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

Hickendorff, M.

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-

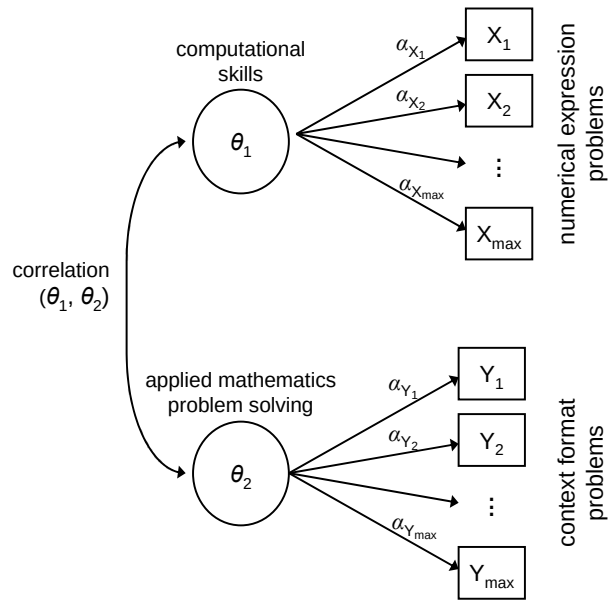


FIGURE 6.1 *Graphical representation of between-item two-dimensional IRT model.*

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

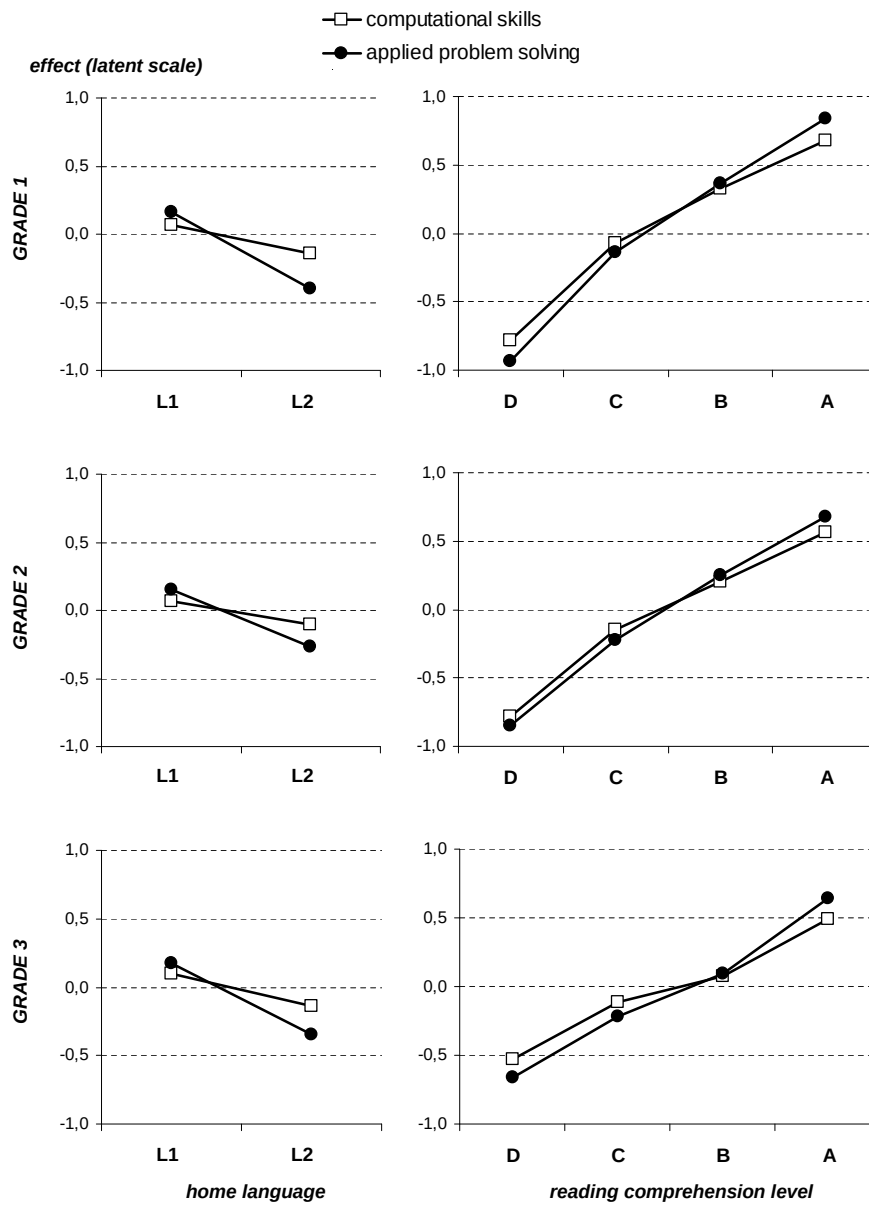


FIGURE 6.2 *Graphical display of home language effects (left plots) and reading comprehension level effects (right plots) for the two ability dimensions, grade 1 (upper part), grade 2 (middle part), and grade 3 (bottom part).*

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

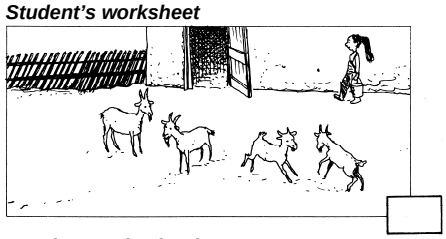
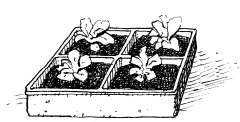
The present findings have implications for testing practices as well as for education. Regarding testing, the current dominance of context format problems in Dutch mathematics competence tests as well as in for example PISA merits critical consideration. We should be well aware that this offers a rather 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 = ___