Educational Research and Evaluation
2002, Vol. 8, No. 2, pp. 197–225

1380-3611/02/0802–197$16.00
© Swets & Zeitlinger

National Assessment of New Standards for Mathematics in Elementary Education in Flanders

Rianne Janssen, Erik De Corte, Lieven Verschaffel, Evelyn Knoors, and Ariane Colémont
Center for Instructional Psychology and Technology and Leuven Institute for Educational Research, University of Leuven, Belgium

ABSTRACT

A baseline assessment of the new curriculum standards of mathematics for elementary education in Flanders is presented and discussed. A total of 24 measurement scales, each representing a cluster of standards, covered the whole mathematics curriculum. A total of 5,763 6th-grade students participated in the study. Performance on each scale was summarized by estimating the proportion of students who were nonmasters, intermediate, and good masters. Overall, the more innovative curriculum standards were mastered worse. Few gender differences in performance were observed.

INTRODUCTION

Since September 1998, new standards for elementary education have become operational in the Flemish part of Belgium (Ministerie van de Vlaamse Gemeenschap, 1997). These standards cover all domains of the curriculum, including mathematics, and state the basic competencies that children should master at the end of primary school (which consists of 6 grades starting at the age of 6 years). For the children from special primary education, the standards have the status of educational goals with no requirement with respect to mastery. It is the first time in the history of education in Flanders that the government has set educational standards. In the past, the three different

Address correspondence to: Rianne Janssen, Leuven Institute for Educational Research, Dekenstraat 2, B-3000 Leuven, Belgium. E-mail: rianne.janssen@ped.kuleuven.ac.be

Manuscript submitted: November 1, 2000.
Accepted for publication: June 14, 2001.

educational sectors (covering community schools, private subsidized schools, and municipal schools, respectively) each developed their own curricula. Since the official introduction of the standards in 1998, these different educational sectors have changed their curricula to bring them in line with the official standards for primary education in Flanders. The development of these standards fits into the international educational reform movement, and is also in line with the growing demand for educational accountability, that is the question of citizens and policymakers whether the substantial investments in education are reaping rewards (National Research Council, 1999).

With respect to mathematics education, the new standards represent a substantial innovation. The most striking novelty is the removal of the concepts, symbols, and representational tools reminiscent of the ''New Math.'' By ''New Math'' we refer to a worldwide reform movement of mathematics education, which started in the 1950s, flourished in the 1960s and 1970s, and continued to have its impact on mathematics teaching in many countries until the 1980s and early 1990s. The prominent components of the international ''New Math'' agenda were the central and foundational role played by set theory, the premature introduction and utilization of formal symbols and other representational tools, the focus on conceptualizing mathematics rather than applying it, and the direction toward the abstract (Verschaffel, 1995). Until the introduction of the standards in 1998, this ''New Math'' approach towards mathematics education continued to dominate the curricula, the textbooks, and the assessment instruments for mathematics education of the different educational sectors in Flanders, although the dissatisfaction with this approach had gradually become more intense and widespread among Flemish mathematics educators. This dissatisfaction partially stemmed from a growing acknowledgment of the problems that (can) occur when pupils and teachers are confronted with a mathematics curriculum that is too heavily based on a disciplinary analysis of the domain and that is too strongly directed toward the abstract, without seriously taking into account the characteristics and the needs of the subjects for whom it is meant and the sociocultural context wherein it is to be implemented (Verschaffel, 1995).

Another factor was the rise of a new approach toward mathematics education in the international scene – sometimes referred to as the realistic, the constructivistic, or the authentic approach – as reflected in curriculum reform documents published over the past years in many countries, such as *The Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics, 1989) in the United States, *Mathematics*

Counts (Cockcroft, 1982) in the United Kingdom, and the Dutch *Proeve van een Nationaal Programma voor het Reken/Wiskundeonderwijs op de Basisschool* (Treffers, De Moor, & Feijs, 1989). According to this new approach, learning mathematics extends far beyond the acquisition of a variety of domain-specific concepts, rules and procedures. It also includes developing a disposition towards mathematics, including the acquisition of higher-order skills like the ability to explore, to conjecture, to reason, to reflect and to communicate mathematically, as well as the sensitivity and the inclination to use these cognitive and metacognitive skills effectively for solving nonroutine problems in mathematics and in daily life. In addition, a mathematical disposition involves developing a positive attitude toward mathematics, seeing mathematics as a powerful way of looking at situations, and becoming confident in one's own mathematical ability. Taking into account this new conception of what constitutes a mathematical disposition, learning mathematics is conceived as an active, constructive, authentic and interactive process in these reform documents. This view on mathematics learning and teaching is accompanied by a reorientation of the content of the elementary school mathematics curriculum, the most prominent characteristic of which is a de-emphasis of mastery of the algorithms for written computation, of the rules for adding, subtracting, multiplying and dividing fractions, of the formula for measuring perimeters, areas and volumes of all kinds of geometrical objects, and so forth, paralleled with an increased attention for aspects such as mental calculation, estimation and number sense, looking for patterns and connections, problem-solving, and so forth. Generally speaking, the Flemish standards are very well in accordance with the new view on the objectives, the content and the teaching/learning processes of mathematics education as expressed in the above-mentioned reform documents, although they go somewhat less far in emphasizing the dispositional view of mathematics and in de-emphasizing some elements of its traditional content. One of the reasons why the developers of the Flemish standards were somewhat reluctant to propose more drastic changes in the mathematics curriculum, were the repeatedly (very) good scores of Flemish pupils on international mathematics tests such as TIMSS (Beaton et al., 1996) and TIMSS Repeat (Mullis et al., 2000).

Implementation of the new standards requires, besides a modified curriculum and appropriate instructional materials and methods, reliable and valid instruments also to assess the degree of attainment of those educational objectives by pupils at the end of elementary school. The present study, which was commissioned by the Department of Education of the Flemish Govern-

ment, aimed at the development of an instrument for the national assessment of pupils' performance on the mathematics standards. Although the main purpose of the study consisted of test construction and test calibration, the scales that were developed were also used as a first, large-scale assessment of the performance of children on the curriculum standards at the end of primary education. Given that the present study took place during the initial stage of the implementation of the new standards, it provides a baseline for future national assessments of mathematics. Moreover, it is also interesting to investigate whether pupils at the end of primary education already master parts of the modified mathematics curriculum at the time of study. Finally, because the assessment carried out covers the whole domain of mathematics, it also offers an excellent opportunity to look for differences in performance between boys and girls within subdomains of mathematics.

This article reports the major findings of the baseline assessment of the curriculum standards of mathematics in Flanders. First, the general framework of the study and the research questions are discussed. This is followed by a description of the method of investigation. The results of the study are presented in the next section. An overall discussion concludes the article.

GENERAL FRAMEWORK AND RESEARCH QUESTIONS

An overview of the curriculum standards of mathematics, together with the design of the measurement scales, opens this section. Afterwards, our approach to set performance standards on the measurement scales is explained, and a short review of the literature of sex differences in primary education within mathematics is given. Finally, the main research questions are presented.

The Curriculum Standards of Mathematics and the Design of the Scales

The curriculum standards of mathematics are officially classified into five categories, which are presented in the first column of Table 1. The curriculum standards of the first three categories refer to different mathematical domains, namely Numbers, Measurement, and Geometry. These domains are further subdivided as to whether declarative or procedural knowledge is involved. For the other two categories, namely Strategies and Problem-solving skills and Attitudes, no further subdivision is made; it is important to mention that the standards in these two categories, so to say, cut across the preceding three

Table 1. Number of Curriculum Standards Within Each Category and Examples of Curriculum Standards.

Category	Declarative	Procedural	Total	Examples: The students . . .
Numbers	11	17	28	*Declarative* . . . can immediately give correct results for sums and substractions up till 10, and for the tables of multiplication and the corresponding division tables till 10. *Procedural* . . . can calculate simple percentages embedded in practical situations.
Measurement	5	7	12	*Declarative* . . . know that 0°C corresponds to the freezing point and that temperatures below this freezing point are indicated with negative numbers. *Procedural* . . . can make meaningful conversions between common measures.
Geometry	3	4	7	*Declarative* . . . can read and apply the symbols for perpendicularity and parallelism. *Procedural* . . . can classify angles and quadrangles (by using the characteristics of sides and angles) and they can draw these geometrical figures themselves.
Strategies and Problem-solving	–	–	4	. . . are able to use the learned concepts, insights, and procedures concerning numbers, measurement, and geometry, as mentioned in the corresponding curriculum standards, in an efficient way in meaningful applications, both inside and outside the classroom.
Attitudes	–	–	4	. . . appreciate mathematics as a dimension of human creativity.

categories, in the sense that children should become able to apply the strategies, skills and attitudes to each of those content domains (Number, Measurement, and Geometry). Table 1 indicates that the total number of curriculum standards in the five categories differ, with the majority of the curriculum standards referring to Numbers. Examples of curriculum standards are given in each row of Table 1. These examples illustrate that the curriculum standards are formulated at different levels of difficulty and generality.

For the compilation of measurement scales for assessing mastery of the curriculum standards, the standards were split in two parts: Numbers and Measurement and Geometry. As will be discussed later, a separate test administration was organized for each part. Within each part, measurement scales were formed by grouping curriculum standards with a common mathematical content. A scale on problem-solving was compiled for each part. A summary of the content of the measurement scales is given in Table 2. The measurement scales contained between 1 and 5 curriculum standards. The further subdivision of Scales 2 and 21 are explained in the Results section.

Mastery Assessment

Mastering an educational standard is commonly considered as an incremental process consisting of a progression along a continuum. This *continuum view* of mastery (Meskauskas, 1976) implies that students can be ordered along a continuously distributed ability dimension. Proficiency is assumed to be cumulative: Students at the high end of the scale generally know and can do all that students at the low end of the scale know and can do, but they also know and can do more than students at lower levels. In line with the continuum view of mastery, it can be expected that the items corresponding to learning tasks of a certain domain can also be ordered along the same continuum as a function of their difficulty. The difficulty of an item can be defined as the amount of proficiency that is necessary to solve the task correctly.

In education, categorical decisions are often made with respect to proficiency, like in pass/fail or mastery/nonmastery classifications. According to the continuum view of mastery, these classifications correspond to different regions of the continuum that are bounded by cutoff points. For example, a cutoff point on the continuum may indicate the point of sufficient proficiency to be classified as a master. The continuum view implies that such classifications are, in principle, arbitrary to a certain extent, as the respondents just below the cutoff score will not differ much in ability from respondents just above the cutoff score. In practice, however, the respondents about whom

Table 2. Overview of the Measurement Scales.

Scale	Content	*CS*[a]	Booklet	*N*	Items	*IR*[a]	Bayesian *p*
Numbers							
1	Mental arithmetic	4	1	683	28	–	.37
2a	Use and display of natural numbers	2	3	579	10	–	.47
2b	Roman numerals	1	3	579	10	1	.20
3	Mathematical notations, operations, and their mutual relations	3	3	579	33	–	.33
4	Fractions and decimal numbers	3	4	585	18	1	.37
5	Natural numbers, fractions, decimal numbers, percentages, and their equivalence	2	5	564	16	–	.45
6	Multiples, divisors, and features of divisibility	3	6	587	21	–	.45
7	Estimation and approximation	3	4	585	19	2	.35
8	Ratios	1	6	587	20	–	.36
9	Written computation with natural and decimal numbers	1	5	564	26	–	.45
10	Percentages	1	2	573	16	–	.52
11	Use of pocket calculator	2	1	683	25	–	.37
12	Problem-solving	5	2	573	22	–	.46
Measurement & Geometry							
13	Problem-solving	4	7	548	24	–	.49
14	Concepts, symbols, notations and conventions about measures	3	8	549	23	–	.51
15	Meaningful conversions of sizes and measures	2	9	554	16	2	.38
16	Choice and use of appropriate reference points	2	10	545	22	–	.46
17	Scale and Average	1	9	554	6	1	.50
18	Perimeter, area, and volume	2	7	548	12	–	.46
19	Clock reading and the monetary system	2	9	554	21	–	.47
20	Dimensions of space and spatial orientation	2	8	549	45	–	.48
21a	Concepts, symbols, notations and conventions about geometry	4	10	545	28	–	.42
21b	Parallelism	1	10	545	7	–	.32
21c	Perpendicularity	1	10	545	5	2	.42

[a] *CS* = Number of curriculum standards, *IR* = Number of items removed.

high-stake decisions have to be made may all be situated far from both sides of the cutoff point.

Given the continuum view of mastery, two approaches are possible to define cutoff points. In an *item-centered* approach, the location of the cutoff is based on judgments about the items, whereas in an *examinee-centered* approach, the cutoff is based on the performance of persons classified in different categories of mastery. Numerous procedures for making mastery classifications have been proposed within each approach. Reviews and discussions can be found in Berk (1986, 1996), Jaeger (1989), and Kane (1994), among others. In this study, we used the Model for Criterion-Referenced Measurement (MCRM) of Janssen, Tuerlinckx, Meulders, and De Boeck (2000). The model encompasses a hierarchical extension of standard models from item response theory (IRT). It is based on two assumptions, namely (1) a probabilistic formalization of the continuum view of mastery, and (2) an item-centered approach to set cutoff points. These two assumptions are now explained consecutively.

In IRT, persons and items are ordered along the same (latent) continuum. This latent continuum is thus defined in two ways: as the dimension on which the persons are located in increasing order of ability, and as the dimension on which the items are located in increasing order of difficulty. The position of a person j is usually denoted by the real-valued parameter θ_j, while the position of an item i is denoted by the real-valued parameter β_i. The position of a person with respect to an item is a measure of his or her proficiency on that item. More specifically, in the basic, one-parameter IRT model (commonly known as the Rasch model), the difference between θ_j and β_i defines the probability of success of person j on item i. This is written formally as

$$P\left(X_{ij} = 1\right) = f\left(\theta_j - \beta_i\right). \tag{1}$$

The function f is an S-shaped function that transforms any value on the real line into a value between 0 and 1. When θ_j equals β_i, the probability of success is .50. When θ_j exceeds β_i, the probability of success is higher than .50; the higher θ_j is in comparison with β_i, the higher the probability of success. When β_i exceeds θ_j, the probability of success is lower than .50; the higher β_i is in comparison with θ_j, the lower the probability of success. Hence, the probability of success on a given item can be interpreted as a measure of proficiency on that item. Individual differences in θ_j can be considered as individual differences in proficiency with respect to the items of the test. When the model in (1) fits a given data set, this implies that individual differences in perfor-

mance on the items can be explained by referring to a single dimension. Differences in β_i among the items refer to differences in difficulty, and, hence, to differences in the amount of θ_j necessary to solve the item correctly with high probability. Fit of the model for a given dataset implies that the items are consistently ordered in difficulty for all persons. The ordering of the items along the latent continuum according to their value of β_i corresponds to the path of continuous progression going from low proficiency to high proficiency.

Hence, standard IRT models, like the one-parameter model presented in (1), can be seen as a probabilistic formalization of the continuum view of mastery. The probability of success on a given *item* is thereby considered as an indicator of the measure of proficiency on that item, as indicated above. In order to get a measure of proficiency on the overall *test*, the MCRM calculates the probability of success on a reference item. The reference item is a fictitious item with a difficulty parameter equal to the average β_i of the items of the test. For mastery classification of test performance, cutoff points on the probability scale for the reference item are used. For example, in the present application we used the cutoff points .50 and .80. Passing the .50 probability indicates that the person has a higher probability to solve the item correctly than to solve the item incorrectly. This can be considered as an initial stage of mastery of the item. Passing the .80 probability indicates that the person's probability of a correct answer is so high that giving an incorrect response is more likely to be interpreted as dependent on all kinds of error fluctuations rather than as a lack of proficiency. On the basis of these cutoffs, each respondent can be classified into one of three categories: nonmastery ($p<.50$), intermediate mastery ($.50 \leq p \leq .80$), and mastery ($p>.80$). Hence, the MCRM follows an item-centered approach to set cutoff points. Note that the .50 boundary between nonmastery and intermediate mastery does not necessarily imply that 50% of the respondents are classified as a nonmaster. A basic property of IRT models is that the estimation of the item difficulty parameters (and, hence, of the position of the reference item) is independent of the ability of the persons in the calibration sample. Hence, the percentage of respondents being classified in the nonmastery category can, in principle, vary between 0 and 100, depending on the location of the respondents on the latent continuum with respect to the reference item. Note also that persons situated at the boundaries of adjacent categories, will not differ much in ability. This is an unavoidable implication of working with a continuum view, as was indicated before.

IRT models are commonly used for national assessment programs, like in the American National Assessment of Educational Progress (e.g., Mislevy,

Johnson, & Muraki, 1992), the Australian Basic Skills Testing program (Masters et al., 1990), or the Dutch National Assessment Program (Verhelst & Eggen, 1991). Apart from their nice conceptual interpretation due to their property of positioning persons and items on the same continuum, IRT models also have several technical advantages, like for scale equating or detecting item bias. The MCRM adds to these advantages an economic way to summarize overall test performance. More technical remarks about the model are given in the Method section.

Performance Differences Between Boys and Girls Within Mathematics

Gender differences in cognitive performance have been studied in great detail mainly in three domains, namely spatial, verbal, and mathematical abilities. For mathematics, the meta-analysis of Hyde, Fennema, and Lamon (1990) continues to be a benchmark study. Regarding children at the end of primary education, their study indicated that girls scored somewhat higher on computation, while there were little or no gender differences in the general knowledge of mathematics or in problem-solving. In the latter domain, clearly better performance by male students emerged from high school onwards. On the basis of an overview of test performance from several national and international surveys, Willingham, Cole, Lewis, and Wilson Leung (1997) came to comparable conclusions. Both studies also found that there is an effect of sample restriction. Gender differences were smallest in samples of the general population. The more selective the group of students, the larger the gender differences. Hence, on the basis of both review studies, and given that we work with a representative sample of sixth-grade students, we can only expect small significant differences for scales involving computation.

However, recent studies by Van den Heuvel-Panhuizen and Vermeer (1999) lead to different predictions. They analyzed performance differences between boys and girls at the end of primary education on the Dutch National Assessment Program. Almost all subscales of the assessments of 1987 and 1992 showed significant better performance by boys. These findings, together with a detailed content analysis on the items for which a better performance was found for one of the sexes, led the authors to the conclusion that boys performed better on items that make a strong appeal to knowledge based on experience with numbers or with measures commonly used in daily life and on items for which a clever strategy is possible. Girls performed better on items involving precise calculations or known standard procedures. The latter

conclusion is in line with the two previous studies, but the conclusion with respect to boys indicates that the better performance by boys on tasks involving some kind of problem-solving may already be found at the end of primary education, and not only from high school onwards.

Research Questions

Taking into account the general framework outlined in the preceding subsections, three main questions were addressed in the present study:

1. Is performance on the different measurement scales in line with a continuum view of mastery, which is evidenced by the fit of a standard IRT model?
2. Are there differences in mastery attainment between the distinct curriculum standards at the initial stage of implementation of the new curriculum? This will be investigated by applying the MCRM.
3. Do boys and girls perform differently on distinct subparts of the curriculum of mathematics?

METHOD

Items

A prerequisite for any large-scale assessment program is the validity of the measurement scales. In the past, many mathematics tests have been criticized with respect to their construct validity. Dominated by the multiple-choice item format, their emphasis was mostly on lower-order procedural skills, and almost exclusively on the product without validating whether this product was a good reflection of the student's underlying problem-solving process. In addition, many test items presented had a decontextualized character (for a critical discussion of traditional tests, see Frederiksen, 1990; Pandey, 1990; Silver & Kenney, 1995). Apart from good construct validity, the items of an assessment instrument should also show good content validity. The summary of test performance given by the MCRM can only be considered generalizable from the sample of the test items to the general domain defined by a particular curriculum standard if the test items are a representative sample from that universe of items.

The item selection for the present study was based on an item pool of 746 items that was generated in a preceding study (De Corte, Verschaffel, Knoors, & Detavernier, 1997). All these items were considered as valid operational-

izations for the corresponding curriculum standard according to experts: Only items with an average score of at least four on a 5-point validity scale were included in the item pool. All items in the item pool were also rated on difficulty. More specifically, experts were asked to rate on a 10-point scale (ranging from 0–10% to 91–100%) the percentage of students who should be able to answer the item correctly if education was provided appropriately in accordance with the new curriculum standards. Finally, apart from the items measuring pure calculating competencies (like in Scales 1 and 9), all mathematical problems were presented with a meaningful context.

From the item pool of the previous project, 340 items were selected for the present study. The following criteria for item selection were used: (1) representativeness and variedness of the set of items as a whole for the curriculum standards; and (2) difficulty level of individual items. Only items with an average rating between 60 and 90% on the scale for expected difficulty were accepted. This implies that the more difficult items (as well as the extremely easy ones) according to the expert ratings were not included. This was done because the curriculum standards were officially conceptualized as minimum educational goals that the majority of children should be able to master. Hence, we opted not to include items that would be too difficult for the majority of children. When compiling the final scales, it turned out that the initial item pool was insufficient. Another set of 54 items was generated by the research team. Due to time constraints, no expert ratings on these items were gathered.

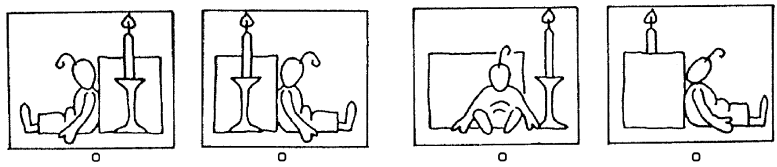
The fact that the items had to be applicable in a large-scale, paper-and-pencil assessment program, gave a certain constraint on possible item formats, such that items requiring an extended response were excluded. In the final item sample, 67% were items with a short-answer format. For 14% of the items, a short-answer format was used for several subquestions to the same problem. Only 11% of the items were multiple-choice: Almost one third of them related to Scale 13 where factual knowledge about measures and measurements was tested by presenting different measures or measurements for a given problem. Finally, 8% of the items, almost all belonging to Scales 7, 12, 13 and 16 where higher-order problem-solving skills were tested, consisted of two questions: one asking for the solution (product) of the problem and one asking for a motivation for this answer or for an explanation of how the answer was derived. The latter required a more extended response. We will refer to this question as the process question. Figure 1 gives an example of an item for each item format.

a. Short-answer format
Ann buys a coat of 4.500 BF for 3.600 BF.
With how many percent is the price reduced?
_____%

b. Short-answer format with several subquestions
Put the following numbers in the table:
250 3564 816 2845 1991 1702
Note: Some numbers may not fit into the table, or may fit in several columns.

divisible through 2	divisible through 3	divisible through 5	divisible through 9	divisible through 10

c. Multiple-choice format
Three of these pictures are made of the same situation.
One picture does not belong here.
Color the round below this picture.



d. Product and process question
Chantal wants to buy a pair of Tiger sneakers and saw these ads in the local paper.

Family Shoe Center

Bottom prices every day

Tiger sneakers
only 1200 BF

Van Dierens shoe shop

This week only

Sales: Tiger sneakers
1100 BF

The Family Shoe Center is within walking distance.
To go to Van Dierens shoe shop Chantal has to take the bus. That would cost 80 BF for a one-way ticket.
If Chantal wants to spend as little money as possible, at which shop should she buy her sneakers?
Answer: _____
Explain why.
Answer: _____

Fig. 1. Examples of an item for each item format.

Test Design

As already mentioned above, the administration of the items was split into two parts. The first part involved the measurement scales on Numbers, spread over six different booklets; the second part contained the scales on Measurement and Geometry, spread over four booklets. In each part, the items were administered according to an incomplete design without overlap among the test booklets. Practical constraints, especially with respect to the required sample size, forced us to use a design with no overlapping items across test booklets.[1] However, such anchor items could have been useful to calibrate item parameters across students and booklets. Each test booklet contained the items of two or three measurement scales representing diverse mathematical contents in order to get booklets that were relatively varied (Table 2). Each test booklet contained about 40 items and was split into two parts in order to allow for a break in the administration of the booklet. The items of the different measurement scales were put in random order within each test booklet.

Within each study, the assignment of booklets to students was based on the spiraling method (Beaton & Zwick, 1992). The spiraling step involved ordering the booklets in a systematic sequence when packages were made for distribution to participating schools. Each package contained an approximately equal number of each booklet. The different booklets were distributed within an assessment session in the order of the package, so that students next to each other did not receive the same booklet. Moreover, only a limited number of students in each class worked on the same booklet. Conversely, each booklet was administered in every participating school, thus increasing sampling efficiency by reducing sampling fluctuations across booklets. Booklet 1 formed an exception to this spiraling method, as will be explained in the Procedure section.

Respondents

Each booklet was administered to a sample of over 500 sixth-graders from normal education (see Table 2 for the precise figures). The respondents were sampled according to a clustered, stratified design: instead of drawing participants individually from the population of Flemish pupils in the sixth grade of

[1]If the total set of items would have been administered according to a design with overlapping (anchor) items, the number of test booklets would have increased, given that the length of a test session and the number of items administered to a single student had to be kept constant. Hence, such a design with overlap would have required a much larger sample of students.

elementary school, a sample of elementary schools was drawn and all sixth-graders of these schools were asked to participate in the test administration. The schools were sampled according to a stratified sampling design with three stratification variables: education sector, province and school size. First, in Flanders, schools belong to three different education sectors representing different organizing authorities: (1) Flemish Community schools, organized in the Autonomous Council for Flemish Community Education, (2) private subsidized schools, organized in the Flemish Federation of Catholic Elementary Schools, and (3) municipal schools, organized in the Educational Secretariat of the Association of Flemish Cities and Municipalities. The proportion of primary school children in sixth grade in these three education sectors is about .14, .64, and .22, respectively. Second, as an index of the geographical location, the population of schools was stratified according to the five Flemish provinces (Antwerp, Limburg, East-Flanders, West-Flanders, and Flemish-Brabant). Finally, as an indicator of school size, the population of schools was stratified as to whether or not more than 180 pupils were enrolled in the six grades of elementary education. The criterion of 180 pupils is an official criterion to determine whether a full-time headmaster can be subsidized.

For both parts of the study, an initial sample of schools was invited by letter to participate. In the first and the second part, this initial sample was respectively three and two times larger than the number of schools needed. Indeed, we did not expect very high response rates, as participation was voluntary; moreover, test administration took place near the end of the school year, which is usually a busy period, especially for sixth graders who then take their final exams. Surprisingly, the response rates were unexpectedly high: 52% for the first part and even 68% for the second part. These high response rates reflect the fact that the curriculum standards were a hot topic in Flemish education at the time of both studies. Because of the high costs of test administration, a number of schools that replied positively to our request were excluded from participation on a random basis. There was no overlap in participating schools between the two studies.

Procedure

Nine out of the 10 test booklets were administered by the teachers in their own class, following thereby detailed written instructions. Each test session took place in the morning during two periods of 90 min, separated by a 15 min break. An exception to this procedure was test booklet 1, which was administered by the researchers in 20 schools as this booklet required special

instructions. Test booklet 1 contained the items related to mental arithmetic (Scale 1) and to the use of a pocket calculator (Scale 11). In order to monitor the presentation time, the mental arithmetic items were presented using a flip chart in front of the classroom. The flip chart was used as well to instruct the pupils about the use of the pocket calculator that was made available to them by the research team; they were also shown on the chart how they had to write down their answers on the items relating to the calculator. The first part of the assessment instrument was administered between mid-April and the end of May 1998, and the second part during 1 month starting from April 15, 1999. Hence, both parts were administered just before and just after the time when the curriculum standards became operational (which was from September 1, 1998).

Item Scoring

All items were scored dichotomously on correctness. For the 31 items with a product and a process question, answers to both questions were first scored dichotomously. The scoring on correctness of the process question required the refinement of the scoring rules on the basis of the answers of the students. Cross-tabulation of the two dichotomous scores and inspection of the answers of the students learned that in all cases the process score was better or necessary to capture the corresponding curriculum standard. In 18 cases, the final item score consisted of the dichotomous process score. In 11 cases, both the process and the product question had to be solved correctly for the final item score to be correct. Finally, two items were deleted from their scale because, given the students' responses, neither the product nor the process score (nor both) was considered to be a valid operationalization of the higher-order thinking skill intended to be measured by the item.

Items with several subquestions were scored in one of the following three ways: (1) a sum score was derived by counting the correct answers to the subquestions; (2) a dichotomous score was given as to whether all or two out of three subquestions were solved correctly; or (3) the item was split in several items by giving a separate, dichotomous score to each subquestion. The choice of which scoring rule was followed for a given item was primarily based on the empirical distribution of students' responses. For example, when all subquestions turned out to be extremely easy, rule (2) was followed. For the present analyses, the sum scores were further dichotomized as to whether the obtained score exceeded half of the maximally obtainable sum. Especially due to the application of rule (3), the total number of items in Table 2 differs from

the previously mentioned number of 394 items (which included the items with different subquestions).

Technical Remarks About the MRCM

Some technical remarks can be made about the MCRM. Firstly, in the basic model used in MCRM the one-parameter model (as presented in equation 1) is extended with an item discrimination parameter. This is called a two-parameter IRT model. The item discrimination parameter is a quantitative index of the degree in which the item is related to the latent continuum, and is conceptually equivalent to a factor loading in factor analysis. Formally, the item discrimination parameter premultiplies the difference between θ_j and β_i in (1) to derive the probability of success on an item. Hence, given the position of two persons on the latent scale, the difference in the probability of success for these two persons will be smaller for an item with a low discrimination parameter than for an item with a high discrimination parameter (but with constant difficulty). Secondly, the difficulty and discrimination parameters of the reference item in the MCRM are set equal to the mean of a normal distribution postulated for each of the two sets of item parameters. The variance of this distribution is also estimated, but this parameter does not affect the probability of success on the reference item. A reference item was estimated for every curriculum standard included in the scale, so that the probability of success to each curriculum standard could be determined. Thirdly, the MCRM is estimated within a Bayesian framework using a Gibbs sampling procedure. Output of the Gibbs sampler was used to derive the posterior distribution of the probability of success on the reference item for each person. These posterior distributions were summarized by calculating over Gibbs draws and over students the proportion of success probabilities in each of the three classes using the .50 and .80 boundaries. Finally, fit of the MRCM was investigated using a posterior predictive check (PPC) based on the χ^2-statistic. The result of a PPC is summarized with a Bayesian p-value, with values below .05 indicating bad fit. For the scales of the present study (Table 2), the Bayesian p-values ranged between .32 and .52, hence, indicating good fit of the MRCM. More information about the MRCM can be found in Janssen et al. (2000) and in Janssen and De Boeck (2000).

Psychometric Analyses

The application of the MCRM to each measurement scale given in Table 2 took place in three steps. Firstly, the one-parameter IRT model was estimated

with the computer program OPLM (Verhelst, Glas, & Verstralen, 1995), which uses conditional maximum likelihood estimation. In this program, fit is investigated both at the item level and at the level of the test on the basis of χ^2-statistics. If necessary to obtain fit, a two-parameter IRT model was estimated, items were deleted, or scales were split (see below). Secondly, the MRCM was applied. Its goodness of fit was investigated and an overview of mastery attainment for the sample of respondents on the scale was made for each curriculum standard of the scale using the .50 and .80 cutoff points. Finally, for measurement scales with more than one curriculum standard, overall performance of the scale was summarized by computing a weighted average of the probabilities of success on the separate curriculum standards, using the number of items measuring each curriculum standard as weights.

Gender Analyses

In order to investigate the possible difference in performance of boys and girls, three kinds of analyses were performed on each scale using the estimated ability parameter coming from the analysis with the one- or two-parameter IRT model as the dependent variable. Firstly, the difference in performance between boys and girls was summarized with the standard mean difference (d). This is a measure of effect size that is comparable across scales, which is commonly used in meta-analysis (Willingham & Cole, 1997). The measure d is defined as the mean for girls minus the mean for boys, divided by the square root of the mean within-sexes variance. Thus, positive values of d represent superior performance of girls and negative values represent superior performance of boys. Secondly, a one-way ANOVA was performed to look for significant differences. The degrees of freedom of the F test were 1 for the numerator and $N - 1$ for the denominator, the value of which can be inferred from Table 2. Finally, the percentage of the variance accounted for by the gender variable was computed as a measure of the importance of a statistical significant effect.

RESULTS

Scale Construction

IRT analyses on the data sets of each measurement scale almost always led quite quickly to a fitting IRT model. For Scale 10, a fitting one-parameter model was found. For the other scales, a two-parameter IRT model was

necessary. The last column in Table 2 shows that only a very small number of items had to be removed in order to obtain fitting scales.

However, for Scales 2 and 21 many items had to be removed before obtaining fit. We therefore performed an item factor analysis on these datasets using the computer program NOHARM (Fraser, 1988). For Scale 2, a two-dimensional solution showed that all items related to the curriculum standard about Roman numerals loaded on a separate dimension in comparison with the other items. When the total item set of the original scale was split accordingly, a fitting two-parameter IRT model was obtained for both subsets (denoted as 2a and 2b in Table 2). Reports from teachers on the test administration taught us that a number of children did not (yet) receive instruction about Roman numerals. Because of this, even the items about Roman numerals that were very easy for children with opportunity to learn in this domain were difficult for children without opportunity to learn. Hence, the items of the total item set of Scale 2 could not be consistently ordered for all respondents. This is a clear violation of a model assumption of the IRT model. As will be shown below, the results on the separate measurement scale constructed for Roman numerals also indicated that the respondents could be split into two groups.

For Scale 21, item factor analysis showed that a three-dimensional solution was necessary: items parallelism and items about perpendicularity each split from the other items concerning factual knowledge about geometry. On the three subsets of items, a fitting IRT model was obtained: a two-parameter IRT model for Scale 21a, and a one-parameter IRT model for Scales 21b and 21c. An explanation for the lack of unidimensionality of the total item set was again found in differences in opportunity to learn. Responses on the items of Scales 21b and 21c showed that a number of children consistently used other symbols for parallelism (e.g., $=$) and perpendicularity (e.g., $\mathbf{T}$ or $\lfloor$) than the official symbols used in the curriculum standards ($//$, $||$, or $\backslash\backslash$, and $\perp$ or $\rfloor\lfloor$, respectively).

Mastery Assessment

Figure 2 gives a schematic representation of the proportion of students in the three mastery categories of nonmastery, intermediate, and good mastery for the measurement scales concerning curriculum standards about Numbers. The abscis denotes the cumulative percentage of students in the three mastery categories. The scales are ordered increasingly on the basis of the number of students in the nonmastery category. Hence, the lower-ordered scales show the best performance. Given this ordening, one can roughly distinguish among

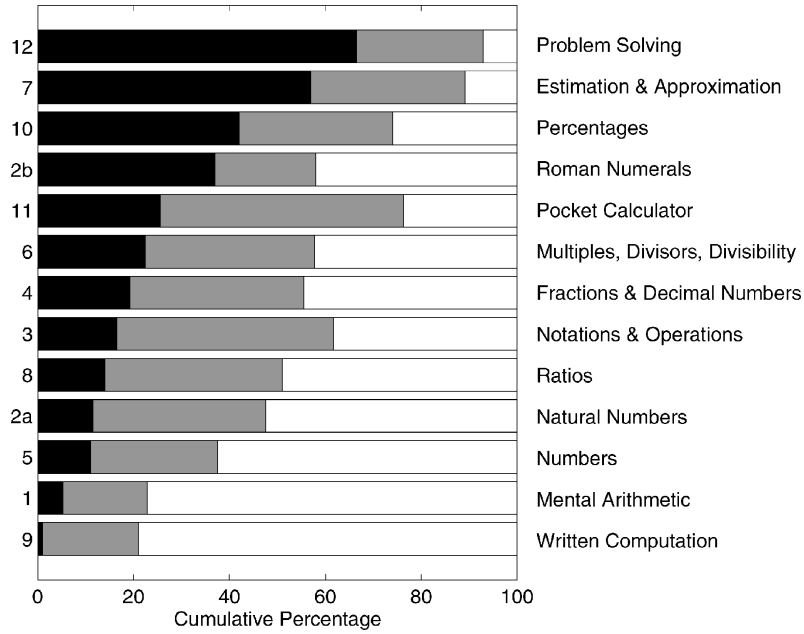


Fig. 2. Schematic representation of mastery attainment for the different measurement scales on Numbers. The dark, gray, and white bars indicate the percentage of students in the categories of insufficient, sufficient, and good mastery, respectively.

four clusters of scales. Firstly, the scales on written computation and mental arithmetic showed very high mastery attainment, with almost 80% of the students in the category of good mastery. Secondly, a group of seven scales can be distinguished with very small differences in the proportion of students in the nonmastery category among adjacent scales. This group contains scales regarding factual knowledge (like the scale on notations and operations) and procedural skills (like the scale on ratios). Scale 11 on the use of a pocket calculator showed relatively few students in the category of good mastery in comparison with the other scales of this group. Thirdly, the scales on Roman numerals and on percentages both showed approximately 40% of the students in the nonmastery category. It is remarkable that the scale on Roman numerals had an almost equal number of students in the category of nonmastery and of good mastery and only a small number of students in the intermediate category. This supports the assumption mentioned above that there may indeed be

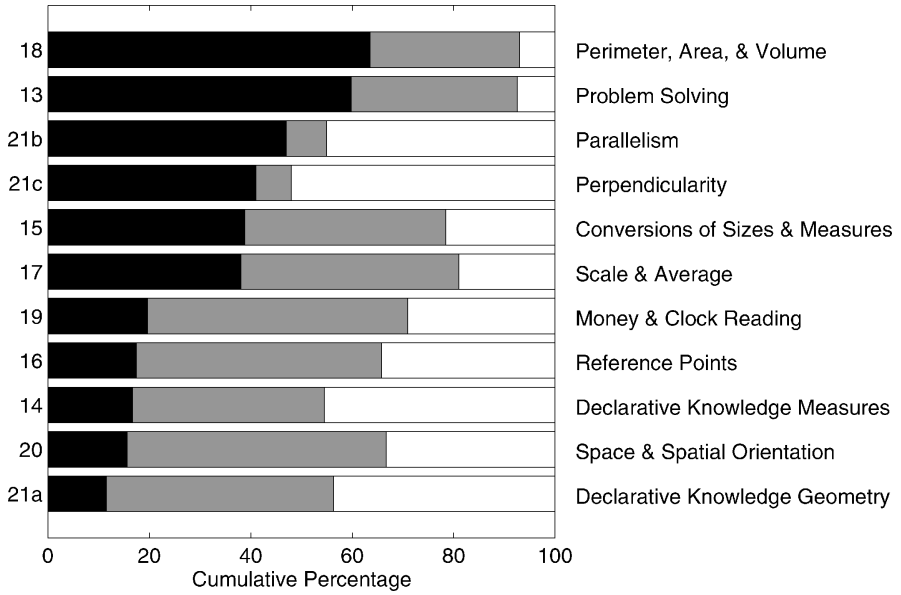


Fig. 3. Schematic representation of mastery attainment for the different measurement scales on Measurement & Geometry. The dark, gray, and white bars indicate the percentage of students in the categories of insufficient, sufficient, and good mastery, respectively.

two groups of students differing in their opportunity to learn the related curriculum standards. Finally, Scales 7 and 12 on estimation and on problem-solving showed a clear majority of the students in the nonmastery category and less than 10% in the category of good mastery.

Figure 3 gives an overview of the performance on the scales on Measurement and Geometry. A global comparison of Figure 3 with Figure 2 shows that the scales on Measurement and Geometry generally had lower mastery attainment than the scales on Numbers: Even the best performing scales on Measurement and Geometry had more students in the category of nonmastery and the students in the highest mastery category generally formed a clear minority. Visual inspection of Figure 3 leads to three clusters of measurement scales with an approximately equal number of students in the nonmastery category. Firstly, the best performances were found on five scales. These scales concern declarative knowledge about measurement and on geometry, on the one hand, and the scales on space, reference points, and money and

clock reading, on the other hand. Secondly, four scales had between 40 and 50% of students in the nonmastery category. These scales are about scale and average, conversions, perpendicularity, and parallelism. The latter two scales show a remarkably small number of students in the intermediate mastery category, possibly evidencing that for the present group of students mastery of these items is all-or-none given opportunity to learn or not. In fact, as was indicated above, a number of students used the wrong symbols. Finally, two curriculum standards showed a clear majority of students in the nonmastery category. These scales were on the higher-order thinking skills of problem-solving (as is the case for Numbers) and on the more complex skills of calculating perimeter, area, and volume.

It is quite remarkable that the scales of both Numbers and Measurement and Geometry were roughly ordered in terms of the complexity of the mathematical skills involved. In general, scales about declarative knowledge and scales involving lower-order mathematical procedures were mastered best. Scales involving more complex mathematical procedures (like calculating percentages or calculating perimeter, area, and volume) and scales involving higher-order thinking skills were mastered worse. The latter scales included the two scales on problem-solving and the scale on estimation and approximation. This is perhaps not at all surprising, as these scales, together with the scale on the use of the pocket calculator, concern curriculum standards that are relatively new for the curriculum of mathematics in Flanders.

These results suggest that at the times of study there were indeed differences among the curriculum standards in their degree of attainment. However, an alternative explanation for the performance differences on the scales might be given in terms of the difficulty of the items of the scales. Maybe the items of the lower-order scales in Figures 2 and 3 formed a selection of the easier items of their domain, while the items of the higher-order scales formed a selection of the more difficult items of their domain. Such a biased selection with respect to item difficulty would also lead to the same results.

However, an argument against this alternative explanation can be found in the item selection procedure, which was based on expert ratings of the expected difficulty level of the items. The majority of the items were selected on the basis of the experts' judgement that the expected percentage of students solving the item correctly would be between 60 and 90% given that education was provided appropriately in accordance with the new curriculum standards. Hence, according to the opinion of the experts, all scales were about equally difficult, which does not imply that the students would master them equally well. This argument

against a biased selection procedure with respect to item difficulty can be supported with the following finding as well. When we calculated for each scale the average difference between the empirical and the expected proportion of correct answers for those items where the latter value was available,[2] this average correlated −.88 with the proportion of students classified in the nonmastery category. The negative sign of the correlation implies that when students performed on the average worse than expected on (in most cases only a subset of) the items of the scale, the MCRM also classified a high number of students in the nonmastery category on the basis of their performance on the total item set. Hence, the strong correlation supports the conclusion that students' performance on the scales is ordered according to differences between experts' expectations and students' performances, and not according to differences in experts' expectations about the difficulty of the scales. For in the latter case, there would not be such a systematic relation between the difference between the empirical and the expected proportion of correct answers on the one hand and the proportion of nonmasters on the other.

Gender Differences

The percentage of boys taking a test booklet varied between 46.5% and 53.2% with a mean of 49.6%. The results of the gender analyses are summarized in Table 3. The standard mean difference showed better performance by the boys for about half of the scales; however, most values of d were close to zero. In fact, only 4 of the 24 standard mean differences (with $|d| > .2$) reached the level of a small effect according to Cohen's (1988) widely used classification. The F tests revealed a significant gender difference for 8 out of the 24 scales. Scales 3, 9, and 11 showed a significant female advantage, as can be inferred from d. These scales concern notations and operations, written computation, and the use of a pocket calculator. Significant male advantages were found for Scales 5 (different kinds of numbers), 8 (Ratios), 10 (Percentages), 16 (Reference points), and 20 (Space and spatial orientation). The percentage of score variation associated with these significant group differences varied between .66% and 4.92% for the scales with a significant F value. Thus, for these scales, and a fortiori for the other scales, individual differences within gender accounted for far most of the variation among the scores.

[2] This is the case for most of the items from the original item pool for which subquestions were not split in separate items and for which there was no significant change in the scoring rule used in the present study.

Table 3. Results of the Gender Analyses: Standard Mean Differences (*d*), *F* statistics, and Percentages of Explained Variance.

Scale	Content	*d*	*F*	% variance
Numbers				
1	Mental arithmetic	−0.08	1.12	.16
2a	Natural numbers	−0.10	1.53	.27
2b	Roman numerals	0.12	2.17	.37
3	Notations & operations	0.16	3.86*	.66
4	Fractions & decimal numbers	−0.11	1.80	.31
5	Numbers	−0.45	29.11**	4.92
6	Multiples, divisors, & divisibility	−0.03	.11	.02
7	Estimation & approximation	−0.14	2.82	.48
8	Ratios	−0.33	16.02**	2.67
9	Written computation	0.19	5.28*	.93
10	Percentages	−0.29	11.80**	2.02
11	Pocket calculator	0.26	11.13**	1.61
12	Problem-solving	0.16	3.51	.61
Measurement & Geometry				
13	Problem-solving	0.08	.97	.18
14	Measurement: declarative knowledge	−0.12	1.97	.36
15	Conversions	−0.02	.08	.02
16	Reference points	−0.21	6.18*	1.13
17	Scale & average	−0.03	.10	.02
18	Perimeter, area, & volume	0.00	.00	.00
19	Money & clock reading	0.09	1.11	.20
20	Space and spatial orientation	−0.17	4.15*	.75
21a	Geometry: declarative knowledge	0.11	1.65	.30
21b	Parallelism	0.14	2.64	.49
21c	Perpendicularity	0.02	.07	.01

*$p<.05$, **$p<.001$.

DISCUSSION

For the assessment of mastery attainment of the new curriculum standards of primary school mathematics in Flanders, a series of measurement scales was constructed, covering the whole mathematical curriculum. When compiling the scales, the validity and expected difficulty level of the items were taken into account. The majority of the items had a short-answer item format. In some cases, a product question was supplemented with or substituted by a process question.

Students' performance on these scales could generally be captured using a one-dimensional IRT model. This indicates that differences among students are consistent among the items, with students of high estimated proficiency performing better on both easy and difficult items in comparison with students of low estimated proficiency. However, two scales formed an exception, as no fitting IRT model was found for the corresponding entire whole item sets. Differences among the students in their opportunity to learn the curriculum standards on Roman numerals and on the use of the correct symbols for parallelism and perpendicularity seemed to cause the inconsistencies in the ordering of the items of the total item set. When the scales where split for the concerning curriculum standards, the students in the no-opportunity-to-learn group were presumably classified as nonmasters. The fact that for both sets of curriculum standards there were relatively few students in the category of intermediate mastery indicates that once appropriate teaching is provided relating to those standards, they are rather easily mastered. However, if a more detailed opportunity-to-learn questionnaire for the teachers would have been included, this interpretation of the results could have been made more strongly.

Clear differences in mastery attainment of the different subdomains of the mathematics curriculum were observed. For both the curriculum standards regarding Numbers and Measurement and Geometry, the performance on the scales was roughly ordered in terms of the complexity of the mathematical skills involved. This ordering may reflect the traditional focus on declarative and lower-order procedural skills in Flemish mathematics; in this respect we remind that at the times of the present study the new standards were not yet (part 1) or just (part 2) implemented. Thus, because in teaching less attention was paid to the newer curriculum standards, lower performance was observed on the corresponding measurement scales. This interpretation of the results relies on the fact that according to experts' ratings of expected difficulty level, performance on all scales should be comparable. However, the judgements of the experts about the expected difficulty level of the items, assuming that good teaching is provided in line with the standards, may also have been somewhat too optimistic, especially for the innovative objectives in the new standards. Therefore, a cross-validation study of the ratings of expected difficulty level by experts after the new standards have been operational for some time may be relevant. Nevertheless, a future national assessment of the mathematics standards will reveal whether or not students' performance has improved in comparison with the baseline assessment derived from the present study.

With respect to gender differences in performance, it was first found that Scales 3, 9, and 11 on notations and operations, written computation and the use of the pocket calculator showed better performance by the girls. All these scales involve careful and precise execution of standard procedures. Hence, these results are in line with a major conclusion of the review studies of Hyde et al. (1990) and Willingham et al. (1997) and with the conclusion of Van den Heuvel-Panhuizen and Vermeer (1999) with respect to girls. Secondly, the better performance by the boys on Scale 20 on space and spatial orientation and on Scale 15 on reference points is probably related to the commonly found male advantage on spatial orientation (e.g., Linn & Petersen, 1985). Finally, the other scales with better performance by the boys partly confirm the conclusion of Van den Heuvel-Panhuizen and Vermeer (1999) that male advantage in problem-solving skills can already be observed at the end of elementary education. Boys performed significantly better on scales where knowledge from experience or a clever strategy might be used like on Scales 8 and 10 on ratios and percentages. However, no significant differences were found on the two scales on problem-solving nor on the scales on estimation and approximation. Moreover, we cannot give a clear explanation for the boys' better performance on Scale 5 on numbers, which showed the biggest effect size of the present study. Hence, a further analysis of the present results at item level may be needed to study the gender differences found in depth. With respect to this, it may be fruitful to leave the current framework of classical gender analyses; instead, one could look for gender-based item and test bias within an IRT framework using test statistics for so-called differential item functioning and differential test functioning.

Finally, we like to stress the importance of the design of large-scale assessment instruments like the ones developed and administered in the present study, in view of changing educational practice. It is well known that the forms and the content of tests have a strong impact on teachers' instruction as well as on students' learning; indeed, the tests convey an implicit message to teachers and students that what is measured is also what is valued educationally. Therefore, it is of great significance in view of the implementation of the new perspective on mathematics education that assessment forms are constructed and regularly applied that are aligned with the current view about the objectives and the nature of mathematics learning and teaching. Moreover, these assessment instruments should have diagnostic power to provide teachers with useful information about students' (lack of) competence as a basis for instructional decision-making, especially for adapting teaching

appropriately to the needs of the learners. In this regard it is important to follow up the initial attempts of this study to construct process questions that provide feedback about students' understanding of basic concepts and about their processes and strategies for solving tasks and problems.

ACKNOWLEDGEMENT

The present paper was supported by OBPWO grants 94.02 and 97.05 of the Department of Education of the Ministery of the Flemish Community (Belgium). The first author is a postdoctoral fellow of the Fund for Scientific Research – Flanders (Belgium).

REFERENCES

Beaton, A.E., Mullis, I.V.S., Martin, M.O., Gonzales, E.J., Kelly, D.L., & Smith, T.A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS).* Chestnut Hill, MA: TIMSS International Study Center, Boston College.

Beaton, A.E., & Zwick, R. (1992). Overview of the National Assessment of Educational Progress. *Journal of Educational Statistics, 17*, 95–109.

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*, 137–172.

Berk, R.A. (1996). Standard setting: The next generation (where few psychometricians have gone before!). *Applied Measurement in Education, 9*, 215–235.

Cockcroft, W.H. (1982). *Mathematics counts*. (Report of the Committee of Inquiry into the Teaching of Mathematics in Schools). London: Her Majesty's Stationery Office.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

De Corte, E., Verschaffel, L., Knoors, E., & Detavernier, O. (1997). *Onderzoek naar de bruikbaarheid van bestaande peilingsinstrumenten voor de evaluatie van de nieuwe eindtermen wiskunde in het lager onderwijs in Vlaanderen* [Research into the usability of existing assessment instruments for the evaluation of the new mathematics standards in Flemish elementary education]. Leuven, Belgium: University of Leuven.

Fraser, C. (1988). *NOHARM: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, NSW: University of New England, Center for Behavioral Studies.

Frederiksen, N. (1990). Introduction. In N. Frederiksen, R. Glaser, A. Lesgold, & M.G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. ix–xvii). Hillsdale, NJ: Lawrence Erlbaum.

Hyde, J.S., Fennema, E., & Lamon, S.J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139–155.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: Macmillan.

Janssen, R., & De Boeck, P. (2000). Exploring the posterior of a hierarchical IRT model for item effects. *Journal of Computational Statistics, 15*, 421–442.

Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–461.

Linn, M.C., & Petersen, A.C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*, 1479–1498.

Masters, G., Lokan, J., Doig, B., Toon, K.S., Lindsey, J., Robinson, L., & Zammit, S. (1990). *Profiles of learning: The basic skills testing program in New South Wales: 1989.* Hawthorn, Victoria, Australia: Australian Council for Educational Research.

Meskauskas, J.A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery in standard setting. *Review of Educational Research, 45*, 133–158.

Ministerie van de Vlaamse Gemeenschap [Ministry of the Flemish Community]. (1997, November). *Basisonderwijs: Ontwikkelingsdoelen en eindtermen. Besluit 27 mei 1997, Decreet 15 juli 1997* [Elementary education: Developmental objectives and curriculum standards. Ministerial Order May, 27, 1997, Decree July 15, 1997]. Brussels, Belgium: Departement Onderwijs, Afdeling Informatie en Documentatie.

Mislevy, R.J., Johnson, E.G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics, 17*, 131–154.

Mullis, I.V.S., Martin, M.O., Gonzales, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J., & Smith, T.A. (2000). *TIMSS 1999: International mathematics report. Findings from IEA's Repeat of Third International Mathematics and Science Study at the eighth grade*. Chestnut Hill, MA: The International Study Center, Lynch School of Education, Boston College.

National Council of Teachers of Mathematics. (1989). *Assessment standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

National Research Council. (1999). *Grading the nations report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.

Pandey, T. (1990). Power items and the allignment of curriculum and assessment. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics* (pp. 39–51). Washington, DC: American Association for the Advancement of Science.

Silver, E.A., & Kenney, P.A. (1995). Sources of assessment information for instructional guidance in mathematics. In T.A. Romberg (Ed.), *Reform in school mathematics and authentic assessment* (pp. 38–86). Albany, NY: State University of New York Press.

Treffers, A., De Moor, E., & Feijs, E. (1989). *Proeve van een nationaal programma voor het reken-wiskundeonderwijs op de basisschool. Deel 1: Overzicht einddoelen* [Toward a National Mathematics Curriculum for the Elementary School. Part 1. Overview of the general goals]. Tilburg, The Netherlands: Uitgeverij Zwijsen.

Van den Heuvel-Panhuizen, M., & Vermeer, H.J. (1999). *Verschillen tussen meisjes en jongens bij het vak rekenen-wiskunde op de basisschool. Eindrapport MOOJ-onderzoek* [Gender differences in arithmetic-mathematics in elementary school. Final Report of the MOOJ-project]. Utrecht, The Netherlands: CD-β press.

Verhelst, N.D., & Eggen, T.J.H.M. (1991). The use of item response theory in the Dutch National Assessment Program. In A.J.M. Luijten (Ed.), *Issues in public examinations* (pp. 143–148). Utrecht, The Netherlands: Lemma.

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). *One-Parameter Logistic Model (OPLM)* [computer program]. Arnhem, The Netherlands: CITO, National Institute for Educational Measurement.

Verschaffel, L. (1995). Visies op reken/wiskundeonderwijs op de basisschool [Perspectives on mathematics education in the primary school]. In L. Verschaffel & E. De Corte (Eds.), *Naar een nieuwe reken/wiskundedidactiek voor de basisschool en de basiseducatie. Deel 1: Achtergronden* [Toward a new approach of mathematics education. Part 1. Theoretical background] (pp. 95–128). Brussel/Leuven, Belgium: Studiecentrum Open Hoger Onderwijs/ACCO.

Willingham, W.W., & Cole, N.S. (1997). Research on gender differences. In W.W. Willingham & N.S. Cole (Eds.), *Gender and fair assessment* (pp. 17–54). Mahwah, NJ: Lawrence Erlbaum.

Willingham, W.W., Cole, N.S., Lewis, Ch., & Wilson Leung, S. (1997). Test performance. In W.W. Willingham & N.S. Cole (Eds.), *Gender and fair assessment* (pp. 55–126). Mahwah, NJ: Lawrence Erlbaum.