
wrong/unclear	.05	.01	.14	.02	.08	.01	.12	.05
N	41	310	98	251	39	314	82	257
Fisher's Exact Test	18.1*		20.0*		21.7*		23.3*	

*$p < .001$

On all 4 items, students who used mental computation in the Choice condition relatively more often applied wrong or unclear procedures on that item in the No-Choice condition than students who used a written procedure in the Choice condition. Schematic high-level chunking was relatively more often used by students who already used a written procedure in the Choice condition. In addition, low-level chunking and high-level chunking without schematic notation were relatively more often used by students applying mental calculation in the Choice condition, except on item 2.

Mental strategies

Table 4.3 displays the distribution of the types of mental computation strategies used on items 1 to 9 presented in the Choice condition, based on the sample of 89 students who were interviewed. A clear difference between items 1 to 4 on the one hand, and items 5 to 9 on the other hand appears. Although on all items high-level chunking carried out mentally was the dominant strategy, this dominance was much more pronounced on items 5 to 9 (92% to 100% high-level chunking) than on items 1 to 4 (56% to 65% high-level chunking). On items 1 to 4, the students interviewed sometimes made a guess or numerical estimate, applied low-level chunking mentally, or used a wrong or unclear strategy. These strategies were almost never observed on items 5 to 9.

Another important observation is that within these mental calculation procedures, compensation strategies were used very often on items 5 and 7: on item 5 in 95% of the 43 mental strategies, and on item 7 in 75% of the 36 mental strategies. These high

TABLE 4.3 *Distribution of mental computation strategies on items in the Choice condition.*

strategy	it. 1	it. 2	it. 3	it. 4	it. 5	it. 6	it. 7	it. 8	it. 9
guess/estimate	.05	.06	.00	.27	.00	.00	.00	.00	.00
low-level	.15	.19	.22	.07	.00	.05	.00	.00	.04
high-level	.65	.56	.56	.56	.98	.95	.97	1.00	.92
wrong/unclear	.15	.19	.22	.10	.02	.00	.03	.00	.04
N (interviewed)	20	48	18	41	43	42	36	58	52

percentages are in contrast with the 13% and 7% of the written strategies on items 5 and 7, respectively. These differences in proportions of use of compensation between written and mental strategies are statistically significant on both items (for item 5, $\chi^2(1, N = 261) = 119.0, p < .001$, and for item 7, $\chi^2(1, N = 238) = 91.4, p < .001$).

4.3.2 Individual differences in strategy choices (second aim)

Thus far, strategy distributions were analyzed per item. In this section, we focus on individual differences in the extent to which students chose a written or a mental solution strategy on the items administered in the Choice condition. In the Choice condition, 38% of the students chose at least once for a mental solution strategy on the first 4 items. Considering all 9 items in the Choice condition, this percentage was 79%.

Multivariate analysis: Latent class models

Since each student chooses a strategy on each of the 9 items administered in the Choice condition, resulting data are multivariate. There are within-subject dependencies between the 9 strategy choice variables, and we need analysis techniques that can take these dependencies or correlations into account. As Hickendorff et al. (2009b) argued, latent variable modeling is appropriate. The latent variable (either categorical or continuous) accounts for the correlations within individuals by mapping the multivariate responses on the latent variable. Individual differences between students are allowed for as well, because each individual is allocated a particular position on the latent variable.

To analyze individual differences in the extent to which students chose written or mental procedures (on the items presented in the Choice condition), we used latent class analyses (LCA) (e.g., Goodman, 1974; Lazarsfeld & Henry, 1968. In LCA, the

latent variable is assumed to be categorical, representing latent (unobserved) classes or subgroups of subjects. The basic latent class model is $f(\mathbf{y}) = \sum_{k=1}^{K} P(k) \prod_i P(y_i|k)$. Classes run from $k = 1, \ldots, K$, and $\mathbf{y}$ is a vector containing the observed data: the strategy categorization (written/mental) on all items $i = 1, \ldots, 9$ presented to the student in the Choice condition. Resulting parameters are the class probabilities or sizes $P(k)$ and the conditional probabilities $P(y_i|k)$. The latter reflect for each latent class k the probabilities of solving item i with a written and a mental strategy, respectively. As latent class software, we used LEM (Vermunt, 1997), a general and versatile program for categorical data analysis. To decide on the number of latent classes underlying the data, we relied on the BIC-criterion. The BIC-value (Bayesian Information Criterion) is a model fit statistic that penalizes the fit of a model with the number of parameters estimated. Lower BIC-values represent a better trade-off between model fit and parsimony of the model.

In addition to searching for subgroups of students who could be characterized by a specific tendency to apply mental calculation on each of the items, we were interested in potential effects of background variables (gender and general mathematics level) on these class sizes. In other words: are the relative sizes of the latent subgroups different for boys and girls, and/or for students with a weak, medium or strong mathematics level? The existence of these kind of class size differences can be tested by inserting covariate(s) in latent class models, meaning that the covariate (such as gender) predicts class membership (Vermunt & Magidson, 2002). The LC-model with one observed covariate z can be expressed as $f(\mathbf{y}|z) = \sum_{k=1}^{K} P(k|z) \prod_i P(y_i|k)$. Now, class probabilities sum to 1, conditional on the level of the covariate, i.e. $\sum_{k=1}^{K} P(k|z) = 1$. The contribution of a covariate can be tested by a Likelihood Ratio (LR) Test statistic Λ, in which the improvement in fit of the model *with* the covariate relative to the fit of the model *without* the covariate is tested for statistical significance.

Model fitting steps

The 9 dependent variables of the LCA were, for each item in the Choice condition, whether it was solved by a written procedure or by mental calculation. Skipped items were coded missing. There were two explanatory variables: gender (2 categories) and general mathematics level (3 categories, based on the student's obtained percentile rank scores on the standardized mathematics test: weak students with a (population)

percentile rank of 33 or below, medium students with a percentile rank between 34 and 66, and strong students with a percentile rank of 67 or higher). Eight students had missing values on either or both of these explanatory variables, and were excluded from these analyses, so $N = 354$.

First, we had to decide how many latent classes to interpret. The model with 3 latent classes had the lowest BIC-value. Taking this 3-class model as the base model, our next step was to assess the effect of gender and general mathematics level (GML) on the sizes of these classes, by including these variables as covariates in the latent class model. Both general mathematics level ($\Lambda = 29.7$, $df = 4$, $p < .001$) and gender ($\Lambda = 38.3$, $df = 2$, $p < .001$) had a significant effect on latent class sizes. Furthermore, gender had a significant effect on the latent class formation independent from GML ($\Lambda = 22.6$, $df = 4$, $p < .001$). Finally, gender and GML did not interact in their prediction of latent class probabilities ($\Lambda = 2.0$, $df = 4$, $p = .74$).

Interpreting the best fitting model

So, we found that a model with 3 latent classes showed the best fit, and gender as well as general mathematics level significantly affected class sizes. First, the latent classes are characterized. Figure 4.2 shows, for each class separately, the probability to use mental calculation on each item. The first class, containing 18% of the students, was characterized by a high probability to apply mental calculation on each item. However, items 1 and 3 were less likely to be solved mentally than the other items. The third class (43% of the students) was characterized by low probabilities to use mental calculation, so these were students who mostly used written procedures. Item 8 had the highest probability to be solved by mental calculation, but this probability was still only .33. The second class (39% of the students) was in between the first and the third class. Students in this class were influenced by item characteristics in their choice between written procedures and mental calculation. On the first 4 items, their probability to use mental calculation did not exceed .17. However, on items 5 to 9 their tendency to use mental calculation was between .55 (item 5) and .85 (item 8).

All three classes showed a similar pattern of strategy choice over the first 4 items, reflecting that students adapted their strategy choices to problem features to some extent. Items 1 and 3 were least likely to be solved by written procedures, followed by items 2 and 4. Items 1 and 3 were whole number division problems, of which item 3 was

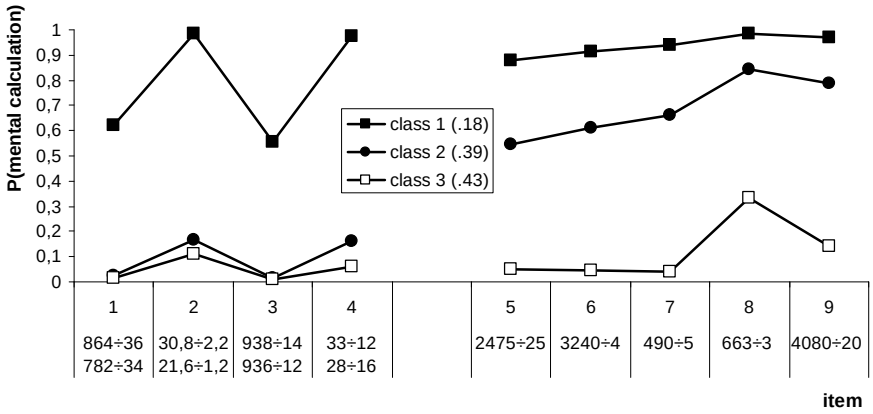


FIGURE 4.2 *Probability of applying mental calculation in 3 latent classes.*

designed to have the highest cognitive demands when doing mental calculation. Items 2 and 4 both dealt with decimal numbers, in the dividend and divisor or in the quotient, respectively. From the remaining 5 items, item 8 triggered mental calculation the most, also for students that were the least inclined to use mental calculation (class 3). This item may have had the least cognitive demands, because 663 can be divided by 3 on a single-digit basis. Although direct mapping of the divisor to the dividend was also possible on items 6 and 8, this may have had a little higher cognitive demands than on item 6, because it was not possible on a single-digit basis, but on a two-digit basis (i.e., 3240 could not be split up in single digits 3, 2, 4, and 0, but instead had to be split up in 32 and 40). Items 5 and 7 were special in the sense that if students did not use a compensation strategy, many steps would have to be taken, requiring high cognitive demands if they would do this mentally. However, when using a compensation strategy, fewer steps need to be taken, which can be well done mentally. So, choosing between mental and written strategies on these items may reflect whether a compensation strategy was used or not.

Table 4.4 shows the effects of gender and general mathematics level (GML), by presenting the class sizes conditional on the combinations of gender and GML[3]. In general, boys were much more likely to be classified in the first class of mainly mental

3 Although gender and GML did not interact in their prediction of class probabilities, class probabilities are still presented for combinations of gender and GML. This was done because boys and girls were not equally spread over the levels of mathematics achievement, which confounded with interpreting the effects of gender and GML separately.

TABLE 4.4 *Estimated class probabilities, conditional on gender and GML. Standard errors (SEs) between brackets.*

gender	GML	class 1 mental	class 2 ment/writ	class 3 written	*N*
	weak	.40 (.09)	.09 (.06)	.51 (.09)	27
boys	medium	.22 (.05)	.50 (.07)	.28 (.05)	66
	strong	.25 (.04)	.52 (.05)	.23 (.05)	100
	weak	.09 (.03)	.06 (.04)	.84 (.05)	49
girls	medium	.06 (.02)	.41 (.07)	.53 (.07)	64
	strong	.07 (.03)	.47 (.08)	.46 (.08)	48
total		.18	.39	.43	1.00

calculation than girls, while the majority of the girls could be characterized as consistently using written procedures. Boys were also slightly more often than girls classified in the class of students that combined written and mental calculation. With regard to general mathematics level, the main differences were found between weak students on the one hand, and medium or strong achieving students on the other hand. These differences predominantly arised in the second and third latent class. About half of the medium and strong students (boys and girls) combined written and mental calculation, while only 6 to 9% of the weak students did so. In contrast to switching between mental and written strategies, weak students were more inclined either to consistently use a written procedure (weak girls) or to consistenly use a mental procedure (weak boys). So, weak students adapted their strategy choices to problem features to a lesser extent than medium and strong students did.

Besides characteristics on the student level, school level characteristics might also have an effect on strategy choices and hence on latent class membership. Since there were 12 schools, entering school as another covariate in the latent class model would result in an unstable model. As an alternative, students were assigned to the class for which they had the highest posterior probability (modal assignment), yielding a classification of students. Relative class sizes differed between the 12 schools: the size of the first latent class (mainly mental) ranged from 0% to 28% between schools, the size of the second latent class (mental/written) ranged from 34% to 53%, and the size of the third latent class (mainly written) ranged from 26% to 55%. However, these school differences were not significant (Fisher's Exact Test $= 29.2$, $p = .12$), and neither were

class size differences with respect to mathematics textbook used (Fisher's Exact Test = 2.8, $p = .59$) or to indicators of social-economical status (Fisher's Exact Test = 3.8, $p = .40$).

4.3.3 Strategy accuracies (third aim)

The lower part of Table 4.1 presents descriptive statistics of the accuracy of each strategy. On the first 4 items in the Choice condition, accuracy of written strategies seems higher than of mental strategies, but this difference will be statistically tested below. Furthermore, comparing the accuracies on a particular item presented in the Choice condition with the accuracies on its parallel version presented in the No-Choice condition does not yield clear differences. However, in the strategy accuracies in the No-Choice condition of Table 4.1, no distinction was made with respect to the type of strategy (mental/written) chosen on the parallel version of that item in the Choice condition. This type of trial (student-by-item combination) information is crucial in the current analyses, because we only expect an accuracy difference between the two conditions on an item for students who chose a mental strategy in the Choice condition *on that particular item*. Only on these student-by-item combinations, strategy use had changed from a mental strategy in the Choice condition to a written strategy in the No-Choice condition.

Multivariate analysis: IRT-models

Strategy accuracy data are also multivarate. Therefore, latent variable modeling is again suitable. To analyze accuracy differences between mental and written strategies, we applied explanatory item response theory (IRT) analyses (De Boeck & Wilson, 2004; Rijmen et al., 2003). In IRT modeling, a continuous latent variable θ is introduced, usually interpreted as the latent ability of each subject. In the most simple IRT measurement model, the Rasch model, the probability of a correct response ($y = 1$) of subject p on item i can be expressed as a logistic (S-shaped) function of the difference between the latent ability of that subject θ_p and the item difficulty β_i, i.e., $P(y_{pi} = 1) = \frac{\exp(\theta_p - \beta_i)}{1+\exp(\theta_p - \beta_i)}$.

Such descriptive or IRT measurement models can be extended by an explanatory part (Rijmen et al., 2003; Wilson & De Boeck, 2004), meaning that covariates or predictor variables are included. These explanatory variables can be subject predictors, item predictors, or subject-by-item predictors. In the present analyses, there were two subject predictors, gender and general mathematics level. There were three predictors on the item level: item number (1 to 9) dummies, parallel item version (a or b), and condition

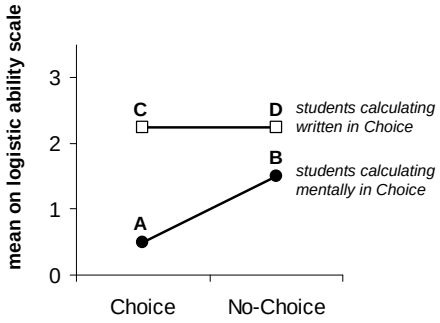


FIGURE 4.3 *Hypothesized group means on logistic latent ability scale for one item pair.*

of administration (Choice/No-Choice). Furthermore, there was one subject-by-item predictor: the strategy a subject used in solving an item (mental or written) in the Choice condition (see Hickendorff et al., 2009b, for further details on using strategy data in explanatory IRT-analyses).

We were interested in the effects of using a mental or a written strategy on the probability to solve an item correctly (i.e., the strategy accuracies), controlled for item difficulty level β_i and potential differences in difficulty level between the two versions of an item. Specifically, we had three hypotheses concerning performance on the item pairs, also presented graphically in Figure 4.3. First, we expected that in the Choice condition, the performance of students using a written strategy would be higher than that of students using a mental strategy, i.e. $\mathsf{C} > \mathsf{A}$ in Figure 4.3 (a between-subjects comparison). Second, we expected that performance of students who calculated mentally on an item in the Choice condition would be higher on the parallel item presented in the No-Choice condition, i.e. $\mathsf{B} > \mathsf{A}$ (a within-subjects comparison). Third, we hypothesized that performance of students who used a written strategy on an item in the Choice condition would be equal to their performance on the parallel item presented in the No-Choice condition, i.e. $\mathsf{D} = \mathsf{C}$ (a within-subjects comparison). In addition, we hypothesized that also on the unpaired items 5 to 9, the performance of students using a written strategy would be higher than that of students using a mental strategy (between-subjects comparisons). All explanatory IRT analyses in the present study were carried out in SAS (SAS Institute, 2002, see also De Boeck & Wilson, 2004).

Model fitting steps

We fitted a series of IRT-models, starting with the general model as illustrated in Figure 4.3. As a first step, several restrictions were imposed to make the model more parsimonious. The effects of condition (Choice/No-Choice) on accuracy were restricted to be equal for items 1 to 4: i.e., the difference between B and A in Figure 4.3 had to be the same for each of the items, and also the D - C difference had to be the same. This restriction resulted in a non-significant loss in model fit ($\Lambda = 3.6$, $df = 6$, $p = .74$), so this was a legitimate simplification of the model. In contrast, restricting the accuracy differences between written and mental strategies in the Choice condition to be equal for all items (i.e., the difference between C and A in Figure 4.3 had to be the same for items 1 to 9) did result in a significant decrease in model fit ($\Lambda = 44.6$, $df = 8$, $p < .001$). So, the differences between chosen mental and written strategies accuracies were item-specific.

As a second step, student characteristics gender and the standardized score on the mathematics achievement test were inserted as predictors in the model. Both variables appeared to have a significant effect on latent ability (for gender, $\Lambda = 8.9$, $df = 1$, $p = .003$; for mathematics achievement, $\Lambda = 176.3$, $df = 1$, $p < .001$). When mathematics achievement was incorporated in the model, adding gender did not have a significant effect on performance anymore ($\Lambda = .1$, $df = 1$, $p = .81$), so gender differences in general mathematics level mediated gender differences in performance on complex division. Mathematics achievement did have a significant effect on performance, as could be expected, and was explored further in the third step. Specifically, two interaction effects of mathematics achievement were tested. It was tested whether the within-subject effect of condition (No-Choice compared to Choice for mental calculators: B - A difference) was dependent on the mathematics level of the student, but this interaction effect was not significant ($\Lambda = .0$, $df = 1$, $p = .90$). However, mathematics achievement and strategy used in the Choice condition (mental/written, between-subjects) did have a significant interaction effect with each other on accuracy ($\Lambda = 17.4$, $df = 1$, $p < .001$). This implied that the item-specific differences between the accuracies of written and mental strategies (the C - A differences) depended on the mathematics achievement level of the student.

Interpreting the best fitting model

Results of the best fitting model are presented graphically in Figure 4.4. We start our interpretation with results on the first hypothesis: the relative accuracies of written and

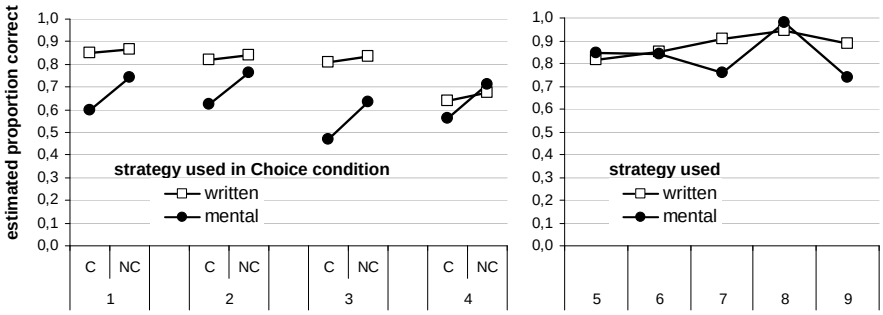


FIGURE 4.4 *Estimated probabilities to solve items 1 to 9 correctly for students at the mean level of mathematics achievement. Left plot: items administered in Choice as well as No-Choice condition, per item students who used mental calculation on that item in the Choice condition are separated from those who used a written procedure. Right plot: items only administered in Choice condition.*

mental strategies on the items in the Choice condition (between-subjects comparison). On more than half of the items, students choosing a written strategy were more accurate than those choosing a mental strategy, but these differences decreased with higher levels of mathematics achievement. For students at the mean level of mathematics achievement, written strategies were significantly more accurate than mental strategies on items 1, 2, 3, 7, and 9. On the remaining items (4, 5, 6, and 8), the accuracy difference between written and mental strategies was not significant for average achievers.

So, although selection effects on these strategy accuracy differences could not be ruled out completely (because they are based on between-subjects comparisons), we could correct for the mathematical achievement level of the students. We found that written strategies were at least as accurate as and usually more accurate than mental strategies. These differences, however, depended on the mathematics achievement level of students, being largest for weak students and smallest for strong students.

Within-subjects comparison of performance on parallel items showed results on the second hypothesis: for those who used a written procedure when they could choose, performance in Choice and No-Choice did not differ significantly from each other (in Figure 4.4, the lines with white squares on items 1 to 4 are almost horizontal). However, as formulated in the third hypothesis, for those who used mental calculation on an item

when they were free to choose, performance was higher in the No-Choice condition on the parallel item (when they were forced to use a written strategy). This difference was present on all four items (for items 1 to 4, the lines with black circles in Figure 4.4 show an ascending trend). The average effect of forcing mentally calculating students to write down their solution strategy is .68 (on the logistic scale), SE = .22, $p = .002$, effect size = .87. This large effect is equivalent to raising the probability of a correct answer from 50% to 66%. Furthermore, the size of the effect did not depend on the student's mathematics achievement level.

4.4 Discussion

The main results of this study were that when students were free to choose how they solved the presented problems on complex division, there were individual differences in these strategy choices. Specifically, about 20% of the students predominantly used mental calculation procedures on all items (more boys than girls), 40% quite consistently used a written procedure (more girls than boys), and the remaining 40% used mental calculation on more easy items but written procedures on the more difficult ones (almost no weak students). There were also individual differences in strategy accuracies. Mental calculation was less accurate than applying a written procedure on several items, especially for students having a weak mathematical level. More importantly, when students who used mental calculation to solve a problem were forced to write down their solution steps on a parallel item, they were usually capable of applying a written procedure, and their performance improved. This effect was unaffected by the level of mathematical achievement of the student. So, the present study showed that mental calculation may be a less accurate strategy than writing down notes or calculations on complex arithmetic problems, not only between but also within students and items. Therefore, encouraging students to make use of scrap paper in solving this type of complex division problem would probably improve performance.

4.4.1 Individual differences in strategy use

There was an association between the type of strategy used in the Choice condition (mental/written) and the types of written strategies used in the No-Choice condition. In particular, mental calculators used less structured written strategies when forced to write down their solution steps than written calculators did. This difference may be the

result of a more general student characteristic affecting strategy choice: the tendency to use algorithmic strategies (perhaps related to a larger emphasis on accuracy) versus the tendency to use more intuitive, less structured strategies. These different tendencies have been found in studies on gender differences in strategy use (Carr & Davis, 2001; Carr & Jessup, 1997; Timmermans et al., 2007). Similar gender differences in strategy choices were also found in the present study: it were mainly boys who relied on mental calculation, girls were much less inclined to do so.

Furthermore, there were differences between items in the extent that they triggered mental calculation, as came forward from the strategy frequencies (Table 4.1) as well as from the latent class analysis (Figure 4.2). Mental calculation was applied less often on the first 4 items than on the remaining 5 items (which had a smaller cognitive demand due to the number characteristics) to some extent. So, students showed that they adapted their strategy choices to problem characteristics, but individual differences were present in the general tendency to apply a mental or a written strategy. In addition, there were differences between students of different levels of mathematical performance in the extent to which they spontaneously adapted their strategy choices to problem features. There were almost no weak students combining mental and written procedures, while almost half of the students with a medium or strong mathematical level did switch to mental calculation on the easier items, adapting to the item characteristics. So, medium and strong students showed some flexibility in their strategy choices, while weak students did not, a result resembling that of Foxman and Beishuizen (2003) and Torbeyns et al. (2006) with mental calculation.

Another interesting finding with respect to the adaptivity of strategy choices concerns application of the compensation strategy on items 5 and 7. The number characteristics of these two items were such, that rounding the dividend would be a very efficient approach. We found only a small proportion of the written strategies making use of compensation, while, in contrast, the majority of the sample of mental strategies involved compensation. A possible explanation for this difference could be that students who were aware of the possibility of compensation given the number characteristics of these items, could solve these items by compensation in their head. In contrast, students unaware of this possibility of compensation required many more solution steps, for which scrap paper would be useful. Another explanation could be that a third variable, such as mathematical insight, influenced both the awareness of the possibility of compensation, and the skills needed to solve these problems mentally. In either way, the fact that those

applying a compensation strategy usually did it mentally, while those who did not use a compensation strategy predominantly used a written strategy, indicates that to some extent an adaptive strategy choice was made.

Another aspect of adaptivity is whether students adapted their strategy choice to their individual strategy accuracies. We can only discuss this topic with respect to accuracy of (forced) written strategies. There is evidence that several students (mainly boys) did not choose adaptively between written and mental calculation, because their performance with mental calculation was lower than when they were forced to write down their solution steps. This performance difference did not depend on the mathematics achievement level of the student, so even high achievers can be said to be unadaptive in this respect. However, it should be noted that adaptiveness in the present study could only be related to accuracy, and not to speed of strategy execution. It could well be that mental calculation is faster than written calculations, and that this plays an important role in students' choices as well.

4.4.2 Methodological considerations

In designing the present study, several methodological choices were made. First, parallel items were used to assess the effect of administration condition. By doing this, it was implicitly assumed that how students solved one version of the item in the No-Choice condition represents how the other item version (presented in the Choice condition) would have been solved in the No-Choice condition, and vice versa. This assumption cannot be tested, because students were never administered the same item version twice. However, because we counterbalanced item version over the administration conditions, we could assess that the parallel item versions did not differ significantly in difficulty level from each other, on any of the four items that had parallel versions.

Second, it was decided not to implement a No-Choice condition in which students would have had to calculate the answer to all items with a mental procedure, because we expected many students to struggle with obligatory mental calculation on these problems with large numbers. As a consequence of having only one No-Choice condition, the difference in accuracy between solving a problem with a mental or a written strategy could not be assessed completely unbiasedly, and therefore, conclusions about adaptivity of strategy choices with respect to individual strategy accuracies could only be drawn on the basis of results for the written strategy. However, we could correct for the students'

level of achievement in mathematics in general, ruling out this source of selection effects. In addition, we were able to assess whether the distribution and accuracy of the written strategies in the No-Choice condition were different for students using a mental or a written strategy when they were free to choose. Another disadvantage of having only one No-Choice condition, which was always preceded by the Choice condition, was that alternative explanations for the positive effect on performance of forcing the mental calculators to use a written strategy could not be ruled out. For example, accuracy may increase just from forcing students to use another strategy, irrespective of the particular strategy used, because they have to be more effortful and deliberate in executing their non-chosen strategy. Another explanation may be that by mental calculation, conceptual knowledge is activated that may have a beneficial effect on subsequent problem solving. Future research should include a No-Choice condition in which students would have to use a mental strategy, with a problem set requiring less cognitive effort.

A third methodological consideration was our procedure of classroom administration of the task, which had several consequences. An advantage was that the testing situation resembled that of the national assessments and classroom practice in general. However, a disadvantage was that it was not possible to gather data on strategy speed, because that would have required individual testing. Therefore, the speed of execution could not be incorporated as a strategy performance component. As a result, the adaptivity of strategy choices could only be assessed with respect to accuracy and not to speed, while speed probably is an important predictor of strategy choice as well. Moreover, students may weigh speed and accuracy differently (i.e., they may hold a different speed/accuracy trade-off), which may be an alternative explanation of found class size differences with respect to gender and mathematics level. Future research should take accuracy as well as speed into account to tap these issues.

Another consequence of classroom administration was that the experimental task could only be followed by the interviews one to several hours later. Robinson (2001) showed that retrospective reporting is a valid measure of cognitive processes in children's subtraction, supporting the validity of our procedure. However, it should be noted that in Robinson's (2001) study, children had to report on their solution strategy immediately after they had stated their answer to an item, on a trial-by-trial basis. In contrast, in the present study students completed all items before they were asked whether they could report how they solved the items. This time lag may have negatively affected the veridicality of the verbal reports, i.e., the reports may not have been accurate descriptions

of the strategy used due to forgetting and fabrication (Ericsson & Simon, 1993; Russo, Johnson, & Stephens, 1989). On the other hand, having the students verbally report only after all items were completed safeguarded against potential reactivity bias, i.e., the strategy choice and accuracies being affected by verbalization requirements (Russo et al., 1989). However, results on the interview data should be interpreted with caution. Future research should test students individually, and let them report on their strategy use immediately after answering each problem.

4.4.3 Educational implications

The most important implication of this study for school practice probably lies in promoting the value of writing down solution steps on more difficult complex arithmetic problems. As noted before, students nowadays are less inclined than students were a decade ago to use a written strategy in solving these kind of problems on complex division (Hickendorff et al., 2009b). In the present study we showed that both in comparisons between as well as within students, mental calculation may be less accurate than written calculation. That raises the question what role school practice plays in the strategy choices students make. It might be that the large emphasis on mental calculation in RME has had the side-effect that some students overuse mental calculation. A recommendation is that teachers emphasize these possible benefits of writing down notes or calculations to their students.

Another interesting finding was the near absence of the traditional algorithm for long division. Apparently, instruction regarding division was completely based on RME principles, at least in the 12 schools that were part of our sample. Instruction in the traditional algorithm has been discredited because it was said to be a mechanistic trick, in which understanding and insight in the numbers and their interrelations are not fostered. However, we argue that it might be possible to build in the traditional algorithm as the optimal form of abbreviation at the end point of the learning trajectory.

APPENDIX 4.A ITEM SET

item	version *a*	version *b*
1	Fanny takes piano lessons. She has to pay 782 euros for 34 lessons. How much does one lesson cost?	Marleen takes piano lessons. She has to pay 864 euros for 36 lessons. How much does one lesson cost?
2	The machine packs up 1.2 kilos of chocolate per minute. How many minutes does it take to pack up 21.6 kilos of chocolate?	The machine packs up 2.2 kilos of candies per minute. How many minutes does it take to pack up 30.8 kilos of candies?
3	FEEDING BOX SUITED FOR 12 COWS The farmer had 936 cows. How many of these feeding boxes does he need?	FEEDING BOX SUITED FOR 14 PIGS The farmer had 938 pigs. How many of these feeding boxes does he need?
4	16 children go to the playground. Together, they have to pay 28 euros. How many euros is that per child?	12 children go to the museum. Together, they have to pay 33 euros. How many euros is that per child?
5	Saskia sells DVDs for € 25 a piece. She received € 2475. How many DVDs did she sell?	
6	A cycle path of 3240 meters will be covered with concrete plates. Each plate is 4 meters long. How many plates are needed for the entire cycle path?	
7	Grandma divides 490 euros among her 5 grandchildren. How many euros does each grandchild get?	
8	The DVD-player costs € 663,-. Jasper pays this amount in 3 times, each time the same amount. How much does he have to pay each time?	
9	The farmer has 4080 liters of milk in his cooling basin. The milk is distributed over milk cans: each milk can is filled with 20 liters of milk. In total, how many milk cans will the farmer fill?	

Note. Items are translated from Dutch. Illustrations of items 5, 6, 8, and 9 are not shown.

CHAPTER 5

Solution strategies and adaptivity in complex division: A choice/no-choice study

This chapter is co-authored by Marije F. Fagginger Auer.

ABSTRACT

The current study systematically investigated mental and written solution strategies for solving complex division problems (e.g., $306 \div 17$), with the main focus on strategy adaptivity. Eighty-six Dutch 12-year-olds were tested using the choice/no-choice design. They first solved division problems in the free strategy choice condition, and consecutively with forced mental and forced written computation in the two respective no-choice conditions. Strategy choice and strategy performance (accuracy and speed) were recorded. Findings showed that mental computation was usually chosen for reasons of speed, while choices for written computation were fit to accuracy characteristics. Moreover, there were group differences regarding gender and mathematics achievement level of the student in the relative preference for accuracy and speed in choosing between mental and written strategies.

5.1 INTRODUCTION

Solution strategies for solving cognitive tasks have been an important psychological research topic. Especially solution strategies for mathematics problems have received considerable attention, since they are interesting both from a cognitive psychological perspective and from the viewpoint of mathematics education. Until recently, these studies were mostly limited to elementary addition, subtraction, and multiplication in the number domain up to 100. In contrast, complex arithmetic that is part of the curriculum of higher grades of primary school (i.e., operations with multidigit numbers for which one may use a written procedure) – and particularly complex division – has not received much research attention. However, systematic studies in complex division are needed, particularly in the Netherlands. Dutch national assessments showed a descending achievement trend on complex arithmetic in general, and on complex division in particular (J. Janssen et al., 2005), and this trend appears to be related to a shift in strategy use from written to mental strategies (Hickendorff et al., 2009b).

Therefore, the present study aims at a systematic investigation of the characteristics of mental and written solution strategies Dutch children at the end of primary school use to solve complex division problems, with a special focus on adaptivity: to what extent do the children choose the strategy (mental or written) with which they perform best? In the remainder of this section, we discuss research into solution strategies including strategy adaptivity, and previous research in the domain of complex division. We end this section with the design and aims of the current study.

5.1.1 Solution strategies

Children and adults know and use multiple strategies to solve cognitive tasks, including mathematics problems. The many studies into solution strategies for elementary addition and subtraction (e.g., Carr & Jessup, 1997; Carr & Davis, 2001; Torbeyns et al., 2002, 2004a, 2005), elementary multiplication (e.g., Anghileri, 1989; Imbo & Vandierendonck, 2007; Lemaire & Siegler, 1995; Mabbott & Bisanz, 2003; Mulligan & Mitchelmore, 1997), and mental multidigit addition and subtraction (e.g., Beishuizen, 1993; Beishuizen et al., 1997; Blöte et al., 2001; Torbeyns et al., 2006) have resulted in near consensus on the strategies used and the characteristics thereof for these mathematical domains.

Research into strategies for solving mathematics problems has been carried out in the field of cognitive psychology and in the field of mathematics education. Cognitive psychology acknowledges that arithmetic performance depends on the type of strategies that a subject uses (Lemaire, 2010). Within the cognitive psychological framework, the work of Siegler and his colleagues has been very influential (e.g., Lemaire & Siegler, 1995; Shrager & Siegler, 1998; Siegler, 1988a, 1988b). Lemaire and Siegler distinguished four dimensions of strategic competence on which individuals may differ: their strategy repertoire (which strategies are used), their strategy distribution (the frequency with which the strategies are used), their strategy efficiency or performance (strategy speed and/or accuracy), and their strategy selection or adaptivity (how strategies are chosen, related to problem and individual strategy characteristics). These four dimensions are central to the current study, with the main focus on the last one: strategy adaptivity.

Cognitive models of the underlying structures and mechanisms of strategy choice or adaptivity have been developed (Shrager & Siegler, 1998; Siegler & Shipley, 1995). In these models, an individual's strategy choice on a particular problem is for the largest part determined by the individual's strategy performance characteristics for that problem. According to Siegler and Lemaire (1997), people tend to choose their strategies adaptively: they choose the fastest and most accurate strategy for a given problem out of their strategy repertoire. Strategy speed and accuracy on a particular task may vary from individual to individual. However, not all research findings support this cognitive claim on the adaptivity of strategy choices. For example, suboptimal strategy choices have been observed in 2-digit addition and subtraction (Torbeyns, De Smedt, et al., 2009b) and in complex division (Hickendorff et al., 2010). Moreover, cognitive models on strategy

choice have been argued to ignore the influence of sociocultural context variables such as sociomathematical norms (Ellis, 1997; Luwel, Onghena, Torbeyns, Schillemans, & Verschaffel, 2009; Verschaffel et al., 2009).

From the perspective of mathematics education, solution strategies are important in the international reform movement (e.g., Kilpatrick et al., 2001) for at least two reasons. First, the didactics for solving complex arithmetic problems have changed, from instructing standard written algorithms to building on children's informal strategies (Freudenthal, 1973; Treffers, 1987, 1993), and mental computation 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). It is worth mentioning at this point that the terms 'adaptivity' and 'flexibility' are used with different meanings by different authors (for a discussion, see Heinze, Star, & Verschaffel, 2009, and Verschaffel et al., 2009). In the present study, adaptivity is defined with respect to both individual strategy performance characteristics and task characteristics, in the following way: to what extent does a child choose the strategy that is the most appropriate or efficient for him or her on a given problem?

Although this conceptualization of strategy adaptivity is used in the literature (e.g., Heinze, Star, & Verschaffel, 2009; Star & Newton, 2009; Torbeyns, De Smedt, et al., 2009b), it is not particularly well-defined, because what constitutes 'appropriate' or 'efficient' is ambiguous. These terms usually refer to the performance of a strategy, but there are at least two components to strategy performance: accuracy and speed. Problems arise when the most accurate strategy on a problem is not the fastest. For example, backup-strategies are slower but can be more accurate than retrieval (e.g., Kerkman & Siegler, 1997; Lemaire & Siegler, 1995; Siegler & Lemaire, 1997), and on complex arithmetic problems it has been suggested that written strategies are more accurate but slower than mental strategies (Hickendorff et al., 2010). In these instances, an adaptive strategy choice is not univocally defined. Some researchers leave the relative importance of accuracy and speed rather unspecified by defining the most efficient strategy as *the fastest and most accurate* (e.g., Lemaire & Callies, 2009; Siegler & Lemaire, 1997). Obviously, such a definition does not accommodate for the situations where one strategy is faster, but another strategy is more accurate. Other researchers defined the most efficient strategy as the one *leading fastest to the correct answer* (e.g., Kerkman & Siegler, 1997; Luwel et al., 2009; Torbeyns, De

Smedt, et al., 2009b; Torbeyns et al., 2004a, 2005, 2006) or the strategy that *produces the most beneficial combination of speed and accuracy* (Verschaffel et al., 2009). Although the latter definitions combine accuracy and speed, in the operationalizing analyses by these researchers accuracy and speed were generally not considered simultaneously but separately instead (an exception is the study by Torbeyns et al., 2005).

So, the relative importance of accuracy and speed plays a role in situations where the most accurate strategy is not the fastest. Moreover, individuals may differ in their relative favoring of accuracy and speed: they may have different speed-accuracy preferences (Ellis, 1997; Phillips & Rabbitt, 1995). In other words, they may differ in which combination of speed and accuracy they find most beneficial. Such considerations have not received much research attention in the research on strategy adaptivity. For elementary cognitive tasks, Siegler (1988a) discusses individual differences in the strength of the confidence criterion (i.e., the certainty required for stating an answer from retrieval), which relates to the individual differences in motivation to make few errors. In such elementary cognitive tasks backup strategies are clearly slower but evenly or more accurate than retrieval. Such clear performance differences do not necessarily exist in more complex cognitive tasks, in which relative strategy accuracy and speed may differ from individual to individual. In the current study, we try to gain insight into different patterns in the relative favoring of strategy accuracy and strategy speed in complex division problem solving. Such insight may have important educational implications, since instruction may be adapted to these individual differences. For instance, students who favor speed over accuracy may be encouraged to work slower but with fewer errors.

5.1.2 Complex division

In the present paper, complex division is defined as division problems in which the quotient is a multidigit number (e.g., $872 \div 4 = 218$)[1], and the divisor may be multidigit too (e.g., $306 \div 17 = 18$). This contrasts with simple division (division problems from the multiplication tables), in which the quotient is a single digit (e.g., $48 \div 8 = 6$; Robinson et al., 2006). Compared to addition, subtraction, and multiplication, the domain of (complex) division is understudied thus far. However, systematic studies on complex division are needed, particularly at the end of primary school in the Netherlands, for at least two reasons.

1 We consider a decimal number (e.g., $34 \div 4 = 8.5$) as multidigit too.

First, the most recent Dutch national assessment showed a large decline in sixth graders (12-year-olds) performance on complex division problems over a period of two decades (J. Janssen et al., 2005). Second, mathematics education reform has had a considerable impact on instruction in complex division. Under the influence of Realistic Mathematics Education (RME) the traditional long division algorithm has disappeared from mathematics textbooks, and has been replaced by more informal strategies based on repeatedly adding or subtracting multiples of the divisor (Freudenthal, 1973; Treffers, 1987). Figure 5.1 presents examples of such repeated addition and subtraction strategies, that differ in their level of abbreviation (i.e., the number of steps taken), see also Hickendorff et al. (2010) and Van Putten et al. (2005). Moreover, the traditional long division algorithm (and its notational form in the Netherlands and the US) is also presented. In addition to this shift in instruction in written strategies, another characteristic of the reform is that mental arithmetic plays a central role in mathematics education (Blöte et al., 2001). In 2004, nearly all Dutch primary schools used mathematics textbooks based on RME principles (J. Janssen et al., 2005), although a return to more traditionally oriented mathematics textbooks has been observed recently (KNAW, 2009).

In a recent study, secondary analyses on the student materials of the two most recent national assessments of 1997 and 2004 were carried out, aiming to relate the achievement decline on complex division to (changes in) the solution strategies used (Hickendorff et al., 2009a, 2009b). Results showed that two changes appeared to have contributed to the decline. First, strategy use had shifted: use of written procedures decreased (attributable to a decrease in the use of the traditional long division algorithm), while an increasing percentage of the students (more boys than girls) predominantly answered without calculations written down on scrap paper. This strategy shift was unfortunate, since answering these problems with a nonwritten strategy was less accurate than using a written strategy. Second, each of the solution strategies yielded less correct answers in 2004 than in 1997, with approximately the same amount of accuracy decrease per strategy. So, the performance decline over time on complex division in the Netherlands seems to be related to a change in strategy choice – in particular to a decrease in the use of written strategies – and to a general decrease in strategy accuracy as well.

These strategy change results of the national assessment data were descriptive by nature and therefore limited in several aspects, among which possible selection effects (cf. Siegler & Lemaire, 1997). That is, because strategy choice is probably influenced by the ability of the student and/or difficulty of the item, the strategy accuracy estimates

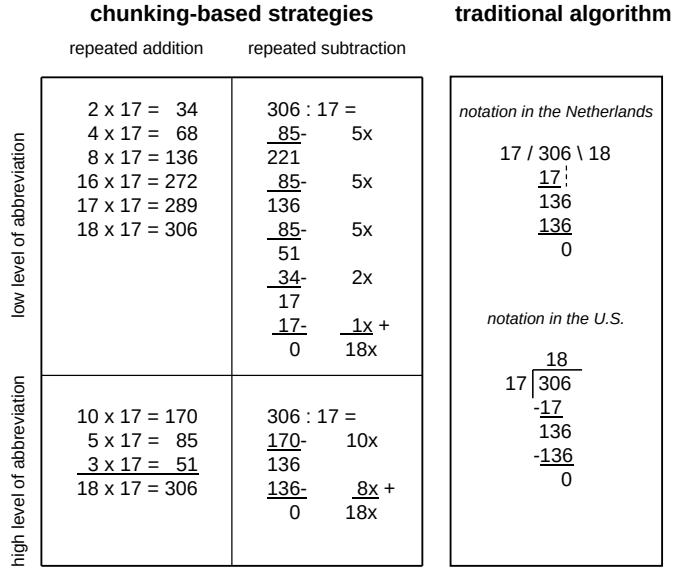


FIGURE 5.1 *Examples of solution strategies for the problem 306 ÷ 17.*

may have been biased by these student and item selection effects. To overcome this, Hickendorff et al. (2010) studied strategy choice in an experimental test design, in which sixth graders had to solve division problems under two conditions: free choice between mental and written computation, and forced written computation. One of the main findings was that accuracy of students who used mental calculation on a particular item in the free choice condition, improved by requiring the use of a written strategy in the forced written condition. So, these findings suggest that these choices for mental strategies were counter-adaptive with regard to accuracy. However, the methodology of this study hampered drawing conclusions on adaptivity rigorously for two reasons. First, data were collected on strategy accuracy but not on strategy speed, so only one aspect of strategy performance could be accounted for. Second, unbiased strategy characteristics (i.e., accuracies) were gathered only for written strategies and not for mental strategies, since it was deemed to be too demanding for a large number of students to solve complex division problems with large numbers with obligatory mental calculation. As a result, unbiased strategy characteristics of only one of the two strategies could be used in

assessing adaptivity.

The current study extends the finding of Hickendorff et al. (2010) in a follow-up experiment in which these two main methodological limitations are overcome. Specifically, we also included speed measures in addition to accuracy data, as well as a condition in which students were forced to use mental computation in addition to a forced-written strategy use condition. In order to prevent students becoming frustrated from having to solve quite difficult problems in their heads, the presents study's division problems were designed to be somewhat less cognitively demanding compared to the ones used in Hickendorff et al. (2010).

5.1.3 The current study

The present study's aim is to systematically investigate the four dimensions of strategic competence in the domain of complex division problem solving (repertoire, distribution, performance, and adaptivity), distinguishing between mental and written computation strategies. Particularly the fourth dimension, adaptivity of the strategy choices, received special attention: to what extent do individual strategy performance characteristics (accuracy and speed) predict the choice of a strategy? We expected that different patterns in preference for accuracy and speed would be present, giving rise to different patterns of strategy adaptivity. Such findings may have implications relevant for educational practices, since students favoring speed over accuracy may require another instructional approach than students favoring accuracy over speed.

The main focus regarding solution strategies was on the distinction between written and mental computation, because secondary analyses on Dutch national assessments at the end of primary school showed that this was a very relevant distinction with respect to the observed decrease of performance over time: use of mental strategies increased over time, but their success rates lagged far behind those of written strategies (Hickendorff et al., 2009b). In the present study, mental computation was defined as carrying out arithmetical operations without the use of any recording devices such as pen and paper, similar to the definitions of for example Hickendorff et al. (2010), Ruthven (1998), Siegler and Lemaire (1997), and Timmermans et al. (2007), but unlike other studies (e.g., Beishuizen et al., 1997; Blöte et al., 2001; Torbeyns, De Smedt, et al., 2009b). Written strategies included all forms of calculations in which some part of the solution process was written down on paper, ranging from only recording intermediate solution

steps to written algorithmic procedures. In order to be able to extend conclusions of the present experiment to the earlier studies on strategy use on complex division problems, participants and problems were chosen to resemble those of the Dutch national assessments. That is, students at the end of primary school (sixth graders) were selected, and we devised division problems presented in a realistic context involving multidigit quotient and/or divisor.

To estimate strategy characteristics in an unbiased manner, we used Siegler and Lemaire's (1997) *choice/no-choice* methodology. Each participant solved three parallel series of division problems under three different conditions. They first solved a series of division problems in the free strategy choice condition, and consecutively solved the two parallel series with forced mental and forced written calculation in the two respective no-choice conditions. From these two no-choice conditions, individual accuracy and speed characteristics of written and mental computation strategies were assessed without selection effects. Several studies in mathematics have applied the choice/no-choice methodology to study solution strategies (for an overview, see Luwel et al., 2009). An important feature of this design is that the adaptivity of strategy choices can be evaluated on an individual level, i.e., whether a subject chooses that strategy that *for him or her* is most efficient.

In addition to assessment of the four dimensions of strategic competence in general, we searched for effects of the student characteristics gender and general mathematics level. Gender differences in strategy use have been reported frequently. For example, girls have been found to have a larger tendency than boys to (quite consistently) use algorithmic strategies instead of using more intuitive, less structured strategies (Carr & Davis, 2001; Carr & Jessup, 1997; Hickendorff et al., 2009b, 2010; Gallagher et al., 2000; Timmermans et al., 2007). In contrast, gender differences in strategy adaptivity not been studied often, thus far. However, the findings of Hickendorff et al. (2010) showed boys making less adaptive strategy choices than girls, at least regarding accuracy, and it was suggested that girls and boys may weigh the importance of accuracy and speed differently. Regarding mathematics achievement level, it has been frequently (but not uniformly, see Torbeyns et al., 2005) reported that students of higher mathematical ability choose more adaptively between strategies than students of low mathematical ability (Foxman & Beishuizen, 2003; Hickendorff et al., 2010; Torbeyns, De Smedt, et al., 2009b; Torbeyns et al., 2002, 2006). So, we expect to find the same pattern in the current study.

5.2 METHOD

5.2.1 Participants

The participants were 86 students in the sixth grade (12-year-olds). They originated from 9 Dutch primary schools located either in the city or in a more rural area. All schools used a mathematics textbook based on RME principles, but they did not use the same textbook. In the sample, there were 43 girls and 43 boys. Also, information about the general mathematical level of the students was obtained: the students most recent level on CITO's Student Monitoring System mathematics test, a national standardized measurement instrument (in which speed of performance is not important) yielding a norm-referenced mathematics score, that we categorized into 2 levels: above the average of the norm group and below the average of the norm group.

5.2.2 Material

Three parallel sets of four complex division problems each were constructed, resulting in a total of 12 problems (see Appendix 5.A). These problems were designed to resemble those that students encounter in their classroom and testing practices. Each problem was presented within a realistic context: a situation that described a hypothetical real life mathematical problem. For each item, three parallel versions were constructed that were as similar as possible to each other with respect to number characteristics and realistic context, but that at the same time would not be perceived as identical problems to prevent practice effects. The 3 parallel sets were counterbalanced over the 3 conditions (choice, no-choice mental, and no-choice written; see below).

The number characteristics of each item set were as follows. In the first item set, the outcome was below 10, but students had to deal with a remainder. In the second item set, numbers were such that a compensation approach (rounding the dividend; e.g., $1089 \div 11 = 1100 \div 11 - 1 = 100$ - $1 = 99$) would be efficient. In the third item set, the dividend and divisor were decimal numbers (while the outcome was not). Finally, in the fourth item set a 3-digit number had to be divided by a 2-digit number, with outcome also a 2-digit number.

5.2.3 Procedure

Per school, six to twelve students were randomly selected for participation. The students were tested individually in a quiet room outside their classroom. They were told that they would be given twelve division problems to solve. Each student was first tested in the free choice condition (C) and then in two no-choice conditions: forced mental calculation (NC-M) and forced written calculation (NC-W). The order of the 2 no-choice conditions was counterbalanced over students. All problems were presented one by one, and solution times were collected with a stopwatch on a trial-by-trial basis. The students received the following instruction: *'With this stopwatch, I will register what time you need to solve the problems, but you can take as much time as you need on each problem.'*

5.2.4 Conditions

The first four division problems were presented in the free choice condition. On these problems, students were free to choose whether they solved them by mental or written calculation. There was a pencil available for the student to use and space for writing down calculations in the booklet. At the end of this first set of problems, the children were asked to report verbally on the strategies they used on the problems that they solved by mental calculation.

In the no-choice mental condition, another parallel set of four problems was presented. The procedure was similar to the choice condition, except for the fact the students could not use paper and pencil in doing their calculations and thus were forced to use mental calculation. In addition, the students were asked to report on their calculation strategy verbally after each problem was solved.

In the no-choice written condition, the final set of four problems was presented. In this condition, students had to write down their calculation procedure and were thus forced to use written calculation.

5.2.5 Responses

For each trial, the following responses were coded: (a) the accuracy of the answer given, (b) the solution time (ST), (c) the main strategy used, mental or written calculation (only in choice condition), and (d) the type of written or mental solution strategy used.

The type of written or mental strategy used was coded to get more insight into the rather broad categorization into mental and written strategies. The types distinguished

were (a) repeated addition or subtraction of multiples of the divisor (see left part of Figure 5.1), (b) traditional algorithm for long division (see right part of Figure 5.1), (c) wrong procedure: e.g., multiplication of dividend and divisor, numerical estimation and splitting up the divisor (e.g., solving $306 \div 17$ by $306 \div 10 + 306 \div 7$), and (d) unclear procedure or student did not remember. The type of written strategy was inferred from the solution steps that were written down. The type of mental strategy was inferred from the verbal reports.

5.2.6 Statistical Analyses

Because item responses were nested in individual students, observations were not independent. To account for these correlated responses we used (generalized) linear mixed (also called hierarchical, multilevel or random effects) regression models (e.g., Hedeker & Gibbons, 2006; Snijders & Bosker, 1999). All estimated models were random intercepts models, in which individual differences were accounted for by the intercept being random over students. The continuous dependent variable 'solution time' was analyzed with linear mixed models using the SAS procedure MIXED. Because solution times deviated from normality (the smallest z-value of skewness of solution times on the 12 problems was 4.32), solution times were log-transformed before entering the analyses (cf. Klein Entink, Fox, & Van der Linden, 2009). The binary dependent variables strategy choice (mental/written) and accuracy (incorrect/correct) were analyzed by mixed binary logistic regression models using the SAS procedure NLMIXED. The statistical significance of predictor effects was tested using a likelihood ratio (LR) test. The LR-test statistic is computed as two times the difference between the log-likelihood of the model with and the model without the predictor effect, and is asymptotically χ^2 distributed with df the number of parameters associated with the predictor effect.

5.3 RESULTS

Preliminary mixed linear (speed) and logistic (strategy choice and accuracy) regression analyses showed that that the three parallel item sets A, B, and C did not differ in proportion mental calculation (choice condition only; LR $= 1.0$, $df = 2, p > .05$) nor in average accuracy (LR $= .0$, $df = 2, p > .05$), but the effect of item set on speed did just reach significance (LR $= 6.1$, $df = 2, p = .048$). Furthermore, the order of the no-choice conditions did not affect accuracy (LR $= .1$, $df = 1, p > .05$) or speed (LR $= .6$,

$df = 1, p > .05$). In the main analyses, data were grouped over the versions of the item set and order of no-choice conditions, but in the speed analyses we statistically controlled for the item set version.

The main results are discussed in three sections: (a) repertoire and distribution of strategies in the choice condition, (b) strategy performance data (accuracy and speed) from the choice as well as from the two no-choice conditions, and (c) results on adaptivity of strategy choices.

5.3.1 Strategy repertoire and distribution in choice condition

Almost half of the students (42 students, 49%) used mental calculation and written calculation at least once. The remaining students used either written calculation on all items (33 students, 38%) or mental calculation on all items (11 students, 13%).

On 67% of all trials a written strategy was chosen. Girls chose a written strategy ($M = 80\%$) significantly more often than boys ($M = 53\%$), LR $= 13.0$, $df = 1, p < .001$. The difference between below-average ($M = 70\%$ written strategies) and above average mathematics achievers ($M = 64\%$ written strategies) was not significant, LR $= .5$, $df = 1, p > .05$. Finally, the four items significantly differed in percentage of written strategies, LR $= 34.3$, $df = 3, p = .001$. Post-hoc pairwise comparisons showed that on item 4 ($M = 84\%$ written strategies) significantly more written strategies were used than on item 3 ($M = 67\%$, $t(85) = 3.20$, $p = .002$), item 2 ($M = 59\%$, $t(85) = 4.40$, $p < .001$), and item 1 ($M = 57\%$, $t(85) = 4.55$, $p < .001$), respectively. All other pairwise comparisons were not significant.

Table 5.1 shows the distribution of each type of strategy used in the choice condition averaged over the four items. The distribution of repeated addition/subtraction strategies and the traditional algorithm on the one hand, and applying the wrong procedure or unclear strategy on the other hand, was significantly different for mental and written strategies (LR $= 15.8$, $df = 1, p < .001$). Within written as well as within mental strategies, repeated addition/subtraction strategies were dominant with 75% and 84% respectively. The traditional algorithm was used very infrequently (on 5% of all trials), and if it was used it was only within written strategies. Furthermore, executing the wrong procedure was more prevalent in mental strategies (17%) than in written strategies (5%). The same pattern holds for unclear strategies: 8% of the mental strategies, and 3% of the written strategies. Another interesting result (not presented in Table 5.1) was on the prevalence

TABLE 5.1 *Distribution of type of strategies used in choice condition.*

type of strategy	within mental strategies		within written strategies		total	
repeated addition/subtraction	86	(75%)	192	(84%)	278	(81%)
traditional algorithm	0	(0%)	18	(8%)	18	(5%)
wrong procedure	19	(17%)	12	(5%)	31	(9%)
unclear	9	(8%)	6	(3%)	15	(4%)
total number of trials	114	(100%)	228	(100%)	342	(100%)

Note. On two trials the student did not give an answer, so the strategy used could not be determined. As a result, the total number of trials equals 342.

of the compensation strategy on item 2 (a strategy in which the dividend was rounded; this was a specific form of the repeated addition/subtraction category): this shortcut strategy constituted 63% of all 35 mental strategies on this item, but only 12% of the 54 written strategies, a significant difference (z-test for proportions = 4.68, $p < .001$).

5.3.2 Strategy performance

Choice condition

Table 5.2 shows strategy performance data (accuracy and speed[2]) in the choice condition, by gender and general mathematics level. On average, the accuracy difference between mental strategies and written strategies was border significant (LR = 3.8, $df = 1, p = .051$). The speed difference was highly significant (LR = 108.8, $df = 1, p < .001$), with mental strategies being faster than written strategies.

Boys and girls did not differ significantly in average total accuracy in the choice condition (LR = .2, $df = 1, p > .05$), nor in accuracy within mental or within written strategies (LR = 3.5, $df = 1, p > .05$). Boys were significantly faster on average (LR =11.3, $df = 1, p < .001$), but within each strategy choice the gender difference in speed was not significant anymore (LR = 2.1, $df = 1, p > .05$). So, difference in strategy choice

2 All speed data presented in the Results are based on all trials (correct and incorrect ones), because we argue that this presents a more complete picture than presenting only speed of correctly executed strategies. However, we also analyzed speed data based on only the correct trials. Results were very similar, with the exception that correct responses were faster.

TABLE 5.2 *Strategy performance in the choice condition, by gender and general mathematics level.*

	accuracy (P (correct))			speed (ST in seconds)		
strategy choice	mental	written	total	mental	written	total
girl	.41	.57	.54	44.0	105.6	93.3
boy	.56	.68	.63	47.4	84.8	67.4
< average math level	.23	.48	.40	66.4	114.1	99.6
> average math level	.77	.77	.77	29.1	78.1	60.3
total	.52	.61	.58	46.4	97.2	80.3

accounted for gender differences in speed in the choice condition: boys were faster on average because they chose fast mental calculation more often than girls.

The effect of general mathematics level of the student was highly significant on average accuracy (LR = 29.0, $df = 1, p < .001$) as well as on average speed (LR = 18.6, $df = 1, p < .001$). Moreover, accuracy differences within mental and written strategy choices were also significant (LR = 32.0, $df = 1, p < .001$), as were speed differences within the two strategies (LR = 20.5, $df = 1, p < .001$). Below-average achievers had a lower proportion correct and were slower than above-average achievers within the two strategies as well as totaled over the strategy choices. The difference in performance between below-average and above-average achievers was the same in mental strategies as in written strategies, since the interaction between general mathematics level and strategy choice was not significant on either accuracy (LR = 3.3, $df = 1, p > .05$) or speed (LR = 2.9, $df = 1, p > .05$).

No-choice conditions

Table 5.3 shows strategy performance data (accuracy and speed) from the two no-choice conditions, by gender and general mathematics level. No-choice condition had a significant effect on accuracy (LR = 11.8, $df = 1, p < .001$) as well as on speed (LR = 46.9, $df = 1, p < .001$). These unbiased strategy performance data thus showed that forced mental strategies were less accurate but faster than forced written strategies.

The accuracy difference between boys and girls within each condition was not significant (LR = 2.8, $df = 1, p > .05$). By contrast, gender did have a significant effect

TABLE 5.3 *Strategy performance in the no-choice conditions, by gender and general mathematics level.*

	accuracy		speed	
	P (correct)		ST in seconds	
condition	NC-M	NC-W	NC-M	NC-W
girl	.46	.55	92.4	92.5
boy	.55	.67	67.2	90.4
< average math level	.25	.43	102.8	115.1
> average math level	.76	.81	55.8	66.9
total	.50	.61	79.6	91.4

on speed (LR = 4.7, $df = 1, p = .031$) with boys being faster than girls. Moreover, the interaction between gender and no-choice condition (mental versus written) was also significant (LR = 6.6, $df = 1, p = .010$). Post-hoc contrasts showed that boys were significantly faster than girls in the no-choice mental condition ($t(85) = 3.06$, $p = .003$), but that the gender difference in speed in the no-choice written condition was not significant ($t(85) = .96$, $p > .05$). Moreover, for boys ($t(85) = 6.97$, $p < .001$) as well as for girls ($t(85) = 3.11$, $p = .003$) forced mental strategies were significantly faster than forced written strategies.[3]

The effect of general mathematics level of the student was highly significant on accuracy (LR = 63.6, $df = 1, p < .001$) as well as on speed (LR = 27.1, $df = 1, p < .001$). Students with below-average mathematics level had a significantly lower proportion correct than above-average achievers, in both no-choice conditions (interaction between mathematics level and no-choice condition on accuracy not significant; LR = 2.3, $df = 1, p > .05$). Regarding speed, below-average achievers were significantly slower than above-average achievers, regardless of the strategy they had to use (interaction between mathematics level and no-choice condition on speed not significant; LR = .9, $df = 1, p > .05$).

3 Although for girls the mean solution times in the two no-choice conditions (92.4s. and 92.5s.) did not seem to differ, these means were influenced by the skewness of the distribution of raw solution times. Log-transformed STs were not affected by skewness, and the mean in the no-choice mental condition was significantly lower (4.14) than the mean in the no-choice written condition (4.36).

5.3.3 Strategy adaptivity

Comparing the strategy performance data from the no-choice conditions with the strategy choice made in the choice condition gives information on the adaptivity of the strategy choice. To what extent was the most appropriate strategy chosen, as evidenced from the individual strategy performance data from the no-choice conditions?

The issue of strategy adaptivity is approached in three ways. In the first two approaches, analyses were done on the item level (aggregating over students), and in the third approach they were done on the student level (aggregating over items) (cf. Luwel et al., 2009). In these latter analyses, group differences with respect to gender and mathematics achievement level were studied as well.

Adaptivity at the item level

In analyzing adaptivity at the item level, we ask the following question: Is the performance difference between forced mental and forced written strategy on an item in accordance with the strategy choice made in the choice condition on the parallel item? First, accuracy and speed are dealt with separately.

For the mental strategy choices in the choice condition, there was no difference in accuracy rates between forced mental ($M = .57$) and forced written computation ($M = .57$) on (the parallel versions of) that item ($t(85) = .00$, $p > .05$). So, regarding accuracy, these mental strategy choices were neither adaptive nor counter-adaptive. In contrast, these mental strategy choices were adaptive to speed: for instances in which a mental strategy was chosen in the choice condition, forced mental strategies were significantly faster ($M = 53.3$ seconds) than forced written strategies ($M = 84.7$ seconds), $t(85) = 8.00$, $p < .001$.

For written strategy choices in the choice condition, the forced mental strategy was significantly less accurate ($M = .48$) than forced written computation ($M = .64$) on the parallel items in the no-choice conditions, $t(85) = 4.07$, $p < .001$. So, the choices for a written strategy were adaptive regarding accuracy. In contrast, these choices were counter-adaptive to speed: for instances in which a written strategy was chosen in the choice condition, forced mental strategies ($M = 93.2$ seconds) were significantly faster than forced written strategies ($M = 94.8$ seconds), $t(85) = -3.31$, $p = .001$.

In short, these separate analyses on accuracy and speed suggest that on average, mental strategy choices seem adaptive to speed considerations as evidenced from the

solution time differences between the no-choice conditions (mental strategies being faster), while there were no differences in unbiased accuracy characteristics. In contrast, written strategy choices seem adaptive to accuracy, but counter-adaptive to speed.

In the next approach, accuracy and speed were combined on a trial-level basis. Following Lemaire and Siegler (1995), an adaptive strategy choice was defined as choosing the strategy that leads the individual fastest to an accurate answer. To operationalize this definition, for each trial in the choice condition we combined accuracy and speed information from the two no-choice conditions, and compared it to the strategy choice made in the choice condition, similar to Imbo and LeFevre (2009).

There were three possible categories. First, a strategy choice was coded as *adaptive* either when (a) a correct answer was obtained on an item with both forced mental and forced written calculation and the fastest of these two strategy was chosen, or (b) when a correct answer was obtained in only one of the no-choice conditions and the strategy that had yielded the correct answer was chosen. For example, for a student who obtained the correct answer in the NC-W condition but an incorrect answer in the NC-M condition, choosing a written strategy in the choice condition was coded as an adaptive strategy choice. For these latter trials, potential differences in speed of the two strategies did not play a role in coding adaptivity: accuracy was deemed more important and therefore decisive. Second, a strategy choice was coded as *counter-adaptive* if (a) a correct answer on an item was obtained with both forced mental and forced written calculation and the strategy that was slowest was chosen, or (b) when the correct answer was obtained in only one of the no-choice conditions and the strategy that had yielded the incorrect answer was chosen. Finally, a strategy choice was coded *indeterminate* when a student answered the item in both no-choice conditions incorrectly. In these instances it is hard to think of adaptivity, since neither of the strategies yielded a correct answer, and hence could never lead to an adaptive choice. Results showed that strategy choices were adaptive on 43% of the items and counter-adaptive on 30% of the trials. In addition, 28% of the strategy choices were indeterminate with respect to adaptivity. Moreover, for each individual student summing the adaptivity scores over the 4 trials in the choice condition, showed that there was substantial variation between students. To illustrate, 51 students (59%) made an adaptive as well as a counter-adaptive choice at least once.

Adaptivity at the student level

In the preceding section results were aggregated over students, obscuring individual variations in accuracy differences and speed differences. In this section, we took these individual differences into account by analyzing adaptivity at the student level. Accuracy and speed were treated separately, by computing the correlation between the frequency of mental calculation of a student in the choice condition and the differences in accuracy (total number correct) and in speed (total log-transformed solution time) between the two strategies from the no-choice conditions (cf. Torbeyns, De Smedt, et al., 2009b).

Spearman's ρ correlation between frequency of mental calculation (choice condition) and the difference in the total number correct between no-choice mental and no-choice written conditions was positive and significant ($\rho = .28$, $df = 84, p = .009$), indicating that students took into account which of the two strategies was most accurate for them. Gender seemed to affect this correlation ($\rho_{\text{girls}} = .40$, $df = 41, p = .008$; $\rho_{\text{boys}} = .26$, $df = 41, p > .05$) but the difference was not significant ($z = .68, p > .05$). Mathematics level did have a significant effect on the correlation ($z = 2.85$, $p = .004$), $\rho_{\text{below}} = .00$, $df = 42, p > .05$; $\rho_{\text{above}} = .56$, $df = 40, p < .001$.

With respect to speed, Spearman's ρ correlation between frequency of mental calculation and differences in solution time between forced mental and written strategies was also significant ($\rho = -.32$, $df = 84, p = .002$). Note that the correlation is negative because solution times are inversely related to speed, so that this result indicates that students took into account their individual strategy speed characteristics. This time, gender had a significant effect on this relation ($z = 2.29$, $p = .022$): $\rho_{\text{girls}} = -.07$, $df = 41, p > .05$; $\rho_{\text{boys}} = -.52$, $df = 41, p < .001$. Although the effect of mathematics achievement level on the size of this correlation was not significant ($z = 1.62, p > .05$), the correlation was not significantly different from zero for below-average achievers ($\rho_{\text{below}} = -.15$, $df = 42, p > .05$), while it was for above-average achievers ($\rho_{\text{above}} = -.47$, $df = 40, p = .002$).

So, for the sample as whole, students seemed to adapt their strategy choices to their individual accuracy and speed characteristics of the two strategies. However, interesting gender differences were found. Girls appeared to fit their strategy choices to accuracy characteristics, ignoring speed characteristics. In contrast, boys showed the opposite pattern, by choosing adaptively regarding speed, and thereby paying less attention to accuracy (although the – nonsignificant – correlation with accuracy difference was

positive and not significantly different from the correlation for girls). Moreover, there were also important differences with respect to mathematics level. Below-average achievers did not significantly fit their strategy choice to either accuracy or speed, while above-average achievers chose significantly adaptive regarding both accuracy and speed.

5.4 DISCUSSION

In this study, mental and written solution strategies on complex division problems were investigated using the choice/no-choice paradigm. Students successively solved three parallel sets of four division problems with free strategy choice, forced mental, and forced written calculation, respectively. Besides assessment of strategy repertoire, distribution, and efficiency, an important focus was on strategy adaptivity. We will first discuss the results on the four dimensions of strategy competence, and then focus on gender differences and mathematics achievement level effects. We will end by discussing cognitive psychological and educational implications of the findings of the current study.

5.4.1 Strategy repertoire, distribution, efficiency, and adaptivity

Concerning strategy repertoire, findings showed that approximately one half of the students used written as well as mental strategies in the choice condition. The majority of the other half used only written strategies, and a small part used only mental strategies. Regarding strategy distribution, the relative frequencies of mental and written strategies in the choice condition showed that each item was solved most frequently by written calculation, but that there were differences in this respect between items. Analysis of the specific types of strategies showed that both written and mental strategies predominantly comprised repeated addition/subtraction. Item 2 deserves special attention because of the possibility of using a compensation strategy (a special case of the repeated addition/subtraction strategy, in which it is possible to take advantage of the closeness of the dividend to a hundredfold of the divisor). The majority of mental strategies involved compensation, while this was not very frequent within written strategies. So, choosing between mental and written computation on that item mainly reflected using the compensation strategy or not: a similar finding to Hickendorff et al. (2010).

Strategy performance in the choice condition showed that, for the sample as a whole, freely chosen mental strategies were evenly accurate but faster than freely chosen written strategies. However, these accuracy data were probably affected by selection effects (cf.

Siegler & Lemaire, 1997), because the accuracy difference between the two no-choice conditions was significant: forced written strategies were on average more accurate than forced mental strategies. Speed differences between the no-choice conditions were congruent with those from the choice condition, with mental strategies being significantly faster than written strategies. Thus, unbiased strategy performance data showed that mental strategies were less accurate but faster than written strategies.

The main focus was on strategy adaptivity, which we approached in three ways. In each approach, we assessed whether the strategy selected in the choice condition was the most 'appropriate' one, as evidenced by the unbiased strategy performance information from the no-choice conditions. First, item-level analysis showed that, on average, mental strategy choices were adaptive to speed, and indifferent to accuracy. In contrast, written strategy choices were fit to accuracy differences, while being counter-adaptive to speed. Second, we combined accuracy and speed on a trial-basis by operationalizing the definition of the 'best' strategy as being the one leading fastest to an accurate answer. On average, on 43% of the trials the 'best' strategy was chosen, and on 30% the best strategy was *not* chosen. On the remaining 28% of the trials there was no correct answer obtained in either of the two no-choice conditions, so the strategy choice in the choice condition could not be scored with respect to adaptivity. Interestingly, more than half of the students made at least one adaptive and one counter-adaptive strategy choice on the 4 trials. Third, student-level analyses showed that the correlations between frequency of use of mental computation on the one hand and unbiased accuracy and speed differences on the other hand were significant and in the expected direction. So, in general, strategy choices seem adaptive both to accuracy and speed. In sum, we found that mental strategies were chosen in trials where they were faster but equally accurate according to the unbiased strategy performance data, while written strategies were chosen on trials on which it was the more accurate (albeit slower) strategy. Combining these findings resulted in the pattern that, on average, students chose adaptively both to accuracy and speed. However, when accuracy and speed were combined to define the optimal strategy, we found that students made a suboptimal strategy choice on a substantial percentage of items (30%).

5.4.2 Gender differences and mathematics achievement level effects

We found interesting gender differences in the dimensions of strategic competence. Regarding strategy choice, boys were more inclined to mental computation than girls, a finding resembling earlier research findings that girls favor structured, algorithmic strategies, while boys tend to use less structured, more intuitive strategies (Carr & Davis, 2001; Carr & Jessup, 1997; Gallagher et al., 2000; Hickendorff et al., 2009b, 2010; Timmermans et al., 2007). In the current study, there were no significant gender differences in strategy accuracy. Regarding speed, girls were slower than boys when they were forced to use mental computation. In all other conditions, gender differences in speed were not significant. Consequently, for boys the speed gains of choosing mental strategies over written ones was larger than for girls, which may partially account for boys' larger inclination of choosing mental strategies. In addition, boys and girls appeared to have different speed-accuracy preferences. Girls appeared to fit their strategy choices to accuracy considerations, ignoring speed, while boys had a preference for speed over accuracy. This may be related to individual differences in the confidence criterion that have been reported in children (Siegler, 1988a, 1988b) and in adults (Hecht, 2006).

Moreover, these found gender differences in strategy choice and strategy adaptivity may be related to the consistent finding that girls have lower levels of confidence with mathematics (Mullis et al., 2008; Timmermans et al., 2007; Vermeer et al., 2000), so they may perform more cautiously than boys and therefore choose accuracy over speed. In line with this reasoning, girls have been found to be less inclined to intellectual risk-taking than boys (Byrnes et al., 1999). In addition, girls tend to be more inclined to (academic) delay of gratification (Bembenutty, 2009; Silverman, 2003), which might partially explain that boys more often choose fast mental calculation over slower but more accurate written computation.

In addition to gender differences, there were mathematics achievement level effects. Below-average and above-average achievers chose mental computation equally often. Performance differences were as could have been expected, with below-average achievers less accurate and slower than above-average achievers, regardless of the strategy they chose (choice condition) or the strategy they had to use (no-choice conditions). Interesting differences were found in strategy adaptivity: below-average achievers did not take either accuracy or speed into account in their strategy choices, while above-average achievers fitted their strategy choices to both components of performance. Therefore, we

argue that the present study's results implies that strategy adaptivity in complex division is currently only attained by the better achieving students, resembling findings of Foxman and Beishuizen (2003), Hickendorff et al. (2010), Torbeyns, De Smedt, et al. (2009b), and Torbeyns et al. (2002, 2006). Further research is needed into whether this important goal of mathematics education reform is feasible or desirable for the below-average achievers. Like Geary (2003), Torbeyns et al. (2006), and Verschaffel et al. (2009), we plead for more research-based evidence for striving for adaptive expertise in mathematics education, especially for average and weaker students.

5.4.3 Methodological considerations

Several methodological considerations of the current study merit attention. First, we focused on distinguishing between between mental strategies, defined as nothing written down on paper, and written strategies in which something was written down on paper, ranging from intermediate answers to procedural algorithms. Note that this is a similar categorization of strategies as in Siegler and Lemaire's (1997) original choice/no-choice study, in which they distinguished between using a calculator, using mental arithmetic, and using pencil and paper (experiment 3). Although this is arguably a rough classification, and other categorizations (for example with respect to the number of solution steps) are thinkable, we chose this strategy split for two reasons. First, earlier studies into strategy use on complex division by Dutch sixth graders showed that both strategy types were used, and that they had large predictive power of the accuracy of solutions (Hickendorff et al., 2009b, 2010). Second, the didactical practice in the Netherlands – with the disappearance of the traditional algorithm and many different informal strategies – leads to obstacles in studying the characteristics of different strategies in a choice/no-choice design. That is, if students are forced to use a particular strategy in the no-choice conditions, they should have those strategies in their repertoire, and many students did not get instruction in for example the traditional algorithm.

Second, the number of items (4 items times 3 conditions) was small, mainly based on practical considerations. Because these kind of complex division problems are quite demanding to solve for sixth graders (as also comes forward from the long solution times, on average 83.7 seconds with peaks to over 500 seconds), we believed it was not practically feasible to administer more than 12 problems to a student in a session. Given this limited number of items, we were unfortunately not able to rigorously analyze

the effect of item features on strategy choice. However, we argue that we were able to study adaptivity, because we did found substantial variation in strategy choice, with half of the students using both strategies on the set of 4 problems in the choice condition. Moreover, the single strategy users were also spread over the two strategies: two-thirds of them consistently used written strategies, while one third consistently used mental strategies. Therefore, even on this small number of problems there were clear differences in strategy choice, that we tried to predict by strategy performance characteristics in order to investigate the issue of adaptivity.

5.4.4 Implications

In conclusion, we found several indications that there are different patterns of adaptivity to strategy accuracy and speed. If we look at the general pattern for the whole sample, it would seem that mental strategies are chosen for reasons of speed, while written strategies are chosen for reasons of accuracy. However, when we look at different subgroups of students (gender, mathematics level), we find that there are different adaptivity patterns. Interestingly, there are students who prefer accuracy over speed, while there are also students showing the opposite pattern. Moreover, the majority of students made both optimal and suboptimal strategy choices on the 4 items.

Individual differences in preference for accuracy and speed are important from a theoretical as well as from practical point of view. Theoretically, they have not been specifically addressed in the cognitive models of strategy choice and adaptivity (Shrager & Siegler, 1998; Siegler & Shipley, 1995), although the concept of different confidence criterion individuals may hold is related. Moreover, future research may investigate whether other individual differences constructs can (partly) account for the found accuracy-speed preferences. The cognitive style of impulsivity-reflection (Kagan, 1966) may very well be related (cf. Siegler, 1988a), with reflectives being slow but accurate, but impulsives being fast but with more errors (Phillips & Rabbitt, 1995). In addition, concepts such as academic delay of gratification (Bembenutty, 2009) and academic risk-taking (Byrnes et al., 1999) may be associated as well.

In addition to task and subject variables, sociocultural context variables may also affect strategy choices (Luwel et al., 2009; Verschaffel et al., 2009). Ellis (1997) pointed out the possibility of (sub)cultural differences in the weights assigned to speed versus accuracy of performance, and the value placed on solutions constructed in the head

versus by means of external aids. For instance, classroom socio-mathematical norms and practices valuing speed over accuracy and/or mental strategies over written ones, may result in students overusing mental strategies at the cost of accuracy. Besides the implication that cognitive models for strategy adaptivity are limited in this respect (see also Ellis, 1997; Verschaffel et al., 2009, there are also important educational consequences. Most educators will agree that being correct is more important than being fast in learning mathematics. However, not all students (particularly boys) seem to reason in that way, and teachers should be aware of that. So, teachers may create a classroom environment in which accuracy is preferred over speed and using an external aid (paper and pencil) is not necessarily less valuable than working in the head. Furthermore, they may explicitly encourage consistently mentally calculation students (especially those with low mathematical ability) to write down their solution procedures, as this would improve their accuracy. Finally, it would be interesting to conduct a similar study in another educational climate with a larger focus on written arithmetic than in the Netherlands and compare the results.

APPENDIX 5.A COMPLETE ITEM SET

Table 5.4 presents the number characteristics of the three parallel sets (A, B, and C) of 4 items each. The realistic contexts in set A were (translated from Dutch):

1. Four children go to an amusement park together. For admission, they have to pay 34 euro in all. How much is that per child?
2. A bookseller has earned 1,089 euro. He sold all his books for 11 euro each. How many books did he sell?
3. Robert is making a fence, which will have a length of 31.2 meters. The planks he uses are 1.2 m long. How many planks will he need for the entire fence?[4]
4. Anne has 304 biscuits. She divides the biscuits over 19 jars. How many biscuits are there in each jar?

The contexts of the items in set B and set C were comparable.

TABLE 5.4 *Number characteristics of the items.*

	item set		
item nr.	A	B	C
1	34 ÷ 4	52 ÷ 8	45 ÷ 6
2	1089 ÷ 11	2450 ÷ 25	1980 ÷ 20
3	31.2 ÷ 1.2	30.8 ÷ 1.1	32.2 ÷ 1.4
4	304 ÷ 19	306 ÷ 17	221 ÷ 13

4 NB. Original item included illustration, making clear that the height of the fence was the height of one plank.

CHAPTER 6

The language factor in assessing elementary mathematics ability: Computational skills and applied problem solving in a multidimensional IRT framework

This chapter has been submitted for publication as Hickendorff, M. *The language factor in assessing elementary mathematics ability: Computational skills and applied problem solving in a multidimensional IRT framework.* A Dutch paper on this study has been published as Hickendorff & Janssen (2009).

I am indebted to Jan Janssen from CITO for collecting the data, Rinke Klein Entink for programming the MCMC-algorithm in R, and Norman Verhelst, Kees van Putten, and Willem Heiser for their helpful suggestions.

ABSTRACT

In this paper, the results of an exploratory study into measurement of elementary mathematics ability are presented. The focus was on the abilities involved in solving standard computation problems on the one hand and problems presented in a realistic context on the other hand. The objectives were to assess to what extent these abilities are shared or distinct, and to what extent students' language level plays a differential role in these abilities. Data from a sample of over two thousand students from first, second, and third grade in the Netherlands were analyzed in a multidimensional item response theory (IRT) framework. The latent correlation between the two abilities (computational skills and applied mathematics problem solving) ranged from .81 in grade 1 to .87 in grade 3, indicating that the abilities are highly correlated but still distinct. Moreover, students' language level had differential effects on the two mathematical abilities: Effects were larger on applied problem solving than on computational skills. The implications of these findings for measurement practices in the field of elementary mathematics are discussed.

6.1 INTRODUCTION

Mathematics education has experienced a large international reform (e.g., Kilpatrick et al., 2001). A general characteristics of this reform is that mathematics education should no longer focus predominantly on decontextualized traditional mathematics skills. Instead, the process of mathematics problem solving and doing mathematics are important educational goals (e.g., National Council of Teachers of Mathematics, 1989, 2000) as is also reflected in large-scale assessment frameworks such as from TIMSS, NAEP, and PISA. Word problems or contextual problems – typically a mathematics structure in a more or less realistic problem situation – serve a central role for several reasons. They may have motivational potential, mathematical concepts and skills may be developed in a meaningful way, and students may develop knowledge of when and how to use mathematics in everyday-life situations (e.g., Verschaffel et al., 2000). Moreover, solving problems in context may ideally serve as tools 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 study (OECD, 2004) included mainly problems in a real-world situation.

In the Netherlands, the reform has gained dominance in mathematics curricula. In 2004, almost all elementary schools used a mathematics textbook based on reform 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 reform-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., 2005; Kraemer et al., 2005) and commonly used student monitoring tests such as CITO's *Monitoring and Evaluation System for primary school students - Arithmetic and Mathematics* also contain predominantly contextual problems. The latter testing system's purpose is to enable teachers to monitor their students' progress in a number of meaningful ways, and it consists of two tests in each school year (midway and at the end) from grade 1 to grade 6. In conclusion, today's Dutch primary school students mathematics education and assessment consists for a large part of problems in (more or less) realistic contexts.

This international shift towards including many or predominantly contextual problems gives rise to two questions. First, to what extent are different abilities involved in solving standard computation problems versus solving contextual problems? This question is important because it will give insight whether the currently used tests that are dominated by contextual problems give rise to the same conclusions on individual or group differences as a test that is dominated by standard computation problems. Second, contextual problems are usually verbal, giving rise to the question what the role of language is. Determining to what extent the student's language level has differential effects on the two abilities is clearly of practical importance, for example in getting more insight in the broadness of the commonly observed performance lag of ethnic minority students for whom the language used at school and in the test is not their first language. Next, these two questions are elaborated further.

6.1.1 Standard computation problems and context format problems

Solving standard computation problems on the one hand and realistic context format problems on the other hand, are likely to involve different aspects of mathematical cognition (e.g., Fuchs et al., 2008). Solving contextual problems is a complex process involving several cognitive processes or phases, as argued by phase-like approaches to mathematical modeling or mathematizing. Only after steps in which a situational

and mathematical model of the problem situation have been formed accurately, computational skill (and carefulness therein) comes into play. Therefore, other factors than 'pure' computational skills are likely to contribute to success in applied mathematics problem solving (Wu & Adams, 2006). Alternative approaches to mathematical modeling in word problem solving are more holistically oriented (e.g., Gravemeijer, 1997a). Ideally, children should approach (unfamiliar) contextual problems as situations to be mathematized, and they should not revert to searching for the application of the appropriate standard procedure. Computational skill is conceived of not so much as a necessary prerequisite of successful applied problem solving, but these two aspects involve separate abilities instead.

Either way – emphasizing problem solving phases or adhering a more holistic approach – it is likely that different abilities are involved in solving standard numerical mathematics problems and context format problems, and that they therefore measure different aspects of mathematics competence. An important question that is addressed in the present study is *to what extent* these abilities are shared or distinct and whether this depends on grade. Similar to the findings of Fuchs et al. (2008), we hypothesize that these are two related but distinct abilities. Furthermore, we expected the relation between these two aspects to increase with age, since students in higher grades have had more years of formal schooling and therefore more developed cognitive schemata to solve word problems (De Corte, Verschaffel, & De Win, 1985; Vicente, Orrantia, & Verschaffel, 2007).

The results on this research question could have implications for theoretical insights into the structure of mathematical competence, but also for mathematics assessment and instruction practices. In particular, information on the extent to which an ability estimate derived from a mathematics test containing almost exclusively problems in a context (as is current practice in the Netherlands) converges with an ability estimate derived from a mathematics test that would contain only standard computational problems may yield practical recommendations for future test construction.

6.1.2 The language factor

A necessary condition for obtaining the correct answer to a contextual problem is that the problem solver accurately understands the problem situation and all relevant parameters to it. Since the problem situation is usually verbal, it is likely that the language level of

the problem solver plays an important role. Supporting the importance of language in word problem solving, research has found that a common source of errors appears to be misunderstanding of the problem situation (Cummins, Kintsch, Reusser, & Weimer, 1988; Wu & Adams, 2006) and that conceptual rewording of word problems facilitated performance (e.g., Vicente et al., 2007).

Ethnic minority students score lower on language ability tests than native students. In addition, they have been consistently found to lag behind in mathematics as well, as has been found in international assessments such as TIMSS (*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. In the US, Abedi and Lord (2001) found that linguistic simplifications of the problem text of NAEP mathematics test items benefited students who were English language learners more than it benefited their proficient English speaking peers. They contended that the use of unfamiliar or infrequent vocabulary and passive voice constructions hampered understanding for certain groups of students. Similarly, Abedi and Hejri (2004) found that the gap between students with limited English proficiency and their proficient peers was larger on linguistically complex items than on noncomplex items, regardless of the item content difficulty. Recently, two Dutch studies investigated this issue in secondary education mathematics. Prenger (2005) found that ethnic minority students were impaired in their understanding of mathematics texts due to their limited vocabulary of typical school words. Similarly, Van den Boer (2003) found that ethnic minority students lagged behind in mathematics achievement as assessed on contextual problems due to hidden language problems, because contextual problems are accompanied by language as well as (mathematical) concepts that need to be interpreted correctly.

The present study extends these previous research findings by addressing the role of language in solving contextual problems for young children (early grades in elementary school) in the Netherlands. We expect that students' language level effects are more profound on the ability to solve contextual problems than on the ability to solve computational problems. Moreover, we expect the language effects to decrease with more years of formal schooling: inexperienced problem solvers rely more heavily on the text because they lack highly developed semantic schemata for word problems (De Corte et al., 1985). So, language level is expected to be more important to understand the

problem situation in lower grades than in higher grades. Of particular importance was whether ethnic minorities (students who spoke a language other than Dutch at home) have a larger performance lag on contextual mathematics problems than on standard computation problems. This would have serious implications for the current testing practices, that focus heavily on contextual problems. Moreover, the role of reading comprehension level is addressed.

6.1.3 The current study

In the current cross-sectional survey students from grade 1, 2, and 3 solved a set of computational problems in addition to a set of contextual problems. The main objectives were to assess to what extent abilities to solve these different types of problems are shared or distinct, and to what extent students' language level plays a differential role in these abilities. To answer these two questions, we used a multidimensional item response theory (MIRT) modeling framework (Reckase, 2009). Specifically, we used between-item or simple structure multidimensional IRT models, in which it is assumed that each item in a test is only related to one of several related subscales that each measure a separate ability dimension (Adams, Wilson, & Wang, 1997).

6.2 METHOD

6.2.1 Participants

Participants were 713 students from grade 1 (average age 6 years), 761 students from grade 2 (average age 7 years), and 753 students from grade 3 (average age 8 years) from 34 different primary schools in the Netherlands. To be able to study language level effects with sufficient power, the schools that were selected had relatively many ethnic minority students. As a consequence, the current sample of schools and students is not entirely representative for the population of Dutch primary schools. Furthermore, we included only the students who completed more than half of the contextual problems and more than half of the numerical expression problems in the analyses. These were 649 students from grade 1 (from 31 schools), 736 students (from all 34 schools) from grade 2, and 664 students (from 31 schools) from grade 3, yielding a effective sample of 2,049 students.

Two types of background information on the students' language level were collected. First, that was the language spoken at home (as reported by the teacher), which we

TABLE 6.1 *Pupil background information: distribution of home language and reading comprehension level.*

	home language			reading comprehension level				
	Dutch	other	?	A	B	C	D	?
grade 1								
frequency	430	215	4	112	130	140	159	108
valid %	66	34		21	24	26	29	
grade 2								
frequency	514	216	6	170	152	177	106	131
valid %	70	30		28	25	29	18	
grade 3								
frequency	454	203	7	171	122	135	116	120
valid %	69	31		31	25	22	21	

classified into Dutch or another language. Almost one-third of the students spoke a language different than Dutch at home, see also Table 6.1. The distribution of home language (Dutch versus other) did not differ significantly by grade, $\chi^2(2, N = 2,032) = 2.3, p = .32$. The most prevalent non-Dutch language was Turkish (over 30%), followed by Moroccan/Arabic (about 10%), Berber/Tamazight (about 10%), and a Dutch dialect such as Friesian (about 10%).

Second, information on each student's reading comprehension level was collected, by gathering the most recent score on CITO's *Monitoring and Evaluation System for primary school students - Reading Comprehension* test. This is a widely used standardized measurement instrument, for which percentile score groups are reported based on a population norm group. We used four percentile groups (quartiles). Level A includes students who scored at or above norm group percentile 75, so these were the top 25%. Level B represents percentile 50-75, level C represents percentile 25-50, and level D represents the bottom 25%. Table 6.1 shows the distribution of students over the different levels of reading comprehension per grade. These distributions – excluding the missing values – differed by grade ($\chi^2(6, N = 1,690) = 33.8, p < .001$): norm-referenced reading comprehension levels of the first graders in the current sample were relatively lower than of the second and third graders in the sample.

TABLE 6.2 *For both subscales, the number of problems per operation, descriptive statistics of the proportion correct scores P (correct), and Cronbach's α.*

	number of problems						*P* (correct)		
	add.	sub.	mult.	div.	combi	total	*M*	SD	α
computational skills									
grade 1	16	15	0	0	0	31	.73	.22	.90
grade 2	15	15	4	0	0	34	.68	.21	.89
grade 3	9	9	10	9	2	39	.75	.18	.89
contextual problem solving									
grade 1	3	8	5	3	3	22	.67	.24	.87
grade 2	4	6	4	4	6	24	.65	.21	.85
grade 3	5	5	5	6	7	28	.69	.22	.88

6.2.2 Material

Each student was administered two types of booklets (collection of multiple items administered in one session): the grade-appropriate regular booklets from CITO's *Monitoring and Evaluation System for primary school students - Arithmetic and Mathematics* and an extra grade-specific booklet that was designed specifically for this study. There were 2 regular CITO booklets for grade 1 (CITO, 2005a) and also 2 regular booklets for grade 2 (CITO, 2005b), and 3 regular booklets for grade 3 (CITO, 2006). All these booklets contained predominantly problems in context format. In contrast, the extra booklet contained only problems in standard computation format (numerical expression only, e.g., $17-5=\ldots$). All problems in the extra booklet required either addition, subtraction, multiplication, division, or a combined operation. In order to make a fair comparison, we selected only those problems from CITO's regular booklets that required one of these four (combined) operations. Therefore, the current analyses are based only on problems requiring either addition, subtraction, multiplication, division, or a combined operation. Moreover, the few problems from CITO's regular booklets that were in numerical expression format were grouped with the extra booklet problems. For both subscales, the number of problems per operation, descriptive statistics of the proportion correct scores, and Cronbach's α are shown in Table 6.2.

All context format problems from CITO's regular booklets included text. In addition, a large majority of the context format problems included an illustration, containing either essential information, duplicate information, or no relevant information at all.

Appendix 6.A shows a sample of problems used.

6.2.3 Procedure

The students completed each of the three (grade 1 and 2) or four (grade 3) different booklets on a different morning. The assessment procedure of CITO's regular booklets (mainly context format problems) differed by grade. In grade 1, each problem text was read aloud by the teacher. In grade 3, students had to read and work trough all problems independently. In the second grade, on one booklet the teacher read out the problem text aloud, while on the other booklet students had to work through the problems independently. The assessment procedure of the extra booklet was equal for each grade: students had to work through the problems independently. After all booklets were administered, the teachers sent in the students' work, and research assistants entered the answers given in a database, and scored them as either correct or incorrect.

6.2.4 Multidimensional IRT models

All statistical analyses were done in a multidimensional IRT modeling framework. For each grade, *descriptive* as well as *explanatory* IRT models were fitted (see also Wilson & De Boeck, 2004). First, we fitted multidimensional descriptive or measurement IRT models, aiming to answer the first research question by obtaining an accurate description of the latent variables involved in solving the two types of mathematics problems and the relation between these latent variables. For the second research question, we added an explanatory part to the IRT models, in which we assessed the (possibly differential) effects of the student's language variables on the latent abilities by means of a latent regression approach.

Measurement MIRT models

Unidimensional IRT models may be generalized to multidimensional IRT (MIRT) models (for a recent review, see Reckase, 2009). In these models, persons are no longer characterized by their position on a single latent variable, but instead by their position on two or more latent variables. If the number of abilities or dimensions is given by m, then each person p is characterized by an ability vector $\boldsymbol{\theta}_p = (\theta_{p1}, \theta_{p2}, \ldots, \theta_{pm})$. The

multidimensional generalization of the 2PL model is:

$$P(X_{ip}=1\,|\boldsymbol{\theta}_p)=\frac{\exp\left(\sum_{k=1}^{m}\alpha_{ik}\theta_{pk}+\delta_i\right)}{1+\exp\left(\sum_{k=1}^{m}\alpha_{ik}\theta_{pk}+\delta_i\right)}. \tag{6.1}$$

Each item is characterized by an intercept δ_i, and by m dimension-specific discrimination parameters α_{ik} ($k=1,\ldots,m$). These discrimination parameters reflect the importance of factor k for solving item i – similar to a factor loading in factor analysis or structural equation modeling. The simplest multidimensional IRT models are simple structure or between-item models (Adams et al., 1997) in which each item is associated with only one of the dimensions, and hence there is only one nonzero element in α_{ik} for each item i. These models are suited if a test is built up of several subtests that are each supposed to measure one ability. In the present application, we used between-item MIRT models with two dimensions or abilities: (a) computational skills: the ability to solve numerical expression format problems, and (b) applied mathematics problem solving: the ability to solve context format problems. Figure 6.1 shows a graphical representation of this two-dimensional model.

MIRT models overcome several shortcomings of applying separate unidimensional IRT scales for each dimension: the intended structure is explicitly taken into account, the relation between the latent dimensions is estimated directly, and it makes use of all available data resulting in more accurate individual ability estimates (Adams et al., 1997). Our main interest lied in the estimate of the latent correlations between the two ability factors. 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 estimating consecutive unidimensional models, or classical test theory approaches that are based on the proportion of items solved correctly.

Explanatory MIRT models

Measurement IRT models (either unidimensional or multidimensional) can be extended by an explanatory part, by estimating the effects of predictor variables on the latent factor(s). These predictors can be either on the person level, item level, or person-by-item level (Rijmen et al., 2003; Wilson & De Boeck, 2004). In the current study, we were interested in the effects of two person-level variables on mathematics ability: students' home language (Dutch or other) and their reading comprehension level (four norm-

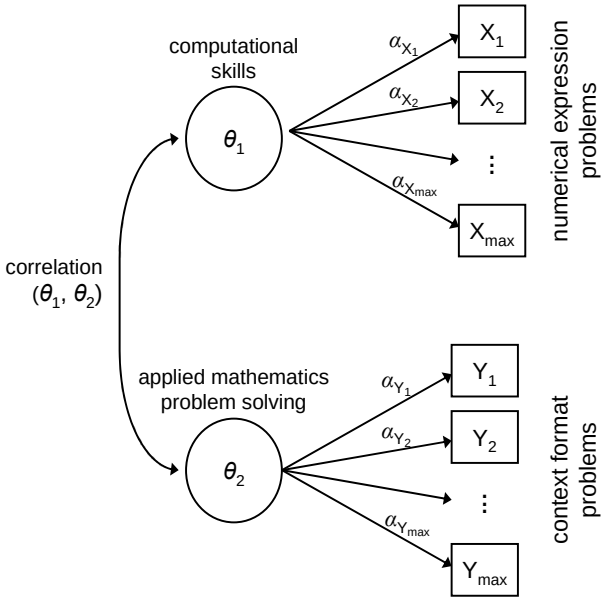


FIGURE 6.1 *Graphical representation of between-item two-dimensional IRT model.*

referenced quartiles). Including person explanatory variables in an IRT model results in a latent regression: the latent person variable θ_p can be considered as being regressed on external person variables. This latent regression can be either univariate (in case of unidimensional IRT models) or multivariate (with multidimensional IRT models) (Von Davier & Sinharay, 2009).

There are three different approaches to assess the effect of external person predictor variables on the ability factor(s) in an IRT framework: a one-step, a two-step, and a three-step approach. The one-step approach involves joint modeling of item parameters and latent regression parameters. The advantage is that measurement error of the item parameters is taken into account, but a disadvantage is that the measurement scale (i.e., the item parameters) depends on the predictor variables included (Verhelst & Verstralen, 2002, for a discussion of this issue in the multidimensional case see Hartig & Höhler, 2008). In the two-step approach this disadvantage is overcome. In the first step the item parameters of the measurement model are estimated. In the second step the item parameters are fixed at their estimated values, and a (univariate or multivariate) latent

regression model is estimated. This approach is commonly employed in large-scale assessment programs, such as NAEP, TIMSS, PIRLS, and PISA (Von Davier & Sinharay, 2009). The three-step approach involves first estimating the item parameters of the IRT model (either unidimensional or multidimensional), next estimating individual person ability scores with item parameters fixed at their estimated value, and finally carrying out a (univariate or multivariate) regression analysis on these ability scores. This approach (which is, strictly speaking, not a *latent* regression analysis) was for example carried out by Hartig and Höhler (2008). A disadvantage is that measurement error of both person and item parameters is not taken into account.

In the present analyses, for each grade separately, we implemented the two-step approach. First, a two-dimensional between-item MIRT measurement model was fit. Next, item parameters α_{ik} and δ_i were fixed to their estimated value, and plugged in as known constants in the multivariate latent regression analyses, by estimating the conditional – given the regression variable(s) – multivariate distribution of $\boldsymbol{\theta}_p$. The effects of dummy-coded home language (2 categories) and reading comprehension level (4 categories) were estimated. Moreover, we tested whether these effects were equal or different for the two latent dimensions.

Model fit

Model fit is approached in two ways. First, by model fit information criteria BIC and AIC in which the statistical fit (log-likelihood, LL) of the model is penalized by the complexity of the model, i.e., the number of parameters P. The BIC is calculated as -2LL $+P\log(N)$, and the AIC as -2LL $+2P$; the BIC values parsimony of the model more than the AIC. Second, likelihood-ratio (LR) tests can be employed to test whether the improvement in fit between two nested models is statistically significant. The LR-test statistic Λ is calculated as two times the difference between the LL-value of the encompassing model and the LL-value of the restricted (nested) model. This statistic is asymptotically χ^2 distributed if the parameter space of the restricted model lies in the parameter space of encompassing model. The number of degrees of freedom (df) equals the difference in df between the two models. The LR-test can be used in models with predictor effects (i.e., explanatory IRT models with a latent regression part): they form the encompassing model; leaving out the regressors creates a restricted model (stating that the explanatory variables have no effect). Furthermore, to test whether a

TABLE 6.3 *Correlations between total number correct scores, latent correlations between computational skills and contextual problem solving, and Likelihood Ratio (LR) test results comparing fit of the one-dimensional (1D) versus the two-dimensional (2D) IRT models.*

	correlation total scores	latent correlation	LR-test (2D vs. 1D) statistic	*p*-value
grade 1	.72	.81 (SE = .02)	305.2	$p < .001$
grade 2	.75	.85 (SE = .02)	199.3	$p < .001$
grade 3	.77	.87 (SE = .02)	171.8	$p < .001$

two-dimensional model (encompassing model) fits better than a unidimensional model (restricted model), one has to take into account that to obtain the unidimensional model the correlation between the dimensions is restricted to one, which is on the boundary of the parameter space. In such situations, the LR-test is no longer χ^2 distributed, but it is asymptotically distributed as a mixture of $\chi^2(1)$ and $\chi^2(2)$ each with probability of .5 (Molenberghs & Verbeke, 2004, p. 136).

Software

All measurement and explanatory MIRT models were estimated in the NLMIXED procedure from SAS (SAS Institute, 2002, see also De Boeck & Wilson, 2004; Rijmen et al., 2003; Sheu et al., 2005). IRT model parameters were estimated by the NLMIXED procedure within a MML formulation, and a (multivariate) normal distribution for the person parameters was assumed. Gaussian quadrature with 20 nonadaptive quadrature points was used for the approximation of the integration, and Newton-Raphson as the optimization method.

6.3 RESULTS

6.3.1 Relationship between the different abilities

To answer the first research question, unidimensional and between-item (also known as simple structure) multidimensional measurement IRT models were estimated. The main results are the size of the latent correlation between the two abilities, and the improvement in model fit by defining two ability dimensions instead of one single dimension, both shown in Table 6.3.

In grade 1, the observed correlation between the proportion correct on numerical expression problems and the proportion correct on contextual problems was .72. The two-dimensional model fits significantly better than the one-dimensional model, as evidenced from the LR-test, as well as from the AIC and BIC criteria (not shown in Table 6.3). Therefore, it seems legitimate to distinguish computational skills (the ability to solve numerical expression problems) from applied mathematics problem solving (the ability to solve context format problems). The latent correlation between these two abilities was .81: obviously very high (and higher than the observed correlation, since it is unaffected by measurement error), but apparently not high enough to consider it as one single ability dimension. To provide a frame of reference for interpreting this size, the latent correlations in PISA 2006 between mathematics and reading was .80, and between mathematics and science .89 (OECD, 2009). So, we would expect a latent correlation of at least .80 between two subscales of mathematics, and the found estimate of .81 is barely higher. The current latent correlation indicates that 65% of the ability variances is shared, while 35% of the variance is unique.

In grade 2, the observed correlation between the proportion correct on numerical expression problems and the proportion correct on contextual problems was .75. Table 6.3 shows that like in grade 1, also in grade 2 the two-dimensional model fits significantly better than the one-dimensional model according to the LR-test. AIC and BIC-criteria were in accordance with this conclusion. The latent correlation between computational skills and applied mathematics problem solving was .85, indicating that 73% of the ability variances is shared, while 27% of the variance is unique.

Finally, in grade 3 the observed correlation between the proportion correct on numerical expression problems and the proportion correct on contextual problems was .77. Table 6.3 shows that, like in grades 1 and 2, also in grade 3 the two-dimensional model fits significantly better than the one-dimensional model as evidenced from the LR-test. In addition, the AIC and BIC criteria also indicate the 2-dimensional model as better fitting. So, again, we can distinguish computational skills from applied mathematics problem solving, as measured by the context format problems (all read independently by the students). The latent correlation between these two abilities was .87, indicating that 76% of the ability variances is shared, while 24% of the variance is unique.

The results thus far quite clearly show that that in each grade, computational skills and applied mathematics problem solving involve highly related but still distinct abilities. This means both dimensions contribute some unique variance to a students' overall

score. Moreover, the relationship between these two abilities seems to increase with grade: the latent correlations increased from .81 to .85 to .87 for grades 1, 2, and 3, respectively.

6.3.2 Language effects

Now that we have established that computational skills and applied mathematics problem solving involve two highly related but distinct abilities, we are moving to the next research question about the role of language. Since students' test scores are determined both by a part that is shared between the two abilities, as well as by unique contribution of each of the abilities, students' language level may affect both parts. This may result in differential effect of language level on the two abilities. Because of their verbal nature, we expected the language level effects to be larger on the ability to solve contextual problems than on the ability to solve computations. It is important to note that the two language predictors – home language and reading comprehension level – were significantly associated with each other (grade 1: $\chi^2(3, N = 540) = 65.5, p < .001$; grade 2: $\chi^2(3, N = 592) = 46.6, p < .001$; and grade 3: $\chi^2(3, N = 544) = 44.9, p < .001$). Not surprisingly, students who spoke a language other than Dutch at home were behind in their reading comprehension level compared to students with Dutch as home language.

Recall that we apply the two-step approach in the explanatory IRT analyses. Per grade, the item parameters (α_{ik} and δ_i) of the two-dimensional models were fixed at their estimated values, and plugged into the multivariate latent regression part as known constants. Several latent regressions were carried out, from which all students with missing values on one or both language predictor variables excluded. The two ability dimensions were scaled with a mean value of 0 and with equal variances. All effects reported are on the logit scale.

Grade 1

In grade 1, 109 students had missing values on one or both predictor variables, so these analyses were based on data of 540 students. Pupils' home language significantly affected overall or average mathematics problem solving ability (LR $= 17.7$, $df = 1$, $p < .001$). Moreover, the difference between Dutch-speaking and other-language speaking children was different for the computational and applied ability dimensions (differential effect significant, LR $= 24.3$, $df = 1$, $p < .001$). The upper left plot of Figure 6.2 graphically

shows that Dutch-speaking students outperformed students with another home language significantly more on the applied dimension (difference on the logit scale = .57, $z = 5.98$) than on the computational dimension (difference = .20, $z = 2.23$).

Similarly, reading comprehension level also had a significant effect on overall mathematics ability (LR = 235.7, $df = 3$, $p < .001$), and this effect was also significantly different for the two ability dimensions (LR = 10.0, $df = 3$, $p < .05$). The upper right plot of Figure 6.2 shows that reading comprehension level had a larger effect on the ability to solve contextual problems than on the computational skills dimension. To illustrate, the difference between students with the highest reading comprehension level A and the lowest level D was significantly larger on the applied dimension (difference = 1.77, $z = 14.84$) than on the computational dimension (difference = 1.46, $z = 12.41$).

Finally, we tested whether the performance lag of non-Dutch speaking students was mediated by their lower reading comprehension level. Statistically controlling for reading comprehension level, the outperformance of students with Dutch as home language disappeared on the applied mathematics dimension (difference = .05, $z = .61$), and even turned into a significant disadvantage on the computational skills dimension (difference = –.28, $z = -3.10$).

Grade 2

In the second grade data, 130 students had missing values on one or both predictor variables and were excluded from the analyses, so 592 students remained. Pupils' home language significantly affected overall mathematics problem solving ability (LR = 10.1, $df = 1$, $p = .001$). In addition, the difference between Dutch-speaking and other-language speaking children was different for the computational and applied ability dimensions (differential effect significant, LR = 14.7, $df = 1$, $p < .001$). The middle left plot of Figure 6.2 graphically shows that Dutch-speaking students outperformed students with another home language significantly more on the applied dimension (difference = .42, $z = 4.54$) than on the computational dimension (difference = .17, $z = 1.86$; home level effect did not reach statistical significance).

Next, there was a significant main effect of reading comprehension level on total mathematics ability (LR = 164.7, $df = 3$, $p < .001$). However this effect was not significantly different for the two dimensions (LR = 6.7, $df = 3$, $p = .08$). The difference between students with reading comprehension A and D on the computational skills

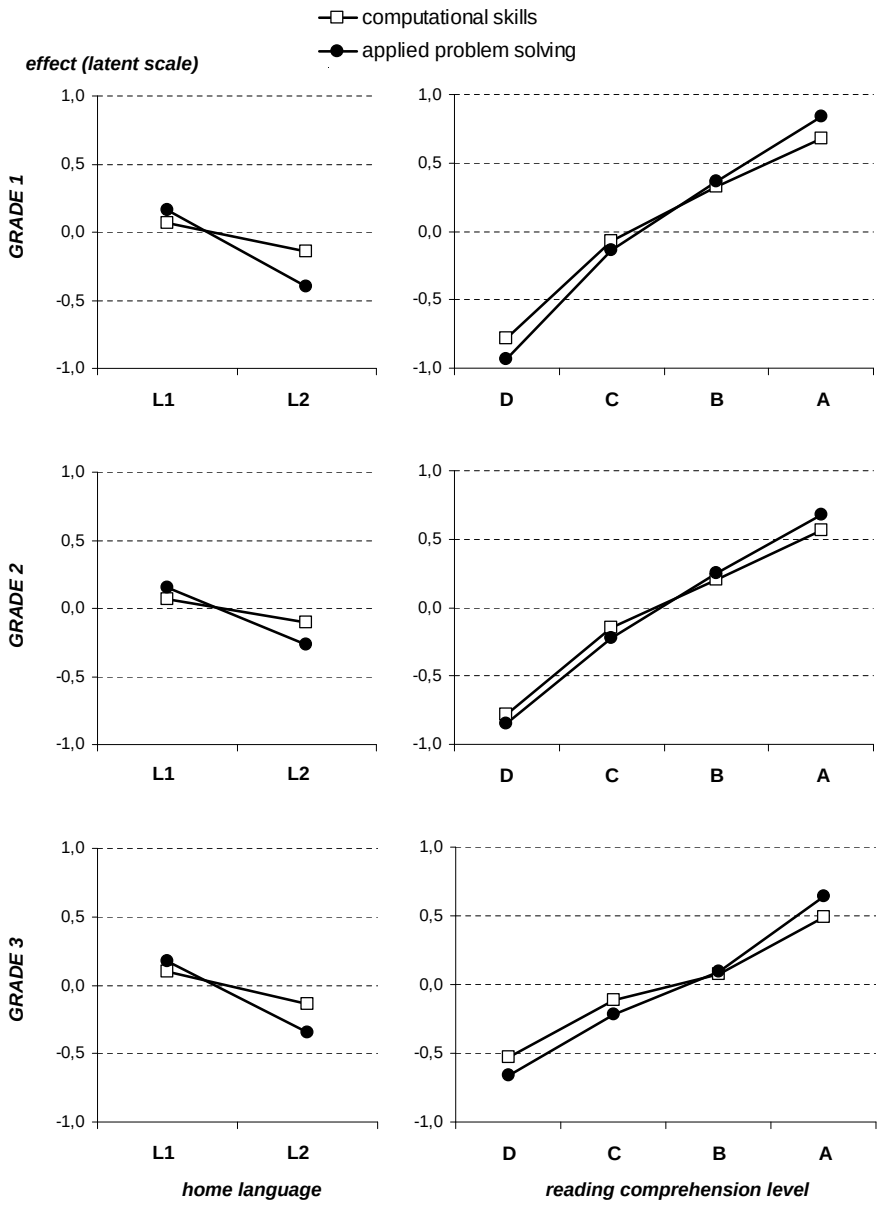


FIGURE 6.2 *Graphical display of home language effects (left plots) and reading comprehension level effects (right plots) for the two ability dimensions, grade 1 (upper part), grade 2 (middle part), and grade 3 (bottom part).*

dimension (difference = 1.35, $z = 11.61$) was nonsignificantly smaller than on the applied problem solving dimension (difference = 1.53, $z = 12.61$), as can be seen from the middle right plot of Figure 6.2.

Finally, statistically controlling for reading comprehension level differences, the home level effects were no longer significantly on applied mathematics (difference = .07, $z = .82$) as well as on the computational skills dimension (difference = –.16, $z = -1.81$). On computation skills, a pattern similar to grade 1 emerged: the (nonsignificant) disadvantage of non-Dutch speaking students reversed to a (nonsignificant) advantage after controlling for reading comprehension level.

Grade 3

In grade 3, 120 students had missing values on one or both predictor variables, so these analyses were based on data of 544 students. Like in grade 1, students' home language had a significant overall effect (LR = 15.3, $df = 1$, $p < .001$), and this effect was significantly different on the two dimensions of math problem solving ability (LR = 16.8, $df = 1$, $p < .001$). The bottom left plot of Figure 6.2 shows that Dutch-speaking students outperformed students with another home language significantly more on the applied dimension (difference = .52, $z = 4.99$) than on the computational skills dimension (difference = .24, $z = 2.24$). Similarly, reading comprehension level also had a significant overall effect (LR = 100.7, $df = 3$, $p < .001$) that was significantly different on the the two ability dimensions (LR = 12.4, $df = 3$, $p = .006$). The difference between students with the highest reading comprehension level A and the lowest level D was significantly larger on the applied dimension (difference = 1.30, $z = 11.35$) than on the computational dimension (difference = 1.02, $z = 9.19$), as is also visible from the bottom right plot of Figure 6.2.

Finally, statistically controlling for reading comprehension level, Dutch-speaking students still significantly outperformed students with another home language on the applied mathematics dimension (difference = .21, $z = 2.10$), but on the computational skills dimension there was no significant difference anymore (difference = $-.03$, $z = -.33$).

Comparison of results by grade

The results for grade 1, 2, and 3, have several things in common. First, students with Dutch as home language significantly outperformed those with another home language on each mathematics dimension in each grade, except on the computational skills dimension in grade 2. Second, this home language effect was not the same for each dimension. As expected, the performance lag of students with a non-Dutch home language was substantially larger on the applied mathematics problem solving ability dimension than on the computational skills dimension. Third, reading comprehension level was positively associated with each mathematics ability dimension in each grade. Also as expected, this reading comprehension effect was larger on the applied mathematics dimension than on the computational skills dimension. Finally, controlling for reading comprehension level, the disadvantage of students with home language other than Dutch was reduced: on the applied mathematics abilities it was either smaller or nonsignificant, while on the computational skills dimensions it was either nonsignificant or had even turned into an advantage.

Looking for a trend in the language level effects between the grades, the following pattern emerges. There was no specific trend in the differences between students with and without Dutch as home language by grade (the respective differences in grades 1, 2, and 3 being .20, .17, and .24 on computational skills, and .57, .42, and .52 on applied problem solving). In contrast, reading comprehension effects seemed to decrease in higher grades: on computational skills the level A versus level D differences decreased from 1.46 to 1.35 to 1.02 between grades 1, 2, and 3, respectively. Similarly, on the contextual problem solving dimension, the differences decreased from 1.77 to 1.53 to 1.30 between the grades. In conclusion, the role of reading comprehension seems to diminish by grade, while the performance lag of students with a non-Dutch home language did not decrease by grade.

6.4 Discussion

A sample of first, second, and third graders with relatively many ethnic minority students solved two sets of mathematics problems: standard computation problems in numerical expression format, and applied problems in context format. Our first research question was on the relationship between the abilities involved in solving the two types of mathematics problems. Evaluating the latent correlation estimates that were between

.81 and .87, we can conclude that two highly related but distinct aspects of mathematical competence are involved. Between 65% and 76% of the variance in overall performance on these problems can be explained by a common ability factor, but the remaining 35% to 24% of the variance are determined by unique contributions of the two dimensions. Moreover, the relationship appeared to get stronger in the higher grades.

Analyses on the second research question on the role of language showed that there were differential effects of both home language and reading comprehension level on the two mathematical abilities. As hypothesized, the effects were larger on applied problem solving than on computational skills in each grade. That is, the performance gap between students who spoke a language other than Dutch at home compared to Dutch-speaking students was larger on the ability to solve contextual problems than on the ability to solve computations. There were no clear grade-specific trends in this differential effect. Reading comprehension level also affected the ability to solve contextual problems to a larger extent than the ability to solve computations. However, the role of reading comprehension seemed to diminish in the higher grades. This may be a result of increased experience in solving word problems that has led to more sophisticated cognitive schemata of older students, so that they to need to rely less on the problem text. Moreover, statistically controlling for reading comprehension level, the performance lag of students with non-Dutch home language (compared to their native peers) was reduced on each dimension by a slightly larger amount on the applied mathematics dimension than on computational skills.

6.4.1 Issues in the multidimensional IRT framework

In this study, we employed a the multidimensional IRT framework. Between-item MIRT models with explanatory variables on both dimensions turned out to be a very useful and flexible approach. However, three issues deserve further attention. First, in the between-item or simple structure MIRT models that were used, each item was assigned a priori to one of the dimensions (Adams et al., 1997; Reckase, 2009). Although this framework was deemed appropriate for the current structure, within-item multidimensionality might provide meaningful results as well. In within-item MIRT models, items can have more than one nonzero discrimination, and hence require multiple latent factors. These multiple factors interact in a compensatory manner: a low level on one factor can be compensated with a high level on the other factor. For example, similar to what Hartig

and Höhler (2008) did on assessment data on reading and listening comprehension in a foreign language, it would be possible to distinguish two dimensions: one general computational skill dimension that affected items of both problem types, and one specific dimension that was only involved in solving context format problems. However, it would be necessary to assume a compensatory mechanism between these two dimensions, which seems unnatural. Furthermore, latent regression analyses, interpretation of the dimensions, and communication with the end-users of the test would be less straightforward. Other alternatives would be to set up a model with noncompensatory dimensions in which an individual must succeed on all subcomponents of item solving (Adams et al., 1997; Embretson & Reise, 2000), or other models from the family of cognitive diagnosis models (e.g., Leighton & Gierl, 2007).

Second, estimating MIRT models in marginal maximum likelihood framework, as was done in SAS PROC NLMIXED (SAS Institute, 2002) is computationally intense and hence very time consuming. The estimation time increases exponentially with number of dimensions, which poses practical limitations on the feasible number of quadrature points one can distinguish, which can affect results (Lesaffre & Spiessens, 2001). Therefore, we investigated whether results were robust against estimation procedure, by implementing two other estimation methods. In a first approach, item parameter were estimated for each dimension separately using conditional maximum likelihood (Verhelst & Glas, 1995) and the latent correlations between the dimensions were estimated, resulting in very similar values as in the present approach (see Hickendorff & Janssen, 2009). 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). Again, results were very similar to the MML-results from SAS, so they seem robust against the estimation procedure used.

Finally, the relation of the currently employed multidimensional IRT framework to a Differential Item Functioning (DIF) approach is worth mentioning. Carrying out DIF-analyses would have been an alternative way to find differential effects of language level on certain problems (such as for example was done by Van Schilt-Mol (2007). However, as noted by several authors (see Embretson & Reise, 2000, p. 262), DIF is usually caused by multidimensionality. If other dimensions than the main ability dimension are involved in an item, and the groups of interest (such as home language groups) differ

on these secondary dimensions, the item will show DIF. In DIF analyses, the secondary dimensions are usually considered as nuisance, and DIF items will be eliminated from the test. As a consequence, the final test will be more homogeneous (i.e., unidimensional), but information on the secondary dimension(s) is lost. Therefore, MIRT modeling is more general than the DIF approach, in the sense that information on all relevant ability dimensions contributing to item responses is retained without making a priori decisions on what the main ability dimension is, and what is considered nuisance.

6.4.2 Recommendations for further research

Several issues regarding the problems included in the current study would require further research. A first issue concerns the number characteristics of the problems. Although both types of problems were on the same content domain (the four basic number operations) in the same number range, the exact numerical properties of the contextual problems and numerical expression problems were not matched. As a consequence, a direct comparison of the difficulty levels of problems with and without context was not possible. It would be very interesting to study this in future research.

A second issue concerns the contexts used. In particular, the level of linguistic demands and the type of context (e.g., the semantic structure or the inclusion of an illustration) varied substantially between the problems, to obtain a broad coverage of applied problem solving reflecting educational practices. Unfortunately, these characteristics were not varied in a systematic way because the test's objective was to monitor the students and not the items. Therefore it was not possible to study effects of context characteristics rigorously. However, it seems very likely that the difficulty of the problem text hampers particularly the students with language difficulties, as suggested by the findings of Abedi and Lord (2001), Abedi and Hejri (2004), Prenger (2005), and Van den Boer (2003). In addition, illustrations can make a difference. Berends and Van Lieshout (2009) reported recently that in their study on grade 3 students, an illustration containing essential information for solving the problem negatively affected performance (accuracy and speed) as compared to problems containing all essential information in the problem text. In secondary education, Van den Boer (2003) reported that ethnic minority students were inclined to interpret the illustration in a wrong way, or ignore it altogether. Van Schilt-Mol (2007) also points out the possibility of wrongly interpreting the illustrations by ethnic minority students, although she observed that these students

devoted *more* attention to illustrations compared to their native peers. Future research is needed to assess to what extent illustrations in context format mathematics problems pose a stumbling block for ethnic minority students.

Another recommendation for future research concerns the fact that the study's findings did not extend beyond grade 3. Since we observed some interesting trends with increasing grades (stronger relationship between computational skills and applied mathematics, diminishing influence of students' reading comprehension level), it would be very interesting to collect similar data in higher grades as well.

6.4.3 Practical implications

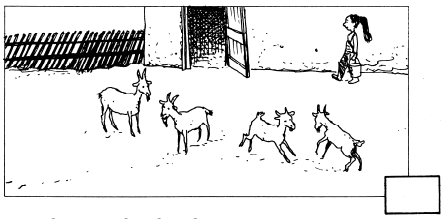
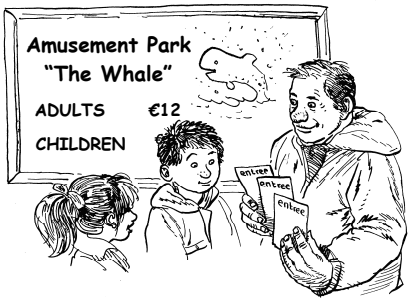
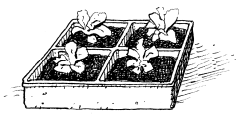
The present findings have implications for testing practices as well as for education. Regarding testing, the current dominance of context format problems in Dutch mathematics competence tests as well as in for example PISA merits critical consideration. We should be well aware that this offers a rather one-sided picture of mathematics competence: the fact that computational skills correlates only .80-.90 with applied problem solving, means that we are missing out on important information provided by administering standard computation problems. In addition, students with low language level score relatively less well on a test that focuses on context format problems compared to a test on computational skills, although this seems to play less a role in the higher grades.

We plead for a separate or embedded mathematics test containing standard numerical expression problems. The total score of such a mixed test would give a more fair representation of the two abilities than the current testing practice does. Alternative to the total score or in addition to the total score, separate subscale scores for computational skill and problem solving skill can be reported (as was also recommended by Fuchs et al., 2008), which may yield diagnostic information on potential remedial an instructional benefit (De la Torre & Patz, 2005). Sinharay, Puhan, and Haberman (2010) showed that caution with reporting subscale scores is needed, however. They have added value over reporting the total score only if the reliability of the subscales is large enough and if the dimensions are sufficiently distinct. These conditions were met in the present application, in which reliabilities of the subtests were at least .85 and two-dimensional models fitted substantially better than unidimensional models. Moreover, in cases were there is essentially one dominant factor or highly correlated dimensions, MIRT modeling

has been shown to yield subscale scores that have improved reliability over unadjusted subscale scores, because the correlational structure is taken into account (De la Torre & Patz, 2005; Stone, Ye, Zhu, & Lane, 2010).

Regarding educational practices, the potential (hidden) language problems of ethnic minority students affecting their mathematics problem solving merit educational attention, in language lessons as well as in mathematics lessons. In addition, a shift of focus of educational assessment towards separate or embedded testing of computational skills might also bring along a shift of focus in educational practice, since assessments signal what is valued and expected in teaching (Greer, 1997). Moreover, a profile of subscores representing different mathematical competencies would yield more fine-grained diagnostic information about a student's specific strengths and weaknesses, which may enable tailoring instruction for students with mathematical difficulties to their specific needs.

APPENDIX 6.A SAMPLE PROBLEMS (PROBLEM TEXTS TRANSLATED FROM DUTCH)

context format	numerical expression format
grade 1 ***Student's worksheet***  ***Teacher reads aloud:*** *"You see 4 goats in the paddock. Inside, 11 goats are having a rest. How many goats live on this children's farm?"*	**grade 1** 5 + 12 = ___ 17 – 5 = ___ 18 – ___ = 10
grade 2  Adults have to pay 12 euros. Children pay only half the price. Father takes his two children to the amusement park. How much does he have to pay in total? ___ euros	**grade 2** 26 + 25 + 27 = ___ 2 x 18 = ___ 58 = 98 – ___
grade 3  One tray contains 4 plants. Joyce buys 12 of these trays. How many plants does that make? ___ plants	**grade 3** 263 + 19 = ___ 487 – ___ = 427 9 x 30 = ___ 36 : 4 = ___

CHAPTER 7

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

This chapter has been submitted for publication as Hickendorff, M. *The effects of presenting multidigit mathematics problems in a realistic context on sixth graders' problem solving.*

This research was supported by CITO, National Institute for Educational Measurement. I would like to thank Suzanne van der Grind and Karlijn Nigg for their help in coding the solution strategies, Ingrid Vriens for her data analyses, Rinke Klein Entink for programming the MCMC-algorithm in R, and Mark de Rooij for his statistical advice.

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 experientally 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

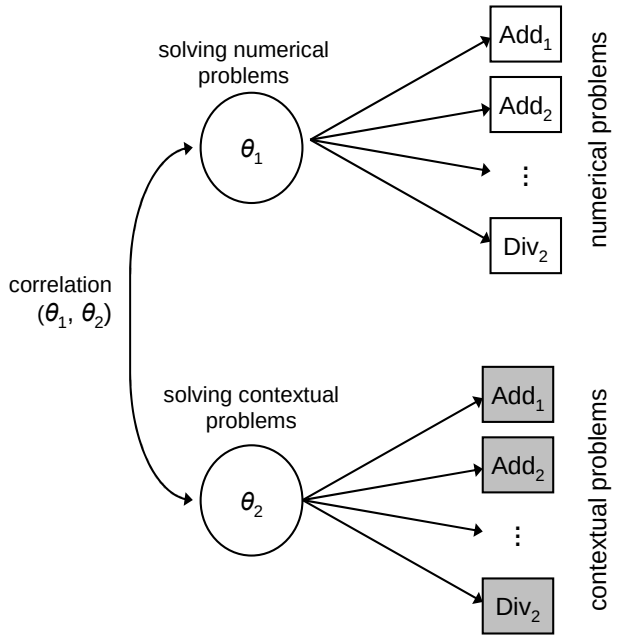


FIGURE 7.2 *Graphical representation of between-item two-dimensional IRT model.*

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

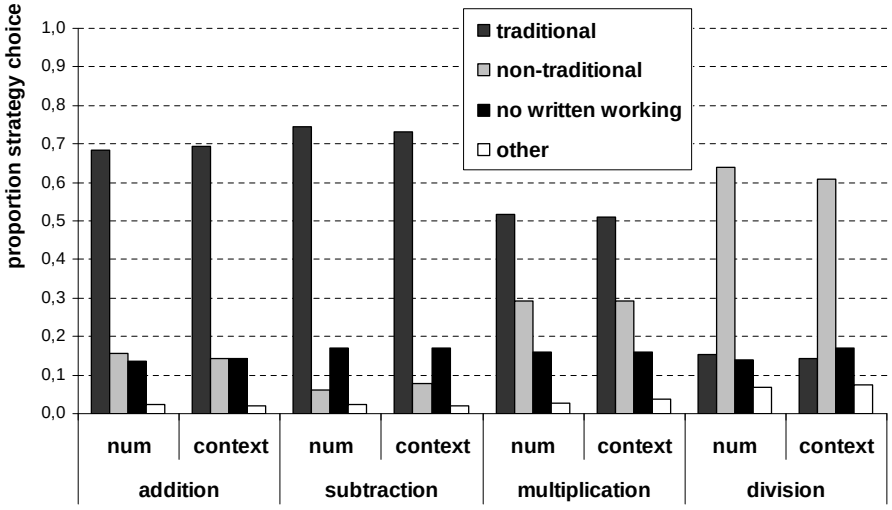


FIGURE 7.3 *Strategy choice proportion of recoded solution strategies on numerical (num) and contextual (context) problems, per operation.*

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.

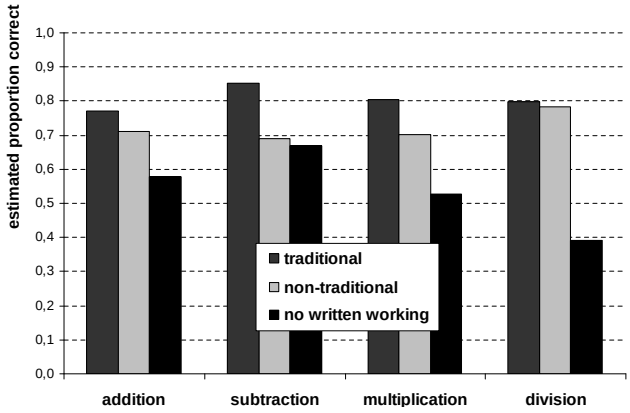


FIGURE 7.4 *Estimated mean accuracy of the three strategies, by operation.*

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

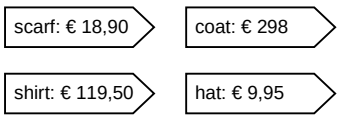
Interestingly, girls outperformed boys on the multidigit arithmetic problems, in contrast to international (PISA, TIMSS) educational assessments findings. Dutch national assessments, however, reported the trend that multidigit operations were the only domain of mathematics on which boys *did not* outperform 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

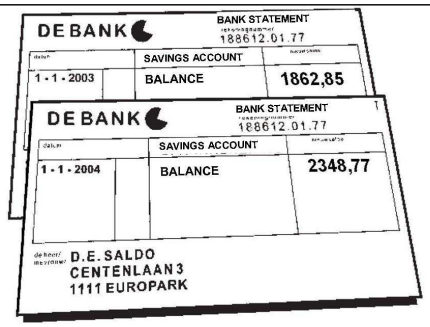
in the current study. A tentative explanatory mechanism may be that girls may better understand the problem text of a contextual problem than boys, because they have been consistently found to have a higher reading ability level than boys in all countries participating in PISA-2009 (OECD, 2010) as well 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 =
 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 =

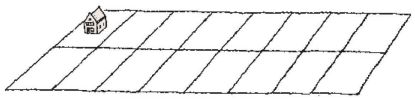
subtraction problems

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" number / price per book / total price **history textbooks** — **32** — [redacted] — **€ 736** The Sunflower has bought 32 new history textbooks. How much is the price per book?	864 : 36 =
 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	subtraction	multiplication	division
①	1 1 677,50 975 + 1652,50	9 10 2 Ø Ø 10 3010 689- 2321	36 27x 252 720+ 972	32/736\23 64 96 96 0
②	677,50 975 + 1500 140 12 0,50 1652,50	3010 689- 3000 -600 -70 -9 2321	36 27x 42 210 120 600+ 972	736 640- 20x 96 96- 3x 0 23x 736 320- 10x 416 320- 10x 96 64- 2x 32 32- 1x 0 23x
③	677,50 + 900 = 1577,50 1577,50 + 70 = 1647,50 1647,50 + 5 = 1652,50	3010 - 600 = 2410 2410 - 80 = 2330 2330 - 9 = 2321	36 x 20 = 720 36 x 7 = 252 720 + 252 = 972	
④	600 + 900 = 1500 70 + 70 = 140 7,50 + 5 = 12,50 1500 + 140 + 12,50 = 1652,50	3000 - 600 = 2400 2400 - 80 = 2320 2320 + 10 - 9 = 2321	30 x 20 = 600 30 x 7 = 210 6 x 20 = 120 6 x 7 = 42+ 972	10 x 32 = 320 20 x 32 = 640 3 x 32 = 96 23 x 32 = 736
⑤		689 + 11 = 700 700 + 300 = 1000 1000 + 2010 = 3010 11 + 300 + 2010 = 2321	27 27 ... 27+ 972	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

CHAPTER 8

General discussion

This thesis opened with the statement that children's mathematical ability is a hotly debated topic. The purpose of the research presented in this thesis was to move beyond personal sentiments and ideological beliefs, by empirically investigating several aspects of primary school students' arithmetic ability in contemporary mathematics education. Specifically, one quantitative research synthesis of performance outcomes of different mathematics programs or curricula, and six empirical research articles that studied determinants of children's mathematical ability, were presented. Starting points for this research were recent developments in mathematics education, in particular the reform movement going by the name of Realistic Mathematics Education (RME), and developments in Dutch primary school students' mathematics performance level, as reported in national and international large-scale assessments.

Chapter 1, presenting a research synthesis of empirical studies (intervention studies and curriculum studies) carried out in the Netherlands that addressed the relation between mathematics instruction or curriculum and students' mathematics performance outcomes, yielded no univocal conclusion. There were few methodologically sound intervention studies comparing different instructional approaches, and the available studies were limited in several aspects such as sample size or content domain. In addition, didactical and instructional aspects were commonly confounded in the programs compared. The curriculum studies, comparing performance outcomes of students who were trained with a specific curriculum, were limited in the amount

of control on the implementation, as well as in correction for confounding variables. So, we may conclude that much is unknown about the relation between mathematics program and performance outcomes. In the remainder of this thesis, attention was therefore refocused to other aspects of students' mathematical ability in contemporary mathematics education, such as solution strategies that students use to solve arithmetic problems, and the effects of presenting mathematics problems in mathematics tests in a realistic context. In these six empirical studies, we aimed to increase our insights in different aspects of primary school students' mathematical ability. In total, data of nearly 5,000 primary school students from grades 1, 2, 3, and 6 were analyzed.

These empirical studies cross the border between the scholar fields of substantive educational and cognitive psychology on the one hand and psychometrics on the other. Several returning themes were solution strategies, individual differences, explanatory variables, and latent variable modeling. Studying *solution strategies* was deemed relevant from an educational psychology perspective, because they are a spearhead of mathematics education reform, as well as from a cognitive psychology perspective where the work of Siegler and his colleagues has initiated a large thread of research into strategic competence and mechanisms of strategy choice. The substantive concepts of *individual differences*, of continuous or of categorical nature, were translated to the psychometric field of *latent variable models*, in particular latent class analysis (LCA) and item response theory (IRT). Finally, incorporating *explanatory variables* in the statistical analyses – among which the latent variable models – made it possible to study differences between groups of students (such as boys and girls), between different types of mathematics problems (such as with and without a context), and between different solution strategies (such as written and mental computation). In all studies, the relevance of the results for educational practice received considerable attention.

In the first two empirical studies (Chapters 2 and 3), secondary analyses of the raw data collected in the Dutch national mathematics assessments at the end of primary school (PPON) were carried out. These studies aimed to get more insight in students' performance level in complex or multidigit multiplication and division, by incorporating information on students' solution strategy use. This performance level was found to decrease over time and to stay far behind educational standards. In the next two empirical studies, new data were collected to study characteristics of written and mental solution strategies in complex division problem solving (such as strategy distribution, accuracy, speed, and adaptivity) in an unbiased manner, by a partial (Chapter 4) and a full

(Chapter 5) choice/no-choice study (cf. Siegler & Lemaire, 1997). The final two empirical studies addressed the effects of presenting mathematics problems in realistic – usually verbal – context, as is common practice in contemporary mathematics instruction and mathematics tests. Both students in the early grades of primary school (Chapter 6) and in the final grade (Chapter 7) were studied.

The remaining part of this discussion is subdivided into two sections. First, the main substantive psychological findings and the (educational and cognitive) implications of the six empirical studies are discussed. Second, we reflect on the statistical modeling approaches used and their contributions to the field of psychometrics.

8.1 Substantive findings

8.1.1 Solution strategies in complex arithmetic problems

Lemaire and Siegler (1995) distinguished four aspects of *strategic competence*: strategy repertoire (which strategies are used), strategy distribution (the frequency with which the strategies are used), strategy efficiency or performance (strategy speed and/or accuracy), and strategy selection or adaptivity (how strategies are chosen, related to problem characteristics and individual strategy characteristics). These aspects, in particular strategy choice or selection and strategy accuracy, are key features in five of the six empirical studies (only Chapter 6 did not address solution strategies). These five studies were all carried out in the domain of complex or multidigit arithmetic with sixth graders (12-year-olds). The main solution strategy categories distinguished in these studies were the traditional standard algorithm that proceeds digit-wise, non-traditional procedures that work with whole numbers, answers without written working, and other strategies (unclear or wrong strategies, and skipped items). A subcategory of the non-traditional strategies are the RME approaches (called *column calculation* by the developers, see Treffers, 1987, and Van den Heuvel-Panhuizen, 2008). These strategies 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).

Complex division (e.g., $432 \div 12$) received most attention in this thesis: All five studies analyzing solution strategies addressed complex division. Division was considered important because the largest performance decrease in the national assessments was observed in this domain (J. Janssen et al., 2005). Moreover, the replacement of the

traditional long division algorithm by the RME-alternative of column calculation (Van den Heuvel-Panhuizen, 2008) – which for division means repeated subtraction of multiples of the divisor from the dividend (see Figure 2.1 in Chapter 2 for an example) – in the learning/teaching trajectories and in the mathematics textbooks makes it a prototype of mathematics education reform. Complex multiplication was addressed in two studies (Chapters 3 and 7), and complex addition and subtraction only in Chapter 7.

Strategy selection in multiplication and division: general patterns and shifts over time

In Chapter 2, students were found to be quite consistent in the type of strategy (traditional, non-traditional, no written working or other) they chose on a set of division problems. However, shifts in the relative frequency of the different strategy choice classes were observed. In line with the disappearance of the traditional division algorithm from the textbooks, the percentage of students predominantly using this strategy decreased between the PPON-assessments of 1997 and 2004. Unexpectedly, however, the percentage of students using predominantly the RME-based repeated subtraction strategy remained about constant. What did increase on the other hand, was the percentage of students consistently answering without any written work, presumably indicative of mental computation (as was supported by findings in Chapter 4). In the other three studies in which solution strategies for division problems were studied (Chapters 4, 5, and 7), the traditional division algorithm was also used rather infrequently, so this appeared to be a robust pattern. Furthermore, Chapter 3 showed that the traditional algorithm was almost exclusively used by students whose teachers instructed it, supporting the influence of the curriculum on students' problem solving behavior. The frequency of using mental computation, however, was more variable over the different studies, and the high frequency found in PPON-2004 (44%) was never matched in later studies. The large frequency found in PPON may thus be considered somewhat exceptional, and it will be interesting to find out whether it carries on in the upcoming subsequent assessment cycle at the end of grade six, for which the data collection is planned to take place in 2011.

In complex or multidigit multiplication, the traditional algorithm (see for example Figure 3.2 in Chapter 3) is still the end point of the contemporary learning/teaching trajectory, contrary to complex division (Van den Heuvel-Panhuizen, 2008). The majority of the sixth grade teachers (88%) also instructed it in PPON-2004 (as the only strategy or

in combination with column calculation), and in the assessments of 1997 and 2004 it was the dominant strategy students used to solve multiplication problems. The dominance of the traditional algorithm in multiplication is supported by the findings of Chapter 7, where more than 50% of the multiplication problems were solved with the traditional algorithm. Like in division, shifts in strategy choice over time between PPONs 1997 and 2004 were found in multiplication too. Similar to division, a decrease in use of the traditional algorithm and an increase in answering without written work were observed. However, this increase in the no written work strategy was smaller in multiplication than it was in division. Moreover, non-traditional multiplication strategies were used more frequently in 2004 than in 1997, in contrast to division where the relative frequency of these strategies remained roughly stable. This latter difference between multiplication and division is striking since one would rather expect to find the opposite pattern, because non-traditional strategies have become the standard approach for division in learning/teaching trajectories, while they are not standard in multiplication.

Strategy accuracy differences

How should we evaluate the decrease in the traditional written algorithm and the increase in using mental computation (multiplication and division) and the increase in non-traditional strategies (multiplication only)? One way to look at this shift is to consider the effects on performance, by comparing the accuracies (probability of a correct answer) of the different strategies. One consistent finding in this thesis was that written computation strategies – including complete solution procedures as well as only intermediate answers – were more accurate than non-written (mental) computation strategies across the operations division (Chapters 2, 4, 5, and 7), multiplication (Chapters 3 and 7) and addition and subtraction (Chapter 7). In other words, the observed shift between 1997 and 2004, showing a decrease in written strategies and an increase in mental strategies in multiplication and division, turned out unfortunate with respect to performance outcomes. Importantly, Chapter 4 showed that forcing students who spontaneously used a mental strategy when solving a complex division problem, to use a written strategy on a parallel problem, improved their performance. Therefore, a reasonable recommendation seems to be that teachers should encourage the use of writing down solution steps or solution strategies, emphasizing the value both in schematizing information and in recording key items (Ruthven, 1998). This may be particularly relevant for boys (who

are more inclined to use mental computation) and for low mathematics performers (who showed the largest performance gap between mental and written strategies). In addition, in Dutch secondary education, it is common practice to evaluate students' entire work, not merely the final answer given. Re-emphasizing the value of written work in primary education may therefore also smoothen the transition to secondary education mathematics.

Another relevant comparison is between the accuracy of traditional and non-traditional strategies. A recurring finding in this thesis was that the traditional algorithm was usually equally accurate in division (Chapter 2 – note however that this only held for low and high achievers; for medium achievers the traditional division algorithm was significantly more accurate than non-traditional strategies – and Chapter 7) and more accurate in multiplication (Chapters 3 and 7) and in subtraction and addition (Chapter 7; for subtraction see also Van Putten & Hickendorff, 2009). Although these non-traditional strategies included a wide range of different approaches, and comparisons were hampered by selection effects because they were based on different students and/or different items (cf. Siegler & Lemaire, 1997), these patterns raise questions on the desirability of learning/teaching trajectory end-points other than the traditional algorithm. Combined with the pattern emerging from the review in Chapter 1 and international reviews (e.g., Kroesbergen & Van Luit, 2003; Swanson & Carson, 1996) that low mathematics achievers benefit from a more directing instruction, these students in particular may need instruction in one standard procedure to solve a problem. We argue that this standard strategy should preferably be the traditional algorithm.

It is important to note that the algorithms can also be learned with insight in what is going on (e.g., Lee, 2007). On a related note, instructing for procedural knowledge – such as skill in the traditional algorithm – does *not* imply that only isolated skills and rote knowledge are developed. As Star (2005) argued, the *knowledge type* distinction in procedural versus conceptual knowledge is perpendicular to the dimension of *knowledge quality* ranging from superficial to deep knowledge. In discussions about mathematics education, however, these two distinctions are often entangled, with conceptual knowledge being considered deep and procedural knowledge considered superficial. The other two combinations, deep procedural knowledge and superficial conceptual knowledge should be recognized as well. We reach a similar conclusion as Gravemeijer (2007) did earlier, that the dual aim of teaching/learning trajectories based on 'column calculation' – attaining *insight in* and *mastery of* standard procedures – is

currently not attained in mathematics education.

The position of the traditional algorithm in the mathematics curriculum has been an object of heated debate. On the one hand, Gravemeijer (2007) for instance made a plea not to focus so much on standard procedures, because they require a large investment of instructional time and practice in order to attain fluent skill. On the other hand, for example Van der Craats (2007) argued that these procedural skills are at the core of mathematics, and should therefore receive much more instruction, drill, and practice, than they receive now. Recent developments in educational policy suggest that basic skills have received renewed attention. For example, a committee has been installed with the mission to define *reference levels* – desired performance outcomes of mathematics education – for several time points in the primary and secondary school years (Expertgroep Doorlopende leerlijnen Taal en Rekenen, 2008). This committee claimed (p. 32-33) that shifts in focus in the mathematics curriculum in the domain of *numbers and operations* are undesirable as long as the general society and educational community have not reached agreement. Currently, fluently solving complex arithmetic problems with standard written procedures are still considered an educational objective (Dutch Ministry of Education, Culture, and Sciences, 2006), so decreased attention for this domain in educational practice may be considered to be unwarranted. An interesting related observation is that 41% of the sixth grade teachers reported instructing the traditional division algorithm in PPON-2004 (as the only strategy or in combination with column calculation), thereby diverging from the *intended curriculum* (Porter, 2006) as formulated in the learning/teaching trajectory (Van den Heuvel-Panhuizen, 2008). Apparently, a substantial minority of the teachers feel that the traditional division algorithm should be included in the mathematics curriculum.

The unexplained part of the performance decrease ...

By taking into account the solution strategies students used, we found a partial explanation of the performance decrease between 1997 and 2004 on multiplication and division problems. That is, a shift in strategy choice, characterized by a decrease in the accurate traditional algorithm and an increase in less accurate mental computation, and in multiplication also an increase in less accurate non-traditional strategies, contributed significantly and substantively to the drop in performance. However, this shift could only partially account for the performance decline: a substantial part that was unaccounted

for remained. That is, within each of the main strategies, the accuracy in PPON-2004 was significantly lower than in PPON-1997. There are no empirical data available in the assessments to study what caused this general accuracy decrease, so we can only revert to more tentative hypotheses, such as the lower value attached to these domains in general, and less opportunity to learn (instruction and practice) in solving these kinds of problems. Evidently, more research is needed.

Adaptive expertise

Related to these above findings on solution strategies is the current aim of mathematics education reform to attain *adaptive expertise*, the ability to solve mathematics problems efficiently, creatively, and flexibly, with a diversity of strategies (Baroody & Dowker, 2003; Torbeyns, De Smedt, et al., 2009b). There are several findings suggesting that students do not make adaptive strategy choices. Most notably, because mental strategies were found to be less accurate than written strategies – both in comparisons *between* and *within* different students and items – the question arises why students choose these 'risky' mental strategies. Chapter 5 suggests that mental computation was mainly chosen for its speed advantage, while the accuracy was considered less important. Moreover, a substantial part of the students did not choose their 'best' strategy – defined as the one leading fastest to an accurate answer – on a problem. These apparent suboptimal strategy choices contrast with predictions from cognitive models on strategy choice (e.g., Shrager & Siegler, 1998; Siegler & Shipley, 1995), that presume that the main determinant of an individual's strategy choice on a particular problem is the individual's strategy performance characteristics for that problem.

These cognitive models are not explicit in the influence of individual differences in the speed-accuracy preferences (the relative weighing of accuracy and speed; Ellis, 1997; Phillips & Rabbitt, 1995) that may cause some students to choose fast but more error-prone mental computation. Furthermore, these models have been argued to ignore aspects of the sociocultural context, such as sociomathematical norms (Ellis, 1997; Luwel et al., 2009; Verschaffel et al., 2009). Ellis pointed out the possibility of (sub)cultural differences in the weights assigned to speed versus accuracy of performance, and the value placed on solutions constructed in the head versus by means of external aids. For instance, classroom socio-mathematical norms and practices valuing speed over accuracy and/or mental strategies over written ones, may result in students overusing

mental strategies at the cost of accuracy. We tentatively argue that due to the importance of mental computation in RME-based mathematics education, the socio-mathematical norms in the classroom are such that mental computation is considered superior to written computation. Although we acknowledge that mental computation is an important competence, we argue that it should not overshadow the competence of using written strategies fluently. A related interesting finding was that on the division problems in PPON-2004, the frequency of answering without written work (as well as of skipping problems entirely) were highest in students whose teacher instructed exclusively the RME-strategy for division, tentatively suggesting students receiving more RME-based instruction valued mental computation over written computation to a larger extent than students who received a more traditional instruction. In multiplication, however, teachers' strategy instruction did not seem to affect the frequency of answering without written work, so the results are not consistent in this respect.

Two patterns found suggest adaptivity in strategy choices to some extent. First, in Chapter 4, individual differences in strategy choices on division problems showed that there were three subgroups of students: students who consistently used written computation, students who consistently used mental computation, and students who switched from written computation on the problems with more difficult number characteristics to mental computation on the problems with easier numbers. The latter group seemed to adapt their strategy choices to the problem characteristics, and thus showed some strategy adaptivity. Second, in Chapters 4 and 5, there were several division problems with number characteristics such that a compensation strategy (rounding the dividend) would be a very efficient approach. Within written strategies, only a small proportion involved this compensation approach, while, in contrast, the majority of the mental strategies involved compensation. Given the fact that the compensation strategy is more efficient, in the sense that it requires fewer computational steps, the finding that students applying a compensation strategy usually did it mentally, while those who did not use a compensation strategy predominantly used a written strategy, is an indication that to some extent an adaptive strategy choice was made.

Also interesting in this respect are findings from another study that was carried out (not included in the current thesis). That study (Hickendorff, 2010c) addressed Dutch sixth graders' use of *shortcut strategies* [in Dutch: *handig rekenen*] on complex arithmetic problems with number characteristics expected to elicit efficient strategies, like indirect addition on subtraction problems, and compensation strategies on multiplication and

division problems. Results showed that such shortcut strategies were used rather infrequently, on between 5% and 20% of the trials, and were equally accurate as non-shortcut strategies. In addition, an explicit hint to *"Solve the problems as clever as possible. Have a close look at the numbers"* hardly increased the frequency of use. These findings thus do not yield much support for the adaptation of strategy choices to problem features, supported by the relatively low frequency of shortcut strategies found in studies with younger children in Belgium, Germany, and the Netherlands (Blöte et al., 2001; De Smedt, Torbeyns, Stassens, Ghesquière, & Verschaffel, 2010; Heinze, Marschick, & Lipowsky, 2009; Selter, 2001; Torbeyns, De Smedt, Ghesquière, & Verschaffel, 2009a; Torbeyns, De Smedt, et al., 2009b; Torbeyns, Ghesquière, & Verschaffel, 2009). As Torbeyns, De Smedt, et al. (2009a, footnote 5) argued, shortcut strategies are not easy strategies, and fluent application requires a sufficient amount of practice. We argue that in current mathematics education, the ease of discovery and application of such strategies, and thereby the efficiency and value of these strategies, may be overrated.

8.1.2 Differences between problems and between students

Problem characteristics

We discuss the effects of two problem characteristics: the operation required (addition, subtraction, multiplication, and division) and the problem format (contextual or numerical problem).

To start with the latter aspect, an often-heard complaint about contemporary mathematics tests is that students with low language or reading skills are disadvantaged by the large number of contextual problems, because it is a necessary condition to understand the problem text to solve the mathematics problem. Findings in this thesis on this issue were mixed: in lower grades (Chapter 6) we found that solving contextual and numerical arithmetic problems involved different abilities (in a technical sense, different individual differences dimensions). Moreover, the performance of students with a lower language level (a non-Dutch home language or low reading comprehension level) lagged behind that of students with a higher language level to a larger extent in solving contextual problems than in solving numerical problems. A direct assessment of whether a context made a problem easier or more difficult, however, was not possible in this study. By contrast, in the study with sixth graders (Chapter 7) it was possible to test this effect directly, and strikingly, hardly any effects of problem format (contextual

versus numerical) were found on performance, strategy choice, and strategy accuracy. Furthermore, the absence of an effect held independently of students' home language and language performance level. The findings of Chapters 6 and 7 taken together suggest that the effects of contexts in mathematics problems decreases with more years of formal schooling, and that the type of contexts used in often-used mathematics tests from CITO do not disadvantage any of the groups of students distinguished at the end of primary school. However, given the findings in the lower grades, more balance between problems with and without a context in mathematics education and in mathematics assessments may be called for, as was also recommended in the KNAW (2009) report.

Regarding differences between problems by operation required (addition, subtraction multiplication, and division), we review the findings of Chapter 7, in which all four operations were studied simultaneously. The following pattern of strategy choices emerged: the frequency of the traditional algorithm was highest for addition and subtraction, lowest for division, and in between for multiplication. This pattern is consistent with the position of the traditional algorithm in the learning/teaching trajectories (Van den Heuvel-Panhuizen, 2008), and also with findings on multiplication and division in the national assessments (Chapter 3). Moreover, addition and subtraction can be considered lower in the arithmetic hierarchy than multiplication and division, because for success in the latter, skill in the former is necessary. Therefore, fluent skill in one standard procedure may be more essential for addition and subtraction than for multiplication and division.

Student characteristics

Throughout this thesis, the effects of the student characteristics gender and general mathematics level on different aspects of mathematical ability (overall performance, strategy choice, strategy accuracy, and strategy adaptivity) have been addressed recurrently.

Gender differences were addressed in five studies (Chapters 2, 3, 4, 5, and 7). Regarding performance, all studies showed slight advantages for girls (usually non-significant, but significant in Chapter 7), which is in contrast with the consistent pattern from national (J. Janssen et al., 2005; Kraemer et al., 2005) and international assessments (Mullis et al., 2008; OECD, 2010) that boys tend to outperform girls on most mathematics domains in the majority of the countries, including the Netherlands. However, complex

arithmetic may be the exception, as (small) girl advantages on these domains were also found in the Dutch national assessments in grade 6. Tentative explanations may be that this domain lends itself pre-eminently for applying structured, algorithmic approaches, something that girls have been found to favor more than boys (Carr & Davis, 2001; Carr & Jessup, 1997; Gallagher et al., 2000; Timmermans et al., 2007).

In line with this reasoning, we found very clear and consistent gender differences in strategy choice on the complex arithmetic problems of all four operations: girls were more inclined to use written strategies, in particular the traditional algorithm, while boys were more inclined to use mental computation. In particular, the observed strategy shift between PPONs 1997 and 2004 in multiplication and division towards an increase in mental computation could even be predominantly attributed to boys.

In none of the studies, gender differences in the accuracy with which these strategies were executed were found, suggesting that the (slight) advantage of girls in performance is mediated by their choice for more accurate strategies. In Chapter 5, strategy speed was addressed, and boys were faster with forced mental computation than girls. Consequently, for boys the speed gains of choosing mental strategies over written ones was larger than for girls, which may partially account for boys' larger inclination of choosing mental strategies. In addition, boys and girls appeared to have different speed-accuracy preferences. Girls appeared to fit their strategy choices to accuracy considerations, ignoring speed, while boys had a preference for speed over accuracy. This may be related to individual differences in the confidence criterion that have been reported in children (Siegler, 1988a, 1988b) and in adults (Hecht, 2006). In addition, girls have been consistently found to have lower levels of confidence with mathematics (Mullis et al., 2008; Timmermans et al., 2007; Vermeer et al., 2000), and as a result may act more cautiously than boys and therefore choose the safety of using slower, well-structured, written strategies. In line with this reasoning, girls have been found to be less inclined to intellectual risk-taking than boys (Byrnes et al., 1999) and more inclined to (academic) delay of gratification (Bembenutty, 2009; Silverman, 2003). All these gender differences together might partially explain that boys more often choose fast mental calculation over slower but more accurate written computation.

In four studies, the effects of students' general mathematics achievement level were studied (Chapters 2, 3, 4, and 5). Not surprisingly, students with higher mathematics level performed better on the complex arithmetic problems overall, had a higher accuracy within each strategy, and were faster within each strategy, than students with lower

mathematics level. An interesting differential effect in strategy accuracy was found: the accuracy advantage of written over mental strategies decreased for students with higher mathematics level, as was found in Chapters 2, 3, and 4. The results regarding differences in strategy choice were somewhat mixed: Chapters 2 and 3 showed that in multiplication and division problems of PPONs 1997 and 2004, students with low mathematics level were more inclined to use a non-written strategy or skip the item than medium and high level students. The latter more often used written strategies, in particular the traditional algorithm. However, in Chapter 5, no differences in the tendency to choose mental strategies as a function of mathematics achievement level were found.

There are clear indications that there are differences in the adaptivity of strategy choices as a function of students' mathematics achievement level. That is, weak students very infrequently classified as 'switchers' (adapting strategy choices to problem characteristics) in Chapter 4, and Chapter 5 showed that below-average achievers did not take either accuracy or speed into account in their strategy choices, while above-average achievers fitted their strategy choices to both performance components. Other studies also reported that students of higher mathematical ability choose more adaptively between strategies than students of low mathematical ability (Foxman & Beishuizen, 2003; Hickendorff, 2010c; Torbeyns, De Smedt, et al., 2009b; Torbeyns et al., 2002, 2006). Similarly, the research synthesis of Chapter 1 showed that low mathematics performers who were instructed in a more free form (i.e., guided instruction) did show a larger strategy repertoire than students who were trained with a more directing instruction, but they did not use this larger repertoire more flexibly or adaptively. In other words, it seems that we did not yet succeed in an instructional approach fostering adaptive expertise for the low mathematics performers. A recommendation may be to devote more educational attention to teaching students to make informed choices for mental or written strategies: when is a mental strategy 'safe' enough, and when is it better to revert to written strategies? Moreover, questions can be raised to the general attainability and feasibility of adaptive expertise for low mathematics performers (see also Geary, 2003; Torbeyns et al., 2006; Verschaffel et al., 2009).

8.2 CONTRIBUTIONS TO PSYCHOMETRICS

In the current thesis, advanced psychometric modeling techniques were used to approach the substantive research questions posed. The most notable application of psychometric modeling of the current thesis was to use *latent variable models* to analyze individual differences between students. Moreover, to move beyond mere measurement of individual differences, the influence of different *explanatory variables* was addressed to study differences between groups of students, between problems with different characteristics, and between solution strategies (student-by-item variables). In short, our approach can be called *explanatory latent variable modeling*. Different aspects are reflected on in the following sections.

8.2.1 Explanatory latent variable modeling

The substantive concept of individual differences was translated to the psychometric field of latent variable models, in particular latent class analysis (LCA) and item response theory (IRT). These models made it possible to analyze complicated data structures consisting of repeated observations (items within students) of dichotomous (correct/incorrect) and/or categorical (solution strategies) measurement level (see Chapter 2).

Latent class analysis (e.g., Goodman, 1974; Lazarsfeld & Henry, 1968) models qualitative (i.e., categorical) individual differences that are measured with categorical observed variables. It is a model-based version of cluster analysis. These models were found to be very useful in analyses of individual differences in strategy choice, searching for latent subgroups of students who are characterized by a specific strategy choice profile over a set of items. To assess the effect of student-level explanatory variables on these latent classes, we included these variables as covariates predicting latent class membership (e.g., Vermunt & Magidson, 2002).

In such an approach, the conditional probabilities (the probability of responding in a particular category on a particular item, given membership of a particular latent class) are unaffected by the covariates, implicitly assuming that the influence of the covariates on the item responses is completely mediated by the latent class variable. This assumption may be relaxed by allowing for direct effects of covariates on observed variables, something that we did not try in the current thesis. Furthermore, in the latent class analyses in the current thesis, there were conditional probabilities for each item separately, making the model quite complex (i.e., with a large number of parameters).

A more parsimonious alternative would be to restrict the conditional probabilities on a set of equivalent items to be equal to each other (Hickendorff, Heiser, Van Putten, & Verhelst, 2008). However, this a a rather stringent assumption. Another approach would be to apply *latent class regression analysis* (e.g., Bouwmeester, Sijtsma, & Vermunt, 2004), in which the effects of particular item features instead of of individual items on strategy choice are modeled. That, however, would require a systematic specification of the features of each item, which is hardly possible in the current empirical studies.

As statistical software to fit the LCA models, we used two programs: LEM (Vermunt, 1997), a general versatile program for the analysis of categorical data, and the `poLCA` package (Linzer & Lewis, 2011, 2010) available in the statistical computing program `R` (R Development Core Team, 2009). With a sufficient number of random starts to avoid locally optimal solutions, these two packages yielded the same results .

Item response theory models (e.g., Embretson & Reise, 2000; Van der Linden & Hambleton, 1997) model quantitative (i.e., continuous) individual differences, and are therefore very suitable to analyze performance. With explanatory IRT-analyses (De Boeck & Wilson, 2004; Rijmen et al., 2003), the effects of explanatory variables at the student level, item level, and student-by-item level, as well as interactions between these variable types, could be studied. For example, in Chapter 7 the interaction effect between the student-level variable *home language* and the item-level variable *problem format* was tested, in order to assess whether problems in a contextual format were relatively more difficult compared to numerical problems for students who did not speak Dutch at home than for their native peers.

Furthermore, the individual differences dimensions need not be one-dimensional. In Chapter 6, in a two-dimensional between-item IRT model (e.g., Adams et al., 1997; Reckase, 2009), two performance dimensions were distinguished. The ability to solve numerical problems and the ability to solve contextual problems appeared to be highly related but still distinct in the lower grades in primary school. In sixth grade, however, these dimensions appeared to be statistically indistinguishable. Given the distinctness of the performance dimensions of solving contextual problems and solving numerical problems in early grades, it would be recommendable to somehow report on these two dimensions separately, because this may yield diagnostic information on potential remedial an instructional benefit (De la Torre & Patz, 2005). In cases were there is essentially one dominant factor or highly correlated dimensions, MIRT modeling has been shown to yield subscale scores that have improved reliability over unadjusted

subscale scores (total scores), because the correlational structure is taken into account (De la Torre and Patz; Stone et al., 2010). However, Sinharay et al. (2010) showed that caution with reporting subscale scores is needed: they have added value over reporting the total score only if the reliability of the subscales is large enough and if the dimensions are sufficiently distinct.

A potentially fruitful alternative to choosing between unidimensional and multidimensional IRT models may be a procedure called *profile analysis* (Verhelst, 2007, in press), that is being used in the most recent edition of CITO's Student Monitoring and Evaluation System (see J. Janssen & Hickendorff, 2009). In this approach, the item parameters of a unidimensional IRT model are estimated. However, different item categories (such as basic skill and applied problem solving) are distinguished. These categories are used in the next step, to determine for each student the deviation of his or her observed response profile on these item categories from the expected response profile under the unidimensional model, with a disparity index. Students (or groups of students) who show large disparities do not respond consistently with the unidimensional model, but show specific strengths and weaknesses on some item categories, *conditional on* their total score. Such deviant profiles may yield valuable diagnostic information for individual students, as well as for groups of students (e.g., different countries).

With respect to the estimation of explanatory IRT models, De Boeck and Wilson (2004) showed that item response models in marginal maximum likelihood (MML) can be formulated in the generalized (non)linear mixed model (GLMM) framework. This formulation makes it possible to use mainstream statistical software platforms, such as the NLMIXED and GLIMMIX procedures from SAS (SAS Institute, 2002), or the `lmer` function from the `lme4` package (Bates & Maechler, 2010) available in the statistical computing environment R (R Development Core Team, 2009), as described in De Boeck et al. (2011). These statistical packages differ in the way they approximate the maximization of the likelihood in parameter estimation (see Equation 2.5 in Chapter 2), and have their own advantages and disadvantages. For example, NLMIXED approximates the integral with a Gauss-Hermite quadrature procedure (numerical integration), and is therefore very accurate with a sufficient number of quadrature points but also very slow, in particular with multidimensional IRT models. That is, the complexity of the estimation problem is exponentially related to the number of dimensions. However, it is the only of the three packages allowing for item discrimination parameters. The `lmer` function approximates the integrand with a Laplace procedure making it very fast, but it

results in slightly biased parameter estimates, in particular for the random effects. An advantage over NLMIXED is that it is possible to estimate models with *crossed* random effects: simultaneous random effects over different modalities, such as individuals and items (De Boeck, 2008). Finally, the GLIMMIX procedure approximates the integrand with quasi-likelihood procedures (PQL or MQL) and produces seriously biased results on the random effect parameters.

In a small comparative study (Hickendorff, 2010a), these three statistical packages were compared. A two-dimensional IRT model was fitted on the correct/incorrect responses of 1546 sixth-graders to the multiplication and division problems of PPON 1997 and 2004. The variance estimates of the first dimension were 1.56 (SE = .19) with NLMIXED, 1.37 (no SE estimated) with `lmer`, and 1.08 (also no SE estimated) with GLIMMIX, illustrating the downward bias of random effects parameter estimates in procedures that approximate the integrand (see also Molenberghs & Verbeke, 2004). The respective latent correlation estimates were .87 (SE = .04), .93, and .96: clearly different from each other. A practical recommendation may be to start model building with `lmer` because of its superior speed and least amount of bias, and re-analyze the final model(s) with NLMIXED for the most accurate results.

One innovative application of explanatory IRT models in the current thesis was to use the solution strategy that a student used to solve a particular item as a student-by-item explanatory variable, as explained in Chapter 2 and done throughout the thesis. By doing so, it was possible to statistically test the difference in accuracy between the strategies while accounting for individual ability differences in overall performance and difficulty differences between problems, something that was not achieved before in studies into solution strategies. The strategy accuracy differences could be modeled to be item-specific (see Figure 2.4 in Chapter 2), or restricted to be equal for some or all items. Although this restriction made the model far more parsimonious and allowed for testing interaction effects between strategies and student-level variables, it was quite a stringent constraint. With the possibility to model crossed random effects in `lmer` (De Boeck et al., 2011), an intermediate alternative seems to be to model the strategy effects as random over items, as was done in Hickendorff (2010a) for the multiplication problems of PPON 1997 and 2004. In that approach, the strategy effects averaged over items are estimated, as well as the variance of this effect over the items. An alternative interpretation is that the item difficulties *per strategy* are modeled as random over items. Furthermore, analyzing the item difficulty per strategy is related to the issue of *differential item functioning* (DIF),

with items not functioning in the same way for different groups of students (in this case, characterized by their strategy choice). Further research is necessary to investigate this approach in more detail.

8.2.2 Final remarks

This thesis concludes with two final remarks. The first one concerns carrying out secondary analyses on data that were collected in large-scale assessments to answer new research questions, as was done in Chapters 2 and 3. These secondary analyses turned out to yield valuable new insights in patterns reported in the national mathematics assessments. There are also other advantages: it is relatively inexpensive because no new data have to be collected, and one can stay close to findings of the original assessments one aims to explain (i.e., they are based on the same problems and same representative sample of students, so these variables cannot confound the results).

However, there are also disadvantages (e.g., Van den Heuvel-Panhuizen et al., 2009), and one major limitation is the fact that the data were collected with a purpose (reporting on the outcomes of the educational system) other than answering the newly posed research questions. One has to make do with what one has. As a consequence, the influence of factors that were not varied systematically, like problem characteristics, cannot be tested directly (Hickendorff et al., 2009a). However, it may yield new hypotheses that can direct new research efforts, as was for example done in the current thesis. Furthermore, we recommend to collect more information in the national assessments on the intended and enacted mathematics curriculum, in order to study the entire chain of curricular materials, teacher interpretation, curricular enactment, and student learning more thoroughly (Hickendorff et al., 2009a; Stein et al., 2007). In addition, we plead for a more multidisciplinary approach in which didactical experts, educational researchers, cognitive psychologists, and experts in educational measurement cooperate to get the most out of large-scale educational assessments.

The second remark concerns the mutual value of crossing the border between psychometrics and psychology. The kind of advanced statistical analyses applied in the current thesis are rather scarce in the field of educational and cognitive psychology. However, we argue that these approaches are better suited to answer the substantive research questions commonly posed in these fields than more traditional analyses such as classical test theory, in particular when it concerns data on solution strategies. In that

respect, psychometrics can advance the field of psychology. This positive influence may also hold in the other direction: psychology may advance the field of psychometrics. As Borsboom (2006) argued, psychometrics has not yet succeeded in getting integrated with mainstream psychology. However, psychometrics is an applied science and it is therefore essential that psychometricians avoid a state of isolation.

References

Abedi, J., & Hejri, F. (2004). Accommodations for students with limited English proficiency in the National Assessment of Educational Progress. *Applied Measurement in Education*, *17*, 371-392.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*, 219-234.

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

Adams, R. J., & Wu, M. L. (2000). *PISA 2000 technical report.* Paris: OECD.

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Albert, J. H. (1992). A Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251-269.

Ambrose, R., Baek, J.-M., & Carpenter, T. P. (2003). Children's invention of multidigit multiplication and division algorithms. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (p. 305-336). Mahwah, NJ: Lawrence Erlbaum Associates.

Anghileri, J. (1989). An investigation of young children's understanding of multiplication. *Educational Studies in Mathematics*, *20*, 367-385.

Anghileri, J., Beishuizen, M., & Van Putten, C. M. (2002). From informal strategies to structured procedures: Mind the gap! *Educational Studies in Mathematics*, *49*, 149-170.

Baroody, A. J., & Dowker, A. (Eds.). (2003). *The development of arithmetic concepts and skills: Constructing adaptive expertise.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bates, D., & Maechler, M. (2010). `lme4`*: Linear mixed modeling using S4 classes.*

(Computer program and manual). Available from `http://cran.r-project.org/web/packages/lme4/index.html`.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541-562.

Beishuizen, M. (1993). Mental strategies and materials or models for addition and subtraction up to 100 in Dutch second grades. *Journal for Research in Mathematics Education*, *24*, 294-323.

Beishuizen, M., Van Putten, C. M., & Van Mulken, F. (1997). Mental arithmetic and strategy use with indirect number problems up to one hundred. *Learning and Instruction*, *7*, 87-106.

Bembenutty, H. (2009). Academic delay of gratification, self-regulation of learning, gender differences, and expectancy-value. *Personality and Individual Differences*, *46*, 347-352.

Berends, I. E., & Van Lieshout, E. C. D. M. (2009). The effect of illustrations in arithmetic problem-solving: Effects of increased cognitive load. *Learning and Instruction*, *19*, 345-353.

Blöte, A., Van der Burg, E., & Klein, A. S. (2001). Students' flexibility in solving two-digit addition and subtraction problems: Instruction effects. *Journal of Educational Psychology*, *93*, 627-638.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425-440.

Bouwmeester, S., Sijtsma, K., & Vermunt, J. K. (2004). Latent class regression analysis to describe cognitive developmental phenomena: An application to transitive reasoning. *European Journal of Developmental Psychology*, *1*, 67-86.

Buijs, C. (2008). *Leren vermenigvuldigen met meercijferige getallen* [Learning to multiply with multidigit numbers]. Utrecht, The Netherlands: Freudenthal Institute for Science and Mathematics Education.

Byrnes, J. P., Miller, D. C., & Shafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin*, *125*, 367-383.

Campbell, J. I. D., Fuchs-Lacelle, S., & Phenix, T. L. (2006). Identical elements model of arithmetic memory: Extensions to addition and subtraction. *Memory & Cognition*, *34*, 633-647.

Carr, M., & Davis, H. (2001). Gender differences in arithmetic strategy use: A function of skill and preference. *Contemporary Educational Psychology*, *26*, 330-347.

Carr, M., & Jessup, D. L. (1997). Gender differences in first-grade mathematics strategy

use: Social and metacognitive influences. *Journal of Educational Psychology*, *89*, 318-328.

CITO. (2005a). *Leerling- en onderwijsvolgsysteem, rekenen-wiskunde groep 3* [Monitoring and evaluation system for primary pupils – Arithmetic and Mathematics, grade 1]. Arnhem, The Netherlands: CITO.

CITO. (2005b). *Leerling- en onderwijsvolgsysteem, rekenen-wiskunde groep 4* [Monitoring and evaluation system for primary pupils – Arithmetic and Mathematics, grade 2]. Arnhem, The Netherlands: CITO.

CITO. (2006). *Leerling- en onderwijsvolgsysteem, rekenen-wiskunde groep 5* [Monitoring and evaluation system for primary pupils – Arithmetic and Mathematics, grade 3]. Arnhem, The Netherlands: CITO.

CITO. (2007). *Terugblik en resultaten 2007. Eindtoets basisonderwijs* [Retrospect and results of the End of Primary School Test 2007]. Arnhem, The Netherlands: CITO.

CITO. (2009). *Terugblik en resultaten 2009. Eindtoets basisonderwijs* [Retrospect and results of the End of Primary School Test 2009]. Arnhem, The Netherlands: CITO.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, *20*, 405-438.

De Boeck, P. (2008). Random item models. *Psychometrika*, *73*, 533-559.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., et al. (2011). The estimation of item response models with the `lmer` function from the `lme4` package in R. *Journal for Statistical Software*, *39*(12), 1-28.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach.* New York, NY: Springer.

De Brauwer, J., & Fias, W. (2009). A longitudinal study of children's performance on simple multiplication and division problems. *Developmental Psychology*, *45*, 1480-1496.

De Corte, E., Verschaffel, L., & De Win, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology*, *77*, 460-470.

De la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of*

Educational and Behavioral Statistics, *30*, 295-311.

De Smedt, B., Torbeyns, J., Stassens, N., Ghesquière, P., & Verschaffel, L. (2010). Frequency, efficiency and flexibility of indirect addition in two learning environments. *Learning and Instruction*, *20*, 205-215.

Dias, J. G., & Vermunt, J. K. (2006). Bootstrap methods for measuring classification uncertainty in latent class analysis. *COMPSTAT 2006 - Proceedings in Computational Statistics, part I*, 31-41.

Dutch Inspectorate of Education. (2008). *Basisvaardigheden rekenen-wiskunde*. [Basic mathematics abilities.] Den Haag, The Netherlands: Inspectie van het Onderwijs.

Dutch Ministry of Education, Culture, and Sciences. (1998). *Kerndoelen basisonderwijs 1998* [Educational standards for Dutch primary education]. Den Haag, The Netherlands: Ministerie van OCW.

Dutch Ministry of Education, Culture, and Sciences. (2006). *Kerndoelen basisonderwijs* [Educational standards for Dutch primary education]. Den Haag, The Netherlands: Ministerie van OCW.

Ellis, S. (1997). Strategy choice in sociocultural context. *Developmental Review*, *17*, 490-524.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (rev. ed.). Cambridge, MA: MIT Press.

Expertgroep Doorlopende leerlijnen Taal en Rekenen. (2008). *Over de drempels met rekenen. Consolideren, onderhouden, gebruiken en verdiepen*. [Crossing the thresholds with mathematics. Strengthen, maintain, use, and deepen.] Enschede, The Netherlands: Expertgroep Doorlopende leerlijnen Taal en Rekenen.

Fennema, E., & Franke, M. L. (1992). Teachers' knowledge and its impact. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (p. 147-164). New York: MacMillan.

Fischer, G. H. (1987). Linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374.

Foxman, D., & Beishuizen, M. (2003). Mental calculation methods used by 11-year-olds in different attainment bands: A reanalysis of data from the 1987 APU survey in the UK. *Educational Studies in Mathematics*, *51*, 41-69.

Freudenthal, H. (1973). *Mathematics as an educational task*. Dordrecht, The Netherlands:

Reidel.

Freudenthal, H. (1991). *Revisiting mathematics education. China lectures.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., et al. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, *98*, 29-43.

Fuchs, L. S., Fuchs, D., Stuebing, K., Fletcher, J. M., Hamlett, C. L., & Lambert, W. (2008). Problem solving and computational skill: Are they shared or distinct aspects of mathematical cognition? *Journal of Educational Psychology*, *100*, 30-47.

Fuson, K. C., Wearne, D., Hiebert, J. L., Murray, H. G., Human, P. G., Olivier, A. I., et al. (1997). Children's conceptual structures for multidigit numbers and methods of multidigit addition and subtraction. *Journal for Research in Mathematics Education*, *28*, 130-162.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology*, *75*, 165-190.

Geary, D. C. (2003). Arithmetical development: Commentary on chapters 9 through 15 and future directions. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (p. 453-464). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Gilmore, C. K., & Bryant, P. (2006). Individual differences in children's understanding of inversion and arithmetical skill. *British Journal of Educational Psychology*, *76*, 309-331.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215-231.

Gravemeijer, K. (1997a). Commentary. Solving word problems: A case of modelling? *Learning and Instruction*, *7*, 389-397.

Gravemeijer, K. (1997b). Instructional design for reform in mathematics education. In M. Beishuizen, K. Gravemeijer, & E. C. D. M. Van Lieshout (Eds.), *The role of contexts and models in the development of mathematical strategies and procedures* (p. 13-34). Utrecht, The Netherlands: Freudenthal Institute.

Gravemeijer, K. (2007). Reken-wiskundeonderwijs anno 2007 - tussen oude waarden en nieuwe uitdagingen [Mathematics education in the year 2007 - between old values and new challenges]. *Reken-wiskundeonderwijs: onderzoek, ontwikkeling, praktijk*, *26*(4), 3-10.

Gravemeijer, K., & Doorman, M. (1999). Context problems in realistic mathematics education: A calculus course as an example. *Educational Studies in Mathematics*, *39*, 111-128.

Gravemeijer, K., Van den Heuvel-Panhuizen, M., Van Donselaar, G., Ruesink, N., Streefland, L., Vermeulen, W., et al. (1993). *Methoden in het reken-wiskundeonderwijs, een rijke context voor vergelijkend onderzoek.* [Methods in mathematics education, a rich context for comparative research.] Utrecht, The Netherlands: Freudenthal Instituut.

Greer, B. (1997). Modelling reality in mathematics classrooms: The case of word problems. *Learning and Instruction*, *7*, 293-307.

Harskamp, E. G. (1988). *Rekenmethoden op de proef gesteld.* [Mathematics textbooks put to the test]. Doctoral dissertation, Groningen University, Groningen, The Netherlands.

Harskamp, E. G., & Suhre, C. J. M. (1995). *Hoofdrekenen in het speciaal onderwijs.* [Mental computation in special education]. Groningen, The Netherlands: GION.

Harskamp, E. G., Suhre, C. J. M., & Willemsen, T. F. W. P. (1993). *Remediële rekenprogramma's voor het basisonderwijs beproefd.* [Remedial mathematics programs for primary education put to the test]. Groningen, The Netherlands: GION.

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie / Journal of Psychology*, *216*, 89-101.

Hartzel, J., Agresti, A., & Caffo, B. (2001). Multinomial logit random effects models. *Statistical Modelling*, *1*, 81-102.

Hecht, S. A. (2006). Group differences in adult simple arithmetic: Good retrievers, not-so-good retrievers, and perfectionists. *Memory & Cognition*, *31*, 207-216.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis.* Hoboken, NJ: John Wiley & Sons, Inc.

Heinze, A., Marschick, F., & Lipowsky, F. (2009). Addition and subtraction of three-digit numbers: adaptive strategy use and the influence of instruction in german third grade. *ZDM Mathematics Education*, *41*, 591-604.

Heinze, A., Star, J. R., & Verschaffel, L. (2009). Flexible and adaptive use of strategies

and representations in mathematics education. *ZDM Mathematics Education, 41*, 535-540.

Hickendorff, M. (2010a, October). *Beyond measurement: Applications of explanatory IRT models to primary school mathematics tests.* Paper presented on the 2010 RCEC workshop on IRT and Educational Measurement, University of Twente, The Netherlands.

Hickendorff, M. (2010b). The language factor in elementary mathematics assessments: computational skills and applied problem solving in a multidimensional IRT framework. *Manuscript submitted for publication.*

Hickendorff, M. (2010c, September). *Subtraction by addition and compensation: Results from a study into shortcut strategy use by Dutch sixth graders.* Paper presented on the Advanced Study Colloquium on Mathematical Inversion, Leuven, Belgium.

Hickendorff, M., Heiser, W. J., Van Putten, C. M., & Verhelst, N. D. (2008). Clustering nominal data with equivalent categories. *Behaviormetrika, 35*, 35-54.

Hickendorff, M., Heiser, W. J., Van Putten, C. M., & Verhelst, N. D. (2009a). How to measure and explain achievement change in large-scale assessments: A rejoinder. *Psychometrika, 74*, 367-374.

Hickendorff, M., Heiser, W. J., Van Putten, C. M., & Verhelst, N. D. (2009b). Solution strategies and achievement in Dutch complex arithmetic: Latent variable modeling of change. *Psychometrika, 74*, 331-350.

Hickendorff, M., & Janssen, J. (2009). De invloed van contexten in rekenopgaven op de prestaties van basisschoolleerlingen [The effect of contexts in mathematics items on primary-school pupils' performance]. *Reken-wiskundeonderwijs: onderzoek, ontwikkeling, praktijk*, *28*(4), 3-11.

Hickendorff, M., & Van Putten, C. M. (2010). Complex multiplication and division in Dutch educational assessments: What can solution strategies tell us? *Manuscript submitted for publication.*

Hickendorff, M., Van Putten, C. M., Verhelst, N. D., & Heiser, W. J. (2010). Individual differences in strategy use on division problems: Mental versus written computation. *Journal of Educational Psychology, 102*, 439-452.

Hiebert, J., & Grouws, D. (2007). The effects of classroom mathematics teaching on students' learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (p. 371-404). Charlotte, CT: Information Age Publishing.

Hill, H. C., Sleep, L., Lewis, J. M., & Ball, D. L. (2007). Assessing teachers' mathematical

knowledge: What knowledge matters and what evidence counts? In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (p. 111-155). Charlotte, CT: Information Age Publishing.

Imbo, I., & LeFevre, J.-A. (2009). Cultural differences in complex addition: Efficient Chinese versus adaptive Belgians and Canadians. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 1465-1476.

Imbo, I., & Vandierendonck, A. (2007). Do multiplication and division strategies rely on executive and phonological working memory resources? *Memory & Cognition*, *35*, 1759-1771.

Janssen, J., & Hickendorff, M. (2009). Categorieënanalyse bij de LOVS toetsen rekenen-wiskunde. [Profile analysis in the LOVS mathematics tests]. In M. Van Zanten (Ed.), *Leren van evalueren: De lerende in beeld bij reken-wiskundeonderwijs.* (p. 49-60). Utrecht, The Netherlands: FIsme.

Janssen, J., Van der Schoot, F., & Hemker, B. (2005). *Balans van het reken-wiskundeonderwijs aan het einde van de basisschool 4* [Fourth assessment of mathematics education at the end of primary school]. Arnhem, The Netherlands: CITO.

Janssen, J., Van der Schoot, F., Hemker, B., & Verhelst, N. D. (1999). *Balans van het reken-wiskundeonderwijs aan het einde van de basisschool 3* [Third assessment of mathematics education at the end of primary school]. Arnhem, The Netherlands: CITO.

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. D. Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (p. 189-212). New York: Springer.

Kagan, J. (1966). Reflection-impulsivity: The generality and dynamics of conceptual tempo. *Journal of Abnormal Psychology*, *71*, 17-24.

Keijzer, R. (2003). *Teaching formal mathematics in primary education.* Doctoral dissertation, Freudenthal Instituut, Utrecht, The Netherlands.

Keijzer, R., & Terwel, J. (2003). Learning for mathematical insight: a longitudinal comparative study on modelling. *Learning and Instruction*, *13*, 285-304.

Kerkman, D. D., & Siegler, R. S. (1997). Measuring individual differences in children's addition strategy choices. *Learning and Individual Differences*, *9*, 1-18.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). *Adding it up. Helping children learn mathematics.* Washington, D.C.: National Academy Press.

Klein, A. S. (1998). *Flexibilization of mental arithmetic strategies on a different knowledge*

base: the empty number line in a realistic versus gradual program design. Doctoral dissertation, Leiden University, Leiden, The Netherlands.

Klein, A. S., Beishuizen, M., & Treffers, A. (1998). The empty number line in Dutch second grades: Realistic versus gradual program design. *Journal for Research in Mathematics Education, 29*, 443-464.

Klein Entink, R. H., Fox, J.-P., & Van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika, 74*, 21-48.

KNAW. (2009). *Rekenonderwijs op de basisschool. Analyse en sleutels tot verbetering* [Mathematics education in primary school. Analysis and recommendations for improvement]. Amsterdam, The Netherlands: KNAW.

Kraemer, J.-M., Janssen, J., Van der Schoot, F., & Hemker, B. (2005). *Balans van het rekenwiskundeonderwijs halvewege de basisschool 4* [Fourth assessment of mathematics education halfway primary school]. Arnhem, The Netherlands: CITO.

Kroesbergen, E. H., & Van Luit, J. E. H. (2002). Teaching multiplication to low math performers: Guided versus structured instruction. *Instructional Science, 30*, 361-378.

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

Kroesbergen, E. H., & Van Luit, J. E. H. (2005). Effects of different forms of instruction on acquisition and use of multiplication strategies by children with math learning difficulties. In P. Ghesquière & A. Ruijssenaars (Eds.), *Learning disabilities: A challenge to teaching and instruction* (p. 161-181). Leuven, Belgium: Leuven University Press.

Kroesbergen, E. H., Van Luit, J. E. H., & Maas, C. J. M. (2004). Effectiveness of explicit and constructivist mathematics instruction for low-achieving students in the Netherlands. *Elementary School Journal, 104*, 233-251.

Kuss, O., & McLerran, D. (2007). A note on the estimation of the multinomial logistic model with correlated responses. *Computer Methods and Programs in Biomedicine, 87*, 262-269.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* New York: Houghton-Mifflin.

Lee, J. (2007). Making sense of the traditional long division algorithm. *Journal of Mathematical Behavior, 26*, 48-59.

Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and application.* New York: Cambridge University Press.

Lemaire, P. (2010). Executive functions and strategic aspects of arithmetic performance: The case of adults' and children's arithmetic. *Psychologica Belgica*, *50*, 335-352.

Lemaire, P., & Callies, S. (2009). Children's strategies in complex arithmetic. *Journal of Experimental Child Psychology*, *103*, 49-65.

Lemaire, P., & Lecacheur, M. (2002). Children's strategies in computational estimation. *Journal of Experimental Child Psychology*, *82*, 281-304.

Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: Contributions to children's learning of multiplication. *Journal of Experimental Psychology: General*, *124*, 83-97.

Lesaffre, E., & Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: an example. *Journal of the Royal Statistical Society Series C - Applied Statistics*, *50*, 325-335.

Linzer, D., & Lewis, J. (2010). `poLCA`*: Polytomous variable latent class analysis. R package version 1.2* (Computer program and manual). Available from `http://userwww.service.emory.edu/~dlinzer/poLCA`.

Linzer, D., & Lewis, J. (2011). `poLCA`: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*(10), 1-29.

Lipsey, M. W., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Luwel, K., Onghena, P., Torbeyns, J., Schillemans, V., & Verschaffel, L. (2009). Strengths and weaknesses of the choice/no-choice method in research on strategy choice and strategy change. *European Psychologist*, *14*, 351-362.

Luwel, K., Verschaffel, L., Onghena, P., & De Corte, E. (2003). Analysing the adaptiveness of strategy choices using the choice/no-choice method: The case of numerosity judgement. *European Journal of Cognitive Psychology*, *15*, 511-537.

Mabbott, D. J., & Bisanz, J. (2003). Developmental change and individual differences in children's multiplication. *Child Development*, *74*, 1091-1107.

Mauro, D. G., LeFevre, J.-A., & Morris, J. (2003). Effects of problem format on division and multiplication performance: division facts are mediated via multiplication-based representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 163-170.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: Wiley.

Meelissen, M. R. M., & Drent, M. (2008). *TIMSS-2007 Nederland. Trends in leerprestaties in exacte vakken van het basisonderwijs*. [TIMSS 2007 the Netherlands. Trends in

achievement in mathematics and science in primary education]. Enschede, The Netherlands: Twente University.

Menne, J. (2001). *Met sprongen vooruit. een productief oefenprogramma voor zwakke rekenaars in het getallengebied tot 100 – een onderwijsexperiment.* [A productive training program for low mathematics achievers in the number domain up to 100 – a design experiment]. Doctoral dissertation, Freudenthal Institute, Utrecht, The Netherlands.

Milo, B. F., & Ruijssenaars, A. J. J. M. (2005). Strategy use and math instruction for students with special needs. In P. Ghesquière & A. Ruijssenaars (Eds.), *Learning disabilities: A challenge to teaching and instruction* (p. 183-200). Leuven, Belgium: Leuven University Press.

Milo, B. F., Ruijssenaars, A. J. J. M., & Seegers, G. (2005). Math instruction for students with special educational needs: Effects of guiding versus directing instruction. *Educational & Child Psychology*, *22*, 70-80.

Milo, B. F., Seegers, G., Ruijssenaars, A. J. J. M., & Vermeer, H. (2004). Affective consequences of mathematics instruction for students with special needs. *European Journal of Special Needs Education*, *19*, 49-68.

Molenberghs, G., & Verbeke, G. (2004). An introduction to (generalized (non)linear mixed models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (p. 111-153). New York, NY: Springer.

Mulligan, J. T., & Mitchelmore, M. C. (1997). Young children's intuitive models of multiplication and division. *Journal for Research in Mathematics Education*, *28*, 309-330.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report. Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades.* Boston: Boston College, TIMSS & PIRLS International Study Center.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report. IEA's Progress in International Reading Literacy Study in primary school in 40 countries.* Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: NCTM.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school*

mathematics. Reston, VA: NCTM.

National Mathematics Advisory Panel. (2008). *Foundations for success. The Final Report of the National Mathematics Advisory Panel*. Washington, DC: U.S. Department of Education.

Neuman, D. (1999). Early learning and awareness of division: A phenomenographic approach. *Educational Studies in Mathematics*, *40*, 101-128.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237-257.

OECD. (2004). *Learning for tomorrow's world. First results from PISA 2003*. Paris: OECD.

OECD. (2009). *PISA 2006 technical report*. Paris: OECD.

OECD. (2010). *PISA 2009 results: What students know and can do. Student performance in reading, mathematics and science (volume I)*. Paris: OECD.

Phillips, L. H., & Rabbitt, P. M. A. (1995). Impulsivity and speed-accuracy strategies in intelligence-test performance. *Intelligence*, *21*, 13-29.

Poland, M. (2007). *The treasures of schematising. The effects of schematising in early childhood on the learning processes and outcomes in later mathematical understanding*. Doctoral dissertation, Free University, Amsterdam, The Netherlands.

Poland, M., & Van Oers, B. (2007). Effects of schematising on mathematical development. *European Early Childhood Education Research Journal*, *15*, 269-293.

Porter, A. C. (2006). Curriculum assessment. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (p. 141-160). Mahwah, NJ: Lawrence Erlbaum Associates.

Prenger, J. (2005). *Taal telt! Een onderzoek naar de rol van taalvaardigheid en tekstbegrip in het realistische rekenonderwijs*. [Language counts! A study into the role of linguistic skill and text comprehension in realistic mathematics education]. PhD thesis, University of Groningen, The Netherlands.

R Development Core Team. (2009). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org`

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reys, R. E. (1984). Mental computation and estimation: Past, present, and future. *The Elementary School Journal*, *84*, 546-557.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185-205.

Robinson, K. M. (2001). The validity of verbal reports in children's subtraction. *Journal of Educational Psychology*, *93*, 211-222.

Robinson, K. M., Arbuthnott, K. D., Rose, D., McCarron, M. C., Globa, C. A., & Phonexay, S. D. (2006). Stability and change in children's division strategies. *Journal of Experimental Child Psychology*, *93*, 224-238.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal reports. *Memory & Cognition*, *17*, 759-769.

Ruthven, K. (1998). The use of mental, written and calculator strategies of numerical computation by upper primary pupils within a 'calculator-aware' number curriculum. *British Educational Research Journal*, *24*, 21 - 42.

SAS Institute. (2002). *SAS online doc (version 9)*. Cary, NC: SAS Institute Inc.

Schopman, E. A. M., & Van Luit, J. E. H. (1996). Learning transfer of preparatory arithmetic strategies among young children with a developmental lag. *Journal of Cognitive Education*, *5*, 117-131.

Selter, C. (2001). Addition and subtraction of three-digit numbers: German elementary children's succes, methods and strategies. *Educational Studies in Mathematics*, *47*, 145-173.

Sherin, B., & Fuson, K. (2005). Multiplication strategies and the appropriation of computational resources. *Journal for Research in Mathematics Education*, *36*, 347-395.

Sheu, C.-F., Chen, C.-T., Su, Y.-H., & Wang, W.-C. (2005). Using SAS PROC NLMIXED to fit item response theory models. *Behavior Research Methods*, *37*, 202-218.

Shrager, J., & Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science*, *9*, 405-410.

Siegler, R. S. (1988a). Individual differences in strategy choices: good students, not-so-good students, and perfectionists. *Child Development*, *59*, 833-851.

Siegler, R. S. (1988b). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General*, *117*, 258-275.

Siegler, R. S., & Lemaire, P. (1997). Older and younger adults' strategy choices in multiplication: Testing predictions of ASCM using the choice/no-choice method. *Journal of Experimental Psychology: General*, *126*, 71-92.

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In G. Halford & T. Simon (Eds.), *Developing cognitive competence: New approaches to process modeling* (p. 31-76). Hillsdale, NJ: Erlbaum.

Silverman, I. W. (2003). Gender differences in delay of gratification: A meta-analysis. *Sex*

Roles, *49*, 451-463.

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, *45*, 553-573.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluation. *Educational Researcher*, *37*, 5-14.

Slavin, R. E., & Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence syntheses. *Review of Educational Research*, *78*, 427-515.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling.* London: Sage Publications.

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, *36*, 404-411.

Star, J. R., & Newton, K. J. (2009). The nature and development of experts' strategy flexibility for solving equations. *ZDM Mathematics Education*, *41*, 557-561.

Stein, M. K., Remillard, J., & Smith, M. S. (2007). How curriculum influences student learning. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (p. 319-370). Charlotte, CT: Information Age Publishing.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, *23*, 63-86.

Swanson, H. L., & Carson, C. (1996). A selective synthesis of intervention research for students with learning disabilities. *School Psychology Review*, *25*, 370-392.

Swanson, H. L., & Hoskyn, M. (1998). Experimental intervention research on students with learning disabilities: A meta-analysis of treatment outcomes. *Review of Educational Research*, *68*, 227-331.

Terwel, J., Van Oers, B., Van Dijk, I. M. A. W., & Van Eeden, P. (2009). Are representations to be provided or generated in primary mathematics education? Effects on transfer. *Educational Research and Evaluation*, *15*, 25-44.

Timmermans, R. E., & Van Lieshout, E. C. D. M. (2003). Influence of instruction in mathematics for low performing students on strategy use. *European Journal of Special Needs Education*, *8*, 5-16.

Timmermans, R. E., Van Lieshout, E. C. D. M., & Verhoeven, L. (2007). Gender related effects of contemporary math instruction for low performers on problem solving behavior. *Learning and Instruction*, *17*, 42-54.

Torbeyns, J., De Smedt, B., Ghesquière, P., & Verschaffel, L. (2009a). Acquisition and use of shortcut strategies by traditionally schooled children. *Educational Studies in Mathematics*, *71*, 1-17.

Torbeyns, J., De Smedt, B., Ghesquière, P., & Verschaffel, L. (2009b). Jump or compensate? Strategy flexibility in the number domain up to 100. *ZDM Mathematics Education*, *41*, 581-590.

Torbeyns, J., Ghesquière, P., & Verschaffel, L. (2009). Efficiency and flexibility of indirect addition in the domain of multi-digit subtraction. *Learning and Instruction*, *19*, 1-12.

Torbeyns, J., Verschaffel, L., & Ghesquière, P. (2002). Strategic competence: Applying Siegler's theoretical and methodological framework to the domain of simple addition. *European Journal of Psychology of Education*, *27*, 275-291.

Torbeyns, J., Verschaffel, L., & Ghesquière, P. (2004a). Strategic aspects of simple addition and subtraction: the influence of mathematical ability. *Learning and Instruction*, *14*, 177-195.

Torbeyns, J., Verschaffel, L., & Ghesquière, P. (2004b). Strategy development in children with mathematical disabilities: Insights from the choice/no-choice method and the chronological-age/ability-level-match design. *Journal of Learning Disabilities*, *37*, 119-131.

Torbeyns, J., Verschaffel, L., & Ghesquière, P. (2005). Simple addition strategies in a first-grade class with multiple strategy instruction. *Cognition and Instruction*, *23*, 1-21.

Torbeyns, J., Verschaffel, L., & Ghesquière, P. (2006). The development of children's adaptive expertise in the number domain 20 to 100. *Cognition and Instruction*, *24*, 439-465.

Treffers, A. (1987). Integrated column arithmetic according to progressive schematisation. *Educational Studies in Mathematics*, *18*, 125-145.

Treffers, A. (1993). Wiscobas and Freudenthal: Realistic mathematics education. *Educational Studies in Mathematics*, *25*, 89-108.

Van den Boer, C. (2003). *Als je begrijpt wat ik bedoel. Een zoektocht naar verklaringen van achterblijvende prestaties van allochtone leerlingen in het wiskundeonderwijs* [If you get what I mean. A search for explanations of lagging achievement of non-native students in mathematics education]. Utrecht, The Netherlands: CD-β press.

Van den Heuvel-Panhuizen, M. (Ed.). (2008). *Children learn mathematics*. Rotterdam, the Netherlands: Sense Publishers.

Van den Heuvel-Panhuizen, M., Buys, K., & Treffers, A. (Eds.). (2001). *Kinderen leren rekenen. Tussendoelen annex leerlijnen: Hele getallen bovenbouw basisschool.* [Children learn mathematics]. Groningen, the Netherlands: Wolters-Noordhoff.

Van den Heuvel-Panhuizen, M., Robitzsch, A., Treffers, A., & Köller, O. (2009). Large-scale assessments of change in student achievement: Dutch primary school students' results on written division in 1997 and 2004 as an example. *Psychometrika, 74*, 351-365.

Van der Craats, J. (2007). Waarom Daan en Sanne niet kunnen rekenen [Why Daan and Sanne can't add]. *Nieuw Archief voor de Wiskunde, 8*(2), 132-136.

Van de Rijt, B. A. M., & Van Luit, J. E. H. (1998). Effectiveness of the Additional Early Mathematics program for teaching children early mathematics. *Instructional Science, 26*, 337-358.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York, NY: Springer-Verlag.

Van der Schoot, F. (2008). *Onderwijs op peil? Een samenvattend overzicht van 20 jaar PPON* [A summary overview of 20 years of national assessments of the level of education]. Arnhem, The Netherlands: CITO.

Van Dijk, I. M. A. W., Van Oers, B., Terwel, J., & Van Eeden, P. (2003). Strategic learning in primary mathematics rducation: Effects of an experimental program in modelling. *Educational Research and Evaluation, 9*, 161-187.

Van Luit, J. E. H. (1994). The effectiveness of structural and realistic arithmetic curricula in children with special educational needs. *European Journal of Special Needs Education, 9*, 16-26.

Van Luit, J. E. H., & Naglieri, J. A. (1999). Effectiveness of the MASTER program for teaching special children multiplication and division. *Journal of learning disabilities, 32*, 98-107.

Van Luit, J. E. H., & Schopman, E. A. M. (2000). Improving early numeracy of young children with special educational needs. *Remedial and special education, 21*, 27-40.

Van Putten, C. M., & Hickendorff, M. (2009). Peilstokken voor Plasterk: Evaluatie van rekenvaardigheid in groep 8. [Assessments for Plasterk: Evaluation of mathematical ability in sixth grade]. *Tijdschrift voor Orthopedagogiek, 48*, 183-194.

Van Putten, C. M., Van den Brom-Snijders, P. A., & Beishuizen, M. (2005). Progressive mathematization of long division strategies in Dutch primary schools. *Journal for Research in Mathematics Education, 36*, 44-73.

Van Schilt-Mol, T. M. M. L. (2007). *Differential item functioning en itembias in de Cito-*

Eindtoets Basisonderwijs [Differential item functioning and item bias in CITO's End of Primary School Test]. PhD thesis, Tilburg University, The Netherlands.

Varol, F., & Farran, D. (2007). Elementary school students' mental computation proficiencies. *Early Childhood Educational Journal*, *35*, 89-94.

Verhelst, N. D. (2007). *Profielanalyse met Item Respons Theorie* [Profile analysis with Item Response Theory]. Arnhem, The Netherlands: CITO.

Verhelst, N. D. (in press). Profile analysis: A closer look at the PISA 2000 Reading data. *Scandinavian Journal of Educational Research.*

Verhelst, N. D., & Glas, C. A. W. (1995). The one-parameter logistic model. In G. H. Fischer & I. Molenaar (Eds.), *Rasch models. Foundations, recent developments and applications* (p. 215-237). New York: Springer.

Verhelst, N. D., & Verstralen, H. H. F. M. (2002). *Structural analysis of a univariate latent variable (SAUL)* (computer program and manual). Arnhem, The Netherlands: CITO.

Vermeer, H. J., Boekaerts, M., & Seegers, G. (2000). Motivational and gender differences: Sixth-grade students' mathematical problem solving behavior. *Journal of Educational Psychology*, *92*, 308-315.

Vermunt, J. K. (1997). *LEM 1.0: A general program for the analysis of categorical data.* Tilburg, The Netherlands: Tilburg University.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (p. 89-106). Cambridge, England: Cambridge University Press.

Verschaffel, L. (2009). 'Over het muurtje kijken': achtergrond, inhoud en receptie van het Final Report van het 'National Mathematics Advisory Panel' in de U.S. [Background, content, and reception of the Final Report of the National Mathematics Advisory Panel in the U.S.] *Reken-wiskundeonderwijs: Onderzoek, ontwikkeling en praktijk*, *28*(1), 3-20.

Verschaffel, L., Greer, B., & De Corte, E. (2000). *Making sense of word problems.* Lisse, The Netherlands: Swets and Zeitlinger.

Verschaffel, L., Greer, B., & De Corte, E. (2007). Whole number concepts and operations. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning.* Charlotte, CT: Information Age Publishing.

Verschaffel, L., Luwel, K., Torbeyns, J., & Van Dooren, W. (2009). Conceptualizing, investigating, and enhancing adaptive expertise in elementary mathematics education. *European Journal of Psychology of Education*, *24*, 335-359.

Vicente, S., Orrantia, J., & Verschaffel, L. (2007). Influence of situational and conceptual rewording on word problem solving. *British Journal of Educational Psychology*, *77*, 829-848.

Von Davier, M., & Sinharay, S. (2009). *Stochastic approximation methods for latent regression item response models* (ETS-RR-09-09). Princeton, NJ: Educational Testing Service.

Willemsen, T. F. W. P. (1994). *Remediële rekenprogramma's voor de basisschool* [Remedial mathematics programs for primary education]. Doctoral dissertation, GION, Groningen, The Netherlands.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (p. 43-74). New York, NY: Springer.

Wu, M. L., & Adams, R. J. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal*, *18*, 93-113.

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

Author Index

De Boeck, P., xv, 57, 58, 91, 133, 134, 177, 178, 181, 211–213, 217, 241–243, 249, 258, 264, 274
De Brauwer, J., 82, 249
De Corte, E., xvii, 10, 115, 172, 173, 219, 249, 256, 263
De la Torre, J., 191, 192, 241, 242, 249
De Lisi, R., 251
De Smedt, B., 79, 83, 105, 145–147, 150, 151, 161, 165, 200, 234, 236, 239, 250, 261, 279
De Win, L., 172, 249
Dias, J. G., 92, 250
Doorman, M., 197, 252, 280
Dowker, A., 79, 146, 200, 234, 247, 279
Drent, M., 6, 77, 256
Dutch Inspectorate of Education, 10, 250

Ellis, S., 146, 147, 166, 167, 234, 250, 279
Embretson, S. E., xv, 88, 91, 189, 211, 241, 250, 274
Ericsson, K. A., 141, 250
Expertgroep Doorlopende leerlijnen Taal en Rekenen, 233, 250

Farran, D., 118, 263
Fennema, E., 10, 250
Fias, W., 82, 249
Findell, B., 47, 254
Fischer, G. H., 213, 250
Fletcher, J. M., 251
Flojo, J., 251
Fox, J.-P., 154, 255
Foxman, D., 83, 105, 114, 119, 138, 151, 165, 239, 250
Foy, P., 6, 223, 257
Franke, M. L., 10, 250
Freudenthal, H., xiv, 47, 79, 115, 146, 148, 197, 199, 250, 251, 272
Fuchs, D., 251
Fuchs, L. S., 171, 172, 191, 198, 199, 201, 211, 219, 220, 251
Fuchs-Lacelle, S., 82, 248
Fuson, K., 80, 81, 200, 259
Fuson, K. C., 112, 251

Gallagher, A. M., 83, 105, 113, 151, 164, 202, 222, 238, 251
Geary, D. C., 109, 165, 239, 251
Gersten, R., 11, 33, 251, 275
Ghesquière, P., 79, 83, 113, 208, 236, 250, 261
Gibbons, R. D., 154, 252
Gierl, M. J., 189, 255
Gilmore, C. K., 114, 251
Glas, C. A. W., 72, 189, 212, 248, 263
Globa, C. A., 259
Goodman, L. A., xv, 54, 90, 128, 240, 251, 273
Gravemeijer, K., 29, 47, 115, 172, 197, 232, 233, 251, 252, 280
Greer, B., xvii, 10, 170, 192, 197, 252, 263
Grouws, D., 11, 33, 34, 105, 253

Haberman, S. J., 191, 260
Hambleton, R. K., xv, 88, 91, 211, 241, 262, 274
Hamlett, C. L., 251
Harskamp, E. G., 24, 26, 30, 252
Hartig, J., 179, 180, 188, 252
Hartzel, J., 215, 252
Hecht, S. A., 113, 164, 238, 252
Hedeker, D., 154, 252
Hedges, L. V., 34, 258
Heinze, A., 146, 236, 252
Heiser, W. J., 3, 79, 80, 241, 253
Hejri, F., 173, 190, 202, 220, 247
Hemker, B., xiii, 30, 31, 254, 255

McCulloch, C. E., 58, 256
McGillicuddy-De Lisi, A. V., 251
McLerran, D., 215, 255
Meelissen, M. R. M., 6, 77, 256
Menne, J., 27, 257
Miller, D. C., 109, 248
Milo, B. F., 17–19, 257
Mitchelmore, M. C., 48, 80, 116, 145, 200, 257
Molenberghs, G., 181, 243, 257
Morely, M., 251
Morphy, P., 251
Morris, J., 82, 256
Mulligan, J. T., 48, 80, 116, 145, 200, 257
Mullis, I. V. S., 6, 7, 34, 77, 83, 84, 105, 114, 164, 173, 201, 202, 223, 237, 238, 257
Murray, H. G., 251

Naglieri, J. A., 25, 262
National Council of Teachers of Mathematics, 170, 196, 257
National Mathematics Advisory Panel, 8, 11, 32, 33, 258
Neuman, D., 48, 116, 258
Newton, K. J., 146, 260
Nivard, M., 249
Nye, B., 34, 258

Olivier, A. I., 251
Onghena, P., 115, 146, 256
Orrantia, J., 172, 264

Patz, R. J., 191, 192, 241, 242, 249
Peres, D., 213, 254
Phenix, T. L., 82, 248
Phillips, L. H., 147, 166, 234, 258, 279
Phonexay, S. D., 259
Poland, M., 22, 258
Porter, A. C., 107, 233, 258
Powell, S. R., 251
Prenger, J., 173, 190, 202, 258
Puhan, G., 191, 260

Rabbitt, P. M. A., 147, 166, 234, 258, 279
R Development Core Team, 90, 189, 211, 212, 241, 242, 258
Reckase, M. D., 174, 177, 188, 211, 241, 258
Reise, S. P., xv, 88, 91, 189, 211, 241, 250, 274
Remillard, J., 9, 260
Reusser, K., 173, 249
Reys, R. E., 118, 258
Rijmen, F., xv, 57, 58, 91, 133, 178, 181, 241, 258, 274
Robinson, K. M., 112, 115, 116, 140, 147, 200, 259
Robitzsch, A., 109, 262
Rose, D., 259
Ruesink, N., 252
Ruijssenaars, A. J. J. M., 17, 19, 257
Russo, J. E., 141, 259
Ruthven, K., 73, 120, 150, 231, 259

Schepers, J., 213, 254
Schillemans, V., 146, 256
Schopman, E. A. M., 22, 23, 259, 262
Searle, S. R., 58, 256
Seegers, G., 17, 19, 83, 257, 263
Seethaler, P. M., 251
Selter, C., 236, 259
Shafer, W. D., 109, 248
Sherin, B., 80, 81, 200, 259
Sheu, C.-F., 58, 91, 181, 259
Shipley, C., 80, 113, 145, 166, 234, 259, 279
Shrager, J., 80, 106, 145, 166, 234, 259, 279
Siegler, R. S., xvii, 70, 77–80, 106, 112–114, 145–148, 150, 151, 160, 163–166,

Verschaffel, L., xvii, 8, 10, 34, 79, 83, 109, 113, 115, 146, 147, 165–167, 170, 172, 196, 198, 200, 201, 208, 234, 236, 239, 249, 250, 252, 256, 261, 263, 264, 279, 280
Verstralen, H. H. F. M., 57, 72, 179, 263
Vicente, S., 172, 173, 201, 219, 264
Von Davier, M., 179, 180, 264

Wang, W.-C., 58, 174, 247, 259
Wearne, D., 251
Weimer, R., 173, 249
Willemsen, T. F. W. P., 19, 24, 26, 252, 264
Wilson, D., 11, 14, 256
Wilson, M., xv, 57, 58, 91, 133, 134, 174, 177, 178, 181, 212, 213, 217, 241, 242, 247, 249, 264, 274
Wu, M. L., 172, 173, 178, 198, 201, 211, 247, 264

Xin, Y. P., 9, 264

Ye, F., 192, 260

Zhu, X., 192, 260
Zwitser, R., 249

Summary in Dutch (Samenvatting)

De rekenvaardigheid van Nederlandse leerlingen is een onderwerp van felle debatten. Eén punt van discussie is het *rekenonderwijs* dat de afgelopen decennia een hervorming van internationale reikwijdte heeft ondergaan. Deze vernieuwing kan grofweg omschreven kan worden als het verlaten van de traditionele aanpak waarin leerkrachten rekenkennis en -vaardigheden direct instrueren aan hun leerlingen, die op hun beurt moeten oefenen en 'stampen'. Hier is een aanpak voor in de plaats gekomen waarin de informele voorkennis die leerlingen hebben aan de basis van een leerlijn ligt. Het doel is niet alleen procedurele kennis en vaardigheden bij de leerling te bewerkstelligen, maar vooral ook inzicht, flexibiliteit en creativiteit. Een ander discussiepunt is het *prestatieniveau* van leerlingen in zowel het basisonderwijs als in het voortgezet onderwijs. Op regelmatige basis vinden grootschalige nationale en internationale peilingsonderzoeken van de rekenvaardigheid van leerlingen plaats, waarin wordt gerapporteerd over ontwikkelingen over tijd, vergelijkingen tussen landen en afwijkingen van de onderwijsdoelen die binnen een land gelden. Deze resultaten vormen vaak de aanleiding voor de discussie over het prestatieniveau.

Dit proefschrift – het resultaat van een samenwerking tussen het Instituut Psychologie van de Universiteit Leiden en Cito Instituut voor Toetsontwikkeling – richt zich op de rekenvaardigheid van Nederlandse basisschoolleerlingen. Het startpunt was de resultaten van de meest recente *Periodieke Peiling van het OnderwijsNiveau* (PPON) rekenen-wiskunde aan het einde van het basisonderwijs (groep 8 – internationaal *sixth grade*; leerlingen van 12 jaar oud). Deze peiling is in 2004 uitgevoerd door Cito (J. Janssen et al., 2005; zie ook Van der Schoot, 2008). PPON-2004 was de vierde peilingscyclus, met

eerdere peilingen in 1987, 1992 en 1997; de dataverzameling voor de vijfde cyclus vindt in 2011 plaats. De ontwikkelingen van het rekenvaardigheidsniveau over de periode 1987 tot 2004 zijn uiteenlopend: op sommige domeinen gingen de prestaties vooruit, terwijl ze op andere domeinen juist achteruit gingen. Daarnaast bleek dat het niveau op vrijwel alle onderdelen achterbleef bij de kerndoelen – op sommige onderdelen meer dan op andere. Deze resultaten haalden de kranten en andere media, waarin vervolgens verschillende mensen hun mening hierover gaven. Een terugkerend element hierin is de didactische theorie van het *realistisch rekenen* (RME; zie bijvoorbeeld Freudenthal, 1973, 1991; Treffers, 1993) die de meest invloedrijke theorie in het hedendaagse Nederlandse rekenonderwijs is geworden sinds de jaren '80 en '90 van de vorige eeuw. Het realistisch rekenen roept sterke gevoelens op en heeft uitgesproken vóór- en tegenstanders. In het maatschappelijk debat hebben overtuigingen en op anekdotes gebaseerde persoonlijke sentimenten echter de overhand op robuuste empirische onderzoeksresultaten over wat leerlingen kennen en kunnen op het gebied van rekenen, en over wat de effecten van verschillende rekeninstructiemethoden zijn. Het doel van dit proefschrift is om deze op onderzoek gebaseerde inzichten te verschaffen.

Om een overzicht te geven van wat uit eerder onderzoek bekend is – en van wat *niet* bekend is – over de resultaten van verschillende rekenprogramma's of -curricula wordt als eerste een onderzoekssynthese gepresenteerd van empirische studies die deze vraag probeerden te beantwoorden voor Nederlandse basisschoolleerlingen. Om inzicht te krijgen in de kennis en vaardigheden van leerlingen op het gebied van rekenen zijn de peilingsonderzoeken een rijke bron van gegevens. Deze peilingen zijn echter *surveys* en zijn daardoor beperkt tot beschrijvende analyses. Verklaringen voor gevonden verschillen of veranderingen over tijd vereisen nader onderzoek, en dat is precies wat in het huidige promotieproject is gedaan. Dit proefschrift bevat zes empirische studies, waarin gegevens van in totaal bijna 5000 leerlingen uit groep 3, 4, 5 en 8 zijn geanalyseerd.

In deze empirische studies naar determinanten van rekenvaardigheid op het gebied van optellen, aftrekken, vermenigvuldigen en delen wordt de afstand tussen de wetenschappelijke disciplines van de inhoudelijke psychologie (cognitieve psychologie en onderwijspsychologie) enerzijds en de psychometrie anderzijds overbrugd. Inhoudelijk spelen *oplossingsstrategieën* een hoofdrol in vijf van de zes studies. Strategieën worden belangrijk geacht vanuit de onderwijspsychologie, omdat ze een speerpunt van het realistisch rekenen zijn, maar ook vanuit de cognitieve psychologie waarin mechanismen van strategiekeuze en concepten als strategische competentie (zie bijvoorbeeld Lemaire

en Siegler, 1995) belangrijke onderzoeksgebieden zijn. In de huidige studies werden oplossingsstrategieën als uitkomstvariabelen gebruikt in analyses van determinanten van strategiekeuze, en als verklarende variabelen in analyses van determinanten van rekenprestaties. Gerelateerde terugkerende elementen zijn individuele verschillen in strategiekeuze en in prestaties, en groepsverschillen zoals die tussen jongens en meisjes.

Vanuit een psychometrisch perspectief is het relevant dat de onderzoeksgegevens een complexe structuur hebben, waardoor geavanceerde statistische analyses nodig zijn. In de inhoudelijke disciplines van onderwijspsychologie en cognitieve psychologie zijn dit soort technieken niet erg gebruikelijk. Dat is dan ook de reden waarom dit proefschrift gezien kan worden als een poging tot het samenbrengen van psychologie en psychometrie, zoals bepleit door Borsboom (2006). Een opvallend kenmerk in de huidige studies is dat de data bestaan uit herhaalde observaties: elke leerling beantwoordt meerdere rekenopgaven. Dit leidt tot een gecorreleerde of afhankelijke gegevensstructuur. We betogen dat *latente variabele modellen* geschikt zijn om rekening te houden met deze afhankelijkheden. Eén of meerdere latente variabelen geven de individuele verschillen tussen leerlingen weer, en de afhankelijke responsen binnen elke leerling worden weerspiegeld in een positie op deze variabele(n). Latente variabelen kunnen ofwel categorisch zijn, wanneer ze kwalitatieve individuele verschillen tussen leerlingen modelleren, ofwel continu, wanneer ze kwantitatieve individuele verschillen modelleren. Ook kan de invloed van *verklarende variabelen* (zoals peilingsjaar of het geslacht van de leerling) vastgesteld worden door het effect op de latente variabele te analyseren.

In de onderzoeken zijn de responsen op elke *trial* (wanneer een leerling geconfronteerd wordt met een opgave) van categorisch meetniveau. Meer specifiek zijn er twee typen responsen: de strategie de gehanteerd is om de opgave op te lossen (meerdere ongeordende categorieën) en de goed/fout score van het gegeven antwoord (dichotoom). Om individuele verschillen in strategiekeuze te analyseren gebruiken we latente klasse analyse (LCA; bijvoorbeeld Goodman, 1974; Lazarsfeld & Henry, 1968). In latente klasse modellen wordt aangenomen dat er ongeobserveerde subgroepen (klassen of clusters) van leerlingen zijn die gekarakteriseerd worden door een specifiek patroon van responsen, in dit geval een specifiek patroon van strategiekeuzes. Latente klasse modellen met covariaten (Vermunt & Magidson, 2002) zijn gebruikt om de invloed van leerlingkenmerken op klasselidmaatschap vast te stellen.

Om individuele verschillen in prestatie te analyseren gebruiken we modellen uit

de item-responstheorie (IRT; bijvoorbeeld Embretson & Reise, 2000; Van der Linden & Hambleton, 1997). Hierin wordt de kans op een goed antwoord op een specifieke opgave bepaald door één of meerdere continue latente (vaardigheids)dimensies. Deze IRT-meetmodellen zijn uitgebreid met een verklarend deel, waarin predictoren op persoonsniveau, opgaveniveau of persoons-bij-opgaveniveau worden opgenomen (De Boeck & Wilson, 2004; Rijmen et al., 2003). Een innovatieve toepassing van dergelijke verklarende IRT-modellen wordt gepresenteerd, waarin de strategie die op een opgave is gehanteerd als persoons-bij-opgave predictor wordt opgenomen. Daarmee wordt de strategie-accuratesse gemodelleerd (de kans op een goed antwoord met een bepaalde oplossingsstrategie), statistisch gecorrigeerd voor individuele verschillen in de vaardigheid van de leerlingen die de strategie gebruiken en voor verschillen in moeilijkheidsgraad tussen de opgaven waarop deze strategie gebruikt wordt. Dit was nog niet eerder bewerkstelligd in psychologisch onderzoek naar oplossingsstrategieën.

Hierna volgt een samenvatting per hoofdstuk. Hoofdstuk 1 van dit proefschrift bevat een onderzoekssynthese van resultaten van Nederlandse empirische studies naar de relatie tussen rekendidactiek en rekenvaardigheid. Dit hoofdstuk is gebaseerd op literatuuronderzoek dat is uitgevoerd voor de adviescommissie *Rekenonderwijs op de basisschool*[1] ingesteld door de *Koninklijke Nederlandse Akademie van Wetenschappen* (KNAW), wier rapport in 2009 is uitgekomen. Deze systematische kwantitatieve onderzoekssynthese laat geen eenduidige conclusies over het effect van verschillende rekeninstructiemethoden of rekencurricula toe. Enerzijds zijn er weinig methodologisch degelijk opgezette *interventiestudies* waarin de effecten van verschillende instructie-aanpakken vergeleken worden. De wel beschikbare studies zijn bovendien beperkt in verschillende aspecten, zoals steekproefgrootte of inhoudsdomein. Ook zijn didactische kenmerken en instructiekenmerken vaak met elkaar verweven in de programma's die vergeleken zijn, zodat het onmogelijk is de unieke effecten van verschillende kenmerken vast te stellen. Anderzijds zijn de *curriculumstudies* waarin de uitkomsten van leerlingen die verschillende rekencurricula (rekenmethodes) gevolgd hebben worden vergeleken, beperkt in de mate van controle over de implementatie van het curriculum en in de mogelijk tot het corrigeren voor verstorende variabelen.

Hoewel er dus geen algemene hoofdconclusie getrokken kan worden, zijn er wel wat

1 Ik heb als toegevoegd onderzoeker voor deze commissie gewerkt, en heb in opdracht van de commissie het systematische literatuuronderzoek dat ten grondslag ligt aan hoofdstuk 4 van het rapport (KNAW, 2009) uitgevoerd. Dat vormt de basis voor hoofdstuk 1 van dit proefschrift.

specifieke patronen die uit de bestudeerde onderzoeksresultaten naar voren komen. Ten eerste is het opvallend dat de prestatieverschillen *binnen* een bepaald type instructie-aanpak groter zijn dan *tussen* verschillende aanpakken. Blijkbaar spelen didactische principes een kleinere rol dan de praktische implementatie door de leerkracht en de interactie tussen de leerkracht en de leerling, bevindingen die in overeenstemming zijn met die van bijvoorbeeld Slavin en Lake (2008) in hun grootschalige internationale onderzoekssynthese. Een tweede patroon is dat meer onderwijstijd voor rekenen betere rekenprestaties tot gevolg heeft, wat in overeenstemming is met de stelling van Hiebert en Grouws (2007) dat *opportunity to learn* the belangrijkste voorspeller van prestatie-uitkomsten is. Ten derde is gebleken dat, als de onderwijstijd gelijk is, experimentele rekeninstructieprogramma's die in kleine groepjes leerlingen buiten de klas worden uitgevoerd positieve effecten op de prestaties hebben ten opzichte van de reguliere klassepraktijk, vergelijkbaar met het gerapporteerde positieve effect van *small-group tutoring* door Slavin en Lake (2008). Ten slotte zijn veel van de bestudeerde onderzoeken gericht op zwakke rekenaars. Deze leerlingen lijken minder gebaat bij een vrije vorm van instructie en meer behoefte te hebben aan een sturende rol van de leerkracht in hun leerproces, in overeenstemming met conclusies van internationale reviews (Gersten et al., 2009; Kroesbergen & Van Luit, 2003; Swanson & Carson, 1996; Swanson & Hoskyn, 1998). Veel minder onderzoeksaandacht is uitgegaan naar instructie voor middelmatige en sterke rekenaars. Ook mogelijke differentiële instructie-effecten voor jongens en meisjes zijn weinig onderzocht.

In de overige zes empirische hoofdstukken is de aandacht verlegd van instructie-effecten naar andere aspecten van de rekenvaardigheid van basisschoolleerlingen in het hedendaagse reken-wiskundeonderwijs. In hoofdstuk 2 en 3 zijn secundaire analyses uitgevoerd op het ruwe leerlingmateriaal (de ingevulde opgaveboekjes) van ongeveer 1600 leerlingen die deelnamen aan één van de twee meest recente peilingen in groep acht. Deze analyses hadden als doel meer inzicht te krijgen in de vaardigheid op het gebied van het vermenigvuldigen en delen met grote getallen en kommagetallen door oplossingsstrategieën in beschouwing te nemen. Het prestatieniveau van de groep 8-leerlingen op deze bewerkingen bleek tussen PPON 1997 en 2004 sterk gedaald, en bovendien ver achter te blijven bij de standaarden. Voor beide bewerkingen zijn de strategieën die leerlingen gebruikten om de opgaven op te lossen gecodeerd in vier hoofdcategorieën: het traditionele algoritme, ook wel cijferen genoemd (de staartdeling en het onder elkaar vermenigvuldigen), niet-traditionele strategieën die met gehele

getallen werken, het geven van een antwoord zonder een schriftelijke uitwerking (complete berekening of tussenstappen) te noteren en overige strategieën (onduidelijke of verkeerde aanpakken en overgeslagen opgaven). Een subcategorie van de niet-traditionele strategieën is het zogenaamde *kolomsgewijs rekenen*, een onderdeel van de didactiek van het realistisch rekenen (Van den Heuvel-Panhuizen, 2008; Van den Heuvel-Panhuizen et al., 2001). Het kolomsgewijs rekenen kan als overgang tussen informele aanpakken en het traditionele algoritme gezien worden: er wordt gewerkt met gehele getallen in plaats van individuele cijfers (zoals in informele strategieën) maar de procedure is min of meer gestandaardiseerd (zoals in de traditionele algoritmen).

In hoofdstuk 2 worden ook de latente variabele modellen geïntroduceerd. Latente klasse analyses zijn gebruikt om individuele verschillen in keuze van strategieën in kaart te brengen en om verschuivingen over tijd zichtbaar te maken. Verklarende item-responsmodellen zijn gebruikt om strategie-accuratesse en verschuivingen daarin zo zuiver mogelijk te analyseren. Dit hoofdstuk richt zich inhoudelijk op de oplossingsstrategieën op de deelsommen van PPON 1997 en 2004. Het blijkt dat twee veranderingen hebben bijgedragen aan de prestatiedaling tussen deze twee peilingen. Ten eerste heeft een verschuiving in strategiegebruik plaatsgevonden: in 2004 waren minder leerlingen geneigd de staartdeling te gebruiken dan in 1997, terwijl meer leerlingen geneigd waren deze complexe deelopgaven zonder uitwerking (vermoedelijk uit het hoofd) uit te rekenen. Verrassend genoeg is het percentage leerlingen dat kolomgsgewijs deelde – wat het eindpunt is van hedendaagse leerlijnen en daarmee de vervanging van de staartdeling in het rekenonderwijs – nagenoeg gelijk gebleven. Uit de analyses van de accuratesse van elke strategie blijkt dat het antwoorden zonder uitwerking een stuk minder succesvol was dan de beide schriftelijke strategieën (traditionele en niet-traditionele aanpakken). Deze strategieverschuiving, die overigens vooral op het conto van jongens geschreven moet worden, heeft dus een ongunstig effect op de prestaties gehad. Ten tweede blijkt dat er nog een deel van de prestatiedaling onverklaard blijft na statistische correctie voor deze ongunstige verschuivingen in strategiegebruik. Dit betekent dat de kans op een goed antwoord met een bepaalde strategie op een bepaalde opgave in 2004 lager lag dan in 1997. Deze algemene daling van de prestaties bínnen elke strategie is zorgelijk.

De resultaten op het gebied van vermenigvuldigen zijn tot op zekere hoogte vergelijkbaar met die van delen, zoals blijkt uit hoofdstuk 3. Ook bij vermenigvuldigen is een verschuiving in strategiegebruik gevonden, waarbij er in 2004 minder leerlingen waren die vooral traditioneel cijferend vermenigvuldigden en meer leerlingen die zonder

schriftelijke uitwerking rekenden (opnieuw vooral jongens). Deze toename in het uit het hoofd rekenen is bij vermenigvuldigen echter van kleinere omvang dan bij delen, en wordt bovendien aangevuld met een toename in het gebruik van niet-traditionele strategieën. Ook was het traditionele vermenigvuldigingsalgoritme in beide peilingsjaren de meest voorkomende strategie, wat in overeenstemming is met het feit dat het algoritme nog wel het eindpunt in de leerlijn voor vermenigvuldigen is (dit is dan ook in tegenstelling tot de leerlijn voor delen). Het is daarmee dan ook opvallend te noemen dat juist bij vermenigvuldigen het gebruik van niet-traditionele strategieën is toegenomen over tijd terwijl het bij delen juist constant is gebleven; het tegengestelde patroon had meer in de lijn der verwachting gelegen.

Net zoals bij delen zijn ook bij vermenigvuldigen de schriftelijke strategieën succesvoller dan het antwoorden zonder uitwerking. Daarentegen blijken bij vermenigvuldigen de niet-traditionele strategieën minder succesvol dan het traditionele algoritme, terwijl bij delen deze strategieën ongeveer even succesvol waren (behalve voor de gemiddelde rekenaars). De afname van het traditioneel cijferend vermenigvuldigen ten koste van de toename in niet-traditionele strategieën en het hoofdrekenen hebben dus ook bij vermenigvuldigen een ongunstig effect op de prestaties gehad. In hoofdstuk 3 is ook de invloed van de strategie-instructie door de leerkracht op het strategiegebruik van leerlingen op vermenigvuldigen en delen in PPON-2004 onderzocht. Vooral het gebruik van de staartdeling wordt beïnvloed door de instructie van de leerkracht: vrijwel uitsluitend leerlingen van wie de leerkracht de staartdeling onderwees (en daarmee dus afweek van de rekenmethode) gebruikten de staartdeling. Bij vermenigvuldigen zijn de trends vergelijkbaar, maar wel een stuk minder uitgesproken.

Uit de resultaten van hoofdstuk 2 en 3 is gebleken dat – vooral bij delen – het meest relevante onderscheid in oplossingsstrategieën dat tussen schriftelijke strategieën en hoofdrekenen is. Daarom zijn in de volgende twee empirische studies nieuwe gegevens verzameld om de kenmerken van deze strategieën (zoals de verdeling over de opgaven, accuratesse, snelheid, en adaptiviteit van strategiekeuzes) voor de bewerking delen op een zuivere manier vast te stellen. Daartoe is een gedeeltelijke (hoofdstuk 4) en een complete (hoofdstuk 5) *choice/no-choice* studie (zie Siegler & Lemaire, 1997) opgezet. In het onderzoek in hoofdstuk 4 maakten 362 leerlingen uit groep 8 twee parallelle sets van elk vier deelopgaven (bijvoorbeeld $782 \div 32$) onder twee condities: eerst de ene set in een conditie waarbij ze zelf mochten kiezen of ze een schriftelijke strategie gebruikten of uit het hoofd rekenden (de *choice* of vrije keuze conditie), en vervolgens

de andere set in een conditie waarbij ze verplicht waren schriftelijk te rekenen (de *no-choice* of verplicht schriftelijke conditie). In de vrije keuze conditie waren ook nog vijf andere deelopgaven opgenomen die makkelijker getalskenmerken hadden (bijvoorbeeld $3240 \div 4$), en waarvan daarom verwacht werd dat ze uit het hoofd uit te rekenen waren.

Latente klasse analyse op de strategiekeuzes in de vrije keuze conditie liet zien dat er drie subgroepen van leerlingen onderscheiden kunnen worden: zij die voornamelijk uit het hoofd rekenden (meer jongens dan meisjes), zij die voornamelijk schriftelijk rekenden (meer meisjes dan jongens), en zij die van strategie wisselden door op de moeilijkere opgaven voornamelijk schriftelijk te rekenen en op de makkelijkere opgaven voornamelijk uit het hoofd te rekenen. Deze laatste groep leerlingen vertoonde dus een adaptief en flexibel patroon van strategiekeuzes. Opvallend is dat in deze groep vrijwel geen zwakke rekenaars zaten. Opnieuw blijken spontaan gekozen schriftelijke strategieën succesvoller dan spontaan gekozen hoofdrekenen, net als in hoofdstuk 2. Daarnaast is vastgesteld dat de prestaties van leerlingen die als ze zelf mochten kiezen hoofdrekenden op een bepaalde opgave, verbeterden als ze gedwongen werden schriftelijk te rekenen. Voor leerlingen die in de vrije keuze conditie al schriftelijk rekenden had het gedwongen schriftelijk rekenen geen effect op de prestaties. Het blijkt dus dat schriftelijke strategieën succesvoller zijn dan hoofdrekenen, niet alleen in vergelijkingen tussen opgaven en leerlingen, maar ook bínnen dezelfde opgave en dezelfde leerling. Een hieruit volgende aanbeveling is dan ook dat leerlingen meer aangespoord moeten worden om tussenuitkomsten of complete berekeningen te noteren bij het oplossen van deelopgaven met grote getallen en kommagetallen. Dit lijkt vooral van belang voor jongens, die meer geneigd zijn tot hoofdrekenen dan meisjes, en voor zwakke rekenaars, voor wie een groter verschil in accuratesse tussen de twee strategieën is vastgesteld dan voor gemiddelde en betere rekenaars.

Waarom kiezen leerlingen voor het riskante hoofdrekenen? Deze vraag komt aan bod in hoofdstuk 5, waarin het onderzoek van hoofdstuk 4 op twee manieren wordt aangevuld. Ten eerste is er een derde onderzoeksconditie opgenomen waarin leerlingen verplicht moesten hoofdrekenen, zodat zuiver vastgesteld kan worden hoe leerlingen met deze strategie presteren. Ten tweede is er een tweede prestatie-uitkomst opgenomen: naast het vaststellen of het goede of foute antwoord is gegeven, is ook de tijd opgenomen die leerlingen nodig hadden om tot hun antwoord te komen. Daardoor is het mogelijk niet alleen de accuratesse van beide strategieën te analyseren, maar ook de snelheid in de analyses mee te nemen. Zesentachtig leerlingen uit groep acht maakten dan ook drie

sets van vier opgaven in drie verschillende condities: de vrije keuze conditie conditie, de verplicht hoofdrekenen conditie en de verplicht schriftelijk rekenen conditie.

De resultaten laten zien dat hoofdrekenen vooral gekozen wordt omdat het sneller is dan schriftelijk rekenen, terwijl schriftelijk rekenen vooral werd gekozen omdat het accurater is. In dit onderzoek kan ook de adaptiviteit van strategiekeuzes worden vastgesteld: in hoeverre kiest een leerling op een opgave de strategie die voor hem of haar het beste is? Dit is gerelateerd aan het concept van *adaptive expertise* (Baroody & Dowker, 2003; Torbeyns, De Smedt, et al., 2009b). De resultaten laten zien dat een substantieel deel van de leerlingen niet hun beste strategie – gedefinieerd als de strategie die het snelst tot het goede antwoord leidt – koos. Deze suboptimale strategiekeuzes spreken aannames in cognitieve modellen van strategiekeuze tegen (zoals Shrager & Siegler, 1998; Siegler & Shipley, 1995). Een mogelijke verklaring hiervoor is dat deze modellen niet expliciet zijn over de invloed van individuele verschillen in de voorkeur voor snelheid en accuratesse (Phillips & Rabbitt, 1995), die ervoor kunnen zorgen dat sommige leerlingen het snelle maar foutgevoeliger hoofdrekenen kiezen. Ook de invloed van de socioculturele context wordt onderbelicht in de cognitieve modellen (Ellis, 1997; Luwel et al., 2009; Verschaffel et al., 2009). In de klas kan het bijvoorbeeld zo zijn dat hoofdrekenen meer gewaardeerd wordt dan schriftelijk rekenen.

Verder zijn er interessante verschillen tussen jongens en meisjes gevonden. Opnieuw blijken jongens meer geneigd tot hoofdrekenen dan meisjes, maar er is nu gevonden dat dat gerelateerd lijkt te zijn aan verschillen in de relatieve voorkeur voor accuratesse en snelheid. Meisjes kiezen hun strategie op basis van de accuratesse, daarmee snelheid negerend, terwijl de strategiekeuze van jongens meer gebaseerd is op snelheidsoverwegingen dan op accuratesse. Ook het algemene rekenniveau van de leerling is van invloed. Naast het weinig verrassende patroon dat bovengemiddelde rekenaars beter presteerden, een hogere strategie-accuratesse hadden en met elke strategie sneller rekenden dan ondergemiddelde rekenaars, zijn ook verschillen in de adaptiviteit van strategiekeuzes gevonden. Ondergemiddelde rekenaars pasten hun strategiekeuze zowel niet aan snelheid als aan accuratesse aan, terwijl bovengemiddelde rekenaars hun keuze aan beide componenten aanpasten.

De onderzoeken beschreven in de voorgaande hoofdstukken zijn merendeels gebaseerd op rekenopgaven in een realistische context: een meestal verbale beschrijving van een mathematisch probleem, vaak aangevuld met een illustratie. Dit type opgaven wordt veel gebruikt in het hedendaagse reken-wiskundeonderwijs (zie bijvoorbeeld

Gravemeijer & Doorman, 1999; Verschaffel et al., 2000) en in de rekentoetsen, zoals Cito's Leerling- en Onderwijs Volg Systeem (LOVS) en PPON. De vraag die gesteld kan worden is in hoeverre de dominantie van contextopgaven effect heeft op het toetsen van rekenvaardigheden. Dit is in twee studies onderzocht: in hoofdstuk 6 voor de prestaties van leerlingen in groep drie, vier en vijf, en in hoofdstuk 7 voor de prestaties en strategiegebruik van leerlingen in groep acht. De bevindingen zijn gemengd.

In hoofdstuk 6 wordt een onderzoek beschreven waarin ruim 2000 leerlingen in de onderbouw van het basisonderwijs naast de reguliere LOVS-taken voor hun jaargroep – die voornamelijk uit contextopgaven bestaan – een extra taak met alleen 'kale' getalsopgaven maakten. Met meerdimensionale IRT-modellen is vastgesteld dat twee samenhangende maar wel verschillende vaardigheden ten grondslag liggen aan het oplossen van deze twee typen opgaven. Dit betekent dat leerlingen die goed zijn in het oplossen van opgaven van het ene type niet direct ook goed zijn in het oplossen van het andere type, en dat een profiel van relatieve prestaties nuttig zou kunnen zijn voor de leerkracht. De samenhang lijkt wel toe te nemen met jaargroep, met een latente correlatie van .81 in groep drie oplopend naar .87 in groep vijf. Verder hebben indicatoren van het taalniveau van de leerling (thuistaal en niveau van begrijpend lezen) een verschillend effect op beide vaardigheden. In zijn algemeenheid zijn de effecten van taalniveau groter op de vaardigheid van het oplossen van contextopgaven dan op het oplossen van kale getalsopgaven. De prestatieachterstand van leerlingen die thuis geen Nederlands spreken ten opzichte van zij die dat wel doen is op de contextopgaven een stuk groter dan op de kale getalsopgaven. Dit betekent dat meer balans in de toetsen tussen contextopgaven en kale getalsopgaven het gat tussen leerlingen die thuis geen Nederlands spreken en zij die dat wel doen waarschijnlijk zal verkleinen. Een beperking van deze studie is echter dat het effect van een context in een rekenopgave niet rechtstreeks kan worden vastgesteld, omdat de twee typen opgaven niet vergelijkbaar waren qua getalskenmerken.

Daarom is een ander onderzoek verricht waarin dit wel mogelijk was, ditmaal in groep acht (hoofdstuk 7). In totaal 685 leerlingen maakten een set opgaven op het gebied van optellen, aftrekken, vermenigvuldigen en delen met grote getallen en kommagetallen. Er waren acht paren van opgaven: elk paar bestond steeds uit een contextopgave (ontleend aan PPON) en de kale getalsversie van die opgave: de rekenkundige bewerking, ontdaan van de context. Het effect van de contexten is vastgesteld op twee aspecten van probleemoplossen: de prestaties en het strategiegebruik. In tegenstelling tot in de onderbouw (hoofdstuk 6) blijkt er in groep acht slechts één latente vaardigheidsdimensie

ten grondslag te liggen aan het oplossen van de twee typen opgaven. Verder blijkt dat de contexten de moeilijkheidsgraad van een opgave nauwelijks beïnvloeden: alleen bij de deelopgaven was de opgavevariant met context iets moeilijker dan die zonder context. Ook de strategiekeuze en de strategie-accuratesse worden niet beïnvloed door de aanwezigheid van een context. Een verdere belangrijke bevinding is dat dit alles geldt voor jongens en voor meisjes, voor leerlingen die thuis Nederlands spreken en zij die een andere taal spreken, en voor leerlingen met een verschillend niveau van begrijpend lezen.

Samenvattend zijn dus in groep acht nauwelijks tot geen effecten gevonden van de aanwezigheid van contexten in rekenopgaven op verschillende aspecten van het probleemoplossen. Dit is tegenstelling tot de verwachtingen op basis van hoofdstuk 6 en ook tot vaak gehoorde overtuigingen. In combinatie met de onderzoeksresultaten bij jongere kinderen lijken de bevindingen van hoofdstuk 7 erop te wijzen dat de invloed van contexten in rekentoetsen in de loop van de basisschooltijd afneemt en zelfs verdwijnt.

Ter afsluiting kan geconcludeerd worden dat de in dit proefschrift gerapporteerde onderzoeken elk meer inzicht hebben verschaft in hoe basisschoolleerlingen rekenen in het hedendaagse reken-wiskundeonderwijs. Zonder geavanceerde psychometrische technieken, in het bijzonder latente variabele modellen waarin verklarende variabelen zijn opgenomen, was het niet mogelijk geweest deze resultaten en inzichten te verkrijgen. Het samenbrengen van de psychometrie en de psychologie heeft dan ook zijn nut en noodzaak bewezen.

Curriculum vitae

Marian Hickendorff werd op 1 juni 1981 te Leiden geboren. In 1999 behaalde zij het diploma gymnasium aan het Stedelijk Gymnasium te Leiden. Hierna volgde zij gedurende één jaar de studie Geneeskunde aan de Universiteit Leiden waarvan zij in 2000 de propedeuse cum laude behaalde. Vanaf 2000 volgde zij de studie Psychologie aan de Universiteit Leiden, in de afstudeerrichting Methoden en Technieken van Psychologisch Onderzoek met een specialisatie in Ontwikkelings- en Onderwijspsychologie. Deze dubbele interesse in de methoden en technieken en in de onderwijspsychologie kwam tot uiting in haar afstudeerstage bij Cito Instituut voor Toetsontwikkeling te Arnhem, waar zij meewerkte aan verschillende fasen van toetsontwikkeling op het gebied van rekenen-wiskunde in het primair onderwijs. Voor haar afstudeerscriptie ontving zij de Van de Geer Scriptieprijs 2004-2005 van de opleiding Psychologie. In 2005 rondde zij haar doctoraal Psychologie cum laude af, en werd vervolgens aangesteld als onderzoeksassistent aan de afdeling Methoden en Technieken van het Instituut Psychologie van de Universiteit Leiden, met als doelstelling het schrijven van een onderzoeksvoorstel voor promotieonderzoek. Van 2006 tot 2011 werd zij aangesteld als assistent in opleiding aan de Universiteit Leiden om dit promotieonderzoek uit te voeren, in een samenwerkingsverband met Cito. In 2008 won zij de Promovendiprijs voor het beste artikel in de opleiding Psychologie van het studiejaar 2007-2008. In 2009 was zij als toegevoegd onderzoeker verbonden aan de KNAW-commissie *Rekenonderwijs op de basisschool*. Momenteel is zij aangesteld als universitair docent aan de afdeling Methoden en Technieken van het Instituut Psychologie van de Universiteit Leiden. Daarnaast werkt ze sinds 2006 voor de *Eindtoets Basisonderwijs* van Cito, waarvoor zij toetsopgaven voor het onderdeel rekenen-wiskunde construeert.