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

Hickendorff, M.

CHAPTER 2

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

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

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

ABSTRACT

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

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

2.1 INTRODUCTION

2.1.1 National assessments of mathematics achievement

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

2.1.2 Mathematics education

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

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

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

2.1.3 Complex division

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

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

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

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

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

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

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

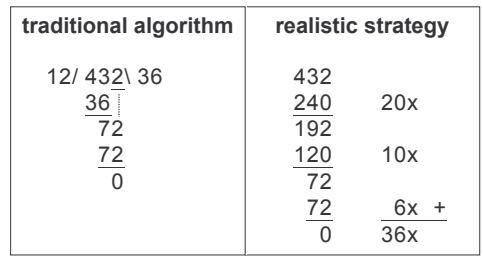

FIGURE 2.1 *Examples of the traditional long division algorithm and a realistic strategy of schematized repeated subtraction for the problem 432 ÷ 12.*

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

2.1.4 Goals of present study

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

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

2.2 METHOD

2.2.1 Sample

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

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

2.2.2 Design of the tests

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

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

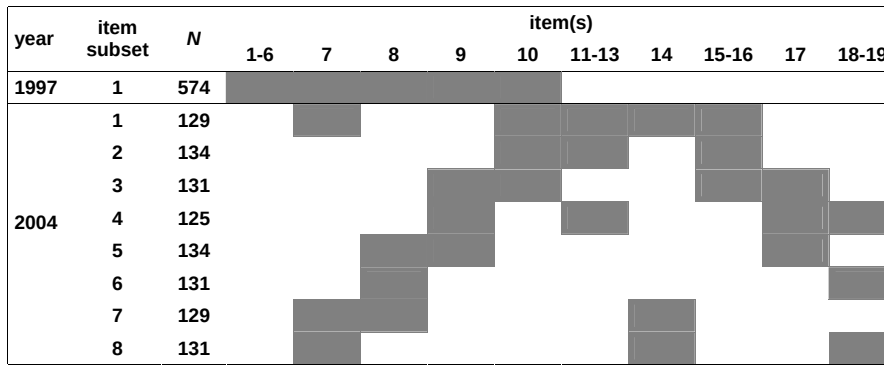

FIGURE 2.2 *Design of the assessments.*

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

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

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

TABLE 2.1 *Specifications of the items.*

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

*Answer that was scored correct, given the item context

2.2.3 Responses

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

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

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

2.2.4 Properties of the data set

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

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

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

TABLE 2.2 *Part of the data set.*

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

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

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

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

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

2.2.5 Latent Class Analysis

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

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

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

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

Predictor effects

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

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

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

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

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

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

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

Software

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

2.2.6 Explanatory IRT

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

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

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

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

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

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

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

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

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

Fitting the models

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

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

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

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

Software

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

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

TABLE 2.4 *Strategy use in proportions.*

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

2.3 RESULTS

2.3.1 Research Question 1

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

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

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

TABLE 2.5 *Latent class models.*

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

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

Effects of predictors on class membership

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

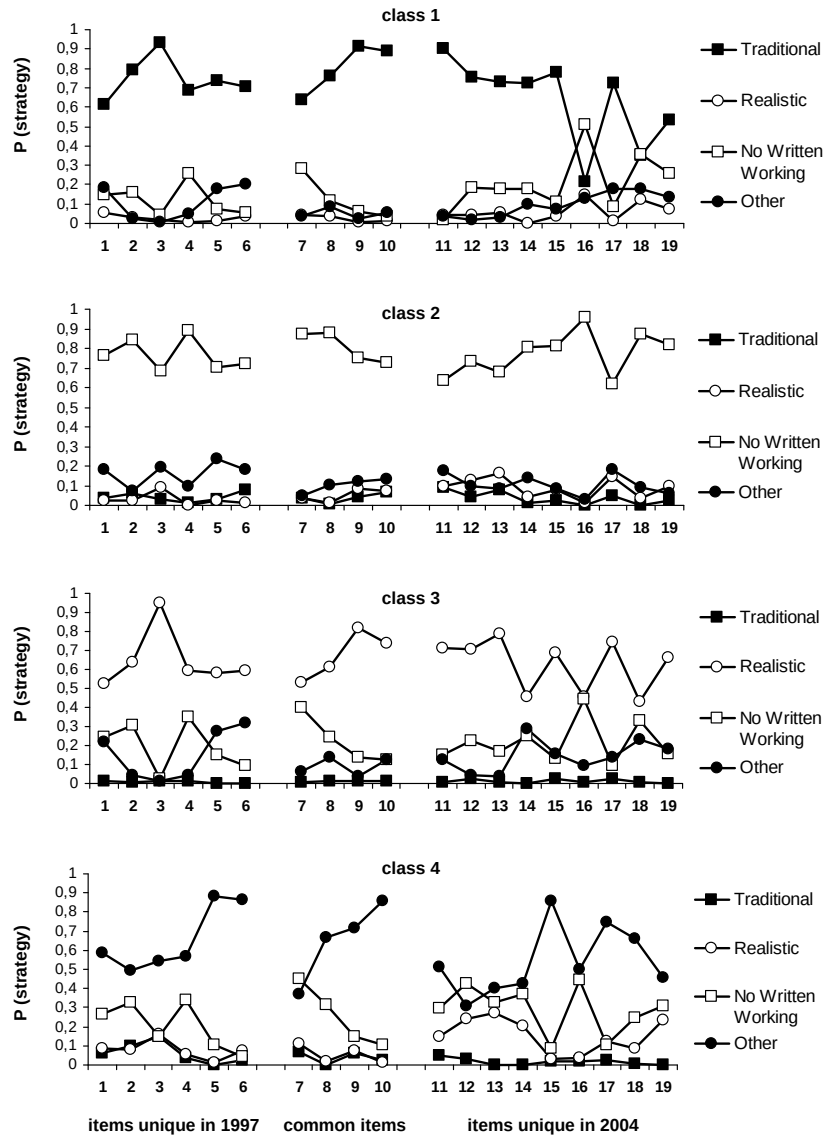

FIGURE 2.3 *Conditional probabilities of the 4-class LC-model.*

TABLE 2.6 *Class sizes in 1997 and 2004.*

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

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

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

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

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

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

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

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

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

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

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

TABLE 2.8 *Explanatory IRT models.*

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

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

2.3.2 Research Question 2

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

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

Strategy effects

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

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

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

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

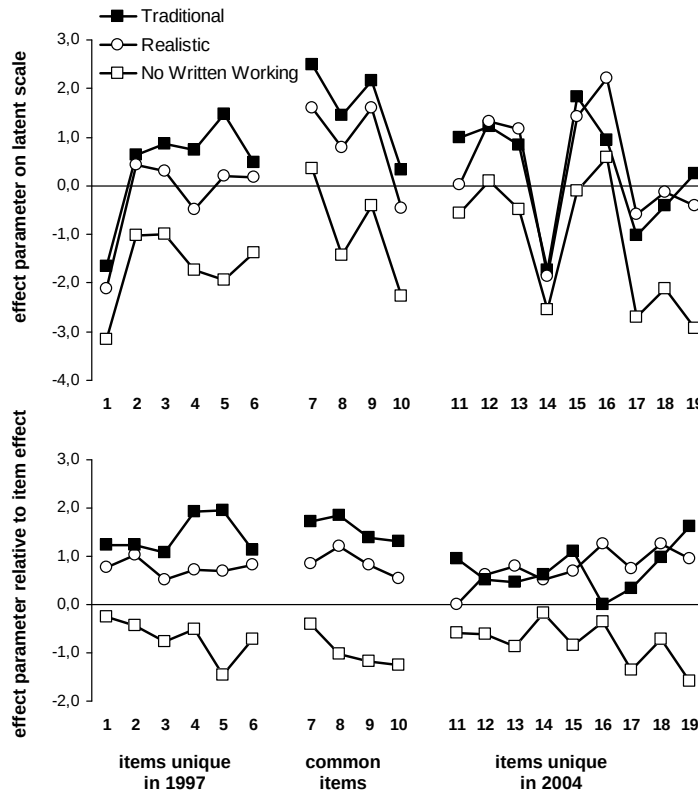

FIGURE 2.4 *Item-specific effect parameters of each strategy, from model M2.*

added.

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

Background variables

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

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

Interpretation of the selected model

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

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

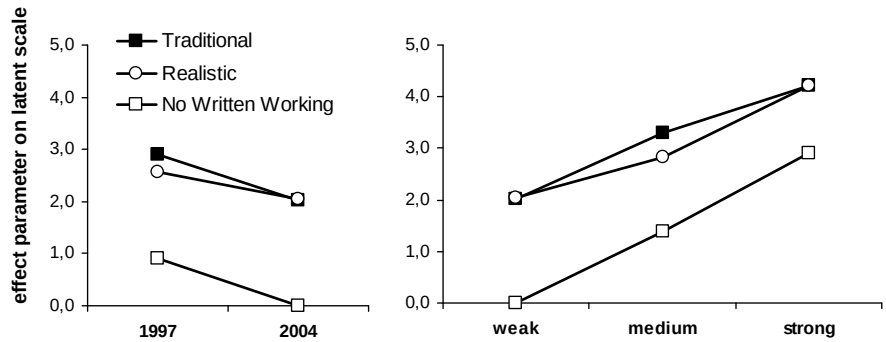

FIGURE 2.5 *Interaction effects of strategy use with year of assessment (left panel) and with general mathematics level (right panel) from model M3b.*

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

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

Conclusions Research Question 2

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

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

2.4 Discussion

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

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

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

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

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

2.4.1 Limitations

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

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

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

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

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

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

2.4.2 Methodological considerations

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

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

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

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

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

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

2.4.3 Educational implications

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

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

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

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