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.