

The Language Factor in Elementary Mathematics Assessments: Computational Skills and Applied Problem Solving in a Multidimensional IRT Framework

Marian Hickendorff

To cite this article: Marian Hickendorff (2013) The Language Factor in Elementary Mathematics Assessments: Computational Skills and Applied Problem Solving in a Multidimensional IRT Framework, *Applied Measurement in Education*, 26:4, 253-278, DOI: [10.1080/08957347.2013.824451](https://doi.org/10.1080/08957347.2013.824451)

To link to this article: <https://doi.org/10.1080/08957347.2013.824451>



Published online: 27 Sep 2013.



Submit your article to this journal [↗](#)



Article views: 1147



View related articles [↗](#)



Citing articles: 1 View citing articles [↗](#)

The Language Factor in Elementary Mathematics Assessments: Computational Skills and Applied Problem Solving in a Multidimensional IRT Framework

Marian Hickendorff

Institute of Psychology, Leiden University, The Netherlands

The results of an exploratory study into measurement of elementary mathematics ability are presented. The focus is on the abilities involved in solving standard computation problems on the one hand and problems presented in a realistic context on the other. The objectives were to assess to what extent these abilities are shared or distinct, and the extent to which students' language level plays a differential role in these abilities. Data from a sample of over 2,000 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 ability dimensions (computational skills and applied mathematics problem solving) ranged from .81 in grade 1 to .87 in grade 3, indicating that the ability dimensions 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.

INTRODUCTION

Mathematics education has experienced great international reform (e.g., Kilpatrick, Swafford, & Findell, 2001). The general characteristic of this reform is that mathematics education should no longer focus predominantly on traditional decontextualized 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 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 can be developed in a meaningful manner; and students may develop knowledge of when and how to use mathematics in everyday-life situations (e.g., Verschaffel, Greer, & De Corte, 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 educational reform has greatly impacted mathematics curricula. The 2004 national assessments showed that almost all elementary schools used a mathematics textbook based on reform principles (Janssen, Van der Schoot, & Hemker, 2005; Kraemer, Janssen, Van der Schoot, & Hemker, 2005), although a return to more traditionally oriented mathematics textbooks has been observed recently (Royal Netherlands Academy of Arts and Sciences, 2009). These reform-based textbooks contain many problems in context, although there are substantial differences in this respect between the different textbooks. To keep up with these developments, Dutch mathematics assessments (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 predominantly contain contextual problems. The latter system, consisting of biannual assessments for grades 1 to 6 with the purpose of enabling teachers to monitor their students' progress, is used by a majority of primary schools. Today's Dutch primary school students mathematics education and assessment thus consist for a large part of problems in (more or less) realistic contexts.

This international shift toward including numerous contextual problems gives rise to two questions. First, to what extent are different abilities¹ involved in solving standard computation problems versus solving contextual problems? This question is important because it provides insight whether the currently used tests that are dominated by contextual problems allow for the same conclusions on individual or group differences as a test consisting predominantly of standard computation problems. Second, contextual problems are usually verbal, giving rise to the question the role language plays in problem solving ability. Determining to what extent students' language level has differential effects on the two ability dimensions is clearly of practical importance, for example in gaining further insight in the broadness of the commonly observed performance lag of language minority students. These two questions are now further elaborated.

Standard Computation Problems and Context Format Problems

Solving standard computation problems on the one hand and realistic context format problems on the other are likely to involve different aspects of mathematical cognition (e.g., Fuchs et al., 2008). There are two different view points on this matter. On the one hand, phase-like approaches to mathematical modeling or mathematizing emphasize that solving contextual problems is a complex process involving several cognitive processes or phases (e.g., as in the model on p. xii of Verschaffel et al., 2000). Only after steps in which a situational and mathematical model of the problem situation have been formed accurately does computational skill come into play. Therefore, other factors than "pure" computational skills are likely to contribute to success in applied mathematics problem solving (Wu & Adams, 2006). On the other hand, solving contextual problems may elicit different strategic approaches than solving computational problems, as for instance has been reported by Koedinger, Alibali, and Nathan (2008) in the domain of algebra. This view is in line with mathematics education reform, in which children

¹With the general term "ability", that is used throughout the article, I refer to a dimension quantifying different levels of performance or achievement, on which individuals may differ.

are supposed to approach (unfamiliar) contextual problems as situations to be mathematized, thereby not reverting to searching for the application of the appropriate formal computational procedure but instead using informal strategies. In this view, computational skill is conceived of not so much as a necessary prerequisite of successful applied problem solving, but rather computational skill and contextual problem solving are expected to involve separate abilities.

Both views have in common that 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. In line with findings of Fuchs et al. (2008), we hypothesize that these are two related but distinct abilities. Furthermore, we expect the relation between these two aspects to increase with age, since students in higher grades have more developed cognitive schemata to solve word problems due to more years of formal schooling (De Corte, Verschaffel, & DeWin, 1985; Vicente, Orrantia, & Verschaffel, 2007).

The results pertaining 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.

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. More generally, content area domains (such as mathematics and science) assessments often confound language skills and academic aptitude (e.g., Solano-Flores & Trumbull, 2003). In science assessments this confounding may even be more pervasive than in mathematics because of the high level of linguistically and culturally dependent content (e.g., Penfield & Lee, 2010). Research in support of the importance of language in mathematics word problem solving has shown that misunderstanding of the problem situation can be a common source of errors (Cummins, Kintsch, Reusser, & Weimer, 1988; Wu & Adams, 2006) and that conceptual rewording of word problems facilitates performance (e.g., Vicente et al., 2007).

The language factor in content-area tests is particularly important in a diverse student population. Ethnic minority students score lower on language tests than native students. In addition, they consistently lag behind in mathematics as well, as has been found in international assessments such as TIMSS (*Trends in International Mathematics and Science Study*; Mullis, Martin, & Foy, 2008) as well as in Dutch national assessments (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 United States, much attention is paid to a similar issue: the assessment of English language learners (ELLs) (e.g., Abedi, Lord, Hofstetter, & Baker, 2000; Penfield & Lee, 2010;

Solano-Flores & Trumbull, 2003). In the domain of mathematics, Abedi and Lord (2001) found that linguistic simplifications of the problem text of NAEP mathematics test items narrowed the performance gap between ELL and non-ELL students. 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 performance gap between ELLs and non-ELLs was larger on linguistically complex items than on noncomplex items, regardless of the item content difficulty. An important distinction to be made is between everyday language and academic language: the contextual application of language, such as the use of nonspecialized academic words and specialized content area words (e.g., Bailey & Butler, 2003; Bailey, Butler, & Sato, 2007). 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 problems in both types of language.

The present study extends these 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. Furthermore, we expect the language effects to decrease with more years of formal schooling since inexperienced problem solvers rely more heavily on the text because they lack highly developed semantic schemata for word problems (De Corte et al., 1985). Therefore, language level is expected to be more important to understand the problem situation in lower grades than in higher grades. Of particular importance is whether language 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. Because this study also included language-free problems (standard computation problems in numerical format), its design can be seen as an extension to the four types of comparisons in the testing of language minorities described by Solano-Flores and Trumbull (2003). That is, linguistic groups were compared not only within one test language, but also on a language-free part of the test in which linguistic complexity is supposed not to affect performance; something that is not possible in science assessments. In addition to home language effects, the role of student reading comprehension level is addressed in the current study.

The Current Study

In the current cross-sectional survey students from grades 1, 2, and 3 solved a set of computational problems in addition to a set of contextual mathematics 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 confirmatory 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).

METHOD

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 a relatively high proportion of 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 an effective sample of 2,049 students that is analyzed by grade.

Two types of background information on the students' language level were collected. First, the language spoken at home (as reported by the teacher), which we classified into Dutch (L1) or another language (L2). Almost one-third of the students spoke a language different than Dutch at home, see also Table 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 distribution of the non-Dutch languages across grades was, in decreasing order of frequency: Turkish (35.7%), Moroccan/Arabic (12.9%), a Dutch dialect or local language such as Friesian (10.9%), Berber/Tamazight (10.2%), East-European languages (5.0%), West-European languages (3.7%), Papiamentu (2.3%), and Sarnami (1.6%); 17.8% of the languages did not fall into these categories. Furthermore, parental educational background differed by student home language: of 96.7% of the L1 students both parents completed at least some track of secondary education, while this was only 56.9% of the L2 students ($\chi^2(1, N = 1,878) = 479.5, p < .001$).

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.

TABLE 1
Pupil Background Information: Distribution of Home Language and Reading Comprehension Level

	Home Language			Reading Comprehension Level				
	Dutch	other	missing	A	B	C	D	missing
<i>Grade 1</i>								
freq	430	215	4	112	130	140	159	108
valid %	66	34		21	24	26	29	
<i>Grade 2</i>								
freq	514	216	6	170	152	177	106	131
valid %	70	30		28	25	29	18	
<i>Grade 3</i>								
freq	454	203	7	171	122	135	116	120
valid %	69	31		31	25	22	21	

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 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.

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 two regular CITO booklets for grade 1 (CITO, 2005a) and also two regular booklets for grade 2 (CITO, 2005b), and three regular booklets for grade 3 (CITO, 2006). All these booklets contained predominantly problems in context format. 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. In contrast, the extra booklet contained only problems in standard computation format (numerical expression only, e.g., $17 - 5 = . . .$). The Appendix shows a sample of problems used.

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 2.

TABLE 2
For Both Subscales, the Number of Problems per Operation, Descriptive Statistics of the Proportion Correct Scores P (Correct), and Cronbach’s α

	<i>Add.</i>	<i>Number of Problems</i>					<i>P (correct)</i>		<i>Cronbach's α</i>
		<i>Sub.</i>	<i>Mult.</i>	<i>Div.</i>	<i>Combi.</i>	<i>Total</i>	<i>M</i>	<i>SD</i>	
<i>Computational skills</i>									
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
<i>Contextual problem solving</i>									
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

Procedure

The students completed each of the three (grade 1 and 2) or four (grade 3) different booklets on separate mornings. 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 through all problems independently. In the second grade, the problems of one booklet was read aloud by the teacher, while the students had to work through the problems independently on the other booklet. The assessment procedure of the extra booklet (numerical problems) was the same 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, scoring them as either correct or incorrect.

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 (also called 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 $\theta_p = (\theta_{p1}, \theta_{p2}, \dots, \theta_{pm})$. The multidimensional generalization of the 2PL model is:

$$P(X_{ip} = 1 | \theta_p) = \frac{\exp(\sum_{k=1}^m \alpha_{ik} \theta_{pk} + \delta_i)}{1 + \exp(\sum_{k=1}^m \alpha_{ik} \theta_{pk} + \delta_i)}. \quad (1)$$

Each item is characterized by an intercept δ_i , and by m dimension-specific discrimination parameters α_{ik} ($k = 1, \dots, m$). 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 suitable for tests comprising several subtests, each aiming to measure one ability. In the present application, we used between-item MIRT models with two correlated 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 1 shows a graphical representation of this two-dimensional model.

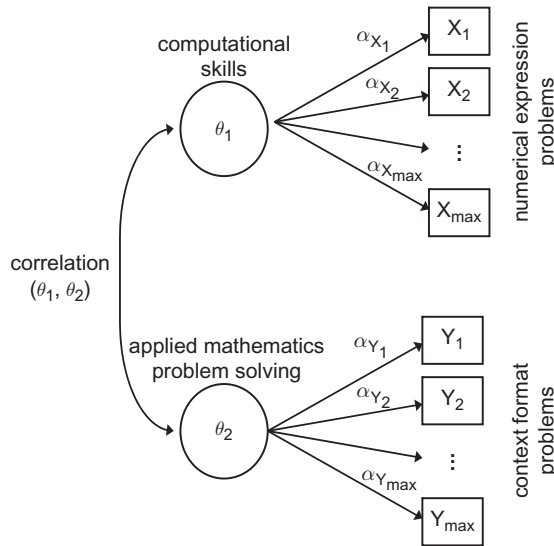


FIGURE 1 Graphical representation of between-item two-dimensional IRT model.

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 pertained to the estimates of the latent correlation 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, Tuerlinckx, De Boeck, & Kuppens, 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 (L1 or L2) and their reading comprehension level (four norm-referenced quartiles). Including person explanatory variables in an IRT model results in a latent regression: the latent person variable θ_p can be considered to be 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, two-step, or 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, we implemented the two-step approach separately for each grade. First, a two-dimensional between-item MIRT measurement model was fit, under the assumption that $(\theta_{p1}, \theta_{p2})$ follows a bivariate normal distribution with mean vector $(0,0)$ and variance-covariance matrix $\begin{pmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{pmatrix}$ to identify the model, as well as to standardize the two scales to the same metric (i.e., $\mu = 0$ and $\sigma = 1$). 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 predictor variable(s)—multivariate distribution of θ_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. In order to retain standardization of the metrics of the two latent dimensions, the (residual) variances of the two dimensions were restricted to be equal, so that reported differences on the two scales have the same meaning in terms of effect sizes (standardized mean difference).

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 used to test whether the improvement in fit between two nested models is statistically significant. The LR-test statistic 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 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, Chen, Su, & Wang, 2005). In this procedure, item parameters are estimated within a marginal maximum likelihood (MML) formulation, assuming a (multivariate) normal distribution for the person parameters. Gaussian quadrature with 20 nonadaptive quadrature points was used for the approximation of the integration, and Newton-Raphson as the optimization method.

RESULTS

Relationship Between the Different Ability Dimensions

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 3.

In grade 1, the correlation between the observed proportions correct on numerical expression problems and on contextual problems was .723. 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 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 estimate between these two abilities was .807 (SE = .021) with a 95% confidence interval of .763–.844, computed as described in Raykov (2012) using the Fisher *r*-to-*z* transformation. This latent correlation is 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

TABLE 3
Correlations Between Proportion 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

	<i>Correlation Total Scores</i>	<i>Latent Correlation</i>	<i>LR-Test (2D vs. 1D)</i>	
			<i>Statistic</i>	<i>p-value</i>
Grade 1	.723	.807 (SE = .021)	305.2	<i>p</i> < .001
Grade 2	.745	.854 (SE = .016)	199.3	<i>p</i> < .001
Grade 3	.769	.872 (SE = .015)	171.8	<i>p</i> < .001

(OECD, 2009). So, we would expect a latent correlation of at least .80 between two subscales of mathematics, and the found estimate of .807 is barely higher. The current latent correlation indicates that 65% of the ability variances is shared, while 35% of the variance is unique. The δ_i -estimates ranged from $-.74$ to 2.32 (with SEs ranging from .09 to .21) indicating that the items were fairly easy, and the α_{ik} -estimates ranged from .77 to 2.30 (with SEs ranging from .11 to .24) indicating that there was substantial variability in the degree to which the items discriminate.

For grade 2, Table 3 shows that, like in grade 1, 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 .854 (SE = .016; 95% CI .818–.883), indicating that 73% of the ability variances is shared, while 27% of the variance is unique. The δ_i -estimates ranged from $-.68$ to 2.98 (with SEs ranging from .09 to .19) and the α_{ik} -estimates ranged from .58 to 1.99 (with SEs ranging from .11 to .19).

Finally, in grade 3 the two-dimensional model again fits significantly better than the one-dimensional model as evidenced from the LR-test (see Table 3). 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 .872 (SE = .015; 95% CI .839–.899), indicating that 76% of the ability variances is shared, while 24% of the variance is unique. The δ_i -estimates ranged from $-.70$ to 3.56 (with SEs ranging from .09 to .24), and the α_{ik} -estimates ranged from .45 to 1.87 (with SEs ranging from .10 to .21).

The results thus far quite clearly show 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, as Figure 2 shows. However, whether these correlation estimates differ significantly from each other cannot be judged directly by these confidence intervals (e.g., Cumming & Finch, 2005). Therefore, for each of the three pairwise differences between grades in the correlation estimates, a 95% confidence interval was computed using the derivation of Zou (2007, eq. 15). Results show that only the interval of the difference in latent correlation estimates of grade 1 and 3 did not contain 0 (.013, .117) and was thus significant at the 5% level. We may conclude that there is a significant increase in the correlation between the performance dimensions of computational skills and applied problem solving between grade 1 and 3.

Language Effects

Now that it is established that computational skills and applied mathematics problem solving involve two highly related but distinct abilities, the next research question about the role of language is addressed. 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

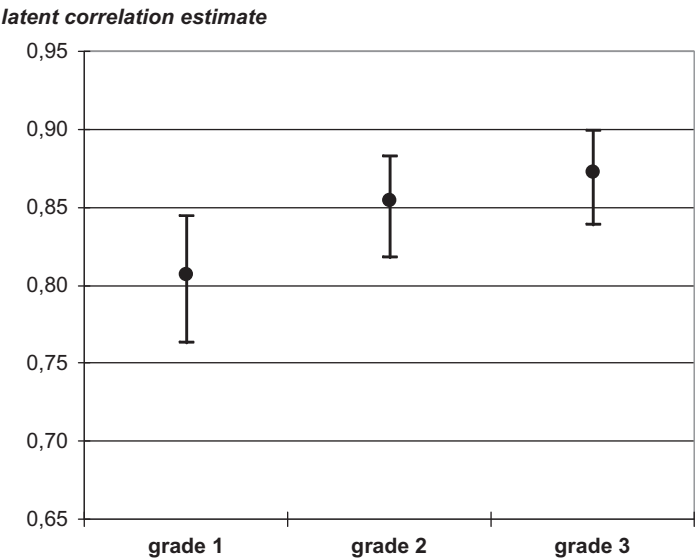


FIGURE 2 95% confidence intervals of estimate of latent correlation between computational skills and applied problem solving.

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, L2 students were behind in their reading comprehension level compared to L1 students.

Recall that we applied 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.² The two ability dimensions were scaled with equal (residual) variances. All effects reported are on the latent scale.

Grade 1. Pupils' home language significantly affected overall or average mathematics problem solving ability ($LR = 17.7, df = 1, p < .001$). Moreover, the difference between L1 and L2 students 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 3 graphically shows that L1 students outperformed students with another home language significantly more on the applied

²The number of missing data on the reading comprehension measure and/or the home language variable was substantial, with 109, 130, and 120 students with missing data on one or both of the predictor variables in grade 1, 2, and 3, respectively. The potential effects of these missing data were investigated by estimating the measurement MIRT models based on the data of students who had non-missing scores on both student background variables, that is with $N = 540$ in grade 1, $N = 592$ in grade 2, and $N = 544$ in grade 5, and comparing these estimates with those of the models based on all students as reported in the previous section. The item parameter estimates δ_i and α_{ik} of these models correlated very highly ($r \geq .97$). That implied that the same measurement model holds for the complete sample as for the sample of students without missing values on the background variables, and the model parameters are thus robust against leaving out students with missing data.

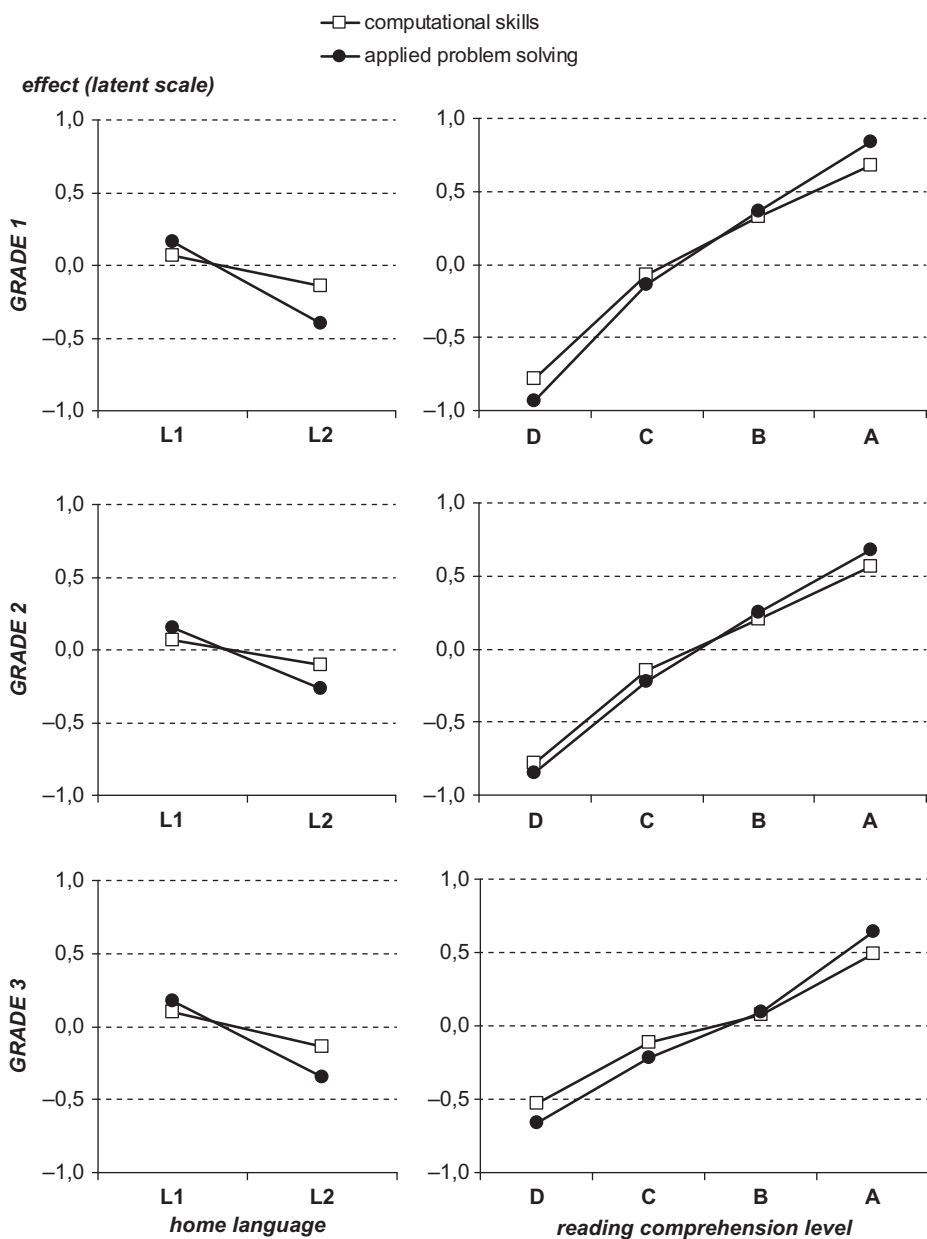


FIGURE 3 Graphical display of home language effects (left plots) and reading comprehension level effects (right plots) for the two ability dimensions, grade 1 (upper part), grade 2 (middle part), and grade 3 (bottom part).

dimension (difference on the latent 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 3 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 L2 students was mediated by their lower reading comprehension level. Statistically controlling for reading comprehension level, the performance lag of L2 students compared to L1 students disappeared on the applied mathematics dimension (difference = .05, $z = .61$), and even turned into a significant advantage of L2 students on the computational skills dimension (difference = $-.28$, $z = -3.10$).

Grade 2. Pupils' home language significantly affected overall mathematics problem solving ability ($LR = 10.1$, $df = 1$, $p = .001$). In addition, the difference between L1 and L2 students 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 3 graphically shows that L1 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 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 3.

Finally, statistically controlling for reading comprehension level differences, the home level effects were no longer significant on applied mathematics (difference = .07, $z = .82$) or on the computational skills dimension (difference = $-.16$, $z = -1.81$). On computation skills, a pattern similar to grade 1 emerged: the (nonsignificant) disadvantage of L2 students compared to L1 students reversed to a (nonsignificant) advantage for L2 students after controlling for reading comprehension level.

Grade 3. 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 mathematics problem solving ability ($LR = 16.8$, $df = 1$, $p < .001$). The bottom left plot of Figure 3 shows that L1 students outperformed L2 students 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 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 3.

Finally, statistically controlling for reading comprehension level, L1 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, L1 students significantly outperformed L2 students on both mathematics ability dimensions 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 L2 students 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 both mathematics ability dimensions in each grade. In addition, this reading comprehension effect was larger on the applied mathematics dimension than on the computational skills dimension, as expected. Finally, controlling for reading comprehension level, the performance lag of L2 students compared to L1 students 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 for L2 students.

Looking for a trend in the language-level effects across the grades, we computed the standardized mean difference between language groups (e.g., between L1 and L2 students) by dividing the estimated difference on the latent scale by the estimated within-group (residual) standard deviation of that scale. From Figure 4, showing these standardized mean differences with respect to home language (upper part) and reading comprehension level A versus D (lower part) by grade, the following pattern emerges. There was no specific trend by grade in the differences between L1 and L2 students on either computational skills or on applied problem solving. By contrast, reading comprehension level effects seemed to decrease in higher grades, judged by the estimates of standardized mean differences between level A and level D students (see lower part of Figure 4). On the computational skills dimension, the 95% CIs of the pairwise difference in these effect sizes between grade 3 and 2 (.03, .76) and between grade 3 and 1 (.24, 1.02) did not contain 0 and were thus significant at the 5% level. On the applied problem solving dimension, only the 95% CI of the difference between grade 3 and 1 in standardized mean difference (.29, 1.10) did not contain 0 and was thus significant at the 5% level. In summary, the effect of reading comprehension diminished between grade 1 and 3 on both dimensions, while the performance lag of L2 students did not decrease by grade.

The influence of linguistic complexity of the problem text. We found that student language variables had a larger effect on contextual problem-solving ability than on numerical problem-solving ability. The contextual problems, however, were very coarsely defined as all problems involving text. To investigate the contextual problems at a more fine-grained level, we also analyzed on which contextual problems the effects of students' language variables were largest, and to what extent that was related to the linguistic complexity of the problem text. Because only the grade 3 tests involved problems that all had to be read by the students independently, we concentrate on the grade 3 data. Specifically, we analyzed *differential item functioning* (DIF) of the contextual problems, with the numerical problems as the anchor-items without DIF. In DIF-analyses, the group of test takers is divided into two (or more) subgroups: the reference

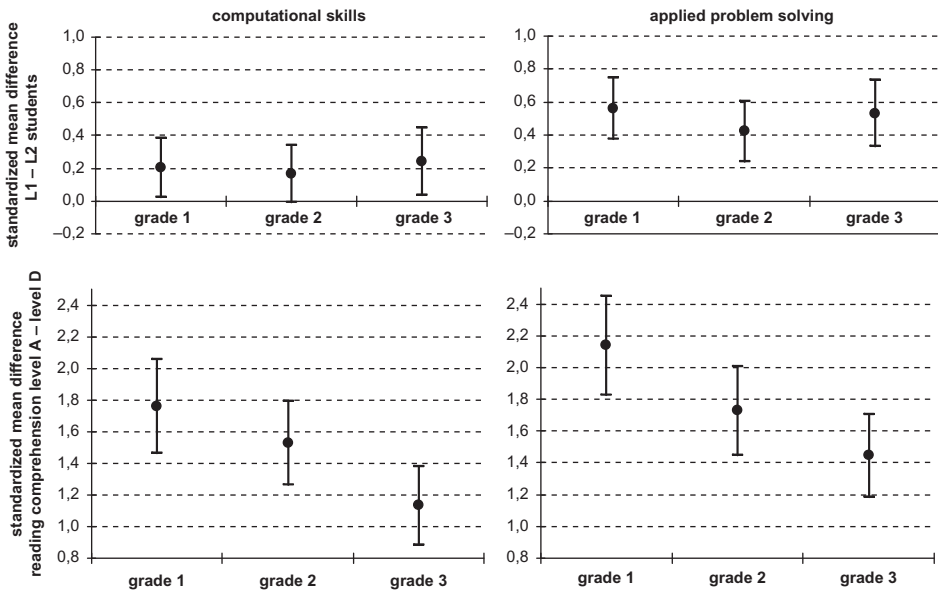


FIGURE 4 95% confidence interval of standardized mean difference between home language groups L1 and L2 (upper part) and reading comprehension level groups A and D (lower part), on computational skills (left plots) and on applied problem solving (right plots).

group, and one or more focal groups. What is tested is whether the item parameters of items that are suspected to function differently for these groups are the same or not. We focus on *uniform* DIF, in which only the item's difficulty parameter can be different for the reference and the focal group. That is, for the reference group, the following measurement model is specified: $P(X_{ip} = 1|\theta_p) = \frac{\exp(\alpha_i\theta_p - \beta_i)}{1 + \exp(\alpha_i\theta_p - \beta_i)}$, with β_i the item difficulty parameter. For the focal group, the shift in difficulty parameter is modeled by the γ_i parameter, as follows:

$$P(X_{ip} = 1|\theta_p) = \frac{\exp(\alpha_i\theta_p - (\beta_i + \gamma_i))}{1 + \exp(\alpha_i\theta_p - (\beta_i + \gamma_i))}.$$

Two separate DIF-analyses were carried out: one with respect to students' home language, and the other with respect to students' reading comprehension level that we recoded into two categories: above median (levels A and B) and below median (levels C and D). First, the DIF-analysis with respect to home language (with L2 students as focal group) showed that 7 of the 28 contextual problems showed significant DIF ($p < .05$), all in the expected direction ($\gamma_i > 0$). That is, conditional on the ability level on the numerical problems, these problems were significantly more difficult for L2 students than for L1 students, with γ_i ranging between .42 (SE = .19) and .85 (SE = .21). The mean difficulty shift over all 28 problems was .25. Second, the DIF-analyses with respect to reading comprehension level (with students with below-median level as the focal group) showed that 10 of the 28 contextual problems showed significant DIF ($p < .05$). Again,

the shift in difficulty on these problems was in the expected direction ($\gamma_i > 0$): conditional on the ability level on the numerical problems, these problems were significantly more difficult for students with below-median reading comprehension level than for students with above median reading comprehension level, with γ_i ranging between .44 (SE = .22) and .94 (SE = .24). The mean difficulty shift over all 28 problems was .35. Furthermore, there were four problems that showed significant DIF in both analyses (i.e., with respect to home language as well as to reading comprehension level). Moreover, across all 28 items, the γ -parameters from the two analyses had a high correlation of $r(28) = .77$ ($p < .001$) meaning that contextual problems that were relatively difficult for L2 students were also relatively difficult for students with below-median reading comprehension level.

Next, we hypothesized that the shift in difficulty parameter for low language-level students was positively related to the linguistic complexity of the problem text. We tested this hypothesis by scoring a measure of text readability (the opposite of linguistic complexity) for all problem texts. This measure was the so-called CLIB-score developed by CITO (Staphorsius, 1994), focusing on reading comprehension (as opposed to technical reading). It is based on four text characteristics: the average word length (in number of characters), the percentage of frequently used words, the type-token ratio (an indicator of vocabulary variation), and the average number of words per sentence. We expected a negative relation between the problem text's readability measure (CLIB-score) and the estimated DIF-parameters. In other words, the extra difficulty for low language level students was expected to be larger in problems with less readable text. Figure 5 shows a tendency to this negative relation between a problem's CLIB-score and the estimated DIF-parameters γ_i for both home language and reading comprehension level. The correlations, however, were not significant, with $r(28) = -.10$ (one-sided $p = .30$) for DIF with respect to home language, and $r(28) = -.17$ (one-sided $p = .19$) for DIF with respect to reading comprehension level. Of the individual text characteristics, the percentage of frequently used words had a significant relation to the item's DIF-parameter with respect to reading comprehension level: $r(28) = -.34$ (one-sided $p = .037$). So, the fewer frequently used words the problem text contained, the larger the extra difficulty of that problem was for students with below-median reading comprehension level.

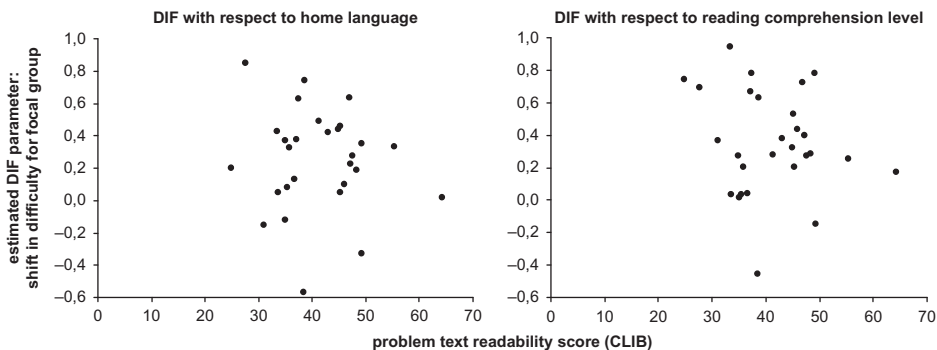


FIGURE 5 Relation between estimated DIF-parameter γ_i and problem text's readability measure for contextual problems in grade 3.

DISCUSSION

A sample of first, second, and third graders with a relatively large proportion of 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 concerned 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 24% to 35% of the variance is determined by unique contributions of the two dimensions. Moreover, the relationship between the two ability dimensions (computational skills and applied problem solving) appeared to get stronger in the higher grades.

Analyses on the second research question pertaining to the role of student language variables 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 for applied problem solving than computational skills in each grade. That is, the performance gap between L1 and L2 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 need to rely less on the problem text. Furthermore, statistically controlling for reading comprehension level, the performance lag of L2 students compared L1 students was reduced on each dimension, and this reduction was slightly larger on the applied mathematics dimension than on computational skills. Finally, we found some indications that in grade 3, the effects of student home language and reading comprehension level were larger on problems with more linguistically complex texts. In particular, the higher the percentage of infrequent words in the problem text, the larger the gap in difficulty level between students with below-median and above-median reading comprehension level.

Issues in the Multidimensional IRT Framework

In this study, we employed a multidimensional IRT framework. Between-item MIRT models with explanatory variables on both dimensions turned out to be a very useful and flexible approach. However, five 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). By contrast, in within-item MIRT models items can have more than one nonzero discrimination, and hence multiple latent factors can be involved in solving a particular problem. For example, similar to what Hartig and Höhler (2008) did on reading and listening comprehension assessment data, it would be possible to distinguish two dimensions: one general computational skill dimension that affects items of both problem types, and one specific dimension that is only involved in solving context format problems. A disadvantage is that it would be necessary to assume a compensatory mechanism between these two dimensions (i.e.,

a low level on one factor can be compensated with a high level on the other factor), which seems unnatural. In addition, for the model parameters to be identified, the general computational skills dimension and the specific applied problem-solving dimension have to be uncorrelated (Hartig & Höhler, 2008). The major disadvantage, however, is that such a within-item model is more complex both in a statistical sense (i.e., more parameters) as well as from a substantive point of view.

We fitted these within-item models to the data, though, and found that they showed a marginal improvement in fit over the between-item models (lower AIC, but higher BIC—although for the grade 2 data the BIC of the within-item model was somewhat lower than of the between-item model). We believe that this does not compensate for the far less straightforward interpretation of the dimensions, latent regression analyses results, and communication with the end-users of the test (cf., Hartig & Höhler, 2008). For these reasons, we considered the between-item model more appropriate for the current structure. 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 & Gier, 2007).

Second, a related issue concerns absolute fit of the 2-dimensional models. Unfortunately, in (M)IRT, there are no overall goodness-of-fit indices such as they exist in factor analysis (Embretson & Reise, 2000). Therefore, we focused on relative fit indices comparing two models, such as BIC and LR-tests. In an attempt to provide some support for the overall fit, we estimated separate unidimensional 2PL models (i.e., one model for the numerical problems, and another model for the contextual problems) with the *ltm*-package in R (Rizopoulos, 2006) and determined item fit. The percentage of items showing misfit at the 5% significance-level was 6.5% (numerical) and 4.5% (contextual) in grade 1, 11.8% and 4.2% in grade 2, and 5.1% and 3.6% in grade 3. Item fit was thus within the expected range (perhaps the numerical problems in grade 2 as the exception), indicating that a one-dimensional IRT model fitted the responses to one type of problems quite well. Because the 2-dimensional models can be conceived as these two separate dimensions ‘glued together’ with the latent correlation between factors, these item fit results provide some support for the fit of the 2-dimensional between-item IRT models.

Third, 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 the 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 consecutively, resulting in very similar values as in the present approach of .83, .85, and .87 for grade 1, 2, and 3, respectively (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 with .81, .86, and .88, respectively, so they seem robust against the estimation procedure used.

Fourth, the issue of the metric of the latent dimensions needs attention. Because item discrimination parameters α_{ik} were allowed to vary across the items, the latent metrics are not a priori equivalent. This equivalence, however, is necessary in order to draw conclusions on differential effects of predictors such as home language. Therefore, in estimating the measurement IRT models, the two latent dimensions were restricted to have the same mean and variance, so that they were standardized in the same way. Furthermore, in estimating the explanatory IRT models (i.e., the multivariate latent regressions), the residual variances of the two dimensions were also restricted to be equal, to retain the equivalence of the metrics. Alternatives to this form of standardization would be either to carry out analyses on estimates of the separate latent abilities (as Hartig and Höhler, 2008 did), or use multidimensional Rasch models in which all item discriminations are equal and the scales therefore have the same logit unit. We tried both alternatives. First, when sample-standardized estimates of the two ability dimensions were used as dependent variables, the same pattern of significant and non-significant differential effects of home language and reading comprehension level emerged (see Hickendorff & Janssen, 2009). Second, multidimensional Rasch models also showed the same pattern of differential effects. So, different modeling approaches resulted in the robust finding that home language and reading comprehension level had larger effects on applied problem solving than on computational skills.

Finally, the relation of the currently employed multidimensional IRT framework to the DIF approach is worth mentioning. In order to find out on which contextual problems student language variables had the largest effects, DIF analyses with the numerical problems as anchor items were carried out. In such an approach it is assumed that all items are on one latent scale (i.e., a unidimensional model). However, the substantial number of contextual problems with significant DIF found is a sign of multidimensionality (see Embretson & Reise, 2000, p. 262). That is, if other dimensions than the single common 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. Usually in DIF analyses such secondary dimensions are considered as nuisance, and DIF items are 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. That was not the aim of the current DIF analyses. Instead, the amount of DIF was deemed substantively interesting problem feature that was analyzed further. In general, we argue that MIRT modeling is broader 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.

Limitations and Recommendations for Further Research

We discuss several limitations in the design of the current study concerning the problems used and the samples investigated, that would require further research. Regarding the mathematics problems used, 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. In addition, the contextual problems included a broader range of operations than the numerical problems (as Table 2 shows) and probably also were of greater cognitive complexity given the fact that some problems involved

combining operations. This also hampers the comparability of problems with and without a context. Future research in which problems are matched for number characteristics and operations, making direct comparisons possible, is therefore called for.

A second issue concerns the contexts used. In particular, to obtain a broad coverage of applied problem solving reflecting educational practices, 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. 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, we tried to measure the linguistic complexity of the problem text by a text readability score, and found some indications that the less readable the problem text, the larger the extra difficulty for language minorities and below-median reading comprehension-level students, in line with findings of Abedi and Lord (2001), Abedi and Hejri (2004), Prenger (2005), and Van den Boer (2003). Although in general the found relation was weak and non-significant, one text variable, the percentage of frequently used words, was significantly related to the extra difficulty for below-median reading comprehension level students, which is an interesting starting point for further research. However, the linguistic complexity as measured, as well as the reading comprehension test, concern everyday language and do not cover academic language. As Bailey and Butler (2003) argued, students with sufficient skills in everyday language do not necessarily have sufficient skills in academic language, and a separate measurement instrument would be needed. That is a complicated matter, however, because of the interwoven nature of language and content-area material (Bailey et al., 2007). But it is clear that the current readability analyses only capture some—and probably not even the most important—aspects of the problem text that hampers L2 students' comprehension of the contextual problems. These results should therefore be considered as a starting point for future research.

Furthermore, 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 pointed 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.

Moreover, for students from diverse backgrounds, cultural aspects in the contexts play a role as well. For instance, the content may be culturally dependent or students may differ in their real-life experiences (Penfield & Lee, 2010). This concerns the cultural validity of assessments. However, as Solano-Flores and Trumbull (2003) argued, it is impossible to construct culture-free tests because of the complex interplay between students' proficiency in their first and second language, their content knowledge, and the linguistic and content demands of the problem. An important advantage of the current study was that there was a language-free part of the test (the computational problems) on which issues of cultural content and linguistic demands could not affect student performance, which could serve as a base for comparisons between student language groups.

With respect to the sample of students, three issues deserve further attention. The first issue is the representativeness of the sample for the population of primary school students. Recall that the schools that were selected had a relatively high proportion of ethnic minority students (in order to have sufficient statistical power to study language level effects) and hence the current sample is not entirely representative for the population of Dutch primary schools and students. Generalizations to this population should therefore be made cautiously. However, the schools in the sample were spread over the entire country and had different percentages of ethnic minority students, so there is substantial variation in these school factors. Furthermore, we could compute population norm scores based on students' performance on the test booklets of CITO's *Monitoring and Evaluation System for Primary School Students—Arithmetic and Mathematics*. Students of all levels were present, although there was a small overrepresentation of the top-quartile students and an underrepresentation of the bottom-quartile students.

A second issue concerns the home language groups distinguished, in particular the category of L2 students. These students were treated as one group, while arguably they are quite heterogeneous in their strengths and weaknesses in Dutch, an issue also receiving attention in research on English language learners in the United States (e.g., Solano-Flores and Trumbull 2003). This heterogeneity not only shows from the fact that there were many different non-Dutch home languages, but also from the large variation in the reading comprehension level of students with non-Dutch home language that was present in the data. Therefore, it is clearly restrictive to analyze these students as one homogeneous group. However, sample size limitations prohibited analyzing specific language groups. Further research, with larger samples, might address specific language groups.

A final limitation is 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.

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 perform relatively less well on a test that focuses on context format problems compared to a test on computational skills.

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). Specifically, such a profile of subscores 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. In addition, because assessments signal what is valued and expected in teaching (Greer, 1997), such a change in educational assessment towards separate or embedded testing of computational skills might also convey renewed attention to these basic mathematical skills in mathematics education, which we believe to be important.

However, Sinharay, Puhon, and Haberman (2010) showed that caution with reporting subscale scores is needed. Subscores 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 where 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).

Although I argue that reporting separate subscores for computational skills and applied problem-solving performance would give a more complete and thereby more balanced picture of students' mathematical competencies, I believe that it may be even more important to address students' academic language skills directly in educational practices. Instead of expecting students to acquire the academic language of different content areas vicariously, teachers should be encouraged to explicitly support its development in their teaching (e.g., Anstrom et al., 2010; Scarcella, 2003). In particular, the potential (hidden) language problems of ethnic minority students affecting their mathematics problem solving should receive attention in language instruction as well as in mathematics education.

ACKNOWLEDGMENTS

The research was supported by CITO, National Institute for Educational Measurement in the Netherlands. I am indebted to Jan Janssen from CITO for collecting the data, Ronald Krom from CITO for his advice on scoring the linguistic complexity of the items, Rinke Klein Entink for programming the MCMC-algorithm in R, and Norman Verhelst, Kees van Putten, Willem Heiser, and Claire Stevenson for their helpful suggestions.

REFERENCES

- Abedi, J. & Hejri, F. (2004). Accommodations for students with limited English proficiency in the National Assessment of Educational Progress. *Applied Measurement in Education*, 17, 371–392.
- Abedi, J. & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, 14, 219–234.
- Abedi, J., Lord, C., Hofstetter, C. & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practices*, 19(3), 16–26.
- Adams, R. J., Wilson, M. & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1–23.
- Adams, R. J. & Wu, M. L. (2000). *PISA 2000 technical report*. Paris: OECD.
- Albert, J. H. (1992). A Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, 17, 251–269.
- Anstrom, K., DiCerbo, P., Butler, F., Katz, A., Millet, J. & Rivera, C. (2010). *A review of the literature on Academic English: Implications for K–12 English Language Learners*. Arlington, VA: The George Washington University Center for Equity and Excellence in Education.

- Bailey, A. L. & Butler, F. A. (2003). *An evidentiary framework for operationalizing academic language for broad application to K–12 education: A design document* (CSE Tech. Rep. No. 611). Los Angeles, CA: University of California, CRESST.
- Bailey, A. L., Butler, F. A. & Sato, E. (2007). Standards-to-standards linkage under Title-III: Exploring common language demands in ELD and science standards. *Applied Measurement in Education*, 20, 53–78.
- Béguin, A. A. & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, 66, 541–562.
- Berends, I. E. & Van Lieshout, E. C. D. M. (2009). The effect of illustrations in arithmetic problem-solving: Effects of increased cognitive load. *Learning and Instruction*, 19, 345–353.
- CITO. (2005a). *Leerling- en onderwijsvolgsysteem, rekenen-wiskunde groep 3* [Monitoring and evaluation system for primary pupils—Arithmetic and Mathematics, grade 1]. Arnhem, The Netherlands: Author.
- CITO. (2005b). *Leerling- en onderwijsvolgsysteem, rekenen-wiskunde groep 4* [Monitoring and evaluation system for primary pupils—Arithmetic and Mathematics, grade 2]. Arnhem, The Netherlands: Author.
- CITO. (2006). *Leerling- en onderwijsvolgsysteem, rekenen-wiskunde groep 5* [Monitoring and evaluation system for primary pupils—Arithmetic and Mathematics, grade 3]. Arnhem, The Netherlands: Author.
- Cumming, G. & Finch, S. (2005). Inference by eye. confidence intervals and how to read pictures of data. *American Psychologist*, 60, 170–180.
- Cummins, D. D., Kintsch, W., Reusser, K. & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, 20, 405–438.
- De Boeck, P. & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.
- De Corte, E., Verschaffel, L. & De Win, L. (1985). Influence of rewording verbal problems on children's problem representations and solutions. *Journal of Educational Psychology*, 77, 460–470.
- De la Torre, J. & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, 30, 295–311.
- Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Fuchs, L. S., Fuchs, D., Stuebing, K., Fletcher, J. M., Hamlett, C. L. & Lambert, W. (2008). Problem solving and computational skill: Are they shared or distinct aspects of mathematical cognition? *Journal of Educational Psychology*, 100, 30–47.
- Greer, B. (1997). Modelling reality in mathematics classrooms: The case of word problems. *Learning and Instruction*, 7, 293–307.
- Hartig, J. & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie/Journal of Psychology*, 216, 89–101.
- Hickendorff, M. & Janssen, J. (2009). De invloed van contexten in rekenopgaven op de prestaties van basisschoolleerlingen [The effect of contexts in mathematics items on primary-school pupils' performance]. *Reken-wiskundeonderwijs: onderzoek, ontwikkeling, praktijk*, 28(4), 3–11.
- Janssen, J., Van der Schoot, F. & Hemker, B. (2005). *Balans van het reken-wiskundeonderwijs aan het einde van de basisschool 4* [Fourth assessment of mathematics education at the end of primary school]. Arnhem, The Netherlands: CITO.
- Kilpatrick, J., Swafford, J. & Findell, B. (2001). *Adding it up. Helping children learn mathematics*. Washington, DC: National Academy Press.
- Koedinger, K. R., Alibali, M. W. & Nathan, M. J. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science*, 32, 366–397.
- Kraemer, J.-M., Janssen, J., Van der Schoot, F. & Hemker, B. (2005). *Balans van het reken-wiskundeonderwijs halvewege de basisschool 4* [Fourth assessment of mathematics education halfway primary school]. Arnhem, The Netherlands: CITO.
- Leighton, J. P. & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and application*. New York, NY: Cambridge University Press.
- Lesaffre, E. & Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Journal of the Royal Statistical Society Series C—Applied Statistics*, 50, 325–335.
- Molenberghs, G. & Verbeke, G. (2004). An introduction to (generalized) (non)linear mixed models. P. De Boeck M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 111–153). New York, NY: Springer.

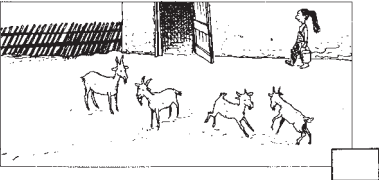
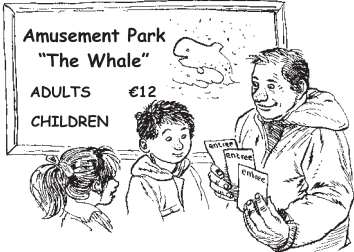

- Mullis, I. V. S., Martin, M. O. & Foy, P. (2008). *TIMSS 2007 international mathematics report. Findings from IEA's Trends in International Mathematics and Science Study at the fourth and eighth grades*. Boston, MA: Boston College, TIMSS & PIRLS International Study Center.
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: NCTM.
- National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.
- OECD. (2004). *Learning for tomorrow's world. First results from PISA 2003*. Paris: Author.
- OECD. (2009). *PISA 2006 technical report*. Paris: Author.
- Penfield, R. D., & Lee, O. (2010). Test-based accountability: Potential benefits and pitfalls of science assessment with student diversity. *Journal of Research in Science Teaching*, 47, 6–24.
- Prenger, J. (2005). *Taal telt! Een onderzoek naar de rol van taalvaardigheid en tekstbegrip in het realistische rekenonderwijs*. [Language counts! A study into the role of linguistic skill and text comprehension in realistic mathematics education]. PhD thesis, University of Groningen, The Netherlands.
- Raykov, T. (2012). Evaluation of latent construct correlations in the presence of missing data: A note on a latent variable modelling approach. *British Journal of Mathematical and Statistical Psychology*, 65, 19–31.
- R. Development Core Team. (2009). *R: A language and environment for statistical computing*. [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org>
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Rijmen, F., Tuerlinckx, F., De Boeck, P. & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, 8, 185–205.
- Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25.
- Royal Netherlands Academy of Arts and Sciences. (2009). *Rekenonderwijs op de basisschool. Analyse en sleutels tot verbetering* [Mathematics education in primary school. Analysis and recommendations for improvement]. Amsterdam, The Netherlands: KNAW.
- SAS Institute. (2002). *SAS online doc (version 9)*. Cary, NC: SAS Institute Inc.
- Scarella, R. (2003). *Accelerating Academic English: A focus on the English Learner*. Irvine, CA: University of California.
- Sheu, C.-F., Chen, C.-T., Su, Y.-H. & Wang, W.-C. (2005). Using SAS PROC NL MIXED to fit item response theory models. *Behavior Research Methods*, 37, 202–218.
- Sinharay, S., Puhan, G. & Haberman, S. J. (2010). Reporting diagnostic scores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, 45, 553–573.
- Solano-Flores, G. & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, 32(2), 3–13.
- Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid. de ontwikkeling van een domeingericht meetinstrument*. Arnhem, The Netherlands: CITO.
- Stone, C. A., Ye, F., Zhu, X. & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, 23, 63–86.
- Van den Boer, C. (2003). *Als je begrijpt wat ik bedoel. Een zoektocht naar verklaringen van achterblijvende prestaties van allochtone leerlingen in het wiskundeonderwijs* [If you get what I mean. A search for explanations of lagging achievement of non-native students in mathematics education]. Utrecht, The Netherlands: CD- β press.
- Van Schilt-Mol, T. M. M. L. (2007). *Differential item functioning en itembias in de Cito-Eindtoets Basisonderwijs* [Differential item functioning and item bias in CITO's End of Primary School Test]. PhD thesis, Tilburg University, The Netherlands.
- Verhelst, N. D., & Glas, C. A. W. (1995). The one-parameter logistic model. In G. H. Fischer & I. Molenaar (Eds.), *Rasch models. Foundations, recent developments and applications* (pp. 215–237). New York, NY: Springer.
- Verhelst, N. D., & Verstralen, H. H. F. M. (2002). *Structural analysis of a univariate latent variable (SAUL)* (computer program and manual). Arnhem, The Netherlands: CITO.
- Verschaffel, L., Greer, B. & De Corte, E. (2000). *Making sense of word problems*. Lisse, The Netherlands: Swets and Zeitlinger.
- Vicente, S., Orrantia, J. & Verschaffel, L. (2007). Influence of situational and conceptual rewording on word problem solving. *British Journal of Educational Psychology*, 77, 829–848.
- Von Davier, M. & Sinharay, S. (2009). *Stochastic approximation methods for latent regression item response models* (ETS-RR-09-09). Princeton, NJ: Educational Testing Service.

Wilson, M. & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp.43–74). New York, NY: Springer.

Wu, M. & Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal*, 18, 93–113.

Zou, G. J. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, 12, 399–413.

APPENDIX: SAMPLE PROBLEMS (PROBLEM TEXTS TRANSLATED FROM DUTCH)

context format	numerical expression format
<p>grade 1</p> <p><i>Student's worksheet</i></p>  <p><i>Teacher reads aloud:</i> "You see 4 goats in the paddock. Inside, 11 goats are having a rest. How many goats live on this children's farm?"</p>	<p>grade 1</p> <p>$5 + 12 = \underline{\hspace{1cm}}$</p> <p>$17 - 5 = \underline{\hspace{1cm}}$</p> <p>$18 - \underline{\hspace{1cm}} = 10$</p>
<p>grade 2</p>  <p>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?</p> <p><u> </u> euros</p>	<p>grade 2</p> <p>$26 + 25 + 27 = \underline{\hspace{1cm}}$</p> <p>$2 \times 18 = \underline{\hspace{1cm}}$</p> <p>$58 = 98 - \underline{\hspace{1cm}}$</p>
<p>grade 3</p>  <p>One tray contains 4 plants. Joyce buys 12 of these trays. How many plants does that make?</p> <p><u> </u> plants</p>	<p>grade 3</p> <p>$263 + 19 = \underline{\hspace{1cm}}$</p> <p>$487 - \underline{\hspace{1cm}} = 427$</p> <p>$9 \times 30 = \underline{\hspace{1cm}}$</p> <p>$36 : 4 = \underline{\hspace{1cm}}$</p>