Educational Research and Evaluation
1999, Vol. 5, No. 2, pp. 85–104

1380-3611/99/0502-0085$15.00

Mathematics Achievement in the Netherlands and Appropriateness of the TIMSS Mathematics Test

Wilmad Kuiper, Klaas Bos, and Tjeerd Plomp
University of Twente, Enschede, the Netherlands

ABSTRACT

Dutch students performed relatively well on the TIMSS mathematics test for population 2, although influential mathematics educators heavily criticised the appropriateness and fairness of the test in terms of the new, application- and inquiry-oriented mathematics curriculum which was implemented beginning in August 1993. This new intended curriculum differs drastically from the previous more formal and abstract curriculum from before 1993. For both policy makers and mathematics and science educators in the Netherlands important questions are the following: how should this discrepancy between students' performance and curricular appropriateness of the test be interpreted, and is there a discrepancy? Via expert appraisal, data have been gathered about the extent to which the TIMSS mathematics items are appropriate for this new intended mathematics curriculum. In addition, teachers were asked to judge the appropriateness of a selection of mathematics items in terms of the implemented curriculum. In this article the outcomes of the analyses with regard to the overlap between, on the one hand, the test and, on the other hand, both the intended curriculum and the implemented curriculum are described. In addition, as far as the intended curriculum is concerned, test-curriculum overlap outcomes are related to students' performances on the test.

INTRODUCTION AND RESEARCH QUESTIONS

The component with the highest and at the same time most challenging profile in any comparative study such as TIMSS is the achievement survey. All participating countries wish to ensure that the achievement items

Address correspondence to: Wilmad Kuiper, University of Twente, Faculty of Educational Science and Technology, Department of Curriculum, P.O. Box 217, 7500 AE Enschede, the Netherlands. Tel.: +31 53 4893757. Fax: +31 53 4893759. E-mail: kuiper@edte.utwente.nl

Manuscript submitted: March 30, 1998
Accepted for publication: January 19, 1999

used in the survey are appropriate for their students and reflect their mathematics and science curricula, enabling their students to give a good account of their knowledge and their ability and ensuring that international comparisons of student achievement will be based on a 'level playing field' insofar as possible (Garden, 1996; Garden & Orpwood, 1996). In TIMSS, like in all IEA studies, an achievement test that matches the content of the mathematics and science curricula of all participating countries was an unattainable goal though. Instead, as a compromise, the development process aimed at a test that was 'equally unfair' to all participants. Due to the number of countries (more than 40), the task of putting together the achievement test for the three TIMSS populations was immense, and took more than 3 years to complete. As is customary in IEA studies, developing the item pools was a co-operative venture involving all countries (including the Netherlands) during the entire process.

From its initial stage of development though, influential mathematics educators in the Netherlands heavily criticised the TIMSS mathematics test for population 2. To their opinion, the test was inappropriate and unfair in terms of the new, application- and inquiry-oriented mathematics curriculum that was implemented beginning in August 1993. Within this context and against the background of the three well-known conceptual levels of curriculum (cf. Editorial in this special issue; see also Beaton, Martin, et al., 1996; Beaton, Mullis, et al., 1996; Robitaille & Maxwell, 1996), the research questions that are addressed in this article are:

(i) How appropriate is the TIMSS mathematics test for population 2 in terms of both the intended and the implemented mathematics curriculum in the Netherlands?
(ii) How do Dutch students perform on the TIMSS mathematics test for population 2 and how should one interpret these students' performances against the background of findings with regard to the curricular appropriateness of the test?

SAMPLE AND RESEARCH GROUP

In the Netherlands – as in all of the other northern-hemisphere countries – the TIMSS data collection took place in spring 1995. A two-stage sample design was used. The first stage involved the selection of schools. Next, within each participating school random procedures were used to select one intact class at the upper target grade and one at the lower

target grade. All the students in those two classes participated in the TIMSS testing (same test for both classes). As part of the second stage the mathematics teachers from the classes involved were selected. In the Netherlands, this approach yielded a stratified representative sample of 95 schools[1], with 2,160 students from secondary 2 (upper target grade) and 2,027 students from secondary 1 (lower target grade) participating in the testing and the filling out of a student background questionnaire. Of the classes tested 155 mathematics teachers filled out a teacher questionnaire.

In the Netherlands, secondary education caters to students aged 12 to 17 or 18 years old. Students may follow one of four main ability tracks (Kuiper & Knuver, 1997):

– Junior secondary vocational or prevocational education, known as VBO. This is a 4-year program, specialising in technical, home economics, commercial, trade or agricultural studies.
– Junior general secondary education, known as MAVO, a 4-year long program.
– Senior general secondary education, known as HAVO. This is a 5-year program, preparing students for higher vocational education.
– Pre-university education, known as VWO. It is a 6-year course, preparing students for university and higher vocational education.

Secondary schools offer various combinations of these tracks. In the sample a distinction has been made in six combinations of tracks. However, at classroom level outcomes are reported for two main groups of students: students in VBO/MAVO and students in HAVO/VWO, both in secondary 1 and 2 (cf. Bos, Kuiper, & Plomp, in preparation).

1. A sample of 95 schools means a school participation rate of 63% after replacement schools were included. Although such a response rate is not uncommon in the Netherlands, it is below the TIMSS sampling participation standard of 85%. Therefore, in the international tables the Netherlands appears as a "country not satisfying guidelines for sample response rates". From an analysis of the quality of the Dutch sample, though, it has been made plausible that the schools that participated in the study were representative for the population. The analysis consisted of a comparison between the schools in the sample and the schools that refused to participate with regard to average school leaving exam scores for mathematics. From a one way analysis of variance per school type it appeared that there were no significant differences in average scores between the two groups of schools.

INSTRUMENTS

Mathematics Test

The TIMSS test for population 2 consisted of 150 mathematics items and 135 science items. Not all students responded to all of these items. To ensure broad subject matter coverage without overburdening individual students, TIMSS used a rotated design that included both mathematics and science items (Adams & Gonzales, 1996; Beaton, Martin, et al., 1996; Beaton, Mullis, et al., 1996). Thus, the same students participated in both the mathematics and science testing. The test consisted of eight booklets, with each booklet requiring 90 minutes of student response time. In accordance with the design, the mathematics and science items were assembled into 26 clusters (labelled A through Z). Cluster A was a core assigned to all booklets. The remaining clusters were assigned to the booklets in accordance with the rotated design so that representative samples of students responded to each cluster. The mathematics test covered six content areas or sub-scales (Table 1). On their turn, these content areas covered 12 content aspects. Performance expectations included: knowing; performing routine procedures; using complex procedures; and solving problems.

Table 1. Content Areas, Content Aspects and Number of Items in the TIMSS Mathematics Test (150 items).

Content areas (*n* items)	Content aspects (*n* items)
Fractions and number sense (51)	Common fractions: meaning and representation (8) Common fractions: operations, relations, properties (14) Decimal fractions (14) Estimation and number sense (15)
Geometry (23)	Congruence and similarity (6) Other geometry (17)
Algebra (27)	Linear equations (10) Other algebra (17)
Data representation, analysis, probability (20)	Data representation and analysis (13) Probability (7)
Measurement (18)	Measurement (18)
Proportionality (11)	Proportionality (11)

About one-fourth of the items were in the free-response format requiring students to generate and write answers. These questions, some of which required extended responses, were allotted approximately one-third of the testing time.

Test-Curriculum Matching Analysis

In order to investigate the appropriateness of the TIMSS mathematics test for the intended curriculum, a Test-Curriculum Matching Analysis (TCMA) was conducted. Two persons who are knowledgeable about the Mathematics 12–16 curriculum (see below) were asked to make a judgement about the relevance of each item to this curriculum. The number of experts involved in this task was relatively small because of international time constraints. The experts received the following instruction:

> You are asked to determine for all of the items, both for the upper target grade (secondary 2) and for the lower target grade (secondary 1), whether the *content* of the item can be supposed to be taught based on the *intended* curriculum (yes/no) *before March 15 1995 to at least 50% of the students in each grade*. You should focus on the content, not on format or difficulty. The intended curriculum has been defined as the Mathematics 12–16 curriculum, as operationalised in the three most widely used mathematics textbooks.

The experts were asked to make a judgement on the curricular appropriateness of each of the 150 mathematics items for each of the two grades as a whole, that is to say not broken down by track combination (VBO/MAVO respectively HAVO/VWO).

Opportunity to Learn

For a selection of items, mathematics teachers indicated whether the content tested had been taught before test administration or not (implemented). Unfortunately, these *Opportunity to Learn* (OTL) data were only collected in the Netherlands. As opposed to TCMA, it was decided to collect these OTL data for only a selection of items in order to avoid overburdening teachers while filling out the questionnaire. For secondary 1 a judgement was made for 32 international items, for secondary 2 for 16 international items (plus 20 national items from a national option mathematics test). Teachers were asked to answer the following question for each item (cf. De Haan, 1992):

> Suppose, you are asked to develop a test that is questioning the mathematics taught to the students of your class thus far, would you consider this item appropriate to be selected for this test, with regard to both content (yes/no) and format (yes/no). It is not the question whether all students in your class will be able to answer the item correctly; important is the consideration whether students did have the opportunity to learn it in this school year or before (secondary and/or primary education). The test consists of multiple choice and open ended questions.

As OTL data were provided by teachers of the classes selected, in this case it is possible indeed to make a breakdown by track combination (VBO/MAVO respectively HAVO/VWO). Criteria for selection of the 32 items for secondary 1 were: (i) representation of each content aspect by at least one item; (ii) both multiple choice and free-response items; (iii) the number of test booklets to which an item had been assigned; and (iv) coverage of part C of the international teacher questionnaire for mathematics. Part C of the teacher questionnaire contained a cluster of OTL questions about 44 stems of items, categorised by content area. The format of this part of the questionnaire was considered as too laborious for teachers, without providing proper OTL data at item level. A more simple format was preferred (see instruction above). In order to make it possible to compare national OTL data (based on the instruction described above) with international OTL data (based on part C of the teacher questionnaire), it was decided to select for each content aspect at least one item from the items in part C of the questionnaire. Thirteen out of the 44 items in part C were selected: two on fractions and number sense; two on geometry; two on algebra; two on data representation, analysis and probability; four on measurement; and one on proportionality. The third criterion has to do with the TIMSS test design: items from clusters that had been assigned to more than one test booklet were selected instead of items from clusters that had been assigned to only one booklet.

The 16 international items that were selected for the OTL analysis at secondary 2 were the items that had been selected as ‘anchor items’ for a national option mathematics test. This test has been administered in secondary 2 classes, only in the Netherlands and along with the TIMSS mathematics test (cf. Kuiper, Bos, & Plomp, 1997). An important criterion for an item to be selected as an anchor item was its perceived appropriateness to the new curriculum. All content areas and aspects were represented by the 16 items selected. Seven items overlapped with the item selection for secondary 1.

PRINCIPAL CHARACTERISTICS OF THE MATHEMATICS 12–16 CURRICULUM

The intended mathematics curriculum for lower secondary education (designated as basic education) is defined in terms of core objectives and has been developed concurrently with the new national examination program for mathematics for VBO and MAVO. The new Mathematics 12–16 curriculum (12–16 refers to students' age cohorts) differs drastically from the previous more formal and abstract curriculum from before 1993. It is based on the central adage for basic education: Application, Skills, and Coherence. The principal characteristics of the new curriculum are the following (Kuiper et al., 1997; Kuiper & Knuver, 1997): (i) mathematics content that appeals to students and that is explicitly linked to real-life situations ('contexts') that challenge students to 'mathematise' and to construct their own solutions in a creative and meaningful way; (ii) an emphasis on reasoning, problem solving and inquiry; and (iii) more coherence across mathematical topics and between mathematics and other subjects.

Changes in emphasis in the new intended curriculum for mathematics in basic education – just as in primary education – show a desire to make mathematics more accessible, interesting, relevant, and meaningful for all students in view of their future (society, education, professional career). The curriculum emphasises learning by doing, particularly in the use of well-chosen concrete materials and in the use of proper problem solving strategies (which is considered as much more important than giving right answers). Logical thinking, reasoning, anticipating, using adequate models, and reflecting on mathematical activities are also emphasised (Kuiper et al., 1997; Kuiper & Knuver, 1997). Significant changes have to do with content, teaching and learning approach, and new methods of assessment. The principal domains are, with between brackets an indication of the intended percentage of allocated time: arithmetic, measurement, and estimation (15%); algebra, relations, graphs, and functions (35%); geometry (35%); statistics and probability (15%); and integrated mathematical activities (aimed at integrating content and skills from the previously mentioned domains into an open, rather substantial investigation, to be conducted individually or in a small group).

Table 2. Judgements of Mathematics Educators on the Appropriateness of the Items from the TIMSS Mathematics Test to the Intended Curriculum by Content Area, Content Aspect, and Grade.

Content areas and content aspects (*n* items) in TIMSS mathematics test	Items appropriate to intended curriculum at secondary 1		Items appropriate to intended curriculum at secondary 2	
	n	%	*n*	%
Fractions and number sense (51)	45	88	46	90
Common fractions: meaning and representation (8)	8	100	8	100
Common fractions: operations, relations, properties (14)	12	86	13	93
Decimal fractions (14)	10	71	10	71
Estimation and number sense (15)	15	100	15	100
Geometry (23)	8	35	14	61
Congruence and similarity (6)	1	17	4	67
Other geometry (17)	7	41	10	59
Algebra (27)	4	15	12	44
Linear equations (10)	1	10	6	60
Other algebra (17)	3	18	6	35
Data representation, analysis, probability (20)	4	20	10	50
Data representation and analysis (13)	4	31	10	77
Probability (7)	0	0	0	0
Measurement (18)	9	50	15	83
Proportionality (11)	5	45	7	64
Total (150)	75	50	104	69

APPROPRIATENESS OF THE TIMSS MATHEMATICS TEST

Intended Curriculum

Table 2 shows the outcomes of TCMA. In this particular case, a breakdown by track combination is not possible. At secondary 1, on the average 50% of the items and for secondary 2 on the average 69% of the items are determined appropriate by the experts. There are large differences in appropriateness across content areas, content aspects, and the two grades. The majority of the items on fractions and number sense are determined appropriate for both secondary 1 (88%) and secondary 2 (90%). For secondary 2, about one-third of the items on proportionality (64%) and geom-

etry (61%) are determined appropriate. The same is true for about half of the items on data representation, analysis and probability (50%) and algebra (44%). For secondary 1, about half of the items on measurement (50%) and proportionality (55%), two-thirds of the items on geometry (65%) and the great majority of the items on data representation, analysis and probability (80%) and algebra (15%) are determined inappropriate. The inappropriateness of all of the seven items on probability is due to the fact that this content aspect is not part of the mathematics textbooks for both secondary 1 and 2.

Implemented Curriculum

Table 3 presents an overview of the average judgements of the teachers about the appropriateness of the items with regard to *content*. As already mentioned, the number of selected items varies per content area and grade. The percentages in the table should be read and interpreted in the following way: 9 out of the 51 items on fractions and number sense have been selected for this OTL analysis at secondary 1; on the average 75% of the teachers determined each of these 9 items appropriate *by content* to be selected for a test for secondary 1 VBO/MAVO students, questioning content taught thus far.

On the average 56% of the teachers at secondary 1 VBO/MAVO consider the content of each of the 32 selected items appropriate. At secondary 1 HAVO/VWO this average percentage is 20 points higher. At secondary 2 VBO/MAVO, the average percentage is 85%, in secondary 2 HAVO/VWO 88%. At the latter track grades, for each of the content areas on the average at least three-quarters of the teachers have a positive judgement about the appropriateness of the content of the items selected. The same is true for secondary 1 HAVO/VWO, with the exception of the four geometry items (each of which is considered appropriate by content by on the average 55% of the teachers) and algebra (70%). At secondary 1 VBO/MAVO the 75% threshold is only met by the items on fractions and number sense. The geometry and algebra items are determined appropriate by content by an on the average relatively low percentage of teachers (38% and 45%).

Due to the small number of items that have been reviewed (especially for secondary 2), one should be very careful with drawing conclusions from these findings. Nevertheless, these figures indicate that, except for secondary 1 VBO/MAVO, the content that has been tested via the TIMSS mathematics test seems to be taught to a large extent before test administration.

Later on the outcomes of the TCMA and OTL analyses will be discussed further.

Table 3. Average Judgements of Teachers on the Appropriateness of the Content of a Selected Number of Items to the Implemented Curriculum, broken down by Content Area and Track.

Content areas (*n* items) in TIMSS mathematics test	Number of items judged (*n*) and average percentage of teachers (*M %*) that considers each item appropriate by content for secondary 1				Number of items judged (*n*) and average percentage of teachers (*M %*) that considers each item appropriate by content for secondary 2			
	VBO/MAVO		HAVO/VWO		VBO/MAVO		HAVO/VWO	
	n	*M %*	*n*	*M %*	*n*	*M %*	*n*	*M %*
Fractions and number sense (51)	9	75	9	88	7	88	7	90
Geometry (23)	4	38	4	55	2	91	2	89
Algebra (27)	6	45	6	70	3	81	3	84
Data representation etc. (20)	5	59	5	77	2	77	2	83
Measurement (18)	5	51	5	76	1	86	1	78
Proportionality (11)	3	55	3	81	1	89	1	97
Total (150)	32	56	32	76	16	85	16	88

MATHEMATICS ACHIEVEMENT

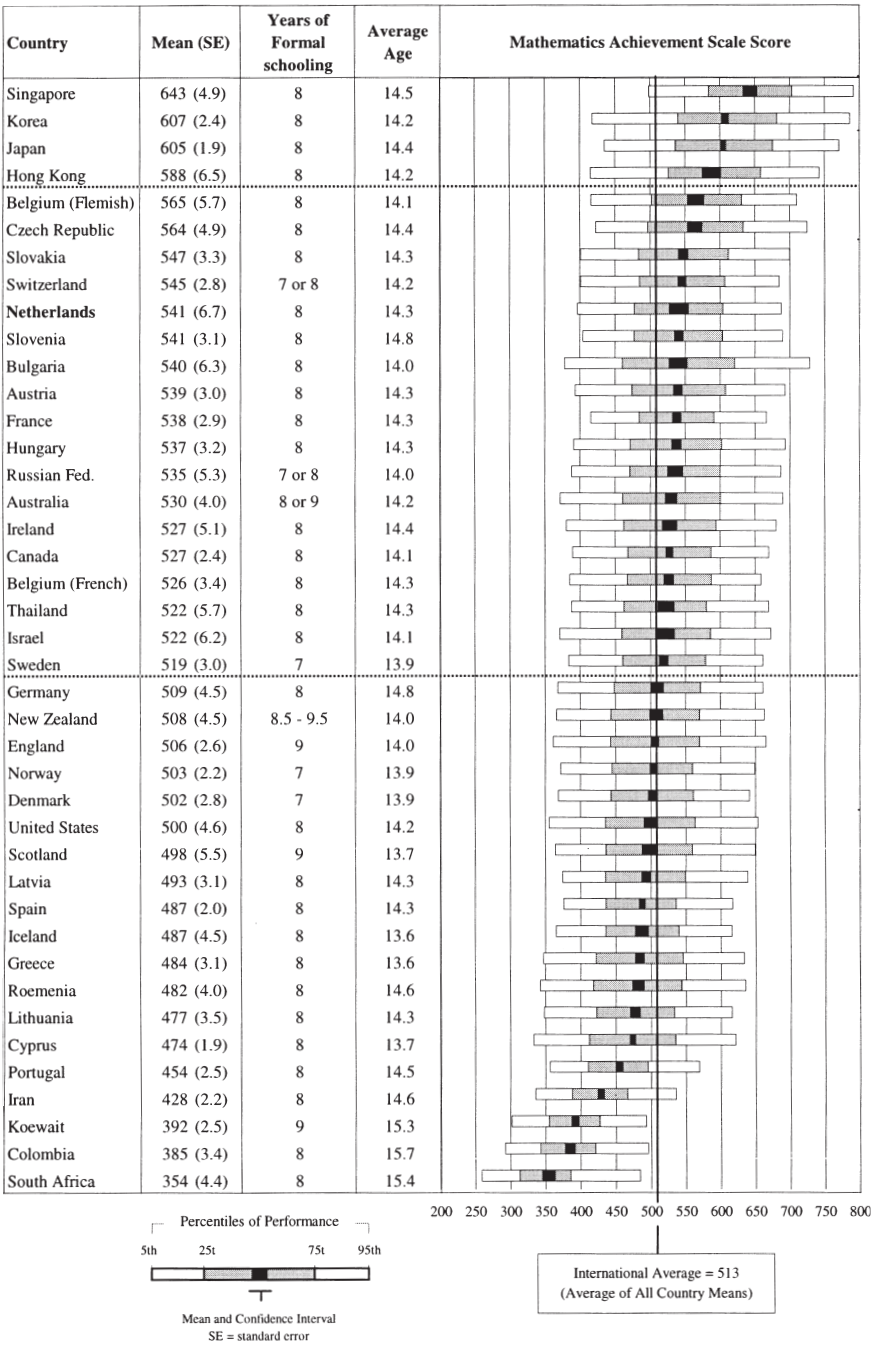
Mean Overall Achievement

Dutch students from both target grades perform relatively well on the TIMSS mathematics test. At the *upper* target grade, the mean overall achievement (in terms of so called plausible values[2]) of students in only Singapore, Korea, Japan and Hong Kong is significantly higher than the mean overall achievement of Dutch students (Table 4; Table derived from Beaton, Mullis, et al., 1996; dark boxes at midpoints of distributions show 95% confidence intervals around average achievement in each country). Dutch students achieve at the same level as students in, among other countries, Belgium-Flemish, Czech Republic, Slovak Republic, Austria,

2. TIMSS used an Item Response Theory (IRT) scaling method (Rasch model) to summarise the achievement results for both grades (cf. Beaton, Martin, et al., 1996; Beaton, Mullis, et al., 1996). The scores were standardised on a scale with a mean of 500 and a standard deviation of 100. Scaling averages students' responses to the subsets of items they took in a way that accounts for differences in the difficulty of those items. The methodology used in TIMSS includes refinements that enable reliable population estimates to be produced even though individual students responded to relatively small subsets of the total mathematics item pool. It allows students' performance to be summarised on a common metric even though individual students responded to different items in the mathematics test. As a consequence, it allows only statements on mathematics achievement of a population of students, not of individual students.

Table 4. Distributions of Mathematics Achievement – Upper Grade (Secondary 2 in the Netherlands).

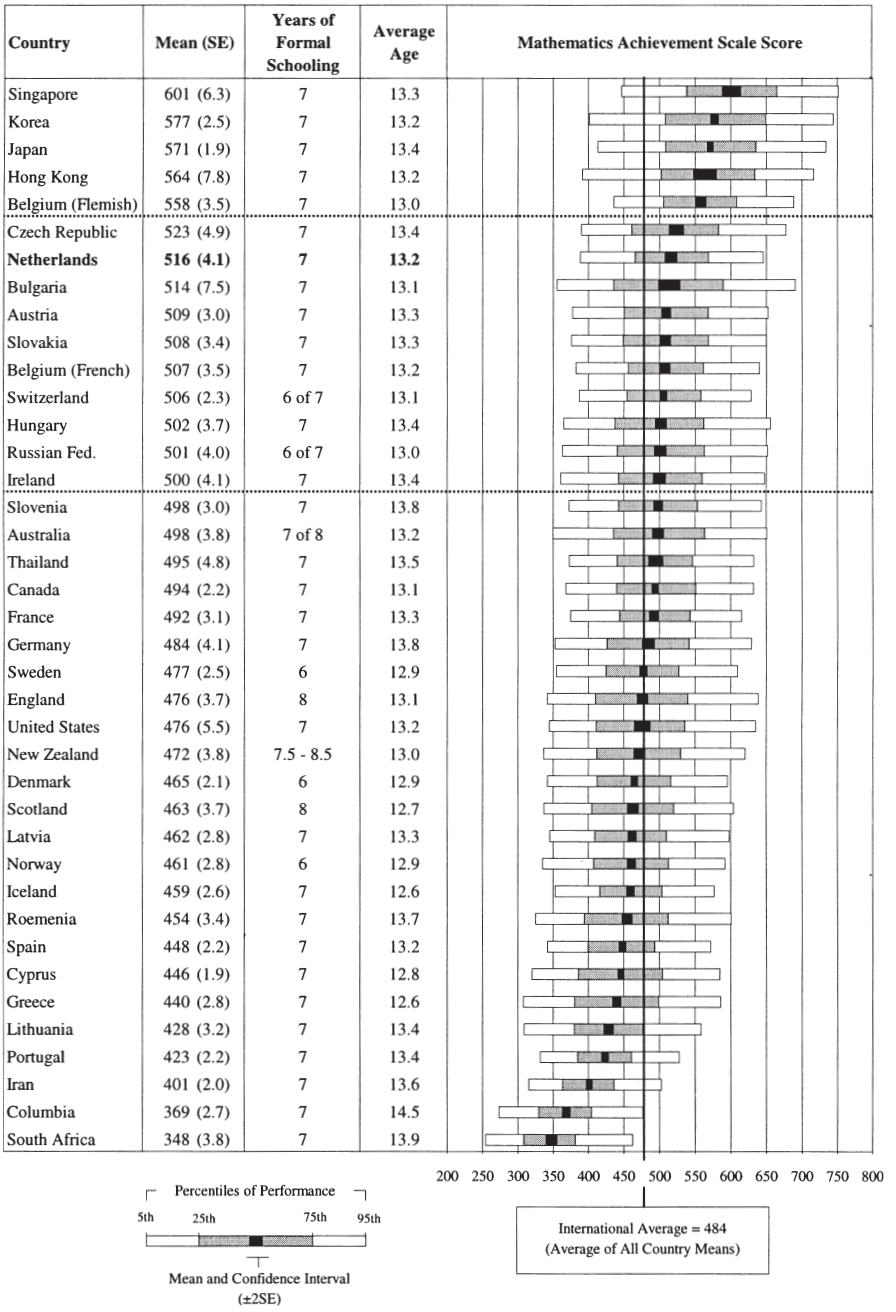
Country	Mean (SE)	Years of Formal schooling	Average Age	Mathematics Achievement Scale Score
Singapore	643 (4.9)	8	14.5	
Korea	607 (2.4)	8	14.2	
Japan	605 (1.9)	8	14.4	
Hong Kong	588 (6.5)	8	14.2	
Belgium (Flemish)	565 (5.7)	8	14.1	
Czech Republic	564 (4.9)	8	14.4	
Slovakia	547 (3.3)	8	14.3	
Switzerland	545 (2.8)	7 or 8	14.2	
Netherlands	541 (6.7)	8	14.3	
Slovenia	541 (3.1)	8	14.8	
Bulgaria	540 (6.3)	8	14.0	
Austria	539 (3.0)	8	14.3	
France	538 (2.9)	8	14.3	
Hungary	537 (3.2)	8	14.3	
Russian Fed.	535 (5.3)	7 or 8	14.0	
Australia	530 (4.0)	8 or 9	14.2	
Ireland	527 (5.1)	8	14.4	
Canada	527 (2.4)	8	14.1	
Belgium (French)	526 (3.4)	8	14.3	
Thailand	522 (5.7)	8	14.3	
Israel	522 (6.2)	8	14.1	
Sweden	519 (3.0)	7	13.9	
Germany	509 (4.5)	8	14.8	
New Zealand	508 (4.5)	8.5 - 9.5	14.0	
England	506 (2.6)	9	14.0	
Norway	503 (2.2)	7	13.9	
Denmark	502 (2.8)	7	13.9	
United States	500 (4.6)	8	14.2	
Scotland	498 (5.5)	9	13.7	
Latvia	493 (3.1)	8	14.3	
Spain	487 (2.0)	8	14.3	
Iceland	487 (4.5)	8	13.6	
Greece	484 (3.1)	8	13.6	
Roemenia	482 (4.0)	8	14.6	
Lithuania	477 (3.5)	8	14.3	
Cyprus	474 (1.9)	8	13.7	
Portugal	454 (2.5)	8	14.5	
Iran	428 (2.2)	8	14.6	
Koewait	392 (2.5)	9	15.3	
Colombia	385 (3.4)	8	15.7	
South Africa	354 (4.4)	8	15.4	



SE = Standard Error

Table 5. Distributions of Mathematics Achievement – Lower Grade (Secondary 1 in the Netherlands).

Country	Mean (SE)	Years of Formal Schooling	Average Age
Singapore	601 (6.3)	7	13.3
Korea	577 (2.5)	7	13.2
Japan	571 (1.9)	7	13.4
Hong Kong	564 (7.8)	7	13.2
Belgium (Flemish)	558 (3.5)	7	13.0
Czech Republic	523 (4.9)	7	13.4
Netherlands	**516 (4.1)**	**7**	**13.2**
Bulgaria	514 (7.5)	7	13.1
Austria	509 (3.0)	7	13.3
Slovakia	508 (3.4)	7	13.3
Belgium (French)	507 (3.5)	7	13.2
Switzerland	506 (2.3)	6 of 7	13.1
Hungary	502 (3.7)	7	13.4
Russian Fed.	501 (4.0)	6 of 7	13.0
Ireland	500 (4.1)	7	13.4
Slovenia	498 (3.0)	7	13.8
Australia	498 (3.8)	7 of 8	13.2
Thailand	495 (4.8)	7	13.5
Canada	494 (2.2)	7	13.1
France	492 (3.1)	7	13.3
Germany	484 (4.1)	7	13.8
Sweden	477 (2.5)	6	12.9
England	476 (3.7)	8	13.1
United States	476 (5.5)	7	13.2
New Zealand	472 (3.8)	7.5 - 8.5	13.0
Denmark	465 (2.1)	6	12.9
Scotland	463 (3.7)	8	12.7
Latvia	462 (2.8)	7	13.3
Norway	461 (2.8)	6	12.9
Iceland	459 (2.6)	7	12.6
Roemenia	454 (3.4)	7	13.7
Spain	448 (2.2)	7	13.2
Cyprus	446 (1.9)	7	12.8
Greece	440 (2.8)	7	12.6
Lithuania	428 (3.2)	7	13.4
Portugal	423 (2.2)	7	13.4
Iran	401 (2.0)	7	13.6
Columbia	369 (2.7)	7	14.5
South Africa	348 (3.8)	7	13.9



SE = Standard Error

France, Hungary and Canada. The mean overall achievement of students in for instance Germany, New Zealand, England, Denmark, and the US is significantly lower than the mean overall achievement of Dutch students. At the *lower* target grade, the mean overall achievement of students in the same four East-Asian countries (Singapore, Korea, Japan and Hong Kong) and also in Belgium-Flemish is significantly higher than the mean overall achievement of Dutch students (Table 5; Table derived from Beaton, Mullis et al., 1996; 95% confidence intervals). At this grade, Dutch students perform at the same level as students from for instance: Czech Republic, Hungary and Russia. Students in Germany, England, Denmark, Canada, the US, New Zealand and Australia perform significantly less well. By way of comparison, the mean overall achievement of Dutch students at the lower target grade is almost equal to the international mean at the upper grade. At both grades, about two-thirds of the students in the Netherlands achieved above the international mean.

Table 6 presents the mean overall achievement for Dutch students at secondary 1 and 2, broken down by track, grade and gender (derived from Kuiper et al., 1997). Achievement is presented in terms of plausible values (on a scale with a standardised average of 500 and a standard deviation of 100). The scores presented in the table are unweighted and, therefore, they deviate a little from the scores presented in the international tables.

The table shows that students at secondary 2 outperformed students at secondary 1. However, the difference is smaller than might be expected. The mean overall achievement for both grades (521 respectively 550) is clearly higher than the international average (484 for grade 7 and 513 for grade 8; see Tables 4 and 5). At both secondary 1 and 2, the mean overall achievement of HAVO/VWO students is higher than the mean overall

Table 6. Mean Overall Achievement for Population 2 Students by Track, Grade and Gender (Unweighted Scores).

Track and gender		Secondary 1 *M (SD)*	Secondary 2 *M (SD)*
VBO/MAVO	Female	480 (67)	504 (71)
	Male	485 (71)	514 (75)
HAVO/VWO	Female	564 (62)	608 (66)
	Male	576 (61)	614 (64)
Total		521 (79)	550 (86)

Note. M (SD) = mean (standard deviation).

achievement of VBO/MAVO students. Noteworthy is the large difference in mean overall achievement between HAVO/VWO students and VBO/MAVO students within the two grades. Apparently, these differences are larger than the differences in mean achievement between grades and also between boys and girls.

Average Achievement in Mathematics Content Areas

Table 7 presents Dutch students' achievement in mathematics content areas. Achievement in content areas is not presented in terms of plausible values (for which too complex and too laborious analyses needed to be conducted, both internationally and nationally) but in terms of average percentages correct on sets of items. The table should be read as follows: at secondary 1, each of the 51 items on fractions and number sense has been answered correctly by on the average 61% of the students.

The table shows that, at secondary 2, the overall average percentage correct is 63%; at secondary 1, this percentage is 58%. As said earlier, a strikingly small difference. When we look at differences between the two grades with regard to average percentages correct to items in the different content areas, then only the average percentages correct to the items on algebra differ substantially (56% at secondary 2 versus 43% at secondary 1). This relatively large difference between the two grades for algebra is probably due to the fact that the content tested is usually only just taught in secondary 2. At both grades, the variation in scores per content area is quite similar.

Both at secondary 1 and 2, students perform relatively poorest on the algebra items and relatively best on data representation, analysis and probability items.

Table 7. Average Percentage Correct by Mathematics Content Areas by Grade (Unweighted Scores).

Content areas (*n* items)	Secondary 1 *M (SD) %* correct	Secondary 2 *M (SD) %* correct
Fractions and number sense (51)	61 (21)	63 (20)
Geometry (23)	55 (18)	62 (17)
Algebra (27)	43 (21)	56 (19)
Data representation etc. (20)	74 (15)	78 (13)
Measurement (18)	60 (23)	65 (22)
Proportionality (11)	56 (21)	56 (19)
Total (150)	58 (22)	63 (20)

Note. M (SD) % correct = average percentage correct (standard deviation).

Table 8. Average Percentage Correct (*M %* correct) for Females and Males (F/M) by Mathematics Content Area and Track-Grade.

Content areas (*n* items)	Secondary 1				Secondary 2			
	VBO/MAVO		HAVO/VWO		VBO/MAVO		HAVO/VWO	
	M % correct		*M %* correct		*M %* correct		*M %* correct	
	F	M	F	M	F	M	F	M
Fractions and number sense (51)	50	52	73	77	53	56	76	81
Geometry (23)	45	46	64	69	51	54	73	76
Algebra (27)	34	33	54	56	45	43	71	72
Data representation etc. (20)	65	66	83	86	69	71	88	90
Measurement (18)	46	47	65	70	50	54	74	76
Proportionality (11)	42	42	65	67	41	43	67	74
Total (150)	47	48	68	71	52	54	75	78

In Table 8, average percentages correct by content areas are further broken down by track and gender. Differences in average percentages correct by gender lie in-between 0% and 7%, with higher average percentages correct for boys than for girls (save the average percentage correct to algebra items at secondary 1 VBO/MAVO, at secondary 2 VBO/MAVO, and at secondary 2 HAVO/VWO). Gender differences have not been tested on significance. This analysis was considered less worthwhile as the international report showed only slight gender differences for the Netherlands and other countries. Even if gender differences in the Netherlands would be significant, differences are small.

MATHEMATICS ACHIEVEMENT AND APPROPRIATENESS OF THE TEST

What do the performances on the mathematics test mean against the background of the findings just described with regard to the appropriateness of the test? Table 9 presents average percentages correct to items that mathematics experts consider appropriate and to items that are considered inappropriate to the new *intended* mathematics curriculum (TCMA outcomes). Unfortunately, such an analysis at the level of the implemented curriculum is not possible due to the small number of items reviewed. At secondary 1, the average percentage correct to the total set of appropriate items is only slightly higher than the average percentage correct to the total set of inappropriate items (61% versus 55%). So, without further analyses we can

say students at secondary 1 seem to perform slightly better on the appropriate items. At secondary 2, the average percentage correct to the total set of appropriate items is exactly the same as the average percentage correct to the total set of items that are considered inappropriate (63%). Looking at the various content areas, Table 9 shows there are three cases in which the average percentage correct to appropriate items hardly differs from the average percentage correct to inappropriate items: proportionality (secondary 1), fractions and number sense (secondary 2), and data representation, analysis and probability (secondary 2). Larger differences – that is a higher average percentage correct to appropriate items – occur at fractions and number sense (secondary 1), geometry (secondary 1 and 2), algebra (secondary 1 and 2), and data representation, analysis and probability (secondary 2). Noteworthy is that the average percentage correct to inappropriate items on data representation, analysis and probability (secondary 1), measurement (secondary 1 and especially secondary 2) and proportionality (secondary 2) is higher than the average percentage correct to appropriate items.

Possibly, some inappropriate items are so easy, that students are able to respond correctly to the item without the content having been taught in school. Another explanation might be that the content tested has been taught already in primary education. Especially with regard to items on probability the latter explanation seems to make sense. A third explanation

Table 9. Average Percentage Correct (*M %* correct) to Appropriate versus Inappropriate Mathematics Items (Intended Curriculum).

Content areas (*n* items)	Secondary 1				Secondary 2			
	Appr. items		*M (SD)* % correct		Appr. Items		*M (SD)* % correct	
	n	*%*	Appr.	Inappr.	*n*	*%*	Appr.	Inappr.
Fractions, number sense (51)	45	88	62 (19)	52 (27)	46	90	63 (20)	62 (25)
Geometry (23)	8	35	66 (13)	49 (18)	14	61	65 (15)	57 (20)
Algebra (27)	4	15	56 (21)	41 (21)	12	44	60 (18)	53 (19)
Data representation etc (20)	4	20	62 (22)	77 (12)	10	50	79 (16)	77 (10)
Measurement (18)	9	50	58 (19)	63 (27)	15	83	62 (22)	82 (10)
Proportionality (11)	5	45	56 (15)	55 (26)	7	64	49 (18)	69 (15)
Total (150)	75	50	61 (19)	55 (24)	104	69	63 (20)	63 (20)

Note. appr. = appropriate; inappr. = inappropriate.

could be that the content of items, notwithstanding their inappropriateness to the intended curriculum, has been taught at school indeed. A fourth possibility is that the implemented curriculum (still) deviates from the intended curriculum. A fifth, more technical, explanation might be that this finding (i.e., higher average percentages correct to inappropriate items) has been influenced by the small number of items involved (see for example the content area proportionality).

By the way, a general consideration that should be taken into account while interpreting the figures presented in Table 9 is that the proportion between the number of appropriate items and the number of inappropriate items for the content areas fractions & number sense, geometry, algebra, and data representation, analysis & probability is rather distorted. For fractions & number sense, for example, 45 (out of 51) items are considered appropriate and only 6 items are judged inappropriate. For those four content areas, this fact may influence the differences between the average percentages correct to the sets of appropriate and inappropriate items.

CONCLUSIONS AND DISCUSSION

First Research Question

The first research question that has been addressed is the following: How appropriate is the TIMSS mathematics test for population 2 in terms of both the intended and the implemented mathematics curriculum in the Netherlands? As far as the relevance of the test to the intended curriculum (Mathematics 12-16) is concerned, we found that at secondary 1, on the average 50% of the items and, at secondary 2, on the average 69% of the items are determined appropriate by the experts. With regard to the appropriateness of the test to the implemented curriculum, there are some indications that, except for secondary 1 VBO/MAVO, the content tested seems to be taught to a large extent.

Comparing the TCMA results (intended curriculum) with the OTL findings (implemented curriculum), it turns out that the judgements of the teachers (implemented) are relatively more positive than the judgements of the mathematics educators (intended). In some cases, the average judgements of teachers are even substantially more positive: geometry (secondary 2); algebra (secondary 1 and 2); data representation, analysis and probability (secondary 1 and 2); and proportionality (secondary 2). Consequently, generally stated the TIMSS mathematics test seems to be more appropriate to the implemented than to the intended curriculum.

Second Research Question

The second research question addressed is: How do Dutch students perform on the TIMSS mathematics test for population 2 and how to interpret these students' performances against the background of findings with regard to the appropriateness of the test? With regard to overall mathematics achievement we have seen that, at both grades, Dutch students perform relatively well. With regard to average percentages correct to items in the different content areas tested, only the average percentages correct to the items on algebra differ substantially between the two grades. Both at secondary 1 and 2, students perform relatively poorest on algebra and relatively best on data representation, analysis and probability.

Looking at average achievement on items that are appropriate and inappropriate to the intended curriculum, we found that, at secondary 1, performance on appropriate items is only slightly better than performance on inappropriate items (average percentage correct 61 versus 55). At secondary 2, the average percentage correct to the total set of appropriate items is exactly the same as the average percentage correct to the total set of items that are considered inappropriate (63%). Unfortunately, such an analysis at the level of the implemented curriculum is not possible because of the small number of items reviewed.

Discussion

Is it legitimate to conclude that the TIMSS mathematics test is more appropriate to the implemented than to the intended curriculum? And how should differences in judgements between teachers and mathematics educators be interpreted? With regard to the latter question, it is not surprising that mathematics educators seem less positive in their judgement than teachers. The Mathematics 12–16 curriculum differs drastically from the previous more formal and abstract curriculum from before 1993. It entails a complex, sweeping and refractory curriculum reform, involving changes in the use of curriculum materials, in the use of teaching strategies, and in teachers' beliefs. Change has to take place in practice along all these three dimensions in order for it to have a chance of affecting the outcome (Fullan, 1991). It is not realistic to assume that this change could have been realised within less than 2 years, that is between August 1993 and test administration (spring 1995). When we relate this compelling sense of reality to the carefully drawn conclusion that the content tested seems to be taught to a large extent (except for secondary 1 VBO/MAVO), then there is sufficient reason to believe that aspects of the old mathematics curriculum were still reflected in the actual teaching practices at the time of testing. Consequently, it is not surprising that teachers have a relatively

positive judgement on the appropriateness of the test and students perform relatively well on the test (which was characterised as 'rather traditional' by influential mathematics educators).

Nevertheless, one should be careful with drawing too firm conclusions. As far as the OTL part is concerned, a first restriction is that teachers have reviewed only a small selection of items. Especially for secondary 2 the number of items reviewed is relatively small (16). In addition, these 16 (anchor) items make up a rather selective set as they were selected because of their supposed appropriateness to the new curriculum. A one-to-one comparison intended versus implemented is only possible for a limited number of items. For secondary 1 slightly more items can be compared. It can be questioned whether the OTL findings would have been the same if teachers had had the opportunity to review all items. Secondly, it is possible some teachers based their judgement about appropriateness of an item also on perceived difficulty of the item. Although there is no evidence for this in this particular case, De Haan (1992) – who developed and tested the test-curriculum overlap instrument we used – points to the possible influence of perceived difficulty on test-curriculum overlap judgements. She found correlations of .50 between estimations by teachers of item difficulty and teachers' judgements on item appropriateness.

With regard to the TCMA part, it should be emphasised that due to time constraints the number of experts involved in this task was relatively small. In addition, making a judgement, based on strict international guidelines, on the curricular appropriateness of the various items for both secondary 1 and 2 as a whole (i.e., not broken down by track) appeared to be rather challenging, at least for part of the items.

In this article we focused on students' outcomes on the TIMSS mathematics test for population 2 against the background of the appropriateness of the test in terms of the intended and the implemented mathematics curriculum. As far as the intended curriculum is concerned, an international Test-Curriculum Matching Analysis was conducted. At the level of the implemented curriculum Opportunity to Learn data were collected using an instrument that was easily and quickly to use by teachers, but unfortunately only on a national basis and only for a selection of items. Taking into account the limitations just mentioned, the findings presented show the importance and relevance of gathering and analysing Test-Curriculum Overlap data, meant as contextual information for interpreting student outcomes in international, comparative studies like TIMSS. A recommendation for future studies, therefore, is to investigate the appropriateness of the whole test for both the intended and the implemented curriculum in all participating countries.

REFERENCES

Adams, J., & Gonzales, E.J. (1996). The TIMSS test design. In M.O. Martin & D.L. Kelly (Eds.), *TIMSS technical report volume 1: Design and development* (pp. 3–1 through 3-36). Boston: Boston College.

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

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.* Boston: Boston College.

Bos, K.Tj., Kuiper, W., & Plomp, Tj. (in preparation). *Third International Mathematics and Science Study (TIMSS). Technische verantwoording van de rapportage over de uitkomsten van het Nederlands aandeel in TIMSS populatie 2 en 3* [Technical report on Dutch participation in TIMSS population 2 and 3]. Enschede: University of Twente.

de Haan, D.M. (1992). *Measuring Test-curriculum overlap*. Enschede: University of Twente.

Fullan, M.G. (1991). *The new meaning of educational change.* London: Cassell.

Garden, R.A. (1996). Development of the TIMSS achievement items. In D.F. Robitaille & R.A. Garden (Eds.), *Research questions and study design. TIMSS Monograph No. 2* (pp. 69–80). Vancouver: Pacific Educational Press.

Garden, R.A., & Orpwood, G. (1996). TIMSS test development. In M.O. Martin & D.L Kelly (Eds.), *Third International Mathematics and Science Study, Technical report. Volume 1* (pp. 2–1 to 2–19). Boston, MA: Boston College.

Kuiper, W.A.J.M., Bos, K.Tj., & Plomp, Tj. (1997). *Wiskunde en de natuurwetenschappelijke vakken in leerjaar 1 en 2 van het voortgezet onderwijs. Nederlands aandeel in TIMSS populatie 2* [Mathematics and the science domains in secondary 1 and 2. Dutch participation in TIMSS population 2]. Enschede: University of Twente.

Kuiper, W.A.J.M., & Knuver, J.W.M. (1997). The Netherlands. In D.F. Robitaille (Ed.), *National contexts for mathematics and science education. An encyclopedia of the education systems participating in TIMSS* (pp. 259–269). Vancouver: Pacific Educational Press.

Robitaille, D.F., & Maxwell, B. (1996). The conceptual framework and research questions for TIMSS. In D.F. Robitaille & R.A. Garden (Eds.), *Research questions and study design. TIMSS Monograph No. 2* (pp. 34–43). Vancouver: Pacific Educational Press.