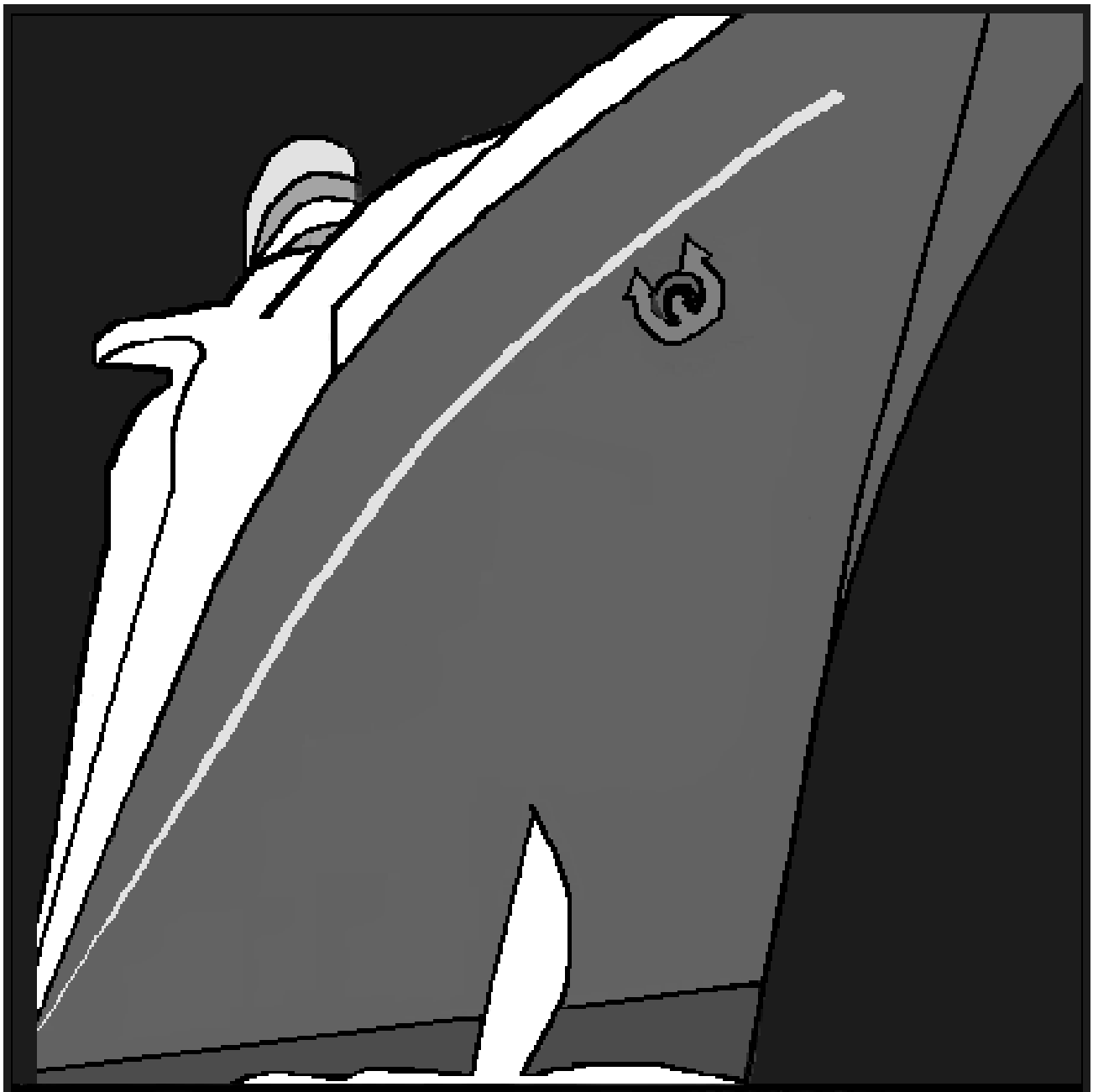


LIKE AN OCEAN LINER CHANGING COURSE

The grade 8 mathematics curriculum in the Netherlands, 1995-2000

PAULINE VOS

Like an ocean liner changing course

The grade 8 mathematics curriculum in the Netherlands, 1995-2000

Pauline Vos

DOCTORAL COMMITTEE

Chairman: Prof. dr. Jules Pieters ▪ University of Twente

Supervisors: Prof. dr. Tjeerd Plomp ▪ University of Twente
Dr. Wilmad Kuiper ▪ University of Twente

Members: Prof. dr. Jill Adler ▪ University of the Witwatersrand
Prof. dr. Jan van den Akker ▪ University of Twente
Prof. dr. Jan de Lange ▪ Freudenthal Institute
Prof. dr. Ina Mullis ▪ IEA International Study Center, Boston College
Prof. dr. Anne van Streun ▪ University of Groningen
Dr. Cees Terlouw ▪ University of Twente

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Vos, Pauline

Like an ocean liner changing course: the grade 8 mathematics curriculum in the Netherlands, 1995-2000
Thesis University of Twente, Enschede – With refs. – With Dutch summary
ISBN 90-365-17-40-0

Cover design: 'Holland-Amerika Lijn' by Willem ten Broek, adapted by Deesign
Lay-out: Sandra Schele
Press: PrintPartners Ipskamp - Enschede

LIKE AN OCEAN LINER CHANGING COURSE

THE GRADE 8 MATHEMATICS CURRICULUM
IN THE NETHERLANDS, 1995-2000

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. F.A. van Vught,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op 26 juni 2002 om 13.15 uur

door

Francisca Pauline Vos

geboren op 16 augustus 1960
te Maurik

Promotor: Prof. dr. Tj. Plomp

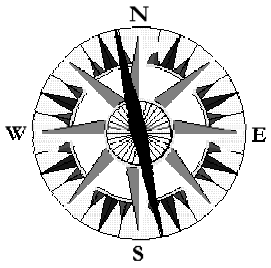
Assistent- promotor: Dr. W.A.J.M. Kuiper

The ocean liner

The ocean liner takes students on an educational journey. The trip can offer splendid views, wide horizons, jumping whales, and stunning sunsets. It is supposed to be a marvellous experience. There can be parties, but there can be seasickness as well.... At the end, after having crossed the wide seas, the passengers enter a new continent, adulthood.

The captain has guided many generations across the sea. He (or she) has many years of experience in steering the vessel along cliffs and sandbanks. His salary is low, considering the enormous responsibility over the passengers and their well being.
New technologies have been installed on the bridge. But the captain was given little time to familiarise himself with them, as the ship has to sail continuously.
The captain is not the only one who determines the track of the ship. Not all circumstances can be controlled. Storms howl, tides fall, and there can be icebergs ahead. Nobody wants to think of disaster. The best coaches are in the sands: the shipowner and the coast guard, giving directions for the intended course.

Table of Contents



APPENDICES

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

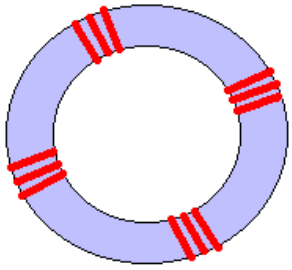
alpha, α	reliability coefficient measuring internal consistency
APS	Algemeen Pedagogisch Studiecentrum - National Center for School Improvement
avg.	average
χ^2-test	test to verify the independency of two independently conducted studies
CIEAEM	Commission Internationale pour l'Étude et l'Amélioration de l'Enseignement des Mathématiques
Cito	Instituut voor Toetsontwikkeling - National Institute for Educational Measurement
CMLW	Commissie Modernisering Leerplan Wiskunde - Committee on Modernisation of the Secondary School Mathematics Curriculum
correlation	indication of the mutuality of two variables
cov.	OTL-covered
COW	Commissie Ontwikkeling Wiskundeonderwijs - Commission for Development of Mathematics Education
FEO, HvU	Faculteit Educatieve Opleiding, Hogeschod van Utrecht - a Pre-Service Training Institute
Fi	Freudenthal Institute
FIMS	First International Mathematics Study
FISS	First International Science Study
GWA	Geïntegreerde Wiskundige Activiteiten - Integrated Mathematical Activities
ICME	International Commission on Mathematical Instruction
IEA	International association for the Evaluation of educational Achievement
IEAP	International Assessment of Educational Progress
intl.	international
instrument	tool for measuring (e.g. questionnaire, test)
inter-test	between the TIMSS Written Test and the TIMSS Performance Assessment
intra-curricular	between the appearances of a curriculum
IOWME	International Organisation for Women and Mathematics Education
IOWO	Institute for Development of Mathematics Education
IPMA	International Programme on Mathematical Attainment
ISC	International Study Center, Lynch School of Education, Boston College
item	small unit in an instrument (e.g. a mathematics problem in a test)

item-curriculum matching index	
	on a nominal scale: yes/no to the question whether the item matches with the intended curriculum; on a ratio scale: percentage of curriculum experts indicating that the item matches with the intended curriculum
Kerndoelen voor de basisvorming	
	core curriculum for junior secondary secondary schools
METRIC	Mathematics Education - Trends in and Relations between the Intended, implemented and attained Curriculum
nom.	on a nominal scale
nominal scale	categorical scale, numerical in name only
NOT	National Option Test
NVvW	Nederlandse Vereniging van Wiskundeleraren - Dutch Association of Mathematics Teachers
OECD	Organization for Economic Co-operation and Development
ordinal scale	ranking scale
OTL	Opportunity To Learn
OW&OC	Research Group on Mathematics Education and Educational Computer Centre
PA	TIMSS Performance Assessment in mathematics for grade 8
PA-1995	TIMSS Performance Assessment in mathematics for grade 8, carried out in 1995 as part of a mathematics and science test
PA-2000	TIMSS Performance Assessment in mathematics for grade 8, carried out in 2000 as part of a mathematics and science test
PISA	Programme for International Student Assessment
PME	International Study Group for Psychology of Mathematics Education
p-value	proportion of students with correct responses to a test item
rat.	on a ratio scale
ratio scale	interval scale with true zero point
rekenen	arithmetic
reliability	evidence of consistency of measurements
replication	repeating a study to see whether the original findings reappear
RME	Realistic Mathematics Education
robust	an effect strong enough to appear under a variety of conditions
sample	subgroup selected from a population in such a way as to accurately represent that population
scatter diagram	plot of scores based on two variables
SE	standard error
sign test	test to compare paired data from two independent samples
SIMS	Second International Mathematics Study

SISS	Second International Science Study
SLO	Stichting Leerplan Ontwikkeling - national institute for curriculum development
TCMA	Test Curriculum Matching Analysis
test-curriculum matching index	percentage of test items matching with the intended curriculum
TIMSS	Third International Mathematics and Science Study
TIMSS-95	TIMSS, carried out in 1995 in approx 40 countries
TIMSS-99	TIMSS, carried out in 1999 in 38 countries (equiv. TIMSS-R)
TIMSS-R	Third International Mathematics and Science Study - Repeat (equiv. TIMSS-99)
trend	changes over time
tst.	OTL-testing
t-test	test to compare the means of two independent samples
UNESCO	United Nations Educational Scientific and Cultural Organization
validity	evidence-based judgement that an instrument measures what it is intended to measure
VOCL-project	Voortgezet Onderwijs Cohorten Leerlingen - Secondary Education Cohorts Students
W12-16	project group for the design of a new mathematics curriculum for junior secondary schools
Wiskunde Werkgroep	Higher Mathematics Workgroup
wiskunde	higher mathematics
WT	TIMSS Written Test in mathematics for grade 8
WT-1995	TIMSS Written Test in mathematics for grade 8, carried outin 1995 as part of a mathematics and science test
WT-1999	TIMSS Written Test in mathematics for grade 8, carried out in 1995 as part of a mathematics and science test

ABBREVIATIONS OF ABILITY TRACKS AT DUTCH SECONDARY SCHOOLS

Vbo	Voorbereidend beroepsonderwijs [pre-vocational education], grade 7-10
Mavo	Middelbaar algemeen voortgezet onderwijs [middle general secondary education], grades 7-10
Havo	Hoger algemeen voortgezet onderwijs [higher general secondary education], grades 7-11
Vwo	Voorbereidend wetenschappelijk onderwijs [pre-university education], grades 7-12



Acknowledgements

~ Izandla ziyahlambana.~

The hands wash each other (collaboration improves results).
(ZULU PROVERB)

The realisation of this dissertation was impossible without the assistance of many people. The following persons have been particularly important:

- My supervisor Tjeerd Plomp and my co-supervisor Wilmad Kuiper, for initiating the research project, guiding me along pitfalls, and sharing their wisdom.

- Klaas Bos, Douwe Kok, Lambrecht Spijkerboer, Rien Steen, and Nellie Verhoef for their readiness in assisting me on methodological and contextual issues.

- Students, teachers and mathematics curriculum experts, for taking part in the study, and to whom I still owe an answer to the question: *"what is in it, for us?"*

- Staff members of the Department of Curriculum at the Faculty of Educational Science and Technology at the University of Twente, for harbouring me during these years.
- René Almekinders, for his patience and the many hilarious hours of discussion.
- Melody Williams, for polishing up the phrasing.
- Linda Odenthal, for sharing hardships.
- Sandra Schele, for the gimmicks and the professional lay-out of this dissertation.

- Deb Andrzejewski, Aleid Diepeveen, Marijke Huitink, Fatima Jawara, P'tje Lanser, Sue Maluwa-Bruce, Nienke Nottelman, Sjoukje Postma, Kees Roozemond, Peter Tabbers, Gerline Wildeman, Anna Zwijnepoel, and the members of Hobihe and Hippo (including the horses), for offering distraction.
- Harry and Winnie Krebbers-Bongaards, for always asking *"why mathematics?"*

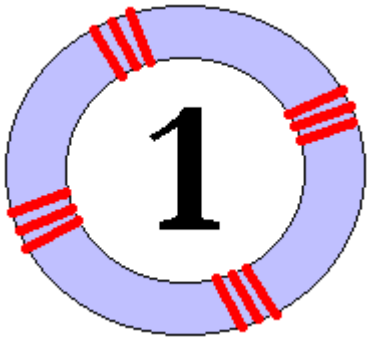
- Miriam Kwari, for making me change the course of my career.

- My family, in-laws, step-family, and friends for continuous tolerance.
- Tin Wolterbeek Muller, for concerns, considerations and care.

- Désiré Baartman, for being present or absent at exactly the right moments, and for an unwarranted belief in me. *Musikanakomana uyu akareba kwazvo. Handei ku vamos.*

Amsterdam/Enschede, spring 2002.

Pauline Vos

Chapter 

Introduction

~ Sormak ayip degil, sormamak ayip.~

It is not disgraceful to ask, it is disgraceful not to ask.
(TURKISH PROVERB)

This first chapter is an introduction to the METRIC study. Section 1.1 provides a first orientation on terminology. Section 1.2 reports on the origins of the study. Section 1.3 explains how the study is structured in different dimensions. Finally, section 1.4 introduces the research problem.

1.1 PREAMBLE

Students' achievements can be unpredictable. In 1995, Dutch 14-year old students were given a mathematics test, the results of which were compared internationally. The test was not considered very appropriate for Dutch students, because it deviated from their curriculum. Nevertheless, the students performed well on this test. Their score was above the international average score. In that same year, 1995, they were also given another international mathematics test. This second test was a performance test. It had a more practical nature and was considered much more appropriate for Dutch students. However, contrary to expectations, the students did **not** perform as expected on that test. Their performance was near the international average.

The notable results were the reason for starting a study on finding explanations for the unexpected phenomena in Dutch mathematics education. This study is described in this dissertation.

The study focused on curriculum as a context for explaining the phenomena. Different appearances of a curriculum were distinguished:

- the *intended* curriculum, at system level (what society expects that students should master),
- the *implemented* curriculum, at school and classroom level (the content as it is interpreted by the teachers and taught to the students),
- the *attained* curriculum, at student level (knowledge, skills and attitudes that students have acquired indeed).

This curriculum framework was developed for international comparative studies conducted by the International Association for the Evaluation of Educational Achievement (IEA). The framework was based on prior fundamental work by Goodlad (Goodlad & Richter, 1966; Robitaille & Garden, 1996; Robitaille et al., 1993; Travers & Westbury, 1989). The framework isolates factors for interpreting students' learning at three different levels: system level (macro level), classroom level (meso level), and the individual student's level (micro level).

At the three levels, the *intended, implemented* and *attained* curriculum can be found. The framework is illustrated in Figure 1.1.

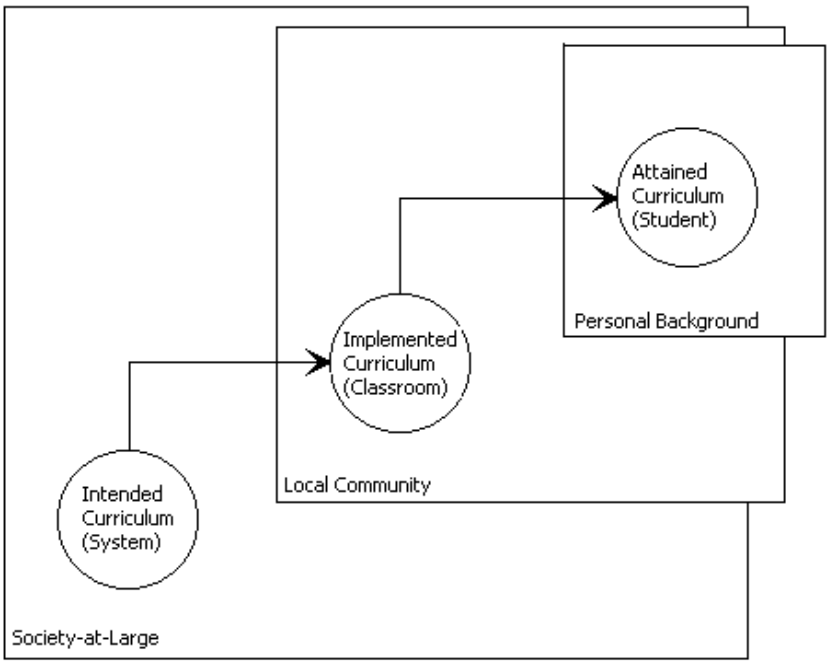


Figure 1.1: IEA's curricular appearances
Source: Robitaille & Garden (1996).

Other authors also use this tripartition for curricular appearances, though sometimes in other terminology. The *intended* curriculum is also referred to as: the 'planned', 'formal', 'expected', 'official', 'explicit' or 'sanctioned' curriculum. The *implemented* curriculum is also referred to as the 'operational', 'enacted', 'taught' or 'implicit' curriculum. The *attained* curriculum is also referred to as the 'experienced', 'realised' or 'learned' curriculum (Howell & Nolet, 2000).
Some authors have provided more detailed curriculum models. For example Goodlad, Klein and Tye (1979) identify five curricular appearances. At system level, they distinguish between (1) the ideal curriculum (the original assumptions, visions and intentions of curriculum developers) and (2) the formal curriculum (textbooks, teacher guides, and exemplary tests). At classroom level, they distinguish between (3) the perceived curriculum (interpretations of the teachers) and (4) the operational curriculum (the instructional process). At student level, they apprehend (5) the experienced curriculum (what students have experienced). Van den Akker (1988, 1998) follows this system, but adds another curriculum appearance at student level: (6) the *attained* curriculum (the learning results of the students).
At student level, further distinctions can be made. Not everything that was learnt will be included in a test or can actually be tested (Hawker & Ollerton, 1999; Leinhardt, 1983). Therefore, one could furthermore distinguish the experienced curriculum and the learned curriculum from (7) the tested curriculum. However, for brevity, this dissertation adheres to the three IEA curriculum appearances as stated before: the intended, implemented and attained curriculum.

The following example may serve as an illustration of the appearances of a curriculum at the macro, meso, and micro level. At the system level, governmental documents may state educational goals for the students of their nation. Thus, in many countries, the intended mathematics curriculum for students at primary schools includes 'doing long division'. Consequently, it is taught to the students. Yet, despite the clarity of the intended curriculum and despite the universality of the mathematical language (Pimm, 1987; Skemp, 1986), the teaching and learning activities surrounding this basic mathematical operation can be strikingly divergent. The interpretation of 'doing long division' differs considerably between countries. The current classroom notations of a simple division exercise in four neighbouring European nations are shown in Figure 1.2.

Question: Divide 544 by 34.

A	B	C	D
34/544\16	544:34=16	544|34	16
34	34	204|16	34)544
204	204	000	-34
204	204		204
0	0		-204
			000

Figure 1.2: Division algorithms from (A) the Netherlands, (B) Germany, also used in the Netherlands, (C) France, and (D) England
Source: Berwald (1988).

In all cases, the answer to the division exercise is identical. However, the different notations in Figure 1.2 show that students are taught to act according to different algorithms. The algorithm from France (C) is clearly shorter in writing. Here, the intermediate subtraction steps are carried out implicitly, requiring more mental arithmetic than the other three algorithms. Therefore, the classroom activities associated with 'doing long division' can differ between countries. In other words, the implemented curriculum of 'doing long division' differs across countries, while the intended curricula state the same goal. As a result, the learning processes associated with the activities can differ as well. It could have different effects on students' outcomes, which is the attained curriculum.

This dissertation describes a study of the intended, implemented and attained Dutch mathematics curriculum at grade 8 level (the second year of secondary education). The students at this level were on average 14 years old. The study covers the last five years of the 20th century (1995-2000). In order to explain the unexpected results, which were described at the beginning of this section, the study compared the attained mathematics curriculum with the intended curriculum and the implemented curriculum. The study is identified by the acronym METRIC. The six letters stand for a study in **M**athematics **E**ducation, investigating **T**rends in, and **R**elations between the **I**ntended, implemented and attained **C**urriculum.

1.2 THE ORIGIN OF THE METRIC STUDY

1.2.1 TIMSS

In 1995, an international comparative research in science and mathematics education was carried out. The name of this research was Third International Mathematics and Science Study (TIMSS). TIMSS was conducted under the auspices of the IEA. After its foundation in 1959, the IEA initiated a series of international comparative studies on reading and literacy, educational use of Information and Communication Technology (ICT), and mathematics and science. One of its founding fathers, Torsten Husén, explained that comparison of educational systems of countries could replace educational experiments, which are often impractical and unethical. The variety of educational systems across the world offers a series of different learning environments that can be studied instead of doing experiments (Husén & Bloom, 1967). The IEA gained experience in large-scale international comparative projects. The successive studies in mathematics and science education were:

- the First International Mathematics Study (FIMS), carried out in 1964,
- the First International Science Study (FISS), carried out in 1970,
- the Second International Mathematics Study (SIMS), carried out in 1981,
- the Second International Science Study (SISS), carried out in 1984, and
- the Third International Mathematics and Science Study (TIMSS), carried out in 1995, merging the two subject areas mathematics and science into one study.

The general aim of TIMSS is the international comparison of curricular outcomes and contexts in mathematics and science education. This should lead to a knowledge base for the analysis of strong and weak points at system, class and student level. TIMSS has been designed to provide educational policy makers, curriculum specialists and researchers with a knowledge base to describe and understand the performance of educational systems from an international perspective. In TIMSS, students' achievement is measured through tests, which are identical in all countries (apart from the language). To interpret achievement outcomes, TIMSS gathers additional information through questionnaires for students, teachers and school principals. The ultimate goal is to find indicators for improving mathematics and science education (Beaton et al., 1996; Mullis et al., 2000).

In 1995, TIMSS defined three populations based on three educational levels. These were at primary school level (Population 1), junior secondary school level (Population 2), and senior secondary school level (Population 3) (Martin, 1996). Population 2 is of interest to the METRIC study. This population was defined as the two adjacent grades (grades 7 and 8) with the largest proportion of 13-year old students at the time of testing. For this population, TIMSS was carried out in 41 countries, one of these being the Netherlands.

In TIMSS, students' achievement was measured through a conventional written, time-restricted mathematics and science test. The METRIC study focuses on the mathematics part only. This part consisted of standard items, testing for a wide range of mathematical abilities, such as knowledge (e.g. characteristics of a parallelogram), procedural skills (e.g. adding fractions) and problem solving (e.g. modelling data). Most items were of the multiple choice format, some items required a short answer, and a few items required an extended answer.
Besides the written test, TIMSS developed another time-restricted test, which had a practical nature comprising hands-on tasks. It was administered at grade 8 level (the upper grade of Population 2). The Netherlands participated in this optional test. This test was called the TIMSS Performance Assessment. As the Performance Assessment will prove important in the forthcoming text sections, it is briefly characterised in the next paragraph. Chapter 4 presents a further elaboration on this test. To make a clear distinction between the two TIMSS tests, from hereon, the written test will be named the TIMSS Written Test.
The TIMSS Performance Assessment complements the TIMSS Written Test. An important reason for measuring students' achievement through this additional test is, that the Written Test mainly tests for content and to a lesser extent for skills. The TIMSS Performance Assessment mainly tests for practical skills. Students are provided with manipulatives (plasticine, magnets, folding paper, adhesive tape, etc.) and instruments (thermometer, ruler, balancing scales, etc.). They are tested through open tasks which require practical skills, such as: designing and executing an experiment, observing and describing observations, using calculators, looking for regularities, finding notations, interpreting their measurements, and so forth. This test was developed from a vision in science education, which seeks coherence between procedural, declarational and conditional cognition. Practicals are no longer seen as mere illustrations of concepts taught. Students are expected to investigate systematically, contrary to cookbook practicals. Seeking and providing explanations for explored phenomena is then part of assessment (Garden, 1999; Harmon et al., 1997; Kind, 1999).

As for mathematics, the TIMSS Performance Assessment can be associated with Gal'perin's view of *learning by doing* in which mental acts (manipulating objects in the mind) develop from material acts (manipulating tangible objects). Piaget (1952) suggested that students need many experiences, including experiences with concrete materials, for learning of mathematical concepts. Skemp's (1986) theories supported the belief that students' interactions with physical objects formed the basis for learning at an abstract level. Yet, manipulatives are mostly used for the demonstration of mathematical concepts, in particular at the primary level (Szendrei, 1996). Many teachers perceive manipulatives suitable for 'fun mathematics', and not related to 'real mathematic' (Moyer, 2002). The use of manipulatives for investigative and assessable mathematical activities in secondary schools is relatively new, and has been advocated by a considerable number of mathematics educators in the Netherlands (Van Dormolen, 1993; Van Dormolen & Zwaneveld, 1992; Verhoef & Bos, 1992).

1.2.2 TIMSS in 1995 in the Netherlands

In 1995, TIMSS was carried out in the Netherlands after a major curriculum change had taken place. Two years earlier, in 1993, a new law on basic education (Wet op de Basisvorming) came into force. It contained a national compulsory curriculum for junior secondary schools. As minimum requirements, *core objectives* (kerndoelen) were formulated for all students. Keywords in this curriculum reform for all subjects were *applications*, *skills* and *coherence* (Van Luyn, 1998).

The mathematics curriculum was adapted, engaging students through real-life contexts and integrated topics for practical, applied skills. It is based on a treatise, which is identified as Realistic Mathematics Education (RME). More details about history, backgrounds and distinguishing features of this new, innovative mathematics curriculum will be given in chapter 2.

The Netherlands measures 200 km (east-west) by 300 km (north-south). In 1995, there were 15 million inhabitants. There were ± 1500 secondary schools with ± 10000 mathematics teachers. There were ± 200000 grade 8 students.

Students' secondary education started a year earlier in grade 7 (=secondary 1). From then onwards, students sat in classes, which were tracked according to abilities, until their final exam in grade 10 (*vbo* and *mavo*), in grade 11 (*havo*), or grade 12 (*vwo*). At grade 8 level, schools often still had combined classes with overlapping ability tracks (e.g. there were grade 8 classes for *vbo/ mavo* or for *havo/ vwo*).

The average class size was 25. The classes for the lower ability tracks *vbo* and *mavo* had a lower number of students (average class size = 22) than the classes for the higher ability tracks *havo* and *vwo* (average class size = 30)

(Inspectie van het Onderwijs, 1998).

Before TIMSS was carried out in 1995, serious doubts arose among mathematics educators in the Netherlands whether the TIMSS Written Test would do justice to the Dutch target population. The TIMSS Written Test reminded the educators of the former abandoned curriculum. It was deemed too reproductive and the items were considered 'bare' without real-life contexts. Curriculum experts also objected to the large number of multiple choice items in the TIMSS Written Test (De Lange, 1997a, 1997b). Tests within the new RME-based curriculum differed considerably from the TIMSS Written Test. It was reasoned that the RME-based curriculum did not prepare students to respond to items such as those from the TIMSS Written Test. Therefore, it was questioned whether Dutch students would be able to display their particular knowledge and skills.

The international curricular diversity was a serious point of concern to the TIMSS International Study Center, which was responsible for the development of the TIMSS Written Test. The goal was to develop an international test, which would be *equally unfair* for all participating countries. Therefore, subject matter specialists from all countries were consulted and they were asked to contribute to the process of test item development (Garden & Orpwood, 1996). A few items were brought in by Dutch mathematics educators. Yet, when scrutinising all mathematics test items in light of the Dutch intended curriculum, it was established that only 71% of the test items matched with the new mathematics curriculum. Most other participating countries in TIMSS had an intended mathematics curriculum that matched with more than 90% of the items (Beaton, Mullis et al., 1996; Kuiper, Bos & Plomp, 1997). Table 1.1 gives the percentages of test items, which matched with the respective intended curricula of the various countries. In the METRIC study, this percentage is termed the *test-curriculum matching index*. It is an indicator of the appropriateness of a test in light of a nation's intended mathematics curriculum. Further information on this concept will be given in chapter 3.

Table 1.1: Selected nations and their test-curriculum matching index of the TIMSS-95 Written Test (percentage of items matching the intended mathematics curriculum)

Nation	test-curriculum matching index (%) *(n=157)*	Nation	test-curriculum matching index (%) *(n=157)*
United States	100	Korea	92
Hungary	100	Canada	91
Latvia (LSS)	99	Singapore	90
Israel	98	New Zealand	90
Germany	96	Romania	88
Lithuania	96	France	86
Australia	95	Belgium (Fl..)	86
Slovak Rep.	94	Denmark	83
Japan	94	England	81
Slovenia	93	South Africa	80
Norway	93	Russian Fed..	78
Czech Rep..	92	Cyprus	76
Hong Kong	92	Bulgaria	74
Iran, Isl. Rep.	92	Netherlands	71

Source: Beaton et al. (1996).

With a lower test-curriculum matching index for the Netherlands, it was expected that students of other countries would outperform Dutch students. However, contrary to the expectations, in 1995 Dutch grade 8 students performed well on the TIMSS Written Test. Their score was significantly above the international average, just below the four Asian top-scoring countries Singapore, Korea, Japan and Hong Kong (Beaton et al., 1996).

Kuiper et al. (1997, 2000) carried out additional research to establish whether the mathematical achievement of Dutch students on the TIMSS Written Test was statistically well measured. They developed an additional test, the National Option Test (NOT). NOT was based on the new RME-based curriculum, containing a selection of items from the TIMSS Written Test considered relevant to the intended curriculum, together with separately developed, RME-based items. All items were selected on the criterion that they matched well with the intended curriculum. As the TIMSS Written Test and NOT contained an overlap of 16 items, the measurement through both tests could be compared. Based on

students' achievement results on both NOT and the TIMSS Written Test, Kuiper et al. concluded that the Dutch students' achievement was well measured through the TIMSS Written Test.

Beaton et al. (1996) carried out another research to investigate whether the TIMSS Written Test did justice to the intended curriculum of various countries. This research, which was named the Test Curriculum Matching Analysis (TCMA), requested curriculum experts of all participating countries to indicate for all test items whether they matched with their intended curriculum. For all countries, the test items were selected that matched with the intended curriculum. Based on the selections, all country's scores were recalculated. As mentioned before, for the Netherlands, 71% of the items in the Written Test were considered to match with the intended curriculum. The score of Dutch students on this set of selected items was exactly equal to their scores on the full test. Somehow Dutch students were knowledgeable about the 29% of test items that were *remote* from their intended curriculum.

From the research through the National Option Test (NOT) by Kuiper et al. (1997, 2000), and through the Test-Curriculum Matching Analysis (TCMA) by Beaton et al. (1996), it was concluded that the achievement of Dutch students was well measured by TIMSS. The portion of 71% of test items that matched with the intended, new curriculum gave students enough room to display their abilities. Moreover, they had the abilities for *transfer* of their knowledge and skills to items that did not match with their intended curriculum. It was assumed that when learning mathematics through real-life contexts and integrated topics, students could also use their abilities to answer items requiring isolated knowledge. An alternative explanation was, that teachers still followed the abandoned curriculum or integrate advanced content, characteristic of higher grades, into their present teaching (Kuiper et al., 1997). However, the two studies on NOT and TCMA only linked the intended and the attained curriculum. Further studies were needed to supplement findings with information on the implemented curriculum.

The TIMSS Written Test and the TIMSS Performance Assessment were administered simultaneously. Contrary to the Written Test, the Performance Assessment seemed to match well with the objectives of the new, context-based curriculum. Dutch mathematics curriculum experts welcomed the majority of the mathematical tasks of the Performance Assessment, because the tasks comprised investigative activities within concrete experiences (Bos et al., 2001; Kuiper et al.,

1997). Students were tested on their flexibility in data modelling and interpreting. These abilities were an integral part of the new mathematics curriculum. According to additional, national research, 88% of the mathematics items in this test were considered to match with the intended curriculum (De Haan, Pakkert & Van der Meij, 1997).

Table 1.2: Achievement results of countries participating in both TIMSS mathematics tests, 1995

TIMSS Written Test Achievement on mathematics items			**TIMSS Performance Assessment Achievement on mathematical tasks**		
	Country	*Scale points*		*Country*	*Avg p-value*
1	Singapore	643	1	Singapore	70
2	Czech Rep	564	2	Switzerland	66
3	Switzerland	545	3	Australia	66
4	Netherlands	541	4	Romania	66
5	Slovenia	541	5	Sweden	65
6	Australia	530	6	Norway	65
7	Canada	527	7	England	64
8	Sweden	519	8	Slovenia	64
	Intl average	*509*	9	Czech Rep	62
9	New Zealand	508	10	Canada	62
10	England	506	11	New Zealand	62
11	Norway	503	12	Netherlands	62
12	USA	502	13	Scotland	61
13	Scotland	498		*Intl average*	*59*
14	Spain	487	14	Iran	54
15	Romania	482	15	USA	54
16	Cyprus	474	16	Spain	52
17	Portugal	454	17	Portugal	48
18	Iran	428	18	Cyprus	44
19	Colombia	385	19	Colombia	37

Source: Bos et al. (2001).

Despite the curriculum experts' positive judgement on the appropriateness of the TIMSS Performance Assessment, the Dutch students did **not** score outstandingly well on this test. Their achievement, measured as average p-value (average percentage correct), was near the international average, while they scored significantly above the international average on the Written Test. Table 1.2 displays the international comparative results of all countries participating in

both tests. Comparing the rankings, most countries hold approximately the same position on both tests (±3 positions). However, there were three exceptions: Romania, the Czech Republic and the Netherlands. The latter took a large nosedive, with a difference of eight positions between the written and the practical test (Bos, Kuiper & Plomp, 2001).

1.2.3 Different discrepancies

The results of Dutch grade 8 students on the Written Test and the Performance Assessment in TIMSS-95 showed a striking discrepancy. On the Written Test, the students scored above the international average and on the Performance Assessment they scored near the international average. This discrepancy is termed as an *inter-test achievement discrepancy*. It is a discrepancy at the level of the attained curriculum.

There was another discrepancy in the TIMSS results. The judgements of the curriculum experts on the appropriateness of the two tests conflicted with students' achievement. The Written Test was judged less appropriate by the experts, while the students attained a high score. For the Performance Assessment, the discrepancy was the other way around. This test was judged appropriate by the experts, while the students attained an average score. Therefore, there was a discrepancy between the intended curriculum and the attained curriculum. This discrepancy will be indicated as an *intra-curricular discrepancy*. It occurred twice: there was an intra-curricular discrepancy on the Written Test. and there was an intra-curricular discrepancy on the Performance Assessment.

All discrepancies are shown in Table 1.3, with the discrepancies in two dimensions, (a) between the tests, and (b) between the intended and the attained curriculum. Vertically, there are the two inter-test discrepancies: (1) the judgement in light of the intended curriculum differed between the tests, and (2) the achievement results differed between the tests. Horizontally, there are the two intra-curricular discrepancies on both tests. With no data available on the implemented curriculum, it is omitted from the table.

Of the four possible discrepancies, there was one discrepancy that did not require an explanation. It was the discrepancy of the experts' judgement on the appropriateness of the two tests. This judgement was taken for granted, being a result of the different characteristics of the two tests. However, the other three discrepancies required an explanation.

Table 1.3: The two TIMSS mathematics tests in light of the Dutch intended and attained curriculum, 1995

Mathematics tests	Match with intended curriculum	Match with attained curriculum
TIMSS Written Test	Low test-curriculum matching index (71%)	Achievement result significantly above international average
TIMSS Performance Assessment	Satisfactory test-curriculum matching index (88%)*	Achievement result near international average

Note: * international comparative data unavailable.

The inter-test achievement discrepancy, at the level of the attained curriculum, revealed that an understanding of students' performance was needed: why did the scores differ between the two tests?
The two intra-curricular discrepancies revealed that an understanding of students' performance was needed, with respect to the intended curriculum: why was students' achievement on both tests not aligned with the experts' judgement?

These questions gave rise to considerations on a number of points. Could it be that the discrepancies were related to the novelty of the RME-based curriculum? The new curriculum was legislated in 1993, to be introduced in yearly phases. In the school year 1993/1994, this new curriculum was only introduced to the grade 7 cohort. In the following school year, 1994/1995, it was introduced to the next level, being grade 8. Hence, one and the same cohort of students was the first one to be introduced to the new curriculum each year. It was this cohort of students that was tested through TIMSS in spring 1995, being in grade 8. Consequently, the TIMSS tests coincided with the first year in which the new curriculum formally had been implemented at grade 8 level. It could be that the mathematics teachers were still unfamiliar with the new curriculum. Perhaps they needed a few years to adapt to the new content. Furthermore, how would they interpret its intentions? The learning process of teachers might need time (Fullan, 1991).
Indicative answers to the questions concerning the implemented curriculum were found, for example, in Kuiper et al. (1997). They reported that in the school year 1994/1995, only half of the Dutch mathematics teachers in grade 8 used textbooks based on the new curriculum. The remaining teachers used textbooks based on the abandoned curriculum. Therefore, in 1995, at the time of the

administration of both the Written Test and the Performance Assessment, the new curriculum was not yet fully implemented. Other indicators of a trailing implementation of the new curriculum are found in the work of De Haan et al. (1997). They reported on Dutch mathematics teachers' judgement on the appropriateness of the mathematics tasks in the TIMSS Performance Assessment of 1995. According to their findings, the teachers had only covered the content of approximately one third of the tasks in their lessons. Therefore, if the mathematics teachers indicated that the mathematics content needed for the Performance Assessment was not covered in their lessons, how could Dutch students then be expected to perform above the international average?
This possible explanation showed that the discrepancy between the intended and the attained curriculum could be closely related to a discrepancy between the intended and the implemented curriculum (i.e. experts' judgement on the appropriateness of the tests versus teachers' judgement on the appropriateness of the tests). It also related to an institutionalisation process with a time dimension. It could be that the discrepancies would decrease after a few years. Therefore, a replication of both the Written Test and the Performance Assessment was imperative for finding possible explanations for the discrepancies. This replication needed to include research at the level of the intended, the implemented and the attained curriculum. The added time dimension offered the advantage of establishing a trend.

1.3 METRIC'S DIMENSIONS

1.3.1 Three dimensions

The following sections explain how the METRIC study was built on three dimensions: (a) a test dimension, (b) a time dimension, and (c) a curriculum dimension. The test dimension concerns two different tests: the TIMSS Written Test and the TIMSS Performance Assessment. The difference between the two tests has already been explained in the previous sections. The time dimension emerged as a result of repeating the tests (see below). The curriculum dimension is based on the IEA curriculum framework, which distinguishes the intended, the implemented and the attained curriculum.

1.3.2 A time dimension

It was a mere coincidence that in 1995, TIMSS came so shortly after the formal introduction of the new core curriculum in 1993. At this time, TIMSS offered a unique opportunity to take a snapshot of students' achievements under the new curriculum. However, as explained before, it was felt that the same measurement at a later stage was also needed to assess whether the new curriculum was being implemented and a trend could be established. This opportunity came when IEA decided to repeat the TIMSS Written Test. This replication study was carried out in 1999. This new project was named TIMSS-99 (or TIMSS-Repeat, TIMSS-R) and thereafter the original TIMSS of 1995 was described as TIMSS-95. TIMSS-99 was conducted for the higher grade of Population 2 (grade 8) with the Written Test only. Thus, a picture as visualised in Table 1.4 emerged. There is a time dimension (a trend from 1995 to 1999) and a test dimension.

Table 1.4: Time and test dimensions in the METRIC study

Mathematics tests	**1995**	**1999**
TIMSS Written Test	Conducted	Conducted
TIMSS Performance Assessment	Conducted	--

Note: Dashes indicate the study was initially not anticipated for.

In the Netherlands, it was felt that repeating the TIMSS Performance Assessment was necessary to study the discrepancies mentioned earlier. Therefore, the METRIC study undertook this task of repeating the TIMSS Performance Assessment to fill the missing cell of Table 1.4. Unfortunately, no other country was interested in a repeat of the TIMSS Performance Assessment, leaving this replication to be a national Dutch study only.

The repeat of the Performance Assessment was conducted in 2000, one year after the repeat of the Written Test. The simultaneous administration of both tests had practical constraints, both for the researchers and the schools. By delaying the repeat of the Performance Assessment by one year, the two test administrations were spread across two years, 1999 and 2000. With a one-year difference, the Performance Assessment of 2000 offered a reliable reference to the Written Test of 1999. There were no reasons to assume that the population of 1999 differed significantly from the population of 2000. The four/five years time span gave teachers ample time to settle to the new intended curriculum.

Thus, there were four data sets in the METRIC study: data for the Written Tests of 1995 and 1999, and data for the Performance Assessments of 1995 and 2000. The tests were abbreviated as WT-1995, WT-1999, PA-1995 and PA-2000. The goal of the METRIC study was to use the data to explore whether the discrepancies, as described before, re-occurred and to find possible explanations for their occurrence.

1.3.3 The curriculum dimension

In the METRIC study, for all four tests, WT-1995, WT-1999, PA-1995 and PA-2000, the following data were gathered:

- curriculum experts were consulted to judge whether the test items matched with the intended curriculum,
- mathematics teachers were consulted to judge whether the test items matched with the implemented curriculum, and
- students' achievement on the tests was taken, representing the attained curriculum.

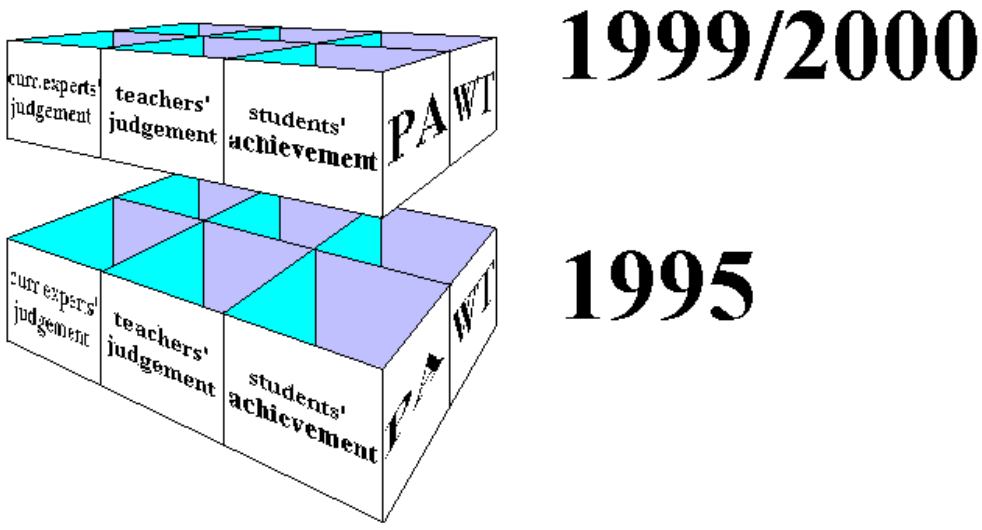


Figure 1.3: Dimensions in data collection in the METRIC study
Legend: PA = Performance Assessment; WT= Written Test.

There were twelve measurements, built on the three dimensions with two tests at two stages and three curricular appearances. The twelve data sets are visualised in Figure 1.3. The time dimension is pictured vertically, the curriculum dimension horizontally from left to right, and the test dimension from front to the rear.

The data were needed for a better understanding of the inter-test achievement discrepancy, as mentioned before. Also, the data were needed for a better understanding of the intra-curricular discrepancies. The discrepancies could occur between the intended, implemented and attained curriculum. Findings of the METRIC study could inform stakeholders about: (1) the reform process of the new Dutch mathematics core curriculum for junior secondary schools in particular, and (2) of curriculum reform processes in general. Additionally, experience would be gained in the sound replication of nation-wide testing, especially pertaining the innovative practical test.

1.4 THE RESEARCH PROBLEM OF THE METRIC STUDY

1.4.1 Problem statement

The problem statement of the METRIC study was formulated against the background of the Dutch TIMSS results in 1995, while anticipating a replication of both the TIMSS Written Test in 1999 and the TIMSS Performance Assessment in 2000. It aimed at establishing a trend, which in turn could enhance the evaluation of the ongoing reform process of the new mathematics curriculum. The problem statement is formulated as follows:

> *To what extent are the inter-test achievement discrepancy and the intra-curricular discrepancies on the TIMSS Written Test and the TIMSS Performance Assessment in 1995 similar in 1999/2000 and to what extent can the discrepancies be explained in light of a trailing implementation of the new mathematics core curriculum for junior secondary schools in the Netherlands?*

This problem lead to two research questions. The first question focuses on the inter-test achievement discrepancy and the second question on the intra-curricular discrepancies.

1.4.2 The first research question

The first research question centres on students' achievement of students on both tests at two stages. In 1995, Dutch students performed relatively well on the Written Test and averagely on the Performance Assessment. The database for this question consists of the results at the level of the attained curriculum. These are illustrated in Figure 1.4.

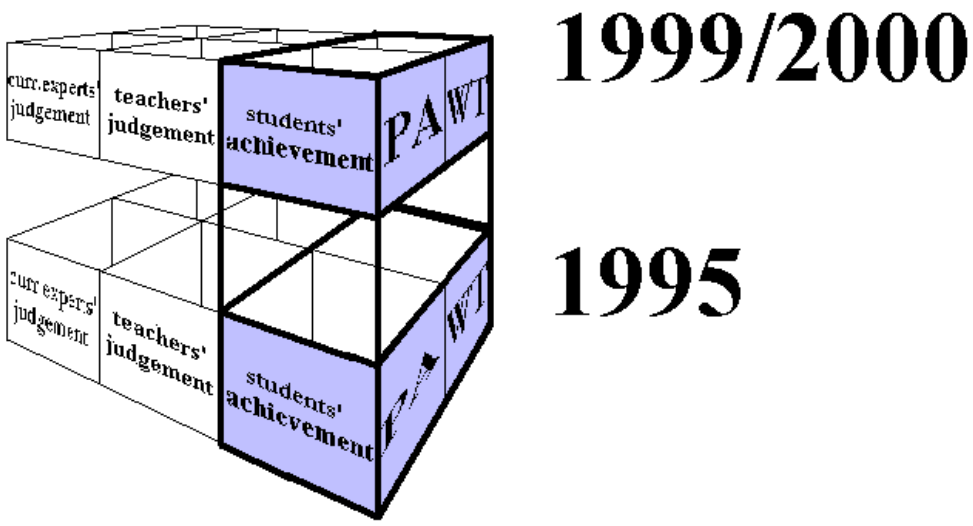


Figure 1.4: Data set for the first research question

There are two dimensions to be considered:

- vertically, a time dimension: a possible trend between 1995 and 1999/2000 and
- to the rear, a test dimension: the TIMSS Written Test and the TIMSS Performance Assessment.

The first question is:

> *To what extent does a repeat of the TIMSS Written Test in 1999 and a repeat of the TIMSS Performance Assessment in 2000, result in an inter-test achievement discrepancy similar to 1995?*

In other words, is students' achievement in 1999/2000 similar to students' achievement in 1995? Do Dutch grade 8 students still perform as well on the TIMSS Written Test as they did in 1995? Have they improved their mathematical practical skills so that their scores on the Performance Assessment will increase? Or do they still underperform on the Performance Assessment in comparison to the Written Test?

1.4.3 The second research question

The second research question focuses on the link between the attained curriculum on the one hand and the implemented and the intended curriculum on the other. The matching of the test with the intended and implemented curriculum serves as the context, in which to interpret students' achievement.

The data set for this question is illustrated in Figure 1.5. For both tests, a distinction is made between the two dimensions:

- vertically, a time dimension (a possible trend between 1995 and 1999/2000), and
- horizontally, a curricular dimension (comparing judgements by curriculum experts and teachers with students' achievements).

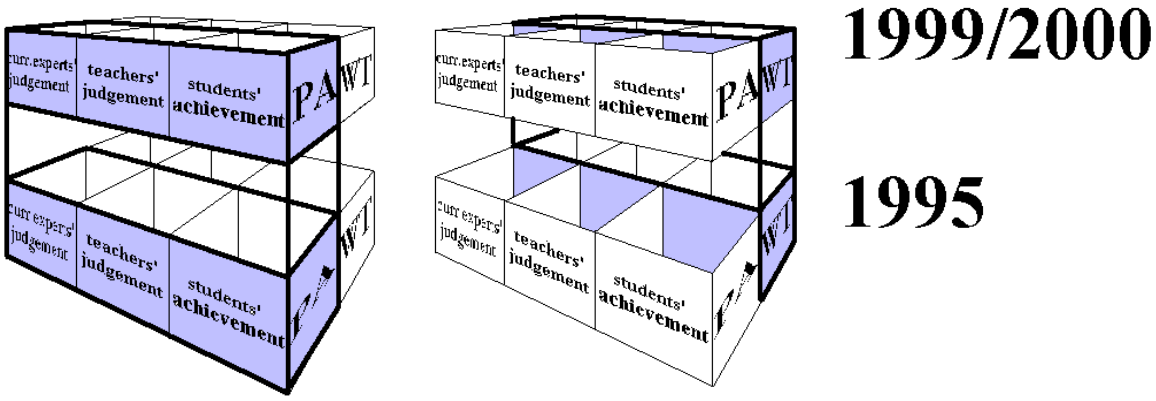


Figure 1.5: Data sets for the second research question

The second research question is:

> *To what extent are students' results on both tests at both periods in time aligned with (a) the appropriateness of the tests in light of the intended curriculum, (b) the appropriateness of the tests in light of the implemented curriculum and (c) possible discrepancies between these?*

The following three sub-questions have emerged from the second research question:

1. *To what extent are students' results on both tests at both periods in time aligned with the appropriateness of the tests in light of the intended curriculum?*
2. *To what extent are students' results on both tests at both periods in time aligned with the appropriateness of the tests in light of the implemented curriculum?*
3. *To what extent are students' results aligned with possible discrepancies between the appropriateness of the tests to (a) the implemented curriculum and (b) the intended curriculum?*

In other words, can the trends in Dutch students' achievements be related to the appropriateness of the tests, as judged by mathematics curriculum experts and mathematics teachers, and to discrepancies between judgements by the two groups?

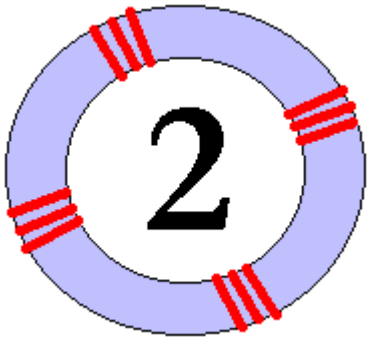
The second research question focuses on intra-curricular discrepancies. Describing discrepancies first requires a description of the separate components between which the discrepancies occur. The description of the components is needed to position the discrepancies. Therefore, initially the data of all twelve measurements will be presented (see chapter 5). At the level of the intended curriculum, four measurements will be presented, comparing the judgements by the curriculum experts in time and between the tests. At the level of the implemented curriculum four measurements will be presented, comparing the judgements by the mathematics teachers in time and between the tests. Finally, at the level of the attained curriculum four measurements will be presented, comparing students' achievements in time and between the tests. The data presentations will consist of descriptive statistics (averages, standard errors and correlations). After the separate descriptions, the intra-curricular linking will take place (see chapter 6). This, again, will be carried out through descriptive statistics (averages, standard errors and correlations).

By answering the research question, the METRIC study aims to explain the occurrence of the observed discrepancies of 1995. The METRIC study explores methods to compare curricular appearances quantitatively. As such, it is an exploratory study on the use of tests for curricular comparison. It is not a curriculum evaluation, although the curricular comparison will shed a light on the new mathematics curriculum for grade 8. In the METRIC study, the intentions of this curriculum play an important role, but they cannot be evaluated. The two available TIMSS tests have not been developed to cover the full range of the innovative Dutch intended mathematics curriculum, including the special characteristic integrating mathematics in context (for more information, see chapter 2). It is beyond the scope of the METRIC study to investigate to what extent the intended, implemented and attained curriculum are covered by the two TIMSS tests. However, the tests do cover a wide range of topics, including abandoned and more advanced content. This will enable the METRIC-study to analyse whether content, which is not relevant to the intended curriculum, is possibly included in the implemented and attained curriculum.

1.5 OVERVIEW OF THE FOLLOWING CHAPTERS

The following chapters will describe the METRIC study. Chapter 2 will elaborate on the background and distinguishing features of the Dutch mathematics curriculum. Chapter 3 presents literature on how to describe, analyse and link curricular appearances. The findings lead to operationalisation of the research questions. The fourth chapter elaborates on the design of the research. This will lead to a description of the database, which contains data from twelve measurements. The data will be described for each curricular appearance separately in chapter 5. The research results will be presented in chapter 6. The study ends in chapter 7, with a discussion of the results of the METRIC study.

Figure 1.6: Student working on the task *Around the Bend* from the TIMSS Performance Assessment

Chapter 

Reforms in Dutch mathematics education

~ Most things are easy to learn, but hard to master. ~
(CHINESE PROVERB)

This chapter describes the context of mathematics education in the Netherlands. The first two sections address the history, which lead to the 1993 core curriculum (sections 2.1 and 2.2). This new curriculum is in the limelight of the METRIC study. Therefore, the remaining part of the chapter (section 2.3) deals with aspects of this new curriculum, such as its content and instructional approach. The chapter ends with a review of evaluation studies on its implementation.

2.1 MANY YEARS WITH LITTLE CHANGE

2.1.1 Early years

Mathematics education has a tradition of thousands of years. In Western culture, during the classical Greek age, a pinch of geometrical and calculating skills sufficed in daily life. In general, people learnt the skills from practice, at the moment when the need arose. However, Plato (4th century BC) recommended a ten-year training in abstract mathematics for future kings. This curriculum comprised geometry, arithmetic, harmonics (the study of structures in music) and astronomy. The Greek philosophers considered every kind of skill connected with daily needs as ignoble and vulgar, and they praised mathematics for its purity. The justification for teaching mathematics was *aesthetic*. It would raise the learners' spirits (Kline, 1953).

For many centuries, the contrast between practical and aesthetical mathematics education persisted. The higher classes had an esoteric activity in studying mathematical deductive reasoning through Euclid's *Elements*. It is the most widely reprinted book after the Bible (Dunham, 1990). It established mathematics as a specialised science, in which axioms and definitions lead to a hierarchy of theorems. For more than two millennia, mathematicians established the truth of theorems through proofs.
On the other hand, the middle and lower classes studied worldly mathematical applications needed in daily life. This practical mathematics did not originate in Greece. The concept of the number zero originated from India, and so did the counting system with base 10. The system of place value originated from Persia. The way numbers were written originated from Arabia.

In the Middle Ages, many Dutch artisans and merchants acquired numerical skills. The first Dutch manuscript, in which calculations with the Hindu-Arabic numerals were taught, dates back to 1445 (Kool, 1999). The mathematics of the Dutch cultural elite was fairly practical as well. This elite consisted of rich, Calvinistic patricians with a pragmatic mind and an aversion of vanity (Schama, 1991). Their mathematics contained many applications on navigation and the architecture of fortifications (Van Maanen, 1987).

2.1.2 At the beginning of the 20th century

In 1901, legislation made six years of education compulsory for all children in the Netherlands, from the age of seven onwards. The mathematical activities at this level encompassed, at the minimum, counting and four basic arithmetic operations (adding, subtracting, multiplying and dividing) with positive numbers. These basic mathematical activities are in the Dutch language conceptualised with the special verb that became a synonym to arithmetic: *rekenen* (to calculate).
The reasons for teaching basic mathematical skills at primary level were never a point of debate. In a nation with a strong trading tradition, it was self-evident that all citizens should have a basic competence in working with numbers. This justification was *utilitarian* with a socio-economic aim.
Despite the practicality of mathematics in the Dutch trading culture, the idea of mathematics' purity, as advocated by the ancient Greeks, persisted in education, even in the Netherlands. It was the guideline for educational standards at

secondary schools. At the *Hogere Burger School* (Higher Citizen's School, established in 1863), which prepared students for the polytechnic school in Delft, abstract mathematics with its theorems and proofs was taught for its beauty of logic (Groen, 2000; Smid, 2000).

In the Netherlands, higher mathematical activities are described by the word *wiskunde*. It is originally a 17th century's expression, which at that time meant 'wise arts' or 'science of surety', depending on the interpretation (Freudenthal, 1991). There is no word available in the Dutch language that unifies *rekenen* (arithmetic) and *wiskunde* (higher mathematics). The sharp separation between the two kinds of mathematics has characterised Dutch mathematics education throughout the 20th century, with *rekenen* being less abstract than *wiskunde*. For many years, *rekenen* was taught at primary schools and *wiskunde* at secondary schools. Yet, the distinction was not always clear. During the 20th century, more mathematical activities (such as calculating with decimals, fractions, percentages, proportions and the metric system) were gradually included into the definition of *rekenen*, only because the topics were introduced to primary school students (Van der Blij & Treffers, 1985). The boundary started to fade further as elementary geometry and algebra were introduced in the curriculum of primary schools. Today, the subject taught at primary schools is a combined subject, called '*rekenen/wiskunde*'.

As said, *wiskunde* was taught at secondary schools. Until 1993, topics from *rekenen* were withheld to be included into the *wiskunde* curriculum. In particular in the two tracks for higher ability students, *havo*[1] and *vwo* (together approximately one third of all students), no consolidation of fractions or proportions would take place, as this was not considered *wiskunde*. Still, all stakeholders agreed that the training of basic computations was very much needed at any secondary school level (Aukema & Jansen, 1992; Van der Blij & Treffers, 1985). In 1993, with the introduction of *rekenen* as integral part of the curriculum of *wiskunde* of secondary schools, confusion arose about the differences between *rekenen* and *wiskunde*. The merging had to be extensively justified (Kok, Meeder, Wijers & Van Dormolen, 1992). Opponents against including *rekenen* into secondary school mathematics argued that *rekenen* was not part of *wiskunde*. They felt that *rekenen* lowered the status of *wiskunde* and therefore had to be fully covered at primary level (Aukema & Jansen, 1992).

[1] For an explanation on the ability tracks of Dutch secondary schools, see the Glossary (page xiii)

2.1.3 The emergence of mathematics education as a specialised research area

During the 20th century, aspects of mathematics education at secondary schools (*wiskunde*) were brought up for discussion. There were controversies about how to introduce topics, the sequencing of content and methods of teaching. The first traceable publication on the teaching of mathematics dates back to 1875 (De Moor, 1999).

The first *modern,* student-centred contribution came in 1924, when Tatiana Ehrenfest-Afanassjewa published a brochure, in which she questioned the teaching of the deductive approach in geometry with its rigour of proof (Ehrenfest-Afanassjewa, 1923). In this publication, she argued that intuitive geometric notions of 12-year old students about their experienced environment should be the basis for learning, and not the statement of Euclid's axioms.

In 1936, an informal discussion forum for mathematics teaching innovation, the *Wiskunde Werkgroep* (Higher Mathematics Workgroup) was established. The members often met at the home of Tatiana Ehrenfest. Among the first participants were not only academic mathematicians, but also teachers and generalists, for example the renowned Dutch educator Philip Kohnstamm. In 1945, after World War II and the persecution of Jews had ended, Hans Freudenthal was able to join. He was a professor in mathematics at the University of Utrecht, specialised in Topology. In 1937, he had published his famous Suspension Theorems. As the years progressed, his interest in the educational aspects of mathematics increased. In 1950, he became chair of the Wiskunde Werkgroep (Van Est, 1993).

The Wiskunde Werkgroep studied the work of Jean Piaget, a Swiss psychologist, who observed how children learn mathematical concepts (e.g. Piaget, 1952). Inspired by his ideas, a general program for research in mathematics education emerged in many countries. The goal of this program was as follows (Schoenfeld, 2000):

1. to understand the nature of mathematical thinking, teaching and learning, and
2. to use such understandings to improve mathematics instruction.

Piaget's influence on the innovative Wiskunde Werkgroep in the Netherlands was obvious, as seen in the development of the Level Theory on stages in learning mathematics by Pierre van Hiele (Van Hiele, 1973; Zwaneveld, 1999). Despite theoretical developments in the 1950's, only negligible changes took

place in classroom practice. The curricular stability is illustrated by the fact that a number of mathematical textbooks, written in the beginning of the century, were still reprinted with minor adaptations until the 1960's (Smid, 2000).
In the 1950's, the general justification for teaching mathematics at Dutch secondary schools was twofold. The aesthetical justification prevailed. Another justification was based on the individual's mental abilities: *doing mathematics would sharpen the mind*. Mathematics was said to have a general enlightening value, because students were taught to think strictly logically. The time had not yet come to recognise the contradiction between the justifications and the reproductive teaching methods. Hans Freudenthal (in a co-publication with Tatiana Ehrenfest) was the first to question this *cerebral* justification in public (De Moor, 1999).

The Dutch developments in mathematics education went hand in hand with international developments. From the 1950's onwards, regular international meetings on mathematics education were organised. In 1950, the Commission Internationale pour l'Étude et l'Amélioration de l'Enseignement des Mathématiques (CIEAEM) was founded. In 1952, the International Commission on Mathematical Instruction (ICME) was constituted to its present formal position (although originally founded in 1908). ICMI is the organiser of the large-scale, quadrennial International Congress on Mathematical Education (ICME). Study Groups have become affiliated to ICMI, such as the International Organisation for Women and Mathematics Education (IOWME) and the International Study Group for Psychology of Mathematics Education (PME).
The emergence of the organisations reflect the growth of mathematics education as an independent, specialised area of research, detached from both mathematics and general pedagogy. This development has been evolving steadily. Two decades later, the growing need for specialised international forums resulted in the founding of scientific journals, such as *Educational Studies in Mathematics* in 1968, and *Journal for Research in Mathematics Education* in 1970.

2.1.4 The year 1959, New Math and Freudenthal

The year 1959 is noteworthy because of two events, which illustrate that Western educational authorities wanted change. During the Cold War in the 1950's, the launch of the spacecraft Sputnik demonstrated colossal technological advancements by the USSR. It came as a shock in Western nations to realise that they were losing supremacy. Policy makers alleged that a nation's educational system could be pivotal for economic (and military) performance.

In 1959, the International Association for the Evaluation of Educational Achievement (IEA) was founded after an idea, which ripened at a UNESCO meeting. IEA was set up as an independent organisation. Its goal was to provide nations with a scientific, comparative basis to evaluate the performances of their educational system. From the 1960's onwards, IEA initiated large-scale international comparative studies in education, with mathematics being one of the subject areas studied. The IEA studies on mathematics education will be described in the next chapter.
In the same year, 1959, the Organisation for European Economic Cooperation (which became the OECD in 1963) organised a groundbreaking seminar in Royaumont (France) on mathematics education for 12-19 year old students. After meetings in 1960 (Zagreb) and 1961 (Paris), the ideology of the attendants bore the name *New Math*. Their aim was to produce new curricula in order (Fehr, 1961, p. 105),

> *"firstly, to provide a better preparation for university studies, secondly, to give all pupils an instrument for use in daily life."*

However, in New Math the difference in Fehr's citation between mathematics-for-the-academics and mathematics-for-all was not discerned. With society's need for engineers, and with academic mathematicians leading the way in mathematics education, the focus came onto the first part. From the mathematicians' perspective, secondary school mathematics was too aesthetical. It differed too much from the mathematics studied at universities. There, calculus and linear algebra had become main content areas, and traditional areas such as Euclidean geometry and number theory lost ground. The academic mathematicians needed school mathematics to be preparatory for university mathematics in the first place, aiming at academic continuity for their prospective students. They wanted the topics *differentiation* and *integration* to be covered at the end of secondary schooling.
Authorised by academic mathematicians, New Math introduced vectors and co-ordinates as the basis for transformation geometry, with rotations and translations instead of Euclid's theorems. Peano's axioms were the basis for algebra, with rules for associativity (a+(b+c)=(a+b)+c), commutativity (a+b=b+a), and distributivity (a(b+c)=ab+ac)). The languages of sets and logic were introduced, with Venn diagrams and concepts such as union, intersection, and relations. These concepts would create a basis for calculus, starting with linear and quadratic functions Statistics and probability were also added to the

curriculum. Those students, who would not study university mathematics, would join in for the first introductory steps to mathematics (together with the future university students) and then drop the subject.
Besides new content, New Math came up with a structuralist approach to the subject in general. Curricula were designed through the structure of mathematical theories, starting from axioms and definitions. The strict logic, including its jargon and nomenclature, would be followed for instruction.

Supported by the authority of academic mathematicians, the New Math movement swept over many countries, deeply affecting mathematics education at all levels. It showed that academic mathematicians were still the first to decide on issues of school mathematics, and not mathematics educators. Also in the Netherlands, New Math gained ground among mathematicians and teachers with an academic background. They felt the need to bridge the gap between secondary school mathematics and the universities. Therefore, in 1961, the Committee on Modernisation of the Secondary School Mathematics Curriculum (CMLW) was installed to develop a new curriculum for secondary schools. In 1968, their work resulted in a reform, introducing new content (sets, functions and relations, vectors, statistics, and probability) to students of all tracks. For senior secondary school level, calculus was the essence.
With compulsory education being raised to 9 years (from the age of 6 onwards), all Dutch students encountered this new content in the curriculum of mathematics. According to the well-known design principle of the teabag (Zwaneveld, 2000), the lower ability tracks (*vbo* and *mavo*) received an extract (of the tea) of the higher tracks. The lower tracks harboured approximately two-thirds of Dutch students. From 1968 onwards, these students encountered the formal bracket notation of sets and the transformations of parabola, as advocated by New Math. The topics would remain in the curriculum for more than 25 years. Until 1995, these topics were nationally examined through exams, which consisted largely of multiple choice questions (Schuring, 2000).

In all countries, debates were held about New Math. In the Netherlands, the opposition, lead by Hans Freudenthal, was very strong. He explained that the logical mathematical structure was the result of a long, tedious mental journey, in which the mathematician had slowly framed chaotic thoughts. Using the mathematical structure in teaching mathematics not only denied the invention process, it also reversed the sequence of activities. Therefore, Freudenthal named

the structuralist approach of New Math an *anti-didactical inversion*. He stated that mathematics was an activity, not a set of facts. According to him, learning should happen as a *re-invention* activity, which imitates the discovery route of the mathematician (Freudenthal, 1973).
Contrary to other countries, no academic mathematician in the Netherlands could accuse the opponents of New Math of ignorance of subject matter. On the contrary, Freudenthal always would lard his publications with highbrow mathematical symbols and formula. He made use of his mathematical learnedness and discussed with the academic mathematicians in their own jargon, as for example in Freudenthal (1973). With him as a decisive factor, the opposition against New Math in the Netherlands was firm. Consequently, mathematics education at Dutch schools did not undergo the full impact of New Math as in other countries. In particular primary education was spared. In secondary schools, the topics of New Math were introduced with the 1968 curriculum, but the structuralist approach was not established firmly in the new curriculum. Immediately at its onset, plans to change the 1968 curriculum immediately came forth. For example, a considerable number of textbook authors started to experiment with innovative approaches, attempting to generate more authentic mathematics education within the margins that the intended curriculum allowed them (e.g. *Exact, Moderne Wiskunde - 4e editie, Passen en Meten, Wageningse Methode, Wiskundelijn)*.

2.2 THREE DECADES WITH MANY CHANGES: THE ESTABLISHMENT OF RME

2.2.1 Features of Realistic Mathematics Education

The previous section ended with the curriculum reform of 1968 for secondary mathematics. The developmental work had been carried out by CMLW, which also insisted on founding a research institute with the mandate to develop national mathematics curricula. In 1971, this resulted in the creation of the Institute for Development of Mathematics Education (IOWO). In 1981, this institute became the Research Group on Mathematics Education and Educational Computer Centre (OW&OC) at Utrecht University. In 1991, this same institute was renamed Freudenthal Institute after its first director. Its present director is Jan de Lange.

Through the creation of the Freudenthal Institute, the Dutch authorities showed their belief in the socio-economic importance of developing mathematics education. The Freudenthal Institute (and its predecessor institutes) guided curriculum reforms in the Netherlands in the last three decades of the 20th century. The Freudenthal Institute developed a new direction for mathematics education. This treatise is generally known as *Realistic Mathematics Education* (RME) and it will be described in this section.

Central idea in RME is that mathematics can best be learnt starting from a concrete, realistic situation that appeals to students. This context should be relevant and challenging (Freudenthal, 1973; Gravemeijer, 1994; Van den Heuvel-Panhuizen, 1996; De Lange, 1987; Treffers, 1987). In the beginning, RME was only associated with primary school mathematics, but after the 1980's, RME became associated with secondary school mathematics as well. RME is characterised by the understanding that mathematics is an integral part of real-life. Thus, mathematics is taught, not for its beauty, but for its applicability. In addition, mathematics is perceived as an activity and not as a set of rules. As such, mathematics is a creative and organising activity in which unknown regularities, relations and structures are discovered.

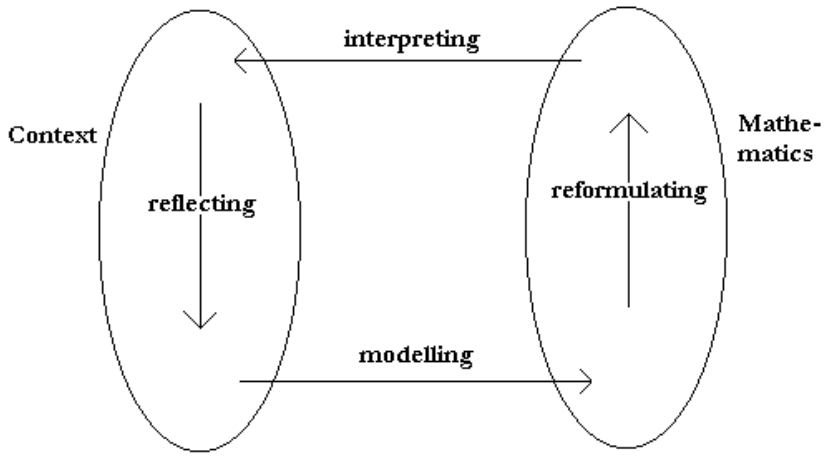


Figure 2.1: Mathematising activities in RME

In RME, the mathematical activities are named *mathematising*. Treffers (1987) discerns two dimensions in mathematisation (see Figure 2.1): vertical and horizontal. The context and mathematics are two different *worlds*, which are connected through mathematising. When starting from a context, the student has to find relations and regularities that result into a mathematical structure, for

example a formula, a graph or a table. This activity is called *modelling* (horizontal mathematisation). Within the mathematical world, the model needs rewriting, restructuring and refining to obtain an answer. This activity is called *reformulating* (vertical mathematisation). These two activities might need to be repeated a number of times, before a sound mathematical answer is reached. Within the mathematical world it is also possible to further generalise the problem and refine mathematical knowledge (not included in the figure).
Going back to the context, the (mathematical) answers need to be *interpreted* (horizontal mathematisation). Finally, the answer will only make sense if it is related to the initial question, as there is need to *reflect* on the process of reaching results (vertical mathematisation).
Contexts can be taken from the physical world, but also from imagined reality (cartoons, games, etc), and even from mathematics itself (Van den Heuvel-Panhuizen, 1996).

An important component in the learning process in RME is to enable students to make mental images. Consequently, lively colourful illustrations of physical edifices and identifiable persons entered Dutch mathematics textbooks from the 1970's onwards. This was also enabled by improved printing technologies.
The following two examples may serve to explain the difference between non-RME and RME. The first example is an item from the TIMSS Written Test (item N13) in which a number has to be substituted into a formula with one variable:

$$\textit{If } x = 2\textit{, what is the value of } \frac{7x+4}{5x-4} \textit{ ?} \_\_\_\_\_$$

In the Dutch RME-based curriculum, similar items on the substitution of a number into a formula can be found, but these will be stated within context. The following item was developed for exactly the same target population as the above item from TIMSS. This item appeared in the National Option Test (NOT) for grade 8, administered in 1995 (Kuiper et al., 1997, 2000). In Appendix A, the full text of the item is included; here is an excerpt:

To calculate for any girl her future length as a grown-up, the school doctor uses the following formula:

$$\textit{Length daughter (in cm)} = \frac{\textit{length father (cm)} + \textit{length mother (cm)} - 12}{2} + 3$$

Danielle's father is 1,82 m tall, her mother is 1,68 m.
How tall will Danielle grow according to the formula?

Thus, a context is created, which approximates the real-life of students (many 14-year old students would like to know how tall they will grow). The context is the prediction of growth. This prediction is modelled (horizontal mathematisation) into a (word-)formula, which contains three variables. After substitution (vertical mathematisation), a result is found. The result of the calculation can be meaningfully interpreted within the context again. It is the predicted length of the daughter.

The example shows how the mathematical content is embedded in context, in order to make mathematics meaningful to the students. Thus, new mathematics curricula were developed by choosing mathematical content that could be useful to students' lives (Van Dormolen, 1999). As a result, the RME-based reforms were focused on the mathematics content and not on the instructional approach (Van Reeuwijk, 1992; Zwaneveld, 1999). However, some authors have tried to link RME with *constructivism*, such as Gravemeijer (1994), De Lange (1987), and Schoenfeld (2000). In constructivism, as opposed to behaviourism, learning is perceived as a process and not as an outcome. It focuses on a learner's ability to mentally construct meaning of their own environment and to create their own learning. Another instructional approach was chosen by Van Streun (1985), who showed that a *heuristic* instructional approach for problem solving could also go hand in hand with the use of realistic contexts in mathematics education. On the other hand, Van Gaans (1991) and Lagerwerf (1994) stated that, besides student-centred approaches, teachers could use traditional, teacher-centred whole-class instruction within RME-based mathematics education. Thus, the instructional approach associated with the RME-based curricula can vary.

2.2.2 Three consecutive nation-wide curriculum reforms

In the Netherlands, from 1971 onwards, each decade had its own curriculum reform with strong influences from the RME treatise.
The first RME-based, national curriculum reform project in the 1970's, *Wiskobas*, centred on primary education (Gravemeijer, 1994; Treffers, 1987, 1993). In this project, straightforward mental arithmetic activities were given much room, such as halving and doubling, which prepared students for multiplication. The development of number sense was supported through estimation and the use of models (e.g. the empty number line). This practice was supported by various theories in cognitive psychology. Through the RME-based curriculum, by delaying the traditional drill in arithmetic, students would gain a *meaningful, conceptual, productive* or *relational* understanding. By starting too early with the drill, only *mechanistic, procedural, reproductive* or *instrumental* understanding would be trained (in the terminology of Baroody & Ginsberg (1986), Gray & Tall (1994), Merrill (1983), and Skemp (1986), respectively).
The second large-scale RME-based curriculum reform took place in the 1980's, focusing on senior secondary education (*HEWET project* and *HAWEX project*). A special curriculum was developed, which was named *Mathematics A*. This curriculum was meant for students, who wanted to study social and economical sciences at university. Many students of this target group failed on the existing mathematics curriculum, as this curriculum was meant for future technical engineers. The low pass rates of many students marked the need for an easier, yet motivating mathematics curriculum. Mathematics A became an applied subject, serving to organise phenomena of the real world. An innovative feature in this mathematics curriculum was the use of networks (graphs) and matrices to solve problems on stock management, biological processes and transport infrastructure. Other topics in Mathematics A were elementary calculus, statistics, probability and linear programming. This project exemplified the link of mathematics with common sense (De Lange, 1987; Freudenthal, 1991).

With each decade having its own curriculum reform, finally in the 1990's, a new curriculum was created to bridge the gap between the two previous reforms. This new curriculum was meant for all students at junior secondary school level (grades 7 and 8) and for the final years of the lower ability tracks (grades 9 and 10 of *mavo* and *vbo*). For the first time in Dutch history, a curriculum for this age group was developed, which was **not** a dilution of the curricula at a higher level.

It was meant as a continuous learning trajectory, extending from primary school onwards, giving all future citizens basic, mathematical abilities.
There were two mottoes associated with this reform: *mathematics-for-all* and *mathematics-as-to-be-useful*. The developers stressed that the usefulness of mathematics should not have its justification from future adult life alone, but the beneficiality should also be experienced during the learning process itself. The intended mathematics curriculum was meant to link with student's interests (Kok et al., 1992). At this stage, the mathematics curriculum in the Netherlands had wandered through the three stages that Marsh and Willis (1995) describe for the development of curricula in the 20th century. They describe how curricula evolve from being subject-centred, to being society-centred, and finally becoming person-centred.

The new curriculum was developed under responsibility of the Commissie Ontwikkeling Wiskundeonderwijs (COW), chaired by Jan de Lange. The COW delegated the developmental work to a team of mathematics educators, the W12-16 team. The team consisted of mathematics teachers, teacher trainers, curriculum developers, mathematicians, representatives from the association of mathematics teachers, and textbook authors, who had gained experience in linking mathematics for junior secondary school level to attractive context. The members endorsed the RME treatise to different extents.
The W12-16 team developed the new curriculum through several cycles, which included field consultations with mathematics teachers, who called for adaptations of the content. In particular, the new approach to algebra raised concern (Aukema-Schepel, 1991). Therefore, the new curriculum for junior secondary school that was created will be indicated in this text as an 'RME-based curriculum' and not as an unmitigated RME curriculum.
Schoemaker (1989) describes how the W12-16 team started to develop the new curriculum in 1987. The team faced two challenges, (1) writing experimental classroom materials without clear objectives, and (2) writing objectives without knowing whether these could lead to workable classroom materials. The materials written were tested at several schools. Each year, a growing number of schools were involved in the experiments.

From 1990 onwards, excerpts of the materials were used in publications and at conferences, organised by the Netherlands Association of Mathematics Teachers (NVvW). For example, in autumn 1990 in 12 regional cities, two-day workshops

were organised. These were attended by more than a thousand mathematics teachers (out of the total of approximately ten thousand mathematics teachers) (Verhage & Wijers, 1991). In the following years, the information conferences were repeated. To further inspire teachers, anecdotal reports of class observations of the experiments were published in the Dutch journals for mathematics education, *Nieuwe Wiskrant* and *Euclides*. In addition, the experimental materials were made available to teachers, together with publications on backgrounds, instructional approaches and collections of test items (Achtergronden, 1992; Van Dormolen, 1993; Van Dormolen & Zwaneveld 1992; Van Gaans, 1991; Kok et al., 1992; Lagerwerf, 1994). Vink et al. (1993) remarked on the large-scale approach of informing teachers, that probably never before, the implementation of a new mathematics curriculum was undertaken so thoroughly.

In 1992, the program for the new curriculum was published and the team of developers, W12-16, was dissolved (Commissie Ontwikkeling Wiskunde-onderwijs, 1992). However, implementation of the curriculum was withheld for another year because of three reasons:

1. to have more time to inform mathematics teachers on the reform,
2. to give commercial textbook publishers time to create books based on the new curriculum, and
3. to make the introduction of the new curriculum coincide with a school-wide reform, which would start in the school year 1993/1994.

The new RME-based curriculum for junior secondary schools complied with parallel developments in other subject areas. New pedagogical insights had asked for reforms of curricula, instruction and school administration. Thus, the introduction of the new mathematics curriculum was assimilated into a school-wide reform to establish the *Basisvorming* (Basic Education), which was legislated in 1993. This reform established an educational basis for all students. A common core curriculum for all subjects was introduced, based on new educational insights. The new curricula for all subjects were characterised by three principal keywords: application, skills and coherence (Van Luyn, 1998; Roelofs, Franssen, Houtveen & Lagerweij, 1999). The keywords structured and unified a pedagogy of *authentic learning* for all subjects. They appropriately suited the new mathematics curriculum and this illustrates how general pedagogy and mathematics education had evolved correspondingly. Yet, the mathematics curriculum sustained a larger overhaul than any other subject did.

In the core curriculum, all subjects were guided by a list of general core objectives (*algemene kerndoelen*). Specific core objectives (*vak-kerndoelen*) were added for single subjects such as mathematics. Depending on students' ability level, the objectives had to be attained after two or three years of secondary education (in grade 8 or in grade 9).

2.3 THE RME-BASED CURRICULUM FOR JUNIOR SECONDARY SCHOOLS

2.3.1 Content areas

This section describes the RME-based curriculum for junior secondary schools (grades 7 and 8). This curriculum has already been introduced in the previous section, as the product of the W12-16 project, which was the last out of three large-scale curriculum innovations in mathematics education in the Netherlands.
Like all subjects of the core curriculum, the intended mathematics curriculum was defined with *kerndoelen* (core objectives) from 1993 onwards. There were general targets, such as: 'being able to gather, describe and arrange information systematically'. The special mathematics core objectives were grouped into four inter-linked content areas (Commissie Ontwikkeling Wiskundeonderwijs, 1992):

- *rekenen* (arithmetic), measuring and estimating;
- algebra, relations, graphs and functions;
- geometry;
- statistics and probability.

After an evaluation, the core objectives were reformulated for a new period of five years (1998-2003), increasing quality and reducing quantity (Mulder, 1996). For mathematics, sub-topics were deleted from the curriculum, such as inter- and extrapolation, solving linear equations and inequalities, Pythagoras' theorem and its applications, two-dimensional transformations and the concept of probability (Ministerie van OW&C, 1998). The four content areas were renamed:

- *rekenen* (arithmetic), measuring and estimating;
- algebraic relations;
- geometry;
- data processing and statistics.

The content areas were explicated as follows (see also Appendix B).

The first area, *rekenen* (arithmetic), measuring and estimating, is an extension of primary school mathematics. It comprises approximation, negative numbers, rules of thumb and conversion of fractions, percentages, square roots and powers into decimal numbers. A smart use of calculators is explicitly included. This content area also contains the application of proportion and scale. As a supportive model for solving problems, the horizontal proportion table is recommended. See Appendix A, for an example of a test item on operating with negative numbers.
The second area, algebraic relations, comprises data handling from real-life situations through the combination of graphs, formula and tables. Students should become able to shift from one representation to another. Against the background of a context, the variables in the formula are given in *words*, as for example in the item on the prediction of the future length as a grown-up, based on the length of the parents. When drawing graphs the axes can be unspecified. Transferring the context into an abstract representation (modelling) and vice-versa (interpretation) is evident. The following exercise shows the graph of the speed of a racing car on a circuit. The speed during the second lap is given, resulting in the fact that the car races with maximum speed across the starting line. The speed is not specified, but is it clear that the top-speed has to be reduced three times during one lap. This is caused by the curves of the circuit. See Figure 2.2. There are five alternative racing circuits given. The multiple choice question asks students to associate the correct circuit to the graph given, thus asking students to make the transfer from context (the racing circuit) to the model (the graph), and back.

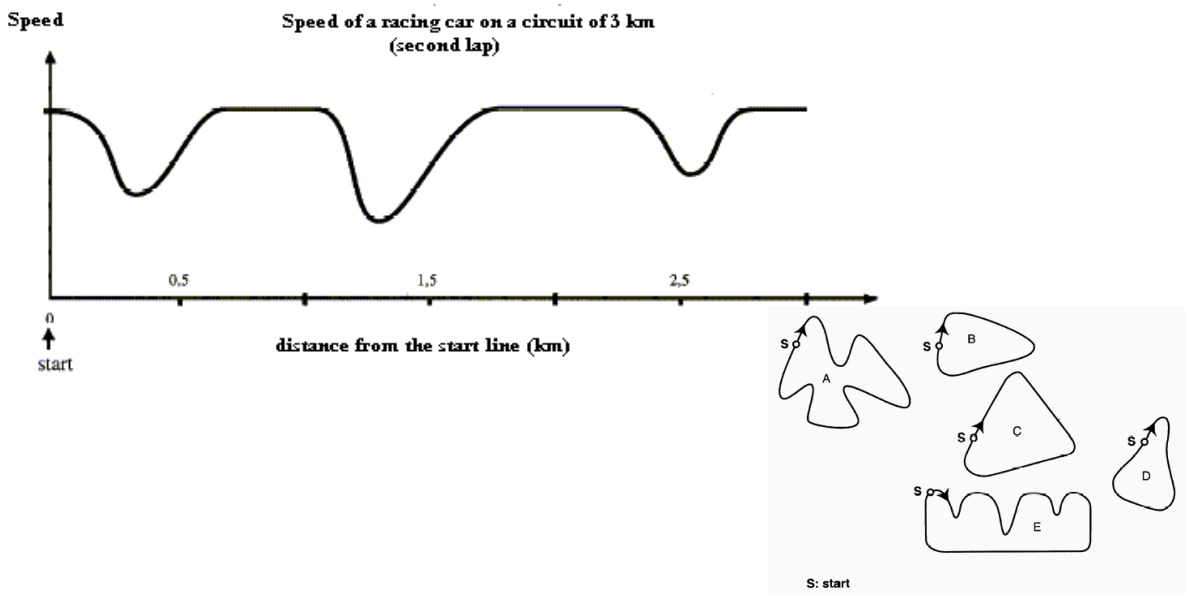


Figure 2.2: Speed of racing car on a circuit, and five possible circuits

correct answer: B

The third area, geometry, departs from students' experiences, as Tatiana Ehrenfest already recommended. This content area aims at the creation of awareness of how the perception of the environment depends on the position of the observer. This *visual geometry* centres on the transfer from the 3-d world to its 2-d representation, and vice-versa. Students will work with location of position, nets and cross section of objects. Through measurement of area and volume, this area overlaps with the first content area of measuring. See Appendix A for an example of a test item on the different views of an object, depending on the position of the observer.
The fourth area is named data processing and statistics. It overlaps with the second area, as graphs and diagrams can also be handled with an algebraic focus. It overlaps with the first area, as the use of calculators can also be extended to the use of computers. This area includes strategies of calculating combinations, for example through the model of a tree diagram.
Besides the general core objectives and the specialised core objectives in the four content areas, in the new mathematics curriculum, 5% of class time was reserved for *integrated mathematical activities* (in Dutch: GWA – geïntegreerde wiskundige activiteiten). According to the new, intended curriculum, these activities consist of investigative projects, which should integrate mathematics with students' daily lives (Commissie Ontwikkeling Wiskundeonderwijs, 1992).

The RME-based curriculum for junior secondary schools broke away from well-established traditions, such as the *parabolica* (Kindt, 2000), a derogatory word for the algebraic rituals that drilled students in meaningless exercises, such as:

- manipulations with symbols: $2p+2q = 2(p+q)$
- drill of algebraic identities: $a^2+2ab+b^2= (a+b)^2$
- completing the squares: $x^2+6x+8= (x+3)^2 -9 +8$.

The algebraic sequence was spread over a few years, finally leading to a procedure for solving quadratic equations needed to draw graphs of parabola.
In the RME-based curriculum, the sequencing of content was not ordered according to mathematical difficulty (see Figure 2.3, diagram left). Instead, through the use of tables and graphical representations, many variations of algebraic relations (inverse, exponential, harmonic, etc) could be included from the start. Gradually all content increased in difficulty, as a spiralling curriculum (see Figure 2.3, diagram right) (Algebragroep W12-16, 1990). Instead of the parabolica drill, solutions to quadratic equations could alternatively be found graphically or through numerical approximation.

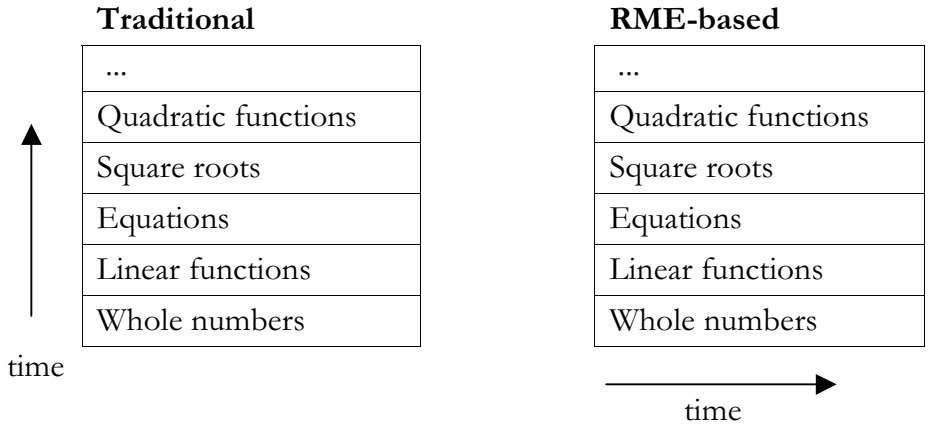


Figure 2.3: Sequencing of content in mathematics curricula
Source: Algebragroep W12-16 (1990).

2.3.2 Assessment in the new curriculum for junior secondary schools

At its introduction in 1993, the new RME-based curriculum was to be concluded by a compulsory test. This test was issued by the National Institute for Educational Measurement (Cito). Teachers were asked to administer the test when they considered their students ready for it.
Through this compulsory assessment, the new curriculum was enforced. The national tests, together with the publications of collections of exemplary test items, sent a strong message towards teachers. The tests assisted teachers in interpreting the program (Hawker & Ollerton, 1999). In practice, the compulsory tests were felt as a straightjacket, being too extensive, too difficult and too time-consuming. As a result, after 1998, the regulations were relaxed. Hereafter, teachers were free in their method to assess whether the minimum requirements had been met (Ministerie van OW&C, 1998).

Tests in the RME-based curriculum have distinguishing features. When comparing internationally, they differ largely from those in other countries., but In Appendix A, three examples have been included. Purely by plain looking at the appearance, without knowledge of the Dutch language, or of mathematics, one will notice the absence of multiple choice questions. There are narrative texts, explaining a context and there are many illustrations. Each test item has a title. Of the examples in Appendix A, one item deals with the application of negative numbers using the context of time differences between countries. Another item deals with the view of a candleholder from several angles. Here, students have to make a transfer from the three-dimensional shape to a two-dimensional representation, and vice-versa.

In RME-based items, several mathematics content areas can be covered in the questions (e.g. geometry, statistics), being integrated by the theme of the context. Any integral test item is expected to keep students' concentration alive for approximately 15 minutes, which is considered short enough for students who are not attracted by the context (Dekker, 1993).

Objections against the RME-based testing firstly focus on the use of context. The context can be unequally comprehensible to students, depending on their gender, cultural background, and so forth. The narrative description of the context can sometimes be problematic for students who learn in a second language, or for students with lower reading abilities. The open character of the questions can be a disadvantage to less proficient writers, but De Lange (1987) found evidence that open questions benefited girls. Other objections focus on the written, time-restricted format. It is alleged that such tests cannot test for all curricular goals. This is a general problem of mathematics curricula with a focus on authentic learning, and not only in RME. As Hawker and Ollerton (1999) state, problematic areas for assessment are those aims, which require students to systematically use mathematics as a tool, aims on awareness, aims on fascination of mathematics, and aims on working co-operatively. This view is extended by Niss (1993), who states that assessment in mathematics education in many countries stayed behind instructional and other reforms. This applies to the RME-based tests just as well.

Although alternative testing methods through interview, portfolio, observation, essay, or take-home tasks were options, assessment based on RME stuck with time-restricted, written tests with short or more extended answers (De Lange, 1987). Only the content of the tests has been adjusted to the new content, with each test item being linked to context. De Lange (1987) explained that, in RME, tests were preferred that can readily be carried out in class. This principle meant that alternative assessment, such as for example portfolio or projects, were rejected as being too time-consuming.

As a result, the two TIMSS tests used in the METRIC study differed from current assessment practice. The Written Test differed because of the large number of multiple choice items, and the absence of context in which mathematical activities were integrated. The Performance Assessment differed because it provided students with manipulatives and asked students to carry out an investigation.

2.3.3 Evaluation studies of the new curriculum for junior secondary schools

Between 1987 and 1992, the new mathematics curriculum for junior secondary schools was developed. In the developmental process, there was little room for evaluation. During the experimental phase, formative evaluation studies were carried out by the developers themselves, observing the reactions of students in the shielded environment of the experimental classrooms. However, as Van Streun (1985, 1990) repeatedly pointed out, summative evaluation studies of new mathematics curricula have been shun, while these could sustain and strengthen the reform process.
Fortunately, the RME-based mathematics curriculum for junior secondary schools was introduced in 1993, as part of a larger reform, which introduced a common core curriculum in all subjects. Anticipating on its legislation, several evaluation studies were planned to monitor the changes (Peschar, 1988). Most of the studies had a school-wide spectrum. They covered administrative and leadership aspects of schools, instructional practices, the introduction of new subjects (technology, science of care) and the merger of physics with chemistry. Only some of the evaluative studies included issues about the new mathematics curriculum. These will be reviewed in this section.

The largest of all evaluation studies was carried out by the Dutch Inspectorate for Education (Inspectie van het Onderwijs, 1999a). During the school year 1997/98, teams of inspectors attended more than 7000 lessons at a sample of 120 schools. The lessons included 670 mathematics lessons in 332 classes (Inspectie van het Onderwijs, 1999b). After its publication, the Inspectorate's report gave rise to many discussions on the intentions of the reform. For example, Van Dormolen (1999), Kuipers (1999), Wijers (2000) and Wijers and Kemme (2000) explained how the initial ideals of W12-16 were diluted, when these were forged into the dry formulation of core objectives.
Another large-scale project for the evaluation of the new core curriculum was the VOCL-project (Voortgezet Onderwijs Cohorten Leerlingen - *secondary education cohorts students*). In this project, cohorts of students were monitored throughout their complete secondary school career. The first cohort consisted of students who started their secondary education in 1989, still before the reform (VOCL-89). This cohort was compared with a second cohort of students, which started their secondary education in 1993, immediately after the reform (VOCL-93).

Several studies draw their data from this longitudinal research (Cremers-van Wees, Akkermans & Brandsma, 1999; Doolaard, Cremers-van Wees & Bosker, 1999; Kuyper, 1999; Van der Werf, Lubbers & Kuyper, 1999).
Besides the two large-scale studies, there were smaller studies. For example, Roelofs (1996) conducted a case study of three schools in their first year after the introduction of the core curriculum. Roelofs, Vermeulen and Houtveen (1998) and Wijers (2000) used interviews with teachers and curriculum experts. Roelofs, Franssen, Houtveen and Lagerweij (1999) and Luyten (2000) carried out other studies on the new mathematics curriculum. Data from TIMSS-95 and TIMSS-99 supplemented additional information on the introduction of the new core curriculum (Bos et al., 1999, 2001; Bos & Vos 2000; Kuiper et al., 1997, 1999, 2000). Finally, Kuiper, Broersma and Van den Akker (2002) tried to synthesise the results of TIMSS and the Inspectorate's report. From these studies, the following picture on the new core mathematics curriculum for Dutch junior secondary schools emerged.

The dissemination and the institutionalisation of a new national curriculum can be measured by asking teachers about the quantity and quality of information they received, and by asking teachers whether they have changed their textbooks (Fullan, 1991). Cremers-van Wees et al. (1999) showed that both in 1994 and 1996, mathematics teachers spent more time on finding informing on the new developments than their colleagues who teach Dutch. This is probably due to the many publications in the field of mathematics education, which forecasted the radical curriculum changes in this particular subject (e.g. Achtergronden, 1992; Van der Blij & Treffers, 1985; Broekman, Spijkerboer & Terlingen, 1991; Commissie Ontwikkeling Wiskundeonderwijs, 1992; Dekker, 1993; Kok et al., 1992; Lagerwerf, 1994; Van Streun, 1990).
Immediately after its legislation in 1993, many mathematics teachers showed enthusiasm about the new curriculum (Cremers-van Wees et al., 1999). In particular, teachers of the lower tracks (*vbo/mavo*) were relieved with the abandonment of abstract concepts (Wijers, 2000). However, Cremers-van Wees et al. (1999) show that the initial delight faded after 1995. This is confirmed by Kuipers (1999), one of the teachers engaged in the testing phase of the new mathematics curriculum. He describes how his enthusiasm dropped, a few years after the introduction of the new curriculum. He asserts that teachers' innovative approaches need continuous maintenance. After a first period of enthusiasm, without stimulus, teachers fall back into their original routine.

Despite the first general positive reception, the reform did not immediately take place at all schools. Kuiper et al. (1997) found that 50% of the students in the first cohort was taught using textbooks based on the former abandoned curriculum. In 1996, still 20% of the mathematics school departments had not replaced their textbooks (Cremers-van Wees et al., 1999). Therefore, it can be concluded that the reform took several years. This slow general adaptation to the new curriculum is also reflected in teachers' attitude towards the test items that were based on the new curriculum. The Dutch Inspectorate submitted items from the newly developed achievement tests to teachers. They assessed whether teachers recognised the content of the items as matching with their lessons. Only 51% of the teachers indicated that the tests fitted their teaching (Inspectie van het Onderwijs, 1999b).

The new RME-based curriculum distinguished itself from the previous curriculum in many ways. Therefore, researchers investigated the typical features, such as the coverage of general core objectives in the mathematics classroom (e.g. conducting a small investigation, taking a stand in a dispute, co-operating, connecting the subject to professional and future life). The Inspectorate considered this coverage as too minimal. They also looked at which specialised mathematical topics were insufficiently covered. These were typically the newer topics, such as data processing and computer usage. The completely new topic in mathematics education, GWA (Integrated Mathematical Activities), which could give room for conducting small-scale investigations, appeared sporadically in the programme (Inspectie van het Onderwijs, 1999b). Kleijne (1999), being himself a mathematics educator and one of the inspectors, additionally noted that teachers made several objections against organising investigative tasks. They perceived these as too time consuming. Moreover, they did not see a point in spending teaching time on the activities, as GWA was not assessed in the national exams.

Some researchers looked at how an average mathematics lesson evolved under the new curriculum and whether changes did occur. They calculated the average percentage of class time being spent in different instructional ways. Unfortunately, the definitions of types of instruction between reports are not fully comparable. Doolaard et al. (1999) differentiated between (a) homework review, (b) demonstration, (c) seatwork, and (d) organisational matters. The Inspectorate's report comprised three categories: (1) whole-class teaching, (2) seatwork and (3) organisational matters. This means that both studies used the

categories 'seat work' and 'organisational matters', but it is difficult to tell whether Doolaards' 'homework review' is included in the Inspectorate's term of 'whole class teaching' or in 'seatwork'. In addition, the methods of research differed. The Inspectorate calculated their averages from classroom inspection, while Doolaard et al. compiled their data from teacher questionnaires. The distribution of class time as given in the two studies is compiled in Table 2.1.
As can be seen from the table, there is possibly a trend towards less whole-class instruction, and towards more time spent on students' individual seatwork. However, as the Inspectorate noted, many mathematics teachers still use a teacher-centred approach (Inspectie van het Onderwijs, 1999b).

Table 2.1: Activities in mathematics class (grade 9) and trends in time spent on these activities (in percentage of a period)

Activity	**1991**	**1995**	**Activity**	**1997/8**
demonstration	28	31	whole class teaching	38
homework review	34	22		
seat work	28	35	seat work	54
organisational matters	9	11	organisational matters	8

Source: Doolaard et al. (1999), Inspectie van het Onderwijs (1999b).

Looking at classroom processes in general, mathematics lessons were well prepared by the teachers and the momentum was good (Inspectie van het Onderwijs, 1999b). Mathematics teachers scored high on the item of 'clear explanations', when compared to teachers of other subjects. In 60% of the mathematics classes, students were activated in their learning (learning to learn). Roelofs (1996) confirms the observation of this phenomenon. He noted that students had more independence and more opportunities to communicate with their peers for solving tasks in mathematics lessons than in English lessons. In addition, students indicated that they acknowledged the relevance of mathematics for later adult life. Yet, he also noticed that mathematics teachers did not utilise this.
Amongst others, Doolaard et al. (1999) established little evidence of group work in mathematics classes. On this issue, Kleijne (1999) cited teachers who did not want to organise group work, because of students' concentration problems. Mathematics teachers generally gave whole-class instruction, and made students work individually. As student in Dutch classrooms generally sit in pairs, teachers often allow them to confer with their direct neighbours (Bos & Vos, 2000).

In the visited classes, little differentiation for abilities was observed. Teachers commented that with the increased streaming, which had coincided with the reform, there was less need for differentiation (Kleijne, 1999). Cremers-van Wees et al. (1999) also noted this trend of increased homogeneity of classes. Before 1993, students in grade 7 would still sit in heterogeneous classes and selection of ability tracks would happen in grade 8 or 9. After 1993, schools tended to group students already from grade 7 onwards according to their ability track (Cremers-van Wees et al., 1999; Doolaard et al., 1999). Therefore, the differentiation in curriculum would primarily be class-related and not student-related. In the homogeneous classes, most mathematics teachers made all students do exactly the same tasks (Cremers-van Wees et al., 1999; Doolaard et al., 1999).

The students' achievements were tested through the compulsory tests, which were specially developed by the National Institute for Educational Measurement (Cito). The tests were based on the new core objectives. The tests capitalised more on the new content areas (visual geometry, data representation) and less on the new general core objectives, which are more difficult to assess in a time-restricted paper-and-pencil test (e.g. conducting a small investigation, taking a stand in a dispute, co-operating, connecting the subject to professional and future life). With the tests, data on the achievement in mathematics from 4388 students (in grade 9) were gathered. According to the Inspectorate, the tests proved too difficult for the students in the lowest track (*vbo*). Students in the two middle tracks (*mavo* and *havo*) did well on the tests, but not on the content area *calculations, measuring and estimating*. Therefore, it was concluded that too few students reached the required minimum targets (Inspectie van het Onderwijs, 1999b).
On the other hand, Dutch students performed well in the international comparative tests of TIMSS-95 and TIMSS-99 (Beaton et al., 1996; Bos & Vos, 2000; Kuiper et al., 1997; Mullis et al., 2000). The TIMSS results were used by politicians to support the success of the reform (Kuiper et al., 2002). There is an apparent contradiction between the unsatisfactory Inspectorate's test results and the satisfactory TIMSS results. However, this could be explained by the fact that TIMSS tested for an accumulation of mathematical knowledge and skills from primary school onwards. The TIMSS Written Test contained a large number of test items, which were irrelevant to the new curriculum (e.g. of primary school level or of the previous curriculum). On the other hand, the mathematics tests, which were used for the Inspectorate's report, were designated especially to test

for the intended, RME-based curriculum at this particular level. However, Kuhlemeijer, Kleijntjes & Van den Bergh (2001) show that the achievement tests for mathematics were far more difficult than the tests for other subjects. Therefore, the different results on the two tests were not necessarily contradictory (Bos & Vos, 2000; Kuiper et al., 2002).
Kuyper (1999) and Van der Werf, Lubbers and Kuyper (1999) have compared the achievement of students before and after 1993. They draw their data from the VOCL-89 and VOCL-93 cohorts, which were tested in their third year of study (respectively in 1992 and 1995). They established that the second cohort scored slightly, though significantly higher than the first cohort did. When differentiating for the four ability tracks (*vbo*, *mavo*, *havo* and *vwo*), they observed that in particular the two lower ability tracks had benefited from the new curriculum and outperformed their counterparts of four years before. The contrary holds for the higher ability tracks: *havo/vwo* students in the '93-cohort scored lower than their counterparts in the '89-cohort. Kuyper (1999) added here, that the pre-reform curriculum was probably more difficult than the new curriculum. He assumes that the new curriculum gave students of the lower ability tracks more confidence. But at the same time the students from the higher tracks learnt less abstract content. Wijers and Kemme (2000) and Goddijn and Kindt (2002), who describe how *havo/vwo* students from grade 9 onwards lack basic algebraic skills, confirm this observation.

From the above, it can be concluded that the institutionalisation of the new RME-based curriculum was a process, taking several years. Some changes were noteworthy. The content changed dramatically with less abstract mathematics and more context-related mathematics. In particular, students from the lower ability tracks benefited from this. Yet, students from the higher ability tracks had a less challenging curriculum and did not gain basic algebraic routine. In addition, the innovative targets in the curriculum on data processing and investigative skills (GWA) needed much more time to become part of the implemented curriculum. As assessment did not support the innovative aspects, teachers were less inclined to focus on them. Another notable, but quiet change could be the fact that teachers withdrew slightly from whole-class teaching and students gained classroom time to work for themselves. However, as the Inspectorate concluded, despite the changed mathematical contents, interaction in Dutch mathematics classrooms remained predominantly traditional.

The METRIC study was carried out against the background of the institutionalisation of the RME-based curriculum. That curriculum was characterised in this chapter. The curriculum reform started in 1993/1994 with grade 7 (the first grade of Dutch secondary education), being implemented in the following grades up to grade 10, every year after. Thus, the measurements for the METRIC study of 1995 were carried out in the first year, in which the curriculum was implemented in grade 8. At that stage, the discrepancies, which were described in chapter 1, occurred.

In the summer of 1997, the first cohort of students, who had learnt mathematics through the new curriculum, had reached their final exams. At that stage, the curriculum had reached the highest level (grade 10), for which it was meant. In that year, 1997, the first national *vbo/mavo* exams in mathematics at all Dutch schools were based on the new curriculum. Based on the practical experiences of four years and on the results of the first exams, the core objectives were reviewed (van der Zwaart & Boertien, 1998). The reviewed core objectives were published in 1998 (Ministerie van OW&C, 1998). Concurrently, many mathematics textbooks were revised. Their first editions were developed, starting from the proposals of W12-16, but now textbook authors were able to build on classroom practice as well (Hoogland, 1998).

The measurements for the METRIC study of 1999/2000 were carried out in the fifth and sixth year in which the curriculum was implemented in grade 8. At this stage, the institutionalisation of the new curriculum was cautiously taking shape. The contours of this process emerge from the evaluation studies, which were described in this chapter. As a consequence of the ongoing curriculum implementation process, the repeat measurements of the METRIC study in 1999/2000 were collected under different circumstances than in 1995.

Chapter 

Representing a mathematics curriculum

~ Kama ukikataa la mkubwa utatembea kutwa nzima.~

If you refuse the elder's advice, you will walk the whole day.
(SWAHILI PROVERB)

This chapter deals with how the intended, implemented and attained mathematics curriculum can be described, compared and analysed. Section 3.1 provides an overview of descriptive methods for the different appearances of a mathematics curriculum. These lead to instruments for conducting measurements for the METRIC study. The second section reviews how researchers have dealt with links between appearances of a mathematics curriculum (section 3.2). Based on the findings, the research questions of the METRIC study are further operationalised.

3.1 DESCRIBING A MATHEMATICS CURRICULUM

3.1.1 Introduction

This chapter offers a research context for the METRIC study. As described in chapter 1, the METRIC study aimed at finding explanations for discrepancies found in the 1995 TIMSS study. Therefore, it replicated the Written Test and the Performance Assessment, in search for trends and relations between the intended, the implemented and the attained mathematics curriculum. The curriculum under study was the mathematics core curriculum for junior secondary schools in the Netherlands in the last decade of the 20th century. The background of this curriculum has been described in the previous chapter.

This current chapter gives an overview of how intended, implemented and attained mathematics curricula can be described. Generally, descriptions of an intended and implemented curriculum are qualitative in nature, while descriptions of an attained curriculum can both be quantitative (e.g. a resulting score) or qualitative (e.g. an achievement profile).
The qualitative descriptions can vary in extensiveness. The descriptions can be narrative and verbose. At the level of the intended curriculum, we can think of enthusiastic visions. At the level of the implemented curriculum, we can think of detailed accounts of classroom interaction. Descriptions can be very concrete, for example by describing exemplary activities such as 'correctly dividing 544 by 34'. At the level of the intended curriculum this can be an exemplary objective, at the level of the implemented curriculum this can be a classroom activity, and at the level of the attained curriculum this can be a test item.
The descriptions of an intended, implemented and attained curriculum often consist of lists of specifications. The specifications can describe a content aspect (e.g. geometry, algebra) or a cognitive aspect (e.g. 'explaining how an answer was found'). The specifications can be more general (e.g. 'linear equations', implying activities without stating these) or more concrete (e.g. 'listing different activities, which are needed to understand, use and solve linear equations').
By structuring the specifications, researchers have developed frameworks to describe, analyse and compare curricula (see e.g. Bloom, 1956; Lapointe, Mead & Askew, 1992; Niss, 1993; Robitaille et al., 1993; Schmidt et al., 1996; Schmidt, Valverde, McKnight, Houand & Wiley, 1997; Travers & Westbury, 1989). The frameworks usually consisted of a content-by-behaviour matrix. One dimension of the matrix is used for content (e.g. mathematical content areas such as algebra or geometry). Another dimension of the matrix is used for skills or performance expectations. The skills can be put into a cognitive hierarchy, from less complex to more complex competencies (Bloom, 1956). For example, the framework underlying TIMSS contains the categories for performance expectations: knowing, performing routine procedures, performing complex procedures and solving problems. The structure chosen in the METRIC study to describe the RME-based curriculum (Appendix B) uses a non-hierarchical list of skills, with the four mathematising activities, which are typical in RME: modelling, reformulating, interpreting and reflecting (see section 2.2.1). In this way, the particular use of contexts in the mathematics curriculum is underlined by the framework. In Appendix B, the RME-based curriculum has been filled into the framework. It completes the framework, and no cell remains empty. The mathematising activities

are well spread over the different content areas. Thus, this framework can help to analyse whether the intended curriculum under study is balanced.
The same framework can be used at the level of the attained curriculum, by spreading out test items over the framework. In this way, it is possible to assess whether a test covers all cognitive domains and, thus, meets different requirements of the intended curriculum. For example, when the items from the TIMSS Written Test were spread over the framework, the result showed that there were few items in the columns for 'interpreting' and 'reflecting', and few items in the row of 'data processing and statistics'. From an RME perspective, the test lacks, for example, items on visual geometry, tree diagrams, networks, and the application of formula in contexts. Similarly, when the items from the TIMSS Performance Assessment were spread over the framework, the result showed that the test was adequately balanced over all rows and columns. However, it was beyond the scope of the METRIC study, to assess whether the two tests covered the aims of the RME-based curriculum. If the study had included that question, the intended curriculum would be the point of departure. Instead, the METRIC study assessed whether the Dutch students' achievement could be aligned with the appropriateness of the tests in light on the intended and implemented curriculum. Therefore, the attained curriculum was the point of departure, and the intended and implemented curriculum were a context to explain the achievement results.

In the subsequent sections, descriptions of each of the three curricular appearances (intended curriculum, implemented curriculum, attained curriculum) will be reviewed. Afterwards, a review will be presented on how the relationships between the appearances can be described, compared and analysed.
International comparative studies, such as the IEA-studies SIMS (Second International mathematics Study) and TIMSS (Third International Science and mathematics Study) made a considerable contribution in the research on mathematics curricula. These studies will be discussed when the occasion occurs.

3.1.2 Describing an intended mathematics curriculum

The intended curriculum is the appearance of a curriculum at system level. It is defined, as what society expects that students should master (Robitaille et al., 1993). An intended curriculum consists of (a) an *ideal* curriculum and (b) a *formal* curriculum (Van den Akker, 1988, 1998; Goodlad, Klein and Tye, 1979).
The ideal curriculum consists of initial beliefs, visions and views. For example, the

innovative RME-based curriculum for Dutch junior secondary schools was initially described in enthusiastic articles in which the curriculum was vaguely outlined (e.g. Algebragroep W12-16, 1990; Van der Blij & Treffers, 1985; Schoemaker, 1989).
The formal curriculum consists of documents and materials that are based on the ideal curriculum, and which are to be used for implementation. According to Kuiper (1993), these can be found at three levels:

- at system level: curriculum documents, describing the program through core objectives, the scheming of time, the assessment procedures, and so forth. These can have a mandatory or a non-mandatory status.
- at school level: planning document, giving a scheme of when, what and how the curriculum is intended to be implemented, including a choice of textbooks to be used.
- at student level: teaching and learning materials, including textbooks.

The formal Dutch mathematics curriculum for junior secondary schools is legislated through a document, stating the *core objectives* for all school subjects. It contains one five-page chapter, listing the core objectives for mathematics. This mathematics chapter contains a section with general aims on attitudes, followed by content sections in which learning objectives are stated as process objectives. All formulations start with the phrase 'students develop a competence to ...' The sections are clustered by content areas, such as geometry, data processing or algebraic relations (Ministerie van OW&C, 1998). The list is structured into a table in Appendix B.

The Dutch mandated document is clearly concise, when comparing internationally. In other countries, for example Slovenia, the governmental document for the mathematics curriculum consists of hundreds of pages, offering an extensive inventory of aims and objectives, supplemented with a list of when and how topics have to be taught and assessed (Schmidt et al., 1997). Because of the variety in intended curricula, Piper (1979) identified different curriculum styles, focusing either on content, processes or contexts. The mandated Dutch document focuses on content (knowledge and skills), and not on how, why, when or where the instruction should be delivered.
The Dutch document stating the *core objectives* has the vagueness innate to policy documents. It was deliberately kept general, giving teachers, schools and textbook authors a framework within which they had freedom of interpretations (Ten Hove & Van der Zwaart, 1993). Additional non-mandatory publications,

for example from curriculum developers or teacher associations, elaborate and clarify this formal curriculum and offer model exercises to clarify particular intentions (Achtergronden, 1992; Commissie Ontwikkeling Wiskundeonderwijs, 1992; Kok et al., 1992; Lagerwerf, 1994; De Lange, 1992). In this way, the use of contexts in mathematics is exemplified. Also, they offer guidelines for possible teaching approaches, for example, on how to discuss with students the mathematical problems that have more than one correct solution. Thus, these publications are the non-mandatory components of the *formal* curriculum.

According to Van den Akker (1988, 1998) and Goodlad et al. (1979), textbooks are an integral part of the formal curriculum, and thus, of the intended curriculum. However, in the Netherlands, the mathematics textbooks show a great variety in content coverage and instructional approaches. For the METRIC study, a strict description was needed of the intended mathematics curriculum. Including the textbooks into the definition of the intended curriculum created ambiguities in the operationalisation of the research questions. If curriculum experts were asked to judge whether items matched with the intended curriculum, the answer could diverge depending on the textbooks used ('according to textbook X, the item is covered; according to textbook Y, the item is not covered'). Therefore, the legislated core objectives, plus the additional materials, such as additional information guides and background brochures, defined the intended mathematics curriculum for junior secondary schools.

Intended curricula are formulated to be implemented in classroom practice. At the same time, researchers have studied intended curricula to unearth their underlying educational assumptions, intentions and designs. For example, Schmidt et al. (1997) internationally compared countries on how much teaching time was devoted to algebra, geometry, etc.
Generic terminology is used, in order to summarise and structure components of an intended curriculum. The terminology identifies content (e.g. algebra, geometry), skills (e.g. 'using routine procedures' and 'using complex procedures'), and so forth. The operationalisation of the terminology hinges on their interpretations, which can be ambiguous (e.g. routine activities can be complex). The ambiguities are especially revealed when it comes to the interpretation of the terminology describing the intended curriculum. This happens when the intended curriculum is linked to the implemented and attained curriculum, or when intended curricula are compared internationally.

At classroom and student level, the boundaries between the categories describing the intended curriculum can become too restrictive. For example, activities on measuring the sides of triangles can be classified as either proportionality or as geometry. Similarly, 'using complex procedures' can be defined as either including or excluding multi-step routine procedures. Thus, the interpretation at classroom and student level can become arbitrary.
When comparing intended curricula internationally, it turns out that the interpretation of content areas and skills is culturally bound. For example, the Dutch document states that *"students can make calculations with proportions and scale"* (Ministerie van OW&C, 1998, p. 40), without explicitly specifying that the mathematical concepts are to be used within a real-life context. This is clarified in some of the additional documents, and is known to all insiders. Thus, the use of generic terminology can encounter dilemmas, as the interpretations are difficult to define.
The enculturation of the description of an intended curriculum is also illustrated by the fact that some features can be considered self-evident. These might not be stated anywhere, not even in the additional documents. For example, the Dutch governmental document does not state that students should have availability over compasses, ruler and protractor (the so-called 'geo-triangle'). This is taken for granted in all documents. Therefore, to describe intended curricula both qualitatively and quantitatively, clarifications by curricular experts are often needed to specify underlying presumptions, which are culturally bound (Mullis et al., 2000; Pepin, 2001).

All generic terminology describing an intended curriculum has its own origins. The curriculum framework used in TIMSS was created to suit many mathematics curricula (Robitaille et al., 1993), but it did not fit the Dutch intended mathematics curriculum. The main reasons were that (1) it made distinctions between content areas (e.g. numbers, proportions), which are not distinguished in the same way in the Dutch mathematics curriculum, and (2) it did not cater for the integration of mathematics in context. Thus, the TIMSS curriculum framework was deemed unfit for use in the METRIC study. Instead, the METRIC study gathered and analysed data at item level, without categorising these into any curriculum framework. In this way, interpretation dilemmas were avoided. The grouping of test items in the METRIC study was according to their match with the intended and implemented curriculum.

Describing an intended mathematics curriculum through the small-scale constituents at student level, the mathematics exercises, is a challenge. Exemplary tasks have proved to be useful in clarifying the purposes of an intended curriculum, in particular in the design phase. They can serve as archetypes for characterising the intentions, as shown for example by De Lange (1987), Treffers (1987) and Van den Heuvel-Panhuizen (1996). Well-chosen tasks can demonstrate views of an intended curriculum concretely, avoiding the vagueness of generic terminology. Additionally, the boundaries of an intended curriculum can also be characterised by those items that are **not** matching with its aims.

Items can help to illustrate curricular characteristics qualitatively by choosing specific examples. However, items can also be used to describe intended curricula more quantitatively, without choosing particular archetypes. The method consists of taking a large set of various items, and then asks stakeholders to indicate whether or not these are covered by the curriculum. This was first done by Stroomberg (1973). His method consisted of asking relevant respondents to look at candidate items, and then judge whether students should master the skills, which were tested by each item. Others then followed this method, among them Beaton et al. (1996), Gulmans, Van Loon and Pelgrum (1981), Husén (1967), Pelgrum, Eggen and Plomp (1983a, 1983b), Mullis et al. (2000) and Travers and Westbury (1989).

With the above-described method, there are two sides to the coin: a curriculum side, and a test side. In the first case, the items can be used to make explicit what is meant by the generic terminology in which the intended curricula are described. For example, using the data from IEA's Second International Mathematics Study (SIMS), Travers and Westbury (1989) established that curricular diversity between countries was considerable. It was wider when taking statistics and geometry items, and less when looking at arithmetic and algebra items. Through this method, it was possible to identify the countries with a stronger New Math tradition. Thus, the items were used to compare and characterise aspects of the intended curricula.

The counter side of this method is the test side. In that case, the judgement of the items yields an indicator of the appropriateness of the test in light of the intended curriculum. In the METRIC study, the resulting percentage of this measurement is termed the *test-curriculum matching index*. It is a percentage, being the portion of test items matching with the **intended** curriculum. For the Netherlands in 1982, the test-curriculum matching index in SIMS was 79%

(Robitaille & Garden, 1989). This figure was lower than in most other countries. More than a decade later, after introduction of the new RME-based curriculum, the test-curriculum matching index of the Written Test in TIMSS-95 dropped to 71% (Beaton et al., 1996). However, the item sets of SIMS and TIMSS-95 differ. This makes the two matching indices of SIMS and TIMSS-95 incomparable.

The test-curriculum matching index yields an indirect indicator of an intended curriculum. The figures of 79% in SIMS and 71% in TIMSS-95 (the Written Test) show that the Netherlands diverged from the international consensus, as the test-curriculum matching index of most other countries was more than 90%. However, the test-curriculum matching index does not show the extent to which an intended curriculum is covered by the test. This is illustrated in Figure 3.1. The shaded areas are defined by the test items. The test can contain items on content areas, which do not match with the intended curriculum. There is the white part of the intended curriculum, which is not covered by the test. This area remains unmeasured. The test-curriculum matching index does not tell how large this 'terra incognita' is, nor what its characteristics are.

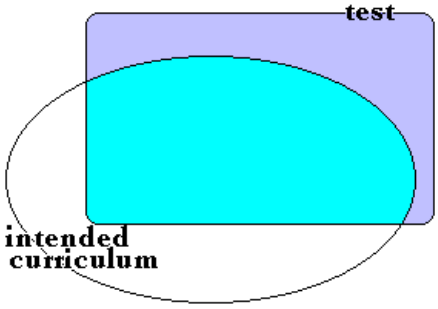


Figure 3.1: Test-curriculum match

The test-curriculum match contains all test items covered by the intended curriculum, whether they are 'near the borderline' or not. The terminology used in the METRIC study will indicate these items as *"items matching with the intended curriculum"* or equivalently, as *"items considered appropriate in light of the intended curriculum"*, or *"items covered by the intended curriculum"*. This terminology is chosen, as some test items in the TIMSS tests were slightly deviant from the Dutch intended curriculum (e.g. because of the multiple choice format, or the lack of context), but still covered. The borderline cases were definitely not RME-based. Therefore, the terminology used by some researchers, such as *"items relevant to the curriculum"* or *"items in the intended curriculum"* (Beaton et al, 1996, p. B-1) was deemed inappropriate for items that matched with the intended curriculum.

In the METRIC study, the proportion of test items matching with the intended curriculum is expressed through a test-curriculum matching index. It differs from:

- the test-curriculum overlap, which is the portion of the test which is covered by the **implemented** curriculum (De Haan, 1992),
- the test coverage index, which is the extent to which the test items cover the intended curriculum of a country (Wolf, 1998),
- the test relevance index, which is the extent to which the test items relate to what is taught according to the intended curriculum (Wolf, 1998), and
- the curriculum coverage index, which is the extent to which a particular intended curriculum matches the more general curriculum formed by the body of content, which might be taught (Wolf, 1998).

The Dutch test-curriculum matching indexes of 79% in SIMS and 71% in TIMSS-95 are indicators of the difference between the Dutch intended mathematics curriculum and the international mainstream. However, the percentage is an aggregated result, which does not reveal any of the special RME characteristics. Therefore, the METRIC study was using the concept of test-curriculum match to gather data on the intended curriculum without aiming at aggregation. The curriculum match was taken at item level. With two tests available, the TIMSS Written Test and the TIMSS Performance Assessment, the total set of items was very heterogeneous, and it probably covered a considerable part of the Dutch intended mathematics curriculum. The items were given to mathematics curriculum experts for judgement. By collecting their judgements on the items in 1995 and 1999/2000, a picture of the appropriateness of both tests in light of the intended curriculum was caught. Indirectly, this would yield a portrayal of the Dutch intended mathematics curriculum at item level as well. The METRIC study used this curriculum match at the level of the intended curriculum for comparison with data from the implemented and attained curriculum.

As explained above, the test-curriculum matching index was an indicator of the test as a whole. However, the METRIC study worked at item level. Therefore, an *item-curriculum matching index* was defined. This rate indicated the match with the **intended** curriculum. This rate was measured on a nominal scale or on a ratio scale. On a nominal scale, the rate was either 'yes' or 'no', indicating whether an individual test item matched with the intended curriculum or not. On a ratio scale, the rate indicated the percentage of experts, who indicated that a test item

matched with the intended curriculum. The ratio scale was introduced in the METRIC study, after Kuiper et al. (1997) noted that carefully selected curriculum experts could disagree on a considerable number of items. Some items were 'on the borderline' of the intended curriculum (i.e. nearly matching with the intended curriculum) and the process to reach consensus on a 'yes' or a 'no' resembled a toss-up. They advised the consultation of a larger number of experts, in order to specify the results as a percentage. With percentages of experts agreeing, the judgement would loose its dichotomy of a yes/no distinction, but gain a rate (a percentage) that identified to what extent items matched with the intended curriculum ('more or less matching'). Therefore, in the METRIC study, the number of experts was increased from three (in 1995 and 1999) to five (in 2000), and the dichotomous yes/no distinction of the item-curriculum matching index for the Written Test was supplemented by a more precise matching index on a ratio scale for the Performance Assessment. This will be further explained in chapter 4 (section 4.3.5).

The instrument used in the METRIC study to measure the match of test items with the intended curriculum had been used in prior studies (e.g. Beaton et al., 1996; Gulmans et al., 1981). The validity and reliability of the instrument depended on several aspects. The objectivity of judgements was increased by asking mathematics curriculum experts to answer the questions independently (i.e. without consulting each other). The reliability of the instrument was considered satisfactory, as Gulmans et al. (1981), using a similar instrument, established a Cronbach Alpha above 0.9.
The validity of judgements depended on the expertise of the experts. They were selected for their acquaintance with the documents describing the intended curriculum. They were well aware of knowledge and skills that students need to complete a certain test item correctly. In this way, the METRIC study gathered judgement data on the two tests in light of the intended curriculum.

3.1.3 Describing an implemented mathematics curriculum

Studying and describing an *implemented* curriculum means studying and describing processes in schools and classrooms (Hoeben, 1993). It can be carried out in various ways. One option is to conduct case studies by direct observation in classroom visits. In particular, the emergence of the educational research direction based on constructivism made many more researchers enter mathematics classrooms (e.g. Clarke, in press; Cobb & Bauersfeld, 1995).

Classroom visits are labour-intensive, and useful to study specific behaviour. But it takes a large number of visits before generalised statements on the implemented curriculum for a whole nation can be made (e.g. Dekker, 1991; Kuiper, 1993; Pepin, 2001). Nevertheless, the Dutch Inspectorate for Education undertook a fascinating exercise to evaluate the new core curriculum for junior secondary schools, by visiting more than 7000 lessons of all subjects in 1997/1998, getting an exhaustive, nation-wide picture of classroom processes (Inspectie van het Onderwijs, 1999a).
The TIMSS Video study also operates at a large, nation-wide scale. It is the first in its kind, documenting national classroom processes in science and mathematics by filming lessons of a representative sample of classes (Kawanaka, Stigler & Hiebert, 1999). A video documentation has a special advantage. The data are re-usable, as observations can be replicated with different research questions in mind. In 1999, the Netherlands was invited to join this study, and the first results are soon to be presented.

In the above named observational studies, the description of classroom processes proves a complex exercise. Classroom processes can take place almost invisibly or simultaneously. The observer (or the cameraman) cannot be an invisible fly-on-the-wall, but becomes part of the classroom processes, making them less genuine. Also, the interpretation of processes by the observer has limited reliability, as it is based on the observers' backgrounds (Kawanaka, Stigler & Hiebert, 1999; Pepin, 2001). To increase reliability, the Dutch Inspectorate made use of duo-observations whereby the independent reports of one lesson by two observers were compared (Inspectie van het Onderwijs, 1999a).

To study the implemented curriculum, alternative research methods replacing classroom observation have been sought. Within TIMSS-95, Schmidt et al. (1996, 1997) studied textbooks, assuming that teachers use these as a foundation of their teaching. Therefore, the textbooks were indicators of *pedagogical flow* at classroom level. As such, the textbooks were used as indicators for the implemented curriculum, although generally, textbooks are perceived as representations at the level of the intended curriculum. Schmidt et al. counted the number of topics introduced and dropped in each school year, introducing the terminology of *average topic duration*. They noticed dramatic differences between nations, with the Dutch mathematics curriculum at junior secondary school having fewer topics than most other countries.

An alternative method to study implemented curricula is the nation-wide study of teachers' logbooks. This was suggested by Alkin (1992) but never carried out. Instead, many researchers have used surveys among teachers. Through questionnaires, they gathered data, which could be considered as indicators of the implemented curriculum. Aspects considered were the mathematical content covered at a certain level, teaching approaches or educational beliefs of teachers. A survey could then be based on any taxonomy (of teaching methods, content covered, attitudes, etc.), on which teachers or school principals were asked to give their judgement.

National surveys have been used in order to research teaching methods, asking mathematics teachers to give a self-report of their pedagogical practices, for example with respect to slow learners or the use of teaching aids (Beaton et al., 1996; Bos & Vos, 2000; Mullis et al., 2000; Travers & Westbury, 1989). Yet, as teachers might tend to give socially desirable answers, the reliability of the answers remain questionable. Also, as Sosniak, Ethington and Varelas (1994) noted from the SIMS data, teachers' opinions show considerable inconsistencies, combining traditional and innovative notions indiscriminately. Therefore, surveys among teachers on their interpretations of the intended curriculum still need further refinement.

Another aspect of the implemented curriculum is content coverage. For example, Freeman et al. (1983), Lapointe et al. (1992), Mullis et al. (2000), and Mehrens and Phillips (1987) used a list of mathematical topics that could have been covered at the particular level of schooling. The general picture emerging from these studies is: diversity between countries, and homogeneity within countries.
The measurement of content coverage by the implemented curriculum is generally associated with the term *Opportunity to Learn* (OTL). This term refers to the rationale that whenever a teacher indicated that he/she has covered certain content, the students have had an opportunity to learn it. Carrol (1963) first introduced the term. With OTL, he meant to measure the time spent on learning a topic (time-on-task). Later, the term OTL changed meaning, becoming a dichotomous variable, indicating whether a teacher had taken the opportunity to teach a certain topic. It does not give an indication of the content that is effectively learnt by the students, as students can also take 'the opportunity to forget' (Van der Linden, 1998).

Instruments for measuring OTL can either be objective-based, topic-based or item-based (De Haan, 1992), although no sources were found in which objective-based OTL studies were reported. Mullis et al. (2000) carried out a topic-based OTL study, which suffers from the ambiguity of the terminology used. It is questionable whether Dutch mathematics teachers interpret the descriptions well. For example, grade 8 mathematics teachers were asked whether they had covered the following topics from algebra:

- number patterns and simple relations,
- simple algebraic expressions,
- representing situations algebraically; formulas,
- solving simple equations, and
- solving simple inequalities.

The difficulty in interpreting the specifications is caused by the fact that the title of the item equals its description. To one mathematics teacher, the topic 'simple algebraic expressions' could have a totally different meaning than to another teacher. First, the description of the topic does not specify for activities (these could be ordering of terms, expanding, rewriting, interpreting the magnitudes, inserting values, classifying, and so forth). For some teachers the topics might mean a mathematical concept, requiring a strict definition. But then the concept of 'simple algebraic equations' should specify for the number or nature of the variables in the expression. Also, the usage of the word 'simple' seems to value certain cognitive limits, as opposed to 'complicated'. Finally, viewed from the Dutch RME-based position, it seems that the use of contexts is limited to the third item only ('representing situations algebraically; formulas'), apparently leaving contexts out from the other topics. Moreover, the sequence of topics implies that the real-life applications are coming after the introduction of the concept. This makes little sense in RME, where the concept is introduced through contexts. Therefore, research on OTL through topic-based instruments could have limited value for certain nations, if the items inhibit interpretation ambiguities within its educational culture.

Instead of using an objective-based or a topic-based approach for studying OTL, an item-based approach can be used. Items offer teachers a concrete cognitive situation, stating implicitly aspects of content and skills. Not only teachers, but also students, can be given a set of test items with the question, whether these

have been covered in their lessons. This method was used by Borg (1979), Comber and Keeves (1973), De Haan (1992), Hardy (1984), Husén (1967), Pelgrum et al. (1983a, 1983b, 1986), and Travers and Westbury (1989). In the studies, teachers were asked whether the content, needed to solve the item, had been taught before test administration or not. In some of the studies, teachers were additionally asked to indicate the time spent on that content, thus differentiating the yes/no question on content coverage.
Travers and Westbury (1989) have argued that the question on content coverage was interfered by teachers' perception of the item's difficulty. They reasoned that teachers would include an estimation of their students' achievement before answering to the coverage question. Therefore, they asked teachers to distinguish between item difficulty and item coverage. Teachers were asked to indicate the coverage, and separately make an estimation of the percentage of their students answering the item correctly.

Considering the interference of coverage with difficulty, De Haan (1992) developed a special instrument to measure OTL, by asking teachers to imagine that they were to set a test for their students, covering all content taught. They were then given a set of items, and asked to judge each item on whether they would include it into their imaginary test. This implied that either the content of the item had been covered in classroom, or that the item was proximate to the content taught and teachers considered his/her students to be 'ready' for the item. In both cases, teachers' answers were an indicator of the item's match with the implemented curriculum. The OTL rate of an item was then calculated as the percentage of teachers, indicating 'yes' on that item. De Haan checked the validity of her instrument, and its reliability and objectivity in a research, in which the obtained OTL rates were compared to

- a detailed OTL questionnaire asking for each item when, how and for which sub-group of students it was covered,
- an estimation by teachers of item difficulty (i.e. teachers' prediction of student scores on the item),
- a textbook analysis (i.e. was the item covered in the textbook used),
- students' own OTL judgement (i.e. was the item covered in class?), and
- students' achievement.

The new OTL instrument correlated well, both with the detailed OTL questionnaire and with the textbook analysis ($r=0.9$ and 0.7 respectively, with $p<0.01$), testifying of the validity of the instrument. The correlation with the

estimated rate of difficulty was 0.6. The correlation with students' own OTL judgement varied between 0.5 and 0.8 (at different measurements). Finally, the correlation with students' score was 0.5.
To check the reliability of answer patterns, De Haan studied the consistency of answers by repeating the OTL measurement later in the school year. By correcting for the additional content taught, the two measurements showed a high consistency (92%) and significant correlation ($r=0.84$, $p<0.01$).
The above-described instrument was used by Kuiper et al. (1997, 2000) in their research on the National Option Test (NOT) and by De Haan et al. (1997) in a research on the TIMSS Performance Assessment. The latter established a Cronbach Alpha varying between 0.70 and 0.92. Therefore, the instrument gave much confidence to be useful for the METRIC study.

With this carefully developed and well-tested instrument, the METRIC study had the availability over a powerful tool to assess the two TIMSS tests, the Written Test and the Performance Assessment, in light of the implemented Dutch mathematics curriculum in grade 8. On each item an OTL rate was gathered, offering insight into whether the teachers considered the item suitable for an imaginary test for their students. The data were then compared to data, which were gathered at the level of the intended and attained curriculum.

3.1.4 Describing an attained mathematics curriculum

The *attained* curriculum is generally associated with the registration of students' achievement outcomes. In many countries, national exams are administered at the end of secondary education. The exams create a barrier for future careers, and create national standards for the educational level at schools. By linking the exams to the *intended* curriculum, a nation can evaluate whether educational aims have been met.
Yet, national exams can only give an indication of an attained curriculum at the end of a process, and not during its course. Therefore, in the Netherlands educational monitoring is carried out by testing samples of students during their course of schooling. Unlike exams, students do not receive special preparation for these tests. Examples of projects in which tests are used for monitoring the attained mathematics curriculum in the Netherlands are the Periodical Measurement of Primary Education (PPON) (Hoeben, 1993; Wijnstra 1990) and the Secondary Education Cohort Students (VOCL) (Van der Werf et al., 1999).

Table 3.1: Achievement results (average p-values) of countries participating in SIMS

	Country	Arith-metic	Algebra	Geometry	Statistics	Measure-ment	Total
1	Japan	60	60	57	71	69	62
2	Netherlands	60	52	53	67	63	58
3	Hungary	57	51	54	61	62	57
4	France	58	55	38	57	60	53
5	Belgium (Fr)	58	51	44	53	57	52
6	Belgium (Fl)	57	51	42	58	58	52
7	Canada (BC)	57	48	42	60	52	51
8	Finland	49	46	45	61	54	50
9	Hong Kong	55	43	43	56	53	49
10	Canada (O)	54	42	42	56	51	49
11	Scotland	51	43	45	60	49	49
12	England	48	39	44	60	48	47
13	Israel	51	46	46	53	47	46
14	New Zealand	46	40	46	58	46	46
15	USA	51	43	38	58	41	45
16	Sweden	43	34	40	60	52	44
17	Thailand	44	38	40	46	49	43
18	Luxembourg	48	34	26	39	52	39
	Intl average	*53*	*45*	*44*	*57*	*54*	*50*

Source: Pelgrum et al. (1986).

National authorities also use international comparative studies to monitor the attained curriculum. The most famous studies are the ones, which were organised under auspices of IEA (International Association for the Evaluation of Educational Assessment). Indicated by their ordinals first, second and third, these are FIMS, SIMS and TIMSS. Other studies monitoring mathematics education internationally are the Programme for International Student Assessment (PISA) (OECD, 2000), International Assessment of Educational Progress (IEAP) (Lapointe et al., 1992), and the International Programme on Mathematical Attainment (IPMA) (Burghes, 1999). Generally, Dutch students perform relatively well in international comparative studies. As an example, in Table 3.1, the achievement results (average p-values) of countries participating in SIMS are given. The data were gathered in 1980-1982. Here, the Dutch grade 7 students scored relatively well in the international comparison, just like the Dutch grade 8 students did on the TIMSS Written Test in 1995 (cf. Table 1.2).

When measuring students' mathematics achievement for a large population, labour-intensive research through observation, interviews or project work cannot be considered. Instead, in all of the projects named before, paper-and-pencil tests have been developed and issued to large samples of students. In particular, multiple choice items have been popular in such studies, because of automated scoring and high reliability. For example, in 1995, TIMSS tested more than half a million students around the world (Beaton et al., 1996). Consequently, millions of answers had to be uniformly assessed on their correctness.

Yet, the use of multiple choice items for measuring students' achievement has come under debate. Van den Bergh (1988), Frary (1985), Resnick and Resnick (1989), and Wiggins (1989) were among the first to criticise the use of standardised, multiple choice tests. These items were associated with low valued factual knowledge, asking for limited thinking processes. Closed questions appear to be easier and less valid than comparable open questions.

Additional assessment methods were needed. The rhetoric against multiple choice tests made almost any test with a few open questions, already non-traditional and modern. The labels used by different authors varied: *performance* assessment or *performance-based* assessment (Burton, 1996; Clarke, 1996; Harmon et al., 1997; Linn & Baker, 1996; Shavelson, 1994; Wolf 1994), *alternative* assessment (Birenbaum & Dochy, 1996), or *authentic* assessment (Darling-Hammond & Ancess, 1996; Glatthorn, 1996; De Lange, 1992; Wiggins, 1989). The descriptions showed considerable overlap. Characterisation of the new assessment practices were given, such as:

- testing through open questions and for higher order skills,
- using a range of methods or approaches,
- making students disclose their own understanding,
- allowing students to undertake practical work,
- asking for performances and products,
- integrating real-life situations,
- integrating school subjects, and
- being as an activity worthwhile for students' learning.

For simplicity, in the following text, the most current term of *performance* assessment will be used if some of the above criteria are met.

In relation to the Dutch new core curriculum, De Lange (1992), Hermans (1992) and Sluijter et al. (1996) also listed the above characteristics. They described new

criteria for testing the new core mathematics curriculum at junior secondary school level. Since the 1990s, none of the exemplary tests for the new mathematics curriculum in the Netherlands does contain multiple choice items. The tests, developed by the curriculum developers of the W12-16 team, and, from 1992 onwards, developed by the National Institute for Educational Measurement (Cito), are written tests containing predominantly short answer questions and a few extended response items asking for a description or an explanation. Their distinguishing feature is that all questions are somehow related to real-life contexts. However, not one of the mathematics tests asks for open investigations, in which students have to design a research, compile data and model them. This is possibly caused by the time-restricted and paper-based format, but the challenge remains to develop hands-on tasks for national performance assessment in mathematics.

With the emergence of performance assessment as an alternative to standardised tests, the question remained whether they would give fair information on students' achievement. Therefore, Baxter and Shavelson (1994) compared the exchangeability of different assessment methods for the subject of science. These were observation, notebook reports, computer simulation, short answer questions and multiple choice questions. They found that observation of a student, carrying out a hands-on, empirical investigation, yielded most detailed information on students' achievement. According to them, notebook reports providing a reasonable 'surrogate'. All other tests failed to approximate the same information.
Therefore, the use of investigational tasks in large-scale assessment such as the TIMSS Performance Assessment, whereby students' notes are scored, is an attractive option with potential for describing an attained curriculum. Still, reliability issues need to be taken care of (Linn & Baker, 1996). This was confirmed by Zuzovsky (1999), who mentioned discrepancies in the coding of students' responses to the free response items. She accounted the disagreement between coders to their experience in teaching the subject. Extensive training of coders could not level out the difference. Another problem faced in the use of performance assessments in international comparative studies is the language. Translation can shift the meaning of words slightly. In particular with open tasks, different wordings between tasks can prompt for significantly different results (Shavelson, Baxter & Pine, 1992). Despite the reliability issues, the TIMSS Performance Assessment was considered relevant for the METRIC study.

Considering the advantages of standardised multiple choice tests for their reliability and performance assessments for their validity, the METRIC study chose best of both worlds by combining the two. In the METRIC study, two tests were used: the TIMSS Written Test and the TIMSS Performance Assessment. The two tests combined a large variety of items, with respect to content, cognitive skills and mathematising activities. Both tests were used in measuring students' achievement, yielding data on the attained curriculum.

Issues pertaining the reliability and validity of the TIMSS Performance Assessment have been reviewed above. Issues pertaining the reliability and validity of the TIMSS Written Test are reviewed below, although not one study could be traced that questioned the reliability of the Written Test. On the contrary, the Written Test has gained a high reputation with respect to reliability (Beaton et al., 1996; Mullis et al., 2000). However, the validity of the Written Test was questioned.

De Lange (1997a, 1997b) questioned the validity of the TIMSS-95 Written Test, mentioning three flaws: (1) there were too many multiple choice items that hardly ever ask for higher order aims; (2) the assessment framework was questionable; and (3) the Written Test did not take enculturation into account. The reasoning for his last statement was based on the league table of country scores. The countries could be grouped by culture in the following sequence: first the Asian countries, followed by small Western European countries, small Eastern European countries, larger Eastern European countries, Commonwealth countries, Scandinavian countries, South European countries, and finally, developing countries.

The statements from De Lange can be countered as follows. First, it needs to be considered that TIMSS was not developed to suit the needs of one particular country, but to make a comparison between countries. Therefore, many items were needed to cover a wide range of mathematics content. The uniform coding of this large number of items across countries was ensured by the multiple choice format. Thus, the format raised the reliability of the international comparison. His remark on the lack of higher order aims is partly right, but does not consider the fact that even multiple choice items can ask for higher order aims. Second, indeed, a number of items proved not to fit well into the framework of Robitaille et al. (1993). However, a disqualified framework does not disqualify the items themselves, nor the test. On the contrary, items that do not fit into the framework can be those that ask for transfer of knowledge and skills across topics or performance expectations, or across both. Therefore, these can be

items that ask for higher order aims. They make the test more varied. Finally, the fact that the countries in the TIMSS league table could be grouped according to culture is a matter of face validity (Krathwohl, 1998). The league tables of the TIMSS Written Test show that culture is an underlying variable in mathematics education. It does not mean that the mathematics test tries for culture as such.

Kuiper et al. (1997, 2000) researched the validity of the TIMSS Written Test for Dutch students. They added a National Option Test (NOT) in the data collection of 1995. This additional test was based on the new RME-based curriculum. Besides 20 newly developed items, it contained 16 items from the TIMSS Written Test. The 16 items were considered relevant to the intended RME-based curriculum, fitting it well when ignoring the multiple choice format. As the 16 items were common in the NOT and in the TIMSS Written Test, the researchers used them as anchors for comparing achievements on the TIMSS Written Test and the National Option Test. Kuiper et al. (1997, 2000) concluded that both tests measured students' achievement along the same scale. Therefore, the validity of the TIMSS Written Test was not dismissed.
The TIMSS Written Test was developed in a careful process in which many mathematical curriculum experts from all participating countries were consulted (Garden & Orpwood, 1996). In this process, Dutch educators brought in some TIMSS items, and other items were almost equivalent to test items in Dutch national tests. These items add a Dutch essence to the test. In hindsight, the curricular validity is also expressed in the judgement by curriculum experts. They indicated that approximately 70% of the items in the Written Test matched with the Dutch intended curriculum for Dutch junior secondary schools. To supplement the Written Test, the METRIC study used the TIMSS Performance Assessment as a complement in validity. This test did not contain multiple choice questions, but asked students to gather data, model these and give explanations for their findings. By combining the two tests, students tested in the METRIC study had to show a broad spectrum of knowledge and skills.

Thus, for every test item in both tests, the METRIC study gathered data with regard to all three curricular appearances. Curriculum experts were consulted to judge the items and indicate their appropriateness to the intended curriculum. Teachers were consulted to judge the items and indicate the OTL. The data were all at item level, enabling a one-to-one mapping with the achievements of the students for comparison with the attained curriculum.

3.2 LINKING CURRICULAR APPEARANCES

The METRIC study was founded on two research questions investigating the discrepancies (1) between students' results on the two TIMSS tests, and (2) between the intended, implemented and attained curriculum. Therefore, a literature study was conducted on associations between the intended, the implemented and the attained mathematics curriculum. Based on the literature, the research questions were further operationalised.
Combining an intended, an implemented and an attained mathematics curriculum, there are four possible combinations to make:

a. linking the intended and implemented curriculum,
b. linking the intended and attained curriculum,
c. linking the implemented and attained curriculum, and
d. linking all three curriculum appearances.

Below, the combinations will each be reviewed in separate sections.

3.2.1 Describing links between an intended and an implemented mathematics curriculum

According to Hoeben (1993), an intended curriculum can be evaluated in two ways, (a) by a description of processes or (b) by a description of effects. In the first case, the intended curriculum is linked to the implemented curriculum. In the second case, the intended curriculum is linked to the attained curriculum. The latter will be summarised in section 3.2.2 below. The first, evaluating an intended curriculum through a description of processes, is a clear example of how the intended and implemented curriculum can be connected.
In the Netherlands, the Inspectorate of Education evaluated the core curriculum for junior secondary schools by focussing on classroom processes. They visited classrooms in 1997/1998, five years after the introduction of the new core curriculum in 1993. The inspectors assessed the implemented curriculum in light of the intended curriculum (Inspectie van het Onderwijs, 1999a). The conclusions of the Inspectorate were, that the innovations had taken place at administrative level, but did not yet appear clearly at instructional level (see section 2.3.3).

There are a few other studies that link the intended and the implemented mathematics curriculum. An example is the SIMS study. Here, the test items were used to compare the coverage by the intended curriculum (the item-curriculum matching index) and the coverage by the teachers (OTL) (Robitaille

& Garden, 1989; Travers & Westbury, 1989). Some countries showed a large discrepancy between the two coverage sets. Other countries showed small discrepancies. As a reason for the large discrepancies, it was noted that many intentions ran ahead of the implementation. In some countries, the SIMS national research centres admitted in hindsight that their country's intended curricula might have been over-ambitious and unrealistic.

In SIMS, the Netherlands was one of the countries with a small discrepancy. The item-curriculum matching indices and the teachers' OTL rates were reasonably at par at that moment. In the year of data collection, 1982, the intended curriculum had been stable for many years. In Table 3.2a, the rates have been aggregated for mathematical content areas. Areas with a high curriculum matching index (i.e. the match with the intended curriculum) also show a high OTL rate (i.e. the match with the implemented curriculum). In all content areas, the matching index of the intended curriculum is higher than the average OTL rates. The area of statistics shows an outlier. This could reflect the delayed implementation of statistics as a topic in junior secondary classrooms. In 1968, this topic was included into the intended curriculum, but it was not assessed at this level (only in grade 10). It is possible that, as a result of the assessment practice, this topic was not covered by the implemented curriculum for grade 7.

In 1999, similar data were collected in TIMSS-99 (the Written Test), with the OTL data being collected as a national option in the Netherlands only (Bos & Vos, 2000). Table 3.2b shows the appropriateness towards the new intended curriculum and the OTL rates from that study.

Table 3.2a: Test-curriculum matching indices and OTL rates in SIMS for the Netherlands

Content area	**Curriculum matching index (%)**	**Average OTL rate**
Arithmetic (62 items)	81	81
Algebra (32 items)	81	73
Geometry (42 items)	71	66
Statistics (18 items)	76	32
Measurement (26 items)	89	82

Source: Robitaille & Garden (1989).

Table 3.2b: Test-curriculum matching indices and OTL rates in TIMSS-99 (Written Test) for the Netherlands

Content area	Curriculum matching index (%)	Average OTL rate
Numbers (51 items)	63	89
Algebra (28 items)	71	70
Geometry (22 items)	59	78
Data repr. (21 items)	86	80
Proportions (12 items)	83	83
Measurement (21 items)	86	89

Source: Bos & Vos (2000).

There are a few differences between the results of SIMS in 1982 and TIMSS in 1999. As described in chapter 2, the intended mathematics curriculum for junior secondary schools had been reformed in 1993. As a result, there are lower test-curriculum matching indices, in particular with respect to numbers, geometry and algebra. On the other hand, the OTL rates in 1999 are high in almost all content areas, comparable to the OTL rates in SIMS (with the exception of statistics). But the Tables 3.2a and 3.2b cannot be compared well. In the first place, the levels differ, with SIMS testing grade 7 and TIMSS-99 testing grade 8. Additionally, the tests in SIMS and TIMSS-99 contained different item sets and a different taxonomy of mathematical content. Still, one might be tempted to say that in Table 3.2b the OTL rates are slightly higher and the test-curriculum matching indices are slightly lower. This discrepancy was object of further study of the METRIC study, as one of the research questions included the issue of a possible discrepancy the appropriateness of the tests to (a) the implemented curriculum and (b) the intended curriculum.

3.2.2 Describing links between an intended and an attained mathematics curriculum

Research, in which an intended mathematics curriculum is linked to the attained curriculum, is usually carried out to assess whether students learn what they should learn, and whether unwanted side effects are avoided. This can be an evaluation of the intended curriculum, taking the shape of a program evaluation describing effects (Hoeben, 1993).

When linking the intended mathematics curriculum with the attained curriculum at a national level, the tests measuring students' achievement are constructed, starting from the intended curriculum. This can be achieved by consulting curricular experts and other stakeholders during the developmental process of the test items, in order to make the test meet the standards of the intended curriculum (Hoeben, 1993; Niss, 1993; Wood, 1991). In the Netherlands, a team of curriculum experts, psychometrists and teachers develops the yearly national exams at the end of secondary schooling. Herein, intended and attained curriculum are clearly linked. In the Netherlands, an average pass rate of 65% for mathematics exams is considered acceptable (Boon, 2001).

A general challenge in linking the intended and the attained curriculum is the mutual coverage of the two. On one hand, there is the coverage of the intended curriculum by test items. The question is to what extent can all aims of an intended curriculum be tested, in particular when it comes to social or investigative attitudes. Many authors have noted this, for example, De Lange (1987), Hawker and Ollerton (1999), Niss (1993), and Shavelson, McDonnell and Oakes (1989). They state that in general, a test is 'less' than an intended curriculum. On the other hand, test items can also ask for knowledge and skills, which are not described in an intended curriculum. This can be the case when an item builds upon presupposed knowledge and skills. For example, an item may require reading skills, before the actual mathematical task can be completed. In that case, a test covers 'more' than the intended mathematics curriculum. Therefore, improvement of test construction is still on the international agenda, in particular when these are based on a modern intended mathematics curriculum (Birenbaum & Dochy, 1996; Burton, 1996; Clarke, 1996; Kilpatrick, 1993).

An intended and an attained curriculum can also be linked through achievement tests that were **not** created to suit a particular intended curriculum. In international comparative studies, one test is issued to all students, irrespective of the intended curriculum of their nation. Through the international tests, new insights came forth on the link between the intended and the attained curriculum. A point of debate was whether a test could be fair to students of any nation, in particular if the test was not suiting their intended curriculum. Freudenthal (1975) already raised this point with the first large-scale IEA-study on mathematics education, FIMS. Therefore, in ensuing studies such as SIMS, TIMSS and IEAP, a forum of international curriculum experts was created to build consensus on the

set of test items (Beaton et al., 1996; Lapointe et al., 1992; Mullis et al., 2000).
In 1995 and in 1999, a special analysis was made in which the appropriateness of the TIMSS Written Test, with respect to the intended curricula, was related to the students' scores. It was named Test Curriculum Matching Analysis (TCMA) (Beaton, 1998; Beaton et al., 1996; Mullis et al., 2000). In all countries, curriculum experts were asked to assess the test items and indicate whether these matched with their respective intended curricula. Then, taking the set of only those items, which matched with the intended curriculum of one particular country, the test scores of all countries were recalculated. Thus, league tables of all participating countries could be established based on the Dutch intended curriculum, or on the Singaporese intended curriculum, or on the Slovenian intended curriculum, and so forth. All resulting country lists turned out to be very similar to the original country list of the overall test. The order of the countries remained virtually the same.
The analysis showed that different item sets had a minimal effect on the general relationship among countries, witnessing of the robustness of the TIMSS Written Test. When omitting items, which were not covered by a country's intended curriculum, this tended to improve the results for that country, but also for other countries. Therefore, the overall pattern remained unaffected. This analysis revealed, "*that different item selections do not make a major difference in how well the countries perform relative to each other*" (Beaton et al., 1996, p. B-5).
Additional secondary analyses based on TCMA showed that students sometimes can underachieve on certain content areas of their intended curriculum. However, the analysis also showed that students could score well on items that are not covered by the intended curriculum. For example, Bos and Vos (2000), Kuiper et al. (1997, 1999) and Sawada (1999) noted that in some cases students scored well on content that was not part of their intended curriculum. In particular, the result of Kuiper et al. (1997, 1999) is noteworthy. They found that Dutch students' achievement on the TIMSS-95 Written Test did not differ significantly between items that matched with the intended curriculum, and items that did not match with the intended curriculum. Thus, students were somehow knowledgeable about content, which did not match with their intended curriculum. This could have different causes: students were able to make a mental transfer from content taught and bridge the gap to new content, or students had learnt the content in other environments (in other subjects or out of school), or the implemented curriculum differed from the intended curriculum (Kuiper et al., 1997).

The METRIC study was also using the TCMA instrument to ask Dutch curriculum experts to judge the test items on their appropriateness to the intended curriculum. This resulted in two disjoint item sets: one set with items matching with the intended curriculum, and a complementary set of items not matching with the intended curriculum. The achievement of students on the two separate sets would create an insight into the link between the intended and the attained mathematics curriculum.

3.2.3 Describing links between an implemented and an attained mathematics curriculum

There are many studies, linking the implemented mathematics curriculum to the attained curriculum. Several of these consist of case studies, in which aspects of classroom processes are related to psychological processes and learning outcomes. Examples of researchers' emphases are teaching style, classroom atmosphere, mathematical approach, or instructional approach (e.g. Baroody & Ginsburg, 1986; Bishop, FitzSimons & Seah, 2001; Boaler, 1997; Ernest, 1988; Gray & Tall, 1994; Piaget, 1952; Pimm, 1987; Skemp, 1986). Other case studies focus on particular student groups, based on classroom heterogeneity, gender, ethnicity, socio-economic status (SES), and so forth. The studies investigate which environments might offer better support for mathematics learning (e.g. Brown, 2000; Shavelson, McDonnel & Oakes, 1989).
In the Netherlands, a few studies have been carried out, linking the implemented curriculum to the attained curriculum, irrespective of the intended curriculum. For example, Dekker (1991) experimented with learning mathematics of 15-year old students in small, mixed-ability groups. Zwaneveld (1999) offered secondary school students the instrument of concept mapping as a cognitive tool to support their learning. Both researchers found that the adaptations in the instructional environment caused minor improvements in the attained curriculum. Other research, using case studies, was carried out by Van den Heuvel-Panhuizen (1996), De Lange (1987), Perrenet (1995), and Schalkwijk (1998).

In general, when linking the implemented to the attained mathematics curriculum, many authors have stated that in general students learn what they have been exposed to (Hiebert, 1999; Husén 1967; Robitaille et al., 1993; Shavelson et al., 1989; Travers & Westbury, 1989). Students will not learn, if they have not had the *opportunity to learn*. Also, the more time is spent on certain content, the better it is mastered (Pelgrum, 1990). For example there is a clear

correlation between the emphasis a nation or school puts on for example algebra and students' mastery of that content (Lapointe et al., 1992; Sawada, 1999). Additionally, there is a correlation between the materials used and students' results (Fey-den Boer, 1999; Van Streun, 1985; Thompson and Senk, 2001).
When comparing the effect of different teaching approaches, pretest-posttest research has proved useful to link the implemented curriculum to the attained curriculum. Boaler (1997) compared two equivalent schools, where totally different approaches to mathematics teaching were practised. She convincingly demonstrated that a modern teaching approach helped students to gain more self-esteem, and to obtain overall higher performances. More recently, Thompson and Senk (2001) compared the achievement of pairs of classes on the consequences of different teaching materials. They established that the student groups, who used new materials, improved their scores on non-traditional items, while they maintained procedural skills on traditional items.
In the Netherlands, Van Streun (1985) used a pretest-posttest research design to study the impact of different materials on the attained curriculum. He gathered data in 21 classes at 7 schools. After students had been tested at the end of grade 8, the classes were randomly allocated to three different textbooks, which were the only materials used throughout their grade 9. The research was controlled for differences between teachers. At the end of their school year in grade 9, after exactly 95 lessons for each class, the students were tested again. The research established a significant difference in the attained curriculum between the three groups, demonstrating obvious benefits of a heuristic approach in mathematics teaching.
Another Dutch study in mathematics education, based on a pretest-posttest design, was reported by Fey-den Boer (1999). She found that achievement of engineering students at university correlated with mathematics textbooks used at secondary school. In this case, the implemented curriculum at secondary schools had an impact on the attained curriculum at a later stage of schooling.

Yet, most of the above studies only characterise processes for particular groups or for particular approaches. They do not depict a national overview. Only Van Streun's research has a scale (21 classes with randomised materials used) that allows for generalisation on a national level.
The only existing large-scale pretest-posttest research in mathematics education at country level was a national option within IEA's SIMS. Six out of the twenty countries participating in SIMS joined this study. Here, a longitudinal design was

used to establish at national level, whether students' learning gain in one year can be ascribed to the content covered within that same year. In general, the answer was affirmative. The more new content was offered, the higher the learning gain (Burstein, 1993).
The Netherlands did not participate in this longitudinal pretest-posttest study of SIMS. The Netherlands only participated in the cross-sectional test in which all cumulative knowledge and skills were tested. Yet, the snapshot data also proved useful to investigate a link between the implemented and the attained mathematics curriculum. Pelgrum et al. (1983a, 1983b) used the SIMS data to analyse the Dutch mathematics curriculum for 13-year old students. They compared the implemented and attained curriculum at topic level and at item level, ignoring the intended curriculum. In Table 3.3a, some of their data are given.

Table 3.3a: OTL data and students' achievement in SIMS for the Netherlands

Content area	**OTL rates** *	**Achievement (avg p-value)**
Arithmetic (62 items)	78	60
Algebra (32 items)	70	52
Geometry (42 items)	64	53
Statistics (18 items)	31	67
Measurement (26 items)	79	63

Source: Pelgrum et al., 1986;

Note: * The OTL rates differ from those in Table 3.2a because of different statistical methods.

When comparing the OTL data with achievement scores at item level, Pelgrum et al. (1983b) found a correlation (r=0.22, n=229). This meant that, in general, items with a high OTL rate also had high students' scores, and conversely, items with low OTL rates had low students scores. However, this correlation of r=0.22 was not considered high. Additionally, Pelgrum et al. asked teachers to predict students' scores. In this way, they made teachers differentiate between 'content taught' (OTL) and 'content learnt' (predicted scores). They found a stronger correlation between teachers' prediction of students' achievement and students' actual achievement (r=0.77, n=229) than between OTL rates and students' achievement. Therefore, they doubted the validity of the OTL instrument used in SIMS. This result lead to further studies conducted by Pelgrum (1990) and De Haan (1992), establishing more reliable and valid instruments for OTL

measurement. The resulting instrument developed by De Haan has been described in the section on OTL (section 3.1.2) and was used in the METRIC study.

More than a decade after SIMS, in TIMSS-99, OTL data and achievement data were compared as well, but only at topic level (Bos & Vos, 2000). The data are displayed in Table 3.3b.

Generally, the achievement data were approximately 20 points lower than the OTL data. It meant that topics with more emphasis in classroom yielded higher student results, and topics with less emphasis in classroom yielded lower student results. The exceptional topic area was data representation, which had comparatively high achievement results compared to the OTL rates.

Table 3.3b: OTL data and students' achievement in TIMSS-99 (Written Test) for the Netherlands

Content area	OTL rates	Achievement (avg p-value)
Numbers (51 items)	89	67
Algebra (28 items)	70	57
Geometry (22 items)	78	61
Data repr. (21 items)	80	78
Proportions (12 items)	83	62
Measurement (21 items)	89	63

Source: Bos & Vos (2000).

The data in Bos and Vos (2000) were aggregated over mathematical content areas. This was not ensued by the METRIC study, because the allocation of items to content areas inhibited arbitrariness in some cases. Therefore, the METRIC study preferred to keep the information on OTL and achievement at item level, and compare these. This yielded information, firstly on the number of teachers who judged each item suitable for an imaginary test, covering all content taught, and secondly, to what extent the students correctly solved it. In this way, the implemented and attained curriculum were linked.

3.2.4 Describing links between all three curriculum appearances

This final section describes research on possible links between an intended, an implemented and an attained mathematics curriculum. However, there are only few studies to be mentioned. One example is the observational study carried out by the Inspectorate for Education (1999a). The Inspectorate set out to explore

the accomplishments of the new core curriculum in classroom practice. Thus, they used the core objectives as benchmark. The bulk of their research is based on classroom observations, in which they assess the implemented curriculum in light of the intended curriculum, however, without perspective on the attained curriculum. However, a small part of their research is based on secondary analyses of test results from the National Institute for Educational Measurement (Cito). Thus, they linked intended, implemented and attained curriculum. Their overall conclusion was, that the intended curriculum did not yet live up to its promise (see section 2.3.3).

Other examples of research on the link between all three curriculum appearances are the IEA studies SIMS, TIMSS-95 and TIMSS-99. Their starting point is the attained curriculum, for which they want to find explanatory parameters at the other two levels. In SIMS, many curricular variables at teacher level were gathered. But the analysis of data proved labour-intensive. Each analysis required its own computer program, which had to be written in computer languages, such as FORTRAN. The main international publications came many years after data collection (Burstein, 1993; Robitaille & Garden, 1989; Travers & Westbury, 1989), drawing attention among researchers to the availability of the data gathered. The large time lapse between data collection, analysis, and presentation did not encourage extensive usage of all available data.

With the emergence of statistical packages, such as SAS and SPSS, researchers gained easier access to data handling, including multi-level analysis. Therefore, TIMSS increased research options. The emphasis moved towards variables of school effectiveness, such as extra-curricular background variables at the macro, meso and micro level (school facilities, teacher qualifications, home situation of students, etc.). Within TIMSS, an abundant amount of data was gathered and made available to researchers. Therefore, many more secondary analyses could emerge within a few years after the data collection (e.g. for the Netherlands: Bos, 2002; Bos et al., 1999, 2001; Bos & Vos, 2000; Kuiper et al., 1997, 1999, 2000; Vos & Bos, 2001a, 2001b; Vos, Kuiper & Bos, 2000;).

Both SIMS and TIMSS gathered curricular data at all three levels, but their approaches differed. In SIMS, data were collected at item level. Therefore, in SIMS, each test item was judged by curricular experts, by teachers and by students (i.e. their achievement). However, not one of the SIMS secondary analyses could be traced, in which all three levels were analysed in conjunction. Pelgrum et al. (1986) focussed on implemented and attained curriculum only.

Robitaille and Garden (1989) focussed on intended, implemented and attained curriculum, but aggregated for five content areas (arithmetic, algebra, geometry, statistics and measurement) without synthesising their data.
In TIMSS, data at all three levels were collected, but with different bases. The data at the intended and the attained curriculum level were item-based, while the data at the implemented curriculum level were topic-based. Possibly, the presumed unreliability of the instrument was a reason for this omission (Pelgrum et al., 1983a). Because of the different bases, the data on the three curriculum appearances are not compatible. Only the data on the intended and the attained curriculum have the same base, thus enabling the Test Curriculum Matching Analysis, which was described in section 3.2.2.
In the Netherlands, the TIMSS database was complemented. As a national option, additional OTL data at item level were collected on the implemented curriculum (Bos & Vos, 2000; Kuiper et al., 1997, 1999). These data were used in the METRIC study, in order to establish a link between intended, implemented and attained curriculum. In this way, the initial ideas from SIMS, to collect data from all three curriculum appearances at item level, was combined with the instrument of De Haan (1992), and attached to the TIMSS studies in 1995 and in 1999. This combination enabled research to find answers to the questions on observed discrepancies between the intended, implemented and attained mathematics curriculum in the Netherlands. In fact, without data at the level of the intended and the implemented curriculum, it is futile to interpret students' achievement results.
This will lead us to the following chapter, in which the research design of the METRIC study will be explained.

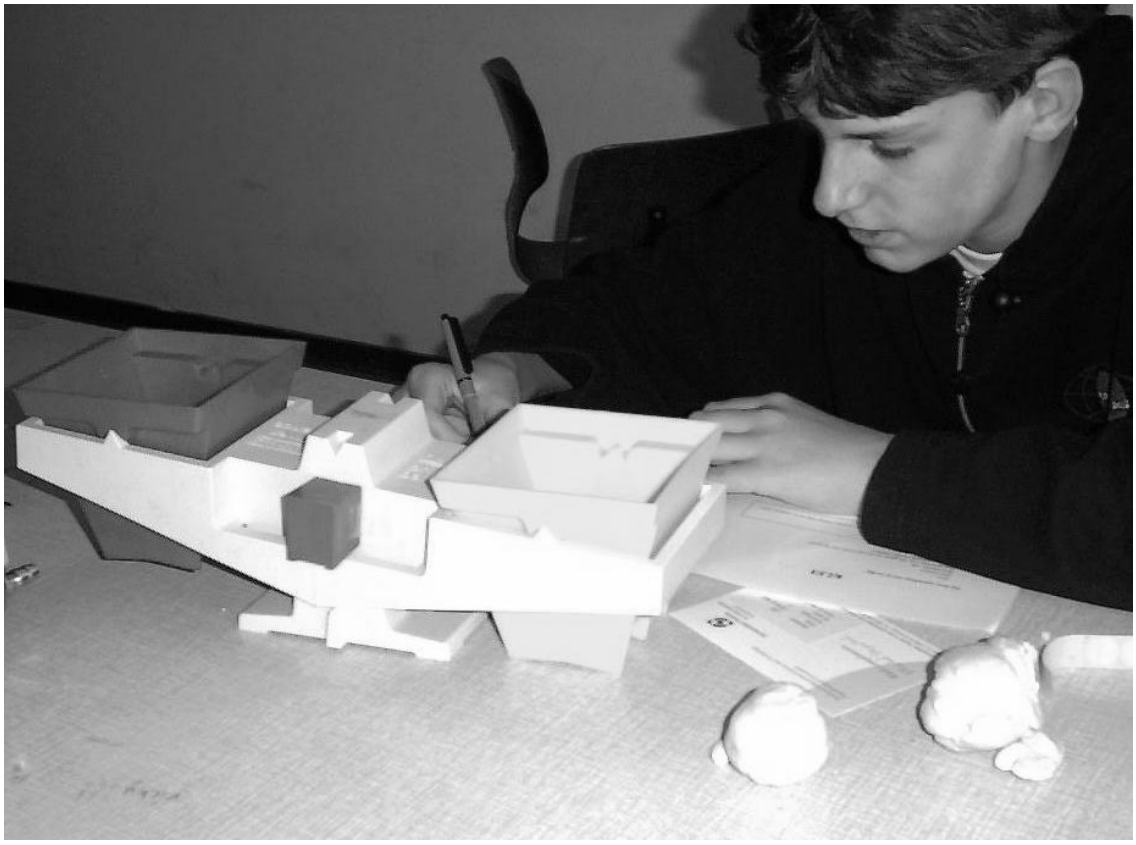


Figure 3.2: Student working on the task *Plasticine* from the TIMSS Performance Assessment

Chapter 

Research design

~ Apprendre il coute, de savoir il vaut.~

To learn is costly, to know is worth it.
(FRENCH PROVERB)

This chapter describes how the research was designed and carried out. After an introduction in section 4.1, section 4.2 presents the participants in the study. The participants were grade 8 students, their teachers and a number of curriculum experts. Section 4.3 describes the instruments used, for each of the participating groups separately. Procedures for data collection and data analysis is dealt with in the last two sections.

4.1 INTRODUCTION

In 1995, the TIMSS Written Test and the TIMSS Performance Assessment resulted in the observation of two discrepancies. First, Dutch students' achievement differed between the two tests in the international comparison. On the Written Test, the achievements were relatively higher than on the Performance Assessment. As explained in chapter 1, this discrepancy was termed the *inter-test achievement discrepancy*. Second, students' achievement on the two tests contrasted with curriculum experts' judgement on the appropriateness of the tests. The experts judged that the Written Test did not match well with the RME-based mathematics curriculum, which stressed the functionality of mathematics. Nevertheless, students' achievement was relatively high in the

international league table. On the other hand, the experts judged that the Performance Assessment matched well with the intended curriculum. Nevertheless, students' achievement was near the international average. This discrepancy between experts' judgement and students' achievement was termed the *intra-curricular discrepancy*.
In order to study the observed discrepancies, and find possible explanations, the METRIC study was initiated. One of the suggested explanations for the discrepancies was, that the intended mathematics curriculum was still new and still needed time to establish itself. The implemented and attained curriculum still showed traces of the abandoned, more theoretical curriculum (Kuiper et al., 1997, 1999).

In chapter 2 of this dissertation, a brief history of the RME-based curriculum was given. One of the central ideas was to offer all students a useful mathematics, integrating mathematics and context. As a result, almost all mathematics exercises in the RME-based curriculum are embedded into real-life settings. They differ from the items in the TIMSS Written Test, which are mostly multiple choice, and 'bare' (without context). They also differ from the hands-on items in the TIMSS Performance Assessment, which require an empirical investigation, while the RME-based assessment practice is paper-and-pencil based.
In the school year 1993/1994, the RME-based intended mathematics curriculum was introduced in annual phases, starting with grade 7. A year later, in 1994/1995, the new curriculum was introduced in grade 8. At the end of that school year, in the spring of 1995, this grade 8 population was tested through the two TIMSS tests. Thus, in 1995, the two TIMSS tests were administered among the first cohort that was taught under the new RME-based curriculum. It was well possible, that teachers and students still needed time to adapt to the innovation. Therefore, the METRIC study aimed at a repeat of both TIMSS tests at a later stage, giving the reform time to settle.

In 1995, many data had been gathered already at the level of intended, implemented and attained curriculum. Considering the motto *'if you want to measure change, do not change the measure'*, the METRIC study kept most instruments intact for the replication of the data collection. In the previous chapter, chapter 3, the instruments used in the METRIC study were reviewed. First, there were the two tests for measuring students' achievement: the TIMSS Written Test and the TIMSS Performance Assessment. Second, the Opportunity To Learn (OTL)

instrument of De Haan (1992) was available for measuring the appropriateness of the tests in light of the implemented curriculum. Finally, an instrument based on the TIMSS Test-Curriculum Matching Analysis (TCMA) was available for measuring the appropriateness of the tests in light of the intended curriculum.

The METRIC study was based on data of two tests at two moments, resulting in four sub-studies. These were the TIMSS Written Test in 1995 and 1999 and the TIMSS Performance Assessment in 1995 and 2000. They were indicated by the abbreviations WT-1995, WT-1999, PA-1995, and PA-2000.

In this chapter, the research design of all four sub-studies WT-1995, WT-1999, PA-1995 and PA-2000 will be described, discerning the measurements at the level of the intended, implemented and attained curriculum. This results in twelve measurements. The measurements were conducted within different contexts. Some measurements were carried out in an international context, under the guidance on the TIMSS International Study Center in Boston. Results of this international data collection were reported in, among others, Beaton et al. (1996), Harmon et al. (1997), and Mullis et al. (2000). For specific Dutch aspects in the international studies, Kuiper et al. (1997, 1999, 2000) reported on WT-1995 and PA-1995. Bos and Vos (2000) reported on the Dutch data concerning WT-1999.
One measurement, the implemented curriculum for WT-1995, took place within the project of the National Option Test (NOT), in which the appropriateness of the test items in WT-1995 were measured through an additional test (Kuiper et al., 1997, 2000). This test consisted of 16 items from WT-1995, supplemented with items that were based on the Dutch RME-based curriculum. Within this project, data on the implemented curriculum were collected for these 16 items from the Written Test. For the remaining items in WT-1995, no data are available for the METRIC study.
The Performance Assessment of 1995, PA-1995, was carried out under the guidance of the TIMSS International Study Center in Boston, but only at the level of the attained curriculum. In the Netherlands, the data set was supplemented with two additional data collections, at the level of the intended and implemented curriculum respectively. De Haan et al. (1997) reported on these data collections.
All remaining measurements were carried out within the METRIC study, to complete the matrix. A scheme of the research context for the data collection is given in Table 4.1.

Table 4.1: Research context of data used in the METRIC study

Sub-study	*Intended curriculum*	*Implemented curriculum*	**A***ttained curriculum*
WT-1995	International	NOT	International
WT-1999	International	METRIC	International
PA-1995	National	National	International
PA-2000	METRIC	METRIC	METRIC

Note: International = under guidance of the International TIMSS Study Center, Boston;
NOT = National Option Test, limited to 16 items only;
National = national supplement to PA-1995.

As many aspects of the data collections of the three sub-studies WT-1995, WT-1999 and PA-1995 have been extensively described in the above named reports, the information on these sub-studies will be kept concise in this current chapter. On the other hand, because of the novelty of the PA-2000, the design and the procedures of this data collection will be described comprehensively.

Of all measurements, those at the level of the attained curriculum were more complex than at the level of the implemented and intended curriculum. The instruments and the test design for measuring students' achievement required development, trialing, rotation across students, coding of students answers, and so forth, before the data were ready for statistical presentation. The research phases will be described in this chapter, with special attention to reliability of instruments. Their validity has been reviewed in the previous chapter (see section 3.1.4).
The measurements at the level of the intended and implemented curriculum were more straightforward, consisting, roughly speaking, of asking the curriculum experts or teachers to judge the appropriateness of the test by replying 'yes' or 'no' to each item. This yielded binary data that were readily collected and processed. In this current chapter, the instruments will be described.

Five measurements were carried out within an international context (see Table 4.1). Three of these were at the level of the attained curriculum (WT-1995, WT-1999, PA-1995) and two were at the level of the intended curriculum (WT-1995, WT-1999). To ensure comparability of data between countries, the International TIMSS Study Center in Boston issued strict protocols for data collection (TIMSS Performance Assessment Administration Manual, 1994; TIMSS-R Survey Operations Manual, 1998).

The international guidelines were used or adapted for the other data collections as follows. For the repeat of the TIMSS Performance Assessment (PA-2000), the guidelines from PA-1995 were re-used (TIMSS Performance Assessment Administration Manual, 1994) to ensure comparability between PA-1995 and PA-2000. For the measurements at the level of the intended curriculum, the protocols from the TIMSS Study Center on WT-1995 and WT-1999 offered guidelines on the selection of experts, on the method of gathering data, and on data processing (TIMSS Test Curriculum Matching Analysis, 1995; TIMSS-R Test Curriculum Matching Analysis, 1998). The protocols were copied for the measurement at the level of the intended curriculum of PA-1995 and PA-2000. Additionally, the protocols were adapted for the measurement of the implemented curriculum. Descriptions will be given in the forthcoming sections.

The dates of data collection are given in Table 4.2. All sub-studies were carried out at the end of the school year, in springtime (i.e. in the months April, May, June). In 1995, the first two sub-studies, WT-1995 and PA-1995, were carried out concurrently. Four years later, due to financial circumstances, the repeat studies could not be carried out in the same year. The repeat of the Written Test was carried out in 1999, while the repeat of the Performance Assessment was carried out in 2000. In both cases, the reform of the intended mathematics curriculum into an RME-based curriculum was several years before and teachers had ample time to adjust to the innovation. The METRIC-study assumed that the implementation of the new curriculum would not differ significantly between 1999 and 2000. Besides, the METRIC study assumed that the achievement of the grade 8 student population would not differ significantly between of 1999 and 2000. Thus, a repeat of the Performance Assessment in 1999 or in 2000 was considered equivalent. The one-year difference between WT-1999 and PA-2000 would not restrict the answering of the research questions.

Before each data collection, field trials were organised, in order to try out instruments and procedures. The dates of the trials are included in Table 4.2. The field trial for PA-2000 was omitted for the following two reasons: (1) no new instruments were introduced, and (2) experience with the testing procedures was already available (e.g. the TIMSS National Research Co-ordinator of 1995 was available upon request to clarify ambiguities). Therefore, a field trial for PA-2000 was discarded as little additional information on the instruments and procedures was expected from it.

Table 4.2: Dates of data collection in the METRIC study

Sub-study	**F***ield Trial*	*Main Study*
WT-1995	Spring 1994	Spring 1995
WT-1999	Spring 1998	Spring 1999
PA-1995	Spring 1994	Spring 1995
PA-2000	--	Spring 2000

Note: Dashes indicate the field trial was not carried out.

*4.2 P*ARTICIPANTS IN THE *METRIC* STUDY

4.2.1 Sample design for the attained curriculum

In the following sections the participating students, teachers and curriculum experts in the METRIC study will be presented. They are the representatives at the level of the attained, implemented and intended curriculum, respectively. First, the selection of students is explained.

All four tests in the METRIC study were administered among a random sample of Dutch students. In the cases of WT-1995 and WT-1999, the samples were drawn in two stages. First, a random sample of schools was drawn, stratified by school size. Second, at each selected school, one grade 8 class was chosen randomly from all grade 8 classes at that school. In the cases of PA-1995 and PA-2000, an additional third step was made. From each selected class, nine students were selected randomly. Consequently, for the Written Test complete classes from each school were tested, while for the Performance Assessment only part of the class population was tested. The considerations at each step in the sampling procedure will be described in the following paragraphs.

Sampling schools

According to the TIMSS protocol for WT-1995 and WT-1999, a random sample of 150 schools was selected from a recently updated governmental list. The 150 schools were invited to participate in the study. Some of the invited schools declined to participate. In those cases, a replacement procedure was applied. Another school was selected, whereby this replacement school had similar characteristics as the withdrawing school (e.g. with respect to school size and their system of grouping students according to ability tracks).
In 1995, from the 150 Dutch schools in the sample for WT-1995, 36 schools

participated (a response of 24%). After replacement of withdrawing schools, a total sample of 94 schools was realised (a response of 63%). This is presented in Table 4.3, which gives the numbers of schools in the school sample, before and after replacement.

The same procedure was applied to WT-1999. Again, a sample of 150 schools was drawn from a recently updated governmental list. However, from the 150 Dutch schools in this sample, two schools had to be annulled, because they had merged with other schools. Therefore, the initial sample size was reduced to 148 schools. Of these, 96 schools participated (a response of 65%). After replacement of withdrawing schools, a total sample of 126 schools was realised (a response of 85%).

Table 4.3: School participation rates in the METRIC study

Sub-study	*Initial sample size*	*Schools participating before replacement (% response rate)*	*Schools participating after replacement (% response rate)*
WT-1995	150	36 (24%)	94 (63%)
WT-1999	148	96 (65%)	126 (85%)
PA-1995	50	18 (36%)	49 (96%)
PA-2000	50	27 (54%)	--

Note: Dashes indicate the replacement procedure was not applied.

For PA-1995, the procedure of sampling schools made use of the fact that WT-1995 was carried out in the same year. Thus, the school sample for PA-1995 was derived from the sample of WT-1995, by drawing a sub-sample. In this way, 50 schools were selected randomly from the realised sample of 94 schools in WT-1995. The 50 schools selected had already participated in WT-1995 and were now additionally invited to participate in PA-1995. Initially, only 18 schools out of these 50 schools committed to participation (a response rate of 36%). After application of the replacement procedure, a sample of 49 schools remained for PA-1995 (a response of 96%).

A similar procedure was followed for PA-2000. Again, the sample of the Written Test was used to create a sub-sample. The list of schools in WT-1999 contained 126 schools. From this list, 50 schools were randomly selected and invited to participate in PA-2000. A sample of 27 out of 50 schools was realised (a response of 54%). Due to financial constraints, the replacement procedure was not carried out in 2000.

Selecting classes

After the selection of schools, an intact class per school was selected. This procedure was identical for WT-1995 and WT-1999: at each selected school, one grade 8 class was selected randomly out of all grade 8 classes of that school. The test was then administered to all students from that class.

The selection of classes for the two test administrations of the Performance Assessment was based on a different method. In 1995, the Performance Assessment was carried out in a subset of schools where the Written Test already had been administered. Each school in PA-1995 had already participated in WT-1995. Thus, the selection of a grade 8 class had already taken place for the Written Test. That same class was also designated for the Performance Assessment.

In 2000, for the repeat of the Performance Assessment this procedure could not be copied, because of the one-year difference with the Written Test. The class, which had participated in WT-1999, had become a grade 9 class in 2000. Therefore, at each school a new grade 8 class had to be selected.

To assure that schools would not push their own preference, the selection pointed at the class that carried the same name as the class that had participated in WT-1999. For example, if the grade 8 class '*2F*' had participated in WT-1999, the new '*2F*' class of 2000 was asked to participate in PA-2000. This procedure worked well at most schools. In a few cases, the replacements were not possible, for example where student numbers had reduced and the class name had become non-existent. In these cases, a class was selected for PA-2000, which had the same ability track characteristics (*vbo*[1], *mavo, havo*, *vwo*) as their counterparts in WT-1999.

Selecting students

For the Written Test, no selection of students was made. Once a class was appointed as the testing class of that school, all students of the class were to participate. This applied to both WT-1995 and WT-2000. As a result, 1984 students from 96 schools participated in WT-1995, and 2957 students from 126 schools participated in WT-1999. See Table 4.4.

[1] For an explanation on the ability tracks of Dutch secondary schools, see the Glossary (page xiii).

Table 4.4: Numbers of students in the METRIC study

Sub-study	**N***umber of schools*	**N***umber of students tested*
WT-1995	94	1984
WT-1999	126	2957
PA-1995	49	437
PA-2000	26	234

When comparing research data on the Written Test of either 1995 or 1999 in the Netherlands, it turns out that the numbers of participating students differ between studies. In other studies, the number of Dutch students participating in WT-1995 or WT-1999 differs from the figures in Table 4.4. For example, Kuiper et al. (1997, 1999) mention 1921 students in WT-1995, instead of the 1984 students tested in the METRIC study. Another example is on WT-1999, where Bos and Vos (2000) mention 2878 students, instead of the 2957 students tested in the METRIC study. The difference in student numbers is caused by differences in research questions across studies. The METRIC study focused on students' achievement and used the maximum number of students participating in the achievement tests. With the Written Test having a break (after 45 minutes of testing), sometimes students had to leave midway the testing session. For WT-1995, there were 1984 students before the break and 1983 students after the break. For WT-1999, the numbers were 2956 and 2957 respectively. These figures are also the numbers as reported in the TIMSS international database (Gonzalez & Miles, 2001).

Other researchers on TIMSS make use of students' achievement together with students' background variables, for example Bos (2002) and Bos and Vos (2000). The additional background data were collected through a questionnaire, administered in a separate session, apart from the achievement test. In order to link background variables to students' performances, the researchers need the overlap of students in all three sessions (two testing sessions plus one questionnaire session). The overlap reduces the number of students and accounts for differences between studies.

As described above, both WT-1995 and WT-1999 were administered among all students from the intact classes selected. This was not the case for the Performance Assessment. Because of its practical nature, abundant physical space was needed for administration. Therefore, the Performance Assessment was administered among only nine students per class. This number was randomly selected from each participating class. Due to circumstances, the selection of students was carried out differently for PA-1995 and PA-2000. First, the selection of nine students from each class for PA-1995 will be described.
The randomisation of students for PA-1995 was coupled to WT-1995. As described before, all students had already participated in WT-1995, before the administration of the PA-1995. The Written Test consisted of eight different test booklets, which were assigned randomly to the students of the class. The students, who had been issued booklet number 1, were first selected for PA-1995. If this selection yielded less than nine students, the students with booklet number 2 were selected. This procedure was continued until nine students were listed. Thus, from 49 schools, 437 students participated in PA-1995 (Kuiper et al., 1997).
The procedure for student selection for PA-1995 was based on a direct link with WT-1995 in same year. This procedure could not be copied for PA-2000, as the class of 2000 had not seen any of the booklets from WT-1999, which was administered in the year before. To randomise the selection of nine students from the participating class, a class list was used, in which the students were ordered by their surnames. The first nine students from this list were selected. If one of the nine students was absent, the tenth student from the list was invited. All other students from the class were excluded from the test.
At one school, the alphabetical class list turned out to be framed in such a way, that the nine best students were participating. This school was removed from the analysis. Consequently, 27 schools participated in the test, but 26 schools were used for analysis. In total, data from 234 (=26x9) Dutch students were used for PA-2000.

4.2.2 Representativeness of the student samples

To verify whether the drawn samples were representative for the whole student population, the samples were compared through background variables, in terms of statistics on gender and ability tracks. The gender ratios of the four samples used in the METRIC study are given in Table 4.5. The fifty/fifty-ratio is approximated in all studies except for PA-95. It is unknown why this imbalance occurred in this test administration (Kuiper et al., 1997).

Table 4.5: Distribution of students' gender in the METRIC study

Sub-study	**F***emale (%)/Male (%)*
WT-1995	49/51
WT-1999	51/49
PA-1995	55/45
PA-2000	50/50

Another background variable checked was students' ability track. In the Netherlands, students are grouped according to ability classes. The grouping prepares students for different final exam programs through different curricula. The lowest ability tracks are *(i)vbo* and *mavo* with final exams in grade 10. The higher ability tracks are *havo* and *vwo* with final exams in grades 11, and 12 respectively. The students tested in the METRIC study were in grade 8. At that stage, all Dutch schools have grouped the students to ability tracks or combinations of tracks (e.g. a combined *havo/vwo* class). For the METRIC study, class teachers were asked to give an indication of the tracks of students. With these data, it could be verified whether the ability tracks were well distributed.

Table 4.6: Distribution of students' ability tracks in the METRIC study

Sub-study	*vbo/mavo (%)/havo/vwo (%)*
WT-1995	59/41
WT-1999	50/50
PA-1995	59/41
PA-2000	54/46
National data at Grade 9*	
1996	61/39
1998	58/42

Source: * Inspectie voor het Onderwijs (1998).

Table 4.6 gives the ratio of ability tracks for the four samples used in the METRIC study. For comparison, national survey data are added. According to these data, approximately 60% of all students are in the lower two ability tracks. The *vbo/mavo* track is slightly underrepresented in WT-1999. In the other three tests, the ratio is sound.

As PA-2000 was new among the sub-studies, its students' sample underwent an extra check for representativeness, besides the check through gender and ability track statistics. For PA-2000, the sample of schools was mapped geographically.

Figure 4.1: Geographical position of 27 Dutch schools participating in PA-2000

Figure 4.1 shows the map of the Netherlands with the position of each participating school indicated by a black square. All provinces were represented in this data collection, with the exception of Flevopolder. Moreover, schools in the large cities of Amsterdam and The Hague are underrepresented. Nevertheless, the geographical diffusion is sufficient.

With all samples drawn randomly, and after the above inspection of the representativeness of the sample, the METRIC study assumed that the data could be generalised within probability margins. However, when very strict margins are considered, the above samples might not fully satisfy. For example, according to the TIMSS procedures of 1995, the Dutch measurements for WT-1995 and PA-1995 did not satisfy the international guidelines for sample participation rates. As a consequence, the results were 'flagged' and reported in a separate section of the international reports (Beaton et al., 1996; Harmon et al., 1997). Compared to international standards, the Dutch response rates were considered too low. With relatively low response rates, it was assumed that there could be a bias in the sample of schools. However, in the Netherlands, there are always low response rates in large-scale, educational studies. For example, in the 1999 PISA study, the response rate after replacement was a mere 27% (OECD, 2001). Compared to that figure, the response rates in the METRIC study (between 54% and 96%) can be considered as relatively high. Additionally, the low Dutch response is by no means related to achievement. Kuiper et al. (1997)

made an analysis of the WT-1995 sample, by comparing the achievement of participating schools with the achievement of refusing schools in national assessments. They concluded that the distribution of achievement of the participating schools did not differ significantly from the achievement of the refusing schools. Therefore, they convincingly demonstrated that the sample was not biased for achievement results.

4.2.3 Sample design for the implemented and intended curriculum

Besides test administration in the METRIC study, additional data were collected at the level of the implemented and intended curriculum. At the level of the implemented curriculum, the mathematics teachers of all participating classes in each of the four sub-studies WT-1995, WT-1999, PA-1995 and PA-2000, were invited to fill in a questionnaire. This questionnaire will be presented in the forthcoming section 4.3.4.

The numbers of mathematics teachers returning the questionnaire and the response rates are given in Table 4.7. For both WT-1995 and WT-1999, the response rates are very satisfying. Teachers' response rate on the Performance Assessments was distinctly lower, in particular in 1995. A reason for this could be, that the teacher questionnaire for PA-1995 was sent by mail, a few months after test administration. For PA-2000, it was handed personally to the teachers, immediately after test administration. The delay and the method of teacher approach could have accounted for the difference in response rates.

Table 4.7: Numbers of mathematics teachers in the METRIC study

Sub-study	***A**bsolute numbers*	*Response rate (%)*
WT-1995	91	97
WT-1999	112	89
PA-1995	20	41
PA-2000	20	74

In Table 4.8 the distribution of schooling tracks and gender are given for the participating mathematics teachers. The distribution of teachers across ability tracks is fairly representative of the general picture (60%-40%). On the other hand, the gender distribution of the teachers is misrepresentative for PA-1995 and PA-2000. It should have approximated the rates 30%-70% and it is unclear why, in 1995 few women, and in 2000 many women returned the questionnaire.

Table 4.8: Distribution of teachers by students' ability tracks and by gender in the METRIC study

Sub-study	*vbo/mavo (%)/havo/vwo (%)*	**F***emale (%)/Male (%)*
WT-1995 *(n=91)*	63/37	22/78
WT-1999 *(n=113)*	54/46	30/70
PA-1995 *(n=20)*	52/48	15/85
PA-2000 *(n=20)*	55/45	45/55

From the analyses, it is clear that the group of responding teachers in the METRIC study is not a random sample of Dutch grade 8 mathematics teachers. However, this is the only representation of teachers available.

For the intended curriculum, in the first three sub-studies (WT-1995, PA-1995 and WT-1999) three mathematics curriculum experts were consulted. These were from a research institute for mathematics curriculum development (the Freudenthal Institute), the National Institute for Educational Measurement (Cito), and an in-service training institute (APS).
However, the number of three was considered low. Already in 1995, the Dutch TIMSS research co-ordinator pleaded for a larger number of experts, in order to obtain a more detailed judgement on the appropriateness of the test items (Kuiper et al., 1997). Therefore, for PA-2000, six curriculum experts were approached. Three were from the same institutes as mentioned before, and three experts were from the national institute for curriculum development (Stichting Leerplan Ontwikkeling), a pre-service training institute (Faculteit Educatieve Opleiding, Hogeschool van Utrecht), and from the Dutch Association of Mathematics Teachers (Nederlandse Vereniging van Wiskundeleraren). Five of them returned the questionnaire.

4.3 I**NSTRUMENTS FOR MEASURING THE ATTAINED, IMPLEMENTED AND INTENDED CURRIULUM**

4.3.1 Introduction

This section describes the instruments for collecting data in the METRIC study. The sequence in the descriptions will be from micro, via meso to macro level. At micro level, achievement tests were administered among the sampled students.

The tests were also used at meso and macro level. The mathematics teachers were asked to assess the match between the tests and the implemented curriculum. The experts were asked to assess the match between the tests and the intended mathematics curriculum.
For the attained curriculum, a description of the instruments of the TIMSS Written Test will be given first. Thereafter, the instruments of the TIMSS Performance Assessment will be described. At the end, the instruments for measuring the appropriateness of the two tests in light of the implemented and the intended curriculum will be described.

4.3.2 The TIMSS **W***ritten Tests of* **1995** *and* **1999**

This section describes the TIMSS Written Tests of 1995 and 1999, which were very much alike. Both tests contained many items, which were administered in a rotational system. First, the test items will be described, and how they were organised into clusters and into test booklets. This will be followed by a description of data collection procedures.

The test items in the Written Test
Both for WT-1995 and WT-1999, the mathematics items were developed in a co-operative venture, involving subject-matter specialists from all participating countries (Beaton et al., 1996; Garden & Orpwood, 1996; Mullis et al., 2000). Countries submitted items and additional items were developed by subject matter specialists from the TIMSS International Study Center to cover a wide range of mathematical topics and performance expectations. All items were pilot-tested.
A mathematics curriculum framework was developed by Robitaille et al. (1993). This framework was used both for the 1995 version as the 1999 version of the Written Test. The framework included five content areas and five cognitive categories, which were dubbed *performance expectations*. The content areas were: fractions and number sense, measurement, data representation, analysis and probability, geometry, and algebra (depending on the report, the content area of proportions is added). The performance expectations were: knowing, using routine procedures, using complex procedures, investigating and solving problems, and communicating and reasoning. The proportional distribution of mathematics items in the TIMSS Written Tests for grade 8 across the categories is given in Tables 4.9a and 4.9b.

Table 4.9a: Content areas of items in WT-1995 and WT-1999

Content area	**%** *of items in* **WT-1995** *(n=150)*	**%** *of items in* **WT-1999** *(n=155)*
Fractions and number sense	34	38
Measurement	12	15
Data repr., analysis and probability	14	13
Geometry	15	13
Algebra	18	22

Table 4.9b: Performance expectations of items in WT-1995 and WT-1999

Performance expectations	**%** *of items in* **WT-1995** *(n=150)*	**%** *of items in* **WT-1999** *(n=155)*
Knowing	22	19
Using routine procedures	25	23
Using complex procedures	21	24
Investigating and solving problems	32	31
Communicating and reasoning	0	2

The TIMSS framework has its limitations (De Lange, 1997; Mullis, et al., 2001). For example, the curricular domains are not well-defined. For example, an item can cover different topic areas (e.g. both numbers and algebra), making it arbitrary in which topic area to categorise it. Similarly, a performance expectation such as 'using routine/complex procedures' is only useful if there is consensus on what to consider 'routine' or 'complex'. From the viewpoint of the Dutch RME curriculum for grade 8, the two Written Tests contain many 'bare' items on fractions and algebra, and there is little emphasis on visual geometry. Additionally, the items hardly focus on the applicability of mathematics in contexts. However, the TIMSS framework ensured that there was a large variety in items. The deviation from the Dutch intended curriculum enabled the METRIC study to assess Dutch students' achievement on 'remote' items. In this way, for example, remnants of the abandoned curriculum were investigated.

In both WT-1995 and WT-1999, the items are grouped into 26 item clusters with alphabetical labels (A, B, C, etc.). Each item is indicated by an alphabetical character and a number, for example N06 is the sixth item in cluster N. After the 1995 international test administration, almost two-thirds of the items were

released for public use. These were the items in the clusters I – Z. In WT-1999, they were replaced by new items. The 48 items in the clusters A-H in WT-1995 were kept secret for re-use in WT-1999. After the 1999 international test administration, another set of items was released. It was the set of 'even' clusters (B, D, F, etc).

Because of the replacement of items, the two Written Tests were not exactly identical. Yet, for the METRIC study, the achievements of students on the two tests needed close comparison, in order to answer the research question on trends in the achievement results. Therefore, comparable items in WT-1995 and WT-1999 were paired. This selection procedure consisted of two steps. First, those items were selected, which were exactly identical in WT-1995 and WT-1999 (48 items from clusters A-H). A second step consisted of comparing the released items in WT-1995 with the replacing items in WT-1999.

The released items were replaced by new items of similar content, format and difficulty (Mullis et al., 2000). A considerable number of the replacement items were 'clones', in which a detail in the original item was altered. For example, item N19 in WT-1995 showed a grid of 24 unit squares and asked the students to *"shade in 5/8 of the unit squares in the grid"*. Its substitute in WT-1999 showed the same grid and asked students to *"shade in 3/8 of the unit squares in the grid"*.

Similarly, the items N13 of WT-1995 and WT-1999 can be compared.

Item N13 in WT-1995 states: *If* $x = 2$*, what is the value of* $\frac{7x+4}{5x-4}$ *?* _____

Item N13 in WT-1999 states: *If* $x = 3$*, what is the value of* $\frac{5x+3}{4x-3}$ *?* _____

Because of the close similarities, the item from WT-1999 was considered a clone of the item from WT-1995. It was assumed that most students, being able to solve the first version, would also be able to solve the clone. Most teachers, indicating that the first item was suitable for inclusion into an imaginary test for their students, would give the same reply to its clone. Most experts, indicating that the first item matches with the intended curriculum, would give the same reply to its clone.

Therefore, the students' scores on items in WT-1995 were considered comparable to scores on cloned items in WT-1999. The same applied to the teachers' and experts' judgements. Thus, the cloning of items proved very useful

for the METRIC study, in order to compare results of WT-1995 and WT-1999 at the level of the attained, implemented and intended curriculum.

In total 96 items in WT-1995 could be identified as having a clone in WT-1999. Together with the 48 identical items (which were kept secret after 1995), this resulted in a set of 144 items on which the measurements on WT-1995 and WT-1999 were well comparable. This set of *comparable* items made up more than 90% of both Written Tests. The remaining 10% of items did not bear resemblance to a counterpart in the other test.

The eight test booklets for the Written Test

The total number of items in WT-1995 was 150, and 155 in WT-1999. This number of items was considered too large to administer to all participating students. It would require more than three hours of testing. Instead, eight different test booklets were created, each requiring 90 minutes to complete. An additional advantage of eight different test booklets was, that the possibility of students' copying could be minimalised.

The items were distributed across the test booklets in a way, that all students answered a representative sample of the items. This was realised through the use of the 26 item clusters, which were rotated over eight test booklets. The item clusters A-R contained a mixture of mathematics and science items, the clusters S-V contained mathematics items only, and the clusters W-Z contained science items only.

The distribution of item clusters over the booklets is given in Table 4.10. The first seven booklets contained four item clusters before the break (after 45 minutes) and three after. Booklet number 8 contained fewer clusters, but these clusters P, Q and R contained a larger number of items.

Table 4.10: Item clusters in the eight test booklets of WT-1995 and WT-1999

Booklet number	*Item clusters (before break – after break)*
1	B, A, C, S – E, I, T
2	C, A, D, W – F, J, X
3	D, A, E, T – G, K, U
4	E, A, F, X – H, L, Y
5	F, A, G, U – B, M, V
6	G, A, H, Y – C, N, Z
7	H, A, B, V – D, O, W
8	B, A, Q – R, P

Cluster A appeared in all booklets and thus, all items in this cluster were given to all students, who were present before the break. Other clusters appeared in one booklet (I, J, K, L, M, N, O, P, Q, R, S and Z), in two booklets (T, U, V, W, X, Y), in three booklets (C, D, E, F, G, and H) or in four booklets (B).

Data collection procedures

After communication with the national TIMSS Research Center, the test booklets were sent to the schools, together with instructions on how to administer the test. The test session took 90 minutes with a short break after 45 minutes. The booklet numbers were assigned randomly to the students. After completion of the tests, the schools sent the materials back to the National Research Center.

In WT-1995 and WT-1999, the numbers of students taking a certain test item differed from item to item, not only because of the rotation of item clusters over the eight test booklets. An additional reason for different student numbers per item was the break during the Written Test after 45 minutes. Some students were absent before or after the break. The numbers of Dutch students during the first and second half of the testing sessions depended on the test booklet given. The numbers are given in Table 4.11.

Table 4.11: Numbers of Dutch students per booklet of WT-1995 and WT-1999

Booklet number	**WT-1995** *(before break – after break)*	**WT-1999** *(before break – after break)*
1	240 - 239	378 - 378
2	233 - 230	371 - 371
3	251 - 250	369 - 369
4	253 - 252	376 - 376
5	252 - 251	367 - 365
6	262 - 259	362 - 362
7	243 - 242	369 - 370
8	250 - 250	365 - 365
Total	1984 - 1983	2957 - 2956

The Tables 4.10 and 4.11 give information on the numbers of students to whom a certain item was administered. For example, the items in cluster S, which is the last cluster before the break in booklet nr.1 (see Table 4.10), was administered in 1995 to 240 Dutch grade 8 students, and in 1999 to 378 students (see Table 4.11). Other items were administered to many more students. For example, in

1995, items in cluster A (present in all booklets before the break) were administered to 1984 students, and to 2957 students in 1999. As a result of the different student numbers per item cluster, the precision of students' results varied between items. The larger the number of students in a sample, the more precise is the measurement for the whole population. This issue will be dealt with in the section on reporting and analysing results (section 4.4.6).
But first, we will turn to a description of the Performance Assessment instruments for measuring the attained curriculum.

4.3.3 The TIMSS Performance **A***ssessment*

This section describes the TIMSS Performance Assessment. The test consisted of twelve hands-on tasks, which were administered in a rotational system. First, the tasks will be described in detail. This will be followed by a description of the paper-based instruments used for testing the tasks. The sections ends with the description of data collection procedures.

The tasks in the Performance Assessment

In the next two sections, an elaborate description of the TIMSS Performance Assessment will be given. This test was used in PA-1995 and exactly replicated in PA-2000. In this current section the tasks will be introduced. Afterwards, the materials associated with the tasks and the testing procedures will be reviewed. The comparability issues between PA-1995 and PA-2000 were complex. These will be described at the very end of the chapter in a separate section (section 4.4.5).

Initially the TIMSS Performance Assessment Committee developed a set of 22 tasks for the subjects mathematics and science. In 1994, the tasks were field-tested in 19 countries. Field test administrators, national research co-ordinators, and subject matter experts evaluated the student results of each task. Consequently, twelve tasks were selected to be included into the TIMSS Performance Assessment. The tasks are named as follows:

Dice	Packaging	Batteries
Calculator	Rubber Band	Pulse
Folding	Shadows	Solutions
Around the Bend	Plasticine	Magnets

There were five tasks with a strong mathematical focus:
The task *Dice* is related to probability: students are given a dice and a transformation rule for each throw (even: plus 2, odd: minus 1). They are asked to throw 30 times, record their findings and explain why one result (the '4') has a higher frequency.
The task *Calculator* is related to number sense: students are given a simple calculator and are asked to discover a pattern in the multiplications of 34x34, 334x334 and 3334x3334. As the calculator holds only eight positions in the display, this is not an obvious task. The second part of the task consists of factorising 455 into two integers between 10 and 50.
The task *Folding* is related to symmetry and spatial abilities: students have to make certain displayed figures by cutting, using a pair of scissors. Because only one cut is allowed for each figure, the paper has to be folded. At the end of the task students have to make a folding plan without actually implementing and testing it.
The task *Around the Bend* is related to scale drawing and finding rules: students are given a cardboard model of a corridor and have to cut rectangles (modelling furniture). By testing which rectangle fits through the corridor, they have to find a rule for the critical lengths.
The task *Packaging* is related to measuring and the design of nets: students are given four table tennis balls and have to design three different boxes to contain the four balls. One design has to be cut, folded and fixed together with the sides exactly fitting the balls.

Besides tasks with a mathematical focus, the test also contains tasks related to science. In the tasks, investigations on science meet with mathematical activities. The tasks are described below, with an emphasis on the mathematical particularities.
For example, the task *Rubber Band* covers the topic of extrapolation. In this task a number of washers are attached to a rubber band. Students have to measure the stretching of the band, related to the number of washers. With only ten washers given, students are asked to predict the length of the rubber band, if twelve washers were attached. This task requires students to analyse the increment of their data.
Another task, named *Shadows*, is related to geometrical transformations. Students are given a torch, a card and a white screen. They have to project a shadow, twice the width of the object, and find a general rule for the distances between torch, card and screen.

The task *Batteries* is related to combinatorics. Students are given four identical batteries and a torch that can only test two batteries simultaneously. They are asked to identify the strong and weak specimen, thus making 2-combinations out of 4. They have to describe the method used.
The task *Plasticine* asks for problem solving heuristics. Students are provided with a two-sided (uncalibrated) balance, two weights (20g and 50g) and a lump of plasticine. They are asked to make smart combinations in order to produce pieces of plasticine of 10g, 15g and 35g. They have to communicate the method used.
The three remaining tasks *Pulse*, *Solutions* and *Magnets* can be identified as mainly biological, chemical and physical tasks, respectively. Still, they are also related to mathematical activities. In the tasks, students have to measure using instruments (stopwatch, thermometer, ruler), make tables to record their findings and analyse trends in the data.

The tasks *Dice*, *Calculator*, *Folding*, *Around the Bend*, *Packaging*, *Plasticine* and *Rubber Band* are included in Appendix C. Further details of all other tasks can be looked up in Harmon et al. (1997) who described the international Performance Assessment of 1995. When this international study was carried out, the team of developers allocated the label 'mathematics' to the first five tasks (*Dice*, *Calculator*, *Folding*, *Around the Bend* and *Packaging*) and the label 'combined mathematics/science' to two tasks (*Shadows* and *Plasticine*). Despite some obvious mathematical aspects, they labelled the other five tasks as science (*Pulse, Magnets*, *Batteries*, *Rubber Band* and *Solutions)*. This separation is maintained in the METRIC study, although the tasks *Batteries* and *Rubber Band* also contain mathematical activities.
According to this subject definition, the science tasks would not be necessary for the METRIC study, which focuses on mathematics. However, the tasks were maintained for two reasons: (a) because of their mathematical aspects, which could clarify transfer of mathematical skills, and (b) because test circumstances had to be kept equal. As many researchers have clarified, the sequence of items in a test can have an impact on performances (Carlsen & Ostrosky, 1992; Hodson, 1984, 1987; Huck & Bowers, 1987; Vos, Kuiper & Bos, 2000; Wood, 1991). Therefore, the full set of test items has to remain intact, including their sequence, in order to keep tests comparable.

The materials in the TIMSS Performance Assessment

The performances of the students on the five mathematics tasks and the two combined science/mathematics tasks of the Performance Assessment were taken as indicator of the attained curriculum. The instruments of PA-1995 were identical to the instruments of PA-2000.

For each tasks structured response sheets were used. They are given in Appendix C in the Dutch version as used in the METRIC study. Initially the sheets were prepared in English and translated into Dutch for PA-1995. The translation effort included: 1) translation of the instruments in accordance with the international guidelines and by using two or more independent translators, 2) consultation with subject matter experts regarding cultural adaptations to ensure that the meaning and difficulty of items had not changed after translation, 3) verification of the quality of the translations by professional translators from an independent translation company, 4) corrections in accordance with the suggestions made, 5) verification that corrections were made, and 6) a series of statistical checks (Harmon et al., 1997).

All response sheets had the same format: an A3-sheet folded into a double A4-sheet. The front-page only carried the title of the task and the logo of the research centre. In this way the testing session could clearly be started by allowing the students to open their sheet. On the second page students would first see an inventory of materials needed (scissors, ruler, etc). In a large frame a summary of the task would be given. Thereafter, numbered items would ask for specific skills and thinking processes. Large white gaps were left for students' answers. All answers had to be filled in on the sheets. At the end of the sheet instructions were given on how to conclude the tasks (e.g. on how to leave the table behind for the next candidate).

The non-paper based instruments in the testing session involved equipment and materials that the students needed for carrying out the tasks. These were developed in accordance with the international procedures. The TIMSS Performance Assessment Administration Manual (1994) contained extensive inventory lists. For each task the equipment and materials were described, including their margins of tolerance. For example, for the thermometers it was indicated that these should measure from 10ºC to 110ºC and could be read to a precision of a degree. Another example is the description of the balance, to be used in the task *Plasticine:*

> *"This may be any kind of simple balance, but it should be accurate. It must not have a scale, that is, not calibrated. Balance it without masses (weights) when setting up the station, and make sure that it does not go out of balance with handling. If it is not possible to obtain a balance, one can be constructed from common materials (coat hanger, plastic cups and string)."* (TIMSS Performance Assessment Administration Manual, 1994, p. 30).

All practical guidelines were meticulously followed in the METRIC study, both for PA-1995 and PA-2000. Within the margins of the guidelines, there were possibilities to make slight adaptations between 1995 and 2000. For example for the task *Plasticine*, in PA-1995 a metal scale balance was used, while in PA-2000 a plastic balance was used (see photograph X). Both instruments were allowed within the above described boundaries of the international guidelines. Through strict boundaries, the guidelines offered a safeguard that the measurements of 1995 and 2000 would be comparable. However, as later will be described, the small adaptations to the equipment turned out to have measurable effects on the results. This will be elaborated in the section on data quality procedures.

Data collection procedures in PA-1995 and PA-2000

The Performance Assessment was administered at each school in a 90 minutes session with nine students and twelve tasks. A matrix design assigned students to a subset of the tasks. Testing time of 90 minutes would not allow all students to do all twelve tasks. Still, achievement data on all tasks were collected. To accomplish this, the TIMSS International Study Center had designed a 'circus' administration system in which the tasks were set up at individual stations arranged in a circuit. Each student visited three stations and took a combination of two short tasks (each 15 minutes) or one long investigation (30 minutes) at each station.

The twelve tasks were presented at nine different stations according to the allocation of Table 4.12. Each station required 30 minutes to complete. Signs labelled the stations: Station A, Station B, etc.

Table 4.12: Assignment of tasks to stations in the Performance Assessment

Station	*Task(s)*
A	Dice ; Pulse
B	Calculator ; Magnets
C	Shadows
D	Folding ; Batteries
E	Rubber Band
F	Packaging
G	Solutions
H	Around the Bend
I	Plasticine

At the beginning of the testing session students received a sequence number, which determined which set of stations he or she would visit. This number was assigned randomly. For this assignment, two rotation schemes were used (Rotation 1 and Rotation 2) as given in Table 4.13. A rotation scheme was allocated randomly to a school. As a result, Rotation 1 was used at 15 of the schools, and Rotation 2 was used at the remaining 12 schools. For example, a student with sequence number 3 at a school with Rotation 1 visited the stations C (*Shadows*), F (*Packaging*) and E (*Rubber Band*) consecutively.

Through this rotational task allocation, possible effects of task interaction were decreased. As described in Vos et al. (2000), the achievement on one task can have an effect on the ensuing task. For example, if all students do task X first and then task Y, it is possible that they gain certain measuring experience in the first task X, which gives them extra skills and raised their achievement on the next task Y. However, by rotating the tasks in different sequences, the predecessor task to task Y would not always be X.

Table 4.13: Assignment of students to a sequence of stations in the Performance Assessment

Students' number	*Rotation 1*	*Rotation 2*
1	A, B, C	A, B, E
2	B, E, D	B, D, G
3	C, F, E	C, A, D
4	D, G, H	D, E, F
5	E, A, G	E, I, H
6	F, H, B	F, H, A
7	G, I, F	G, F, I
8	H, C, I	H, G, C
9	I, D, A	I, C, B

Both in 1995 and 2000, at each selected school trained administrators, external to the schools, carried out the test. The administrators were provided with an administration box, which held all testing materials. It contained an abundance of supplies and spare parts. In this way testing conditions were independent of schools' resources. To ensure uniform testing procedures, several training sessions were conducted to train the administrators. The training included information on:

- the communication with the schools,
- how to set up the stations and prepare equipment,
- how to introduce the testing session to the students,
- which stations to replenish in between testing,
- how to deal with questions of students during the session, and
- how to fill in the administration forms and return all students' work to the research centre.

The Performance Assessment administrator ensured that the students visited the correct stations. After completion of their tasks, students submitted their work booklets to the administrator together with any created products (if applicable). They were taken to the Research Centre for processing (see section 4.4).

4.3.4 Instruments for measuring the implemented curriculum

For the implemented curriculum, the mathematics teachers of the classes tested were asked to judge each test item in light of the content taught in the classrooms. The instrument that was used for the items in the Written Tests differed slightly from the instrument used for the Performance Assessment. Both

were based on the research of De Haan (1992). The instruments are described in the following paragraphs.

Constructing an imaginary test

For the Written Test, both in 1995 and 1999, teachers were asked to imagine that they were to construct a mathematics test for their class. This test was supposed to cover all mathematics content from the past, not only from the current school year but also from prior school years. This assumed that teachers had a close knowledge of the content taught in grade 7 and 8 up to the time of testing, and a reasonable knowledge of the content taught at primary schools. The teachers were then given the test items in the Written Test and asked to indicate for each item whether they found it suitable for their imaginary test. In this way, each item was given an indication on whether students had been given an opportunity to learn the content of the item. Of course, teachers could also include items into their imaginary test if they estimated that the item was a challenge for which their students were 'ready'. In these cases, the METRIC study still considered the item as matching with the implemented curriculum.

The judgement on the appropriateness of the items by the teachers was made two-fold in the following way. Teachers were asked to give an indication on including the item into the imaginary test:

1. when considering the content of the item only, and
2. when considering its format.

The aspect of the format was incorporated into the questionnaire as objections had been raised against the large number of multiple choice items in the Written Test (see chapter 3, section 3.1.4). By separating content from format, teachers could air their objections against the multiple choice format and still give a positive judgement towards the content of the item.

Thus, the teacher questionnaire contained an introduction, followed by the test items as they were given to the students, with two additional icons. For example:

B10. Which of the following numbers is smallest?
A. 0.625 B. 0.25 C. 0.375 D. 0.5 E. 0.125

Content
❑ yes
❑ no

Format
❑ yes
❑ no

The first icon was an indicator, of whether the teacher regarded the item as suitable for inclusion into an imaginary test, when only looking at its content. The second icon was an indicator, of whether or not the teacher would leave the format of the item unaltered. If a teacher ticked 'yes' into the first, and 'no' into the second icon, this meant that the content of the item could be included into the imaginary test, but the format needed editing.
For the METRIC study, the first icon was an indicator of OTL. If ticked 'yes', it indicated that the item matched with the implemented curriculum. This could mean that the content of the item was indeed covered in class, or that the teacher estimated that his/her students were 'ready' for it.

For the Performance Assessment, there was no need to make the distinction between content and format of test items, as there were no multiple choice questions in this test. However, another aspect needed care. The instrument aimed at obtaining an indication of whether students had been given an opportunity to learn the content and skills tested. De Haan (1992) had proposed to ask teachers to judge whether they would include the test items into an imaginary test covering all content taught before the test administration. However, most Dutch mathematics teachers have very little experience with the use of manipulatives in assessment environments. This could make them hesitant to imagine setting a practical test. As a result, the judgement was made two-fold in the following way:

1. Has the content of this item been taught to the class before test administration?
2. Independent of the answer to question 1, assuming that you would set a Performance Assessment, would you include this item?

Thus, the teacher questionnaire associated with PA-1995 and PA-2000 contained an introduction, followed by the seven tasks (five mathematics tasks, two combined tasks). At each test item in the tasks two icons were added, for example at item 1a from the task *Plasticine*:

1a. Use the balance to make a lump of plasticine that weighs 20 g.
- *When you have made the 20 g lump, write 20 g on a coloured label and stick it on the lump. Put the lump in a plastic bag.*

Content covered?
❑ yes ❑ no

Include into a test?
❑ yes ❑ no

The answers of the teachers to both questions were considered relevant to the METRIC study. The first icon would give an indication of content (OTL-covered) and the second icon would give an indication of teachers' estimation whether students were ready for the item (OTL-testing). The latter could mean that the content had not been yet taught, but that the teacher expected transfer. Thus, for the judgement on the appropriateness of the Performance Assessment in light of the implemented curriculum, two different data collections were made.

Reducing the number of items for allocation to the teachers

The above described issues on the instruments for measuring the implemented curriculum pertained to the method of asking a judgement from the teachers. Another issue in this data collection concerned the number of test items that were included in the instruments. The Written Test contained more than 150 items, and it would be too demanding to ask teachers to assess all of them. Therefore, a subset of the item set had to be submitted.

In 1995, 16 items from WT-1995 were submitted to the mathematics teachers. The 16 items were selected as part of a National Option Test (NOT). NOT was developed to assess whether the TIMSS Written Test did justice to Dutch students who had learned mathematics through an internationally deviant curriculum (Kuiper et al., 1997, 2000). NOT was based on the new curriculum, containing items from WT-1995 and items developed separately. The items from WT-1995 were selected on the criterion that the matched well with the intended curriculum. As these items were the intersection of WT-1995 and NOT, they were named *anchor items*. Based on students' achievement on the 16 anchor items in NOT and WT-1995, Kuiper et al. concluded that the Dutch students' achievement was well measured through WT-1995.

All responding teachers judged the 16 anchor items. However, these items were not a random sample representing the full test. They matched well with the intended curriculum, and therefore teachers' judgement could be misrepresentative of their judgement on the appropriateness of the full test.

Therefore, in 1999 all 155 mathematics items in the Written Test were submitted to the mathematics teachers, in such a way that not all teachers had to assess every item. A rotational system was devised, whereby the item set was split into three groups (Group 1, Group 2 and Group 3) of 51 or 52 items. First, the multiple choice items were randomly distributed over the three groups, and then the free response items were randomly distributed over the three groups. In this

way, all groups contained a fair proportion of the multiple choice items. Afterwards, the randomisation was checked by looking at the distribution of content areas in each item group. They are stated in Table 4.14. In each item group the distribution of topics over the whole set of items in WT-1999 is reflected.

The three item groups were randomly allocated to the teachers. Thus, from the 112 responding teachers, 39 judged the items in Group 1, 35 judged the items in Group 2 and 38 judged the items in Group 3.

Table 4.14: Content areas per sub-group of items, in the teacher questionnaire for WT-1999

Content area	**%** *items in* **G***roup 1* *(n=52)*	**%** *items in* **G***roup 2* *(n=52)*	**%** *items in* **G***roup 3* *(n=51)*	**%** *items in* **WT-1999** *(n=155)**
Fractions and numbers	31	35	33	33
Algebra	19	17	18	18
Measurement	13	13	14	14
Data repr., analysis and prob.	13	13	14	14
Geometry	13	13	16	14
Proportions	10	8	6	8
Number of teachers for each group	39	35	38	Total: 112 teachers

Note: * Data differ from Table 4.9a due to different interpretations of item categories.

Thus, in WT-1995, there were few items (the 16 anchor items), which were judged by all 91 mathematics teachers. By contrast, in WT-1999, there were many items, which were each judged by a lower number of teachers (35-39 teachers).

For the Performance Assessment, the selection of items, on which a teacher's judgement was asked, was straightforward. All items across the five mathematics tasks and the two combined mathematics/science tasks were submitted to the teachers, both for PA-1995 and PA-2000. However, in some cases the items were considered to be identical from an OTL point of view. In those cases teachers were asked to judge items together. For example, in the task *Plasticine* the test items had the sequence: create a lump of 20g - explain how you worked - create a lump of 10g - explain how you worked - create a lump of 15g - explain how you worked – and so forth. In this case, the items asking for an explanation

were grouped, and teachers were asked to judge the items as if they were one and the same. Consequently, the number of items per task differs between the instruments for the attained curriculum and the implemented curriculum.

4.3.5 Instruments for measuring the intended curriculum

For the intended curriculum, mathematics curriculum experts were given all mathematics items, both from the Written Test and from the Performance Assessment. They were given the test items exactly as these were given to the students. On a separate form all items were listed. On the form, the experts had to indicate for each test item whether its content matched with the intended curriculum, by either ticking a 'yes' or a 'no'. The intended curriculum was defined as the educational aims and objectives, which were stated as core objectives (*kerndoelen*) in the official educational documents for grade 8. An item was considered to match with the intended curriculum if the content tested was covered by the intended curriculum of at least 50% of the students at the time of testing (a few months before the end of the school year). With the core objectives being different for the tracks (*vbo*, *mavo*, *havo*, *vwo*), the experts were asked to interpolate between *mavo* and *havo*. The experts were explicitly asked to exclude the format (multiple choice) from their considerations and to look purely at the content of the items (Beaton, 1998).
During judgement, the experts encountered dilemmas. The dilemmas pertained (1) the use of the list of core objectives and (2) the nature of the test items. A number of experts indicated that the core objectives were only an indication of the intended curriculum. The core objectives were deliberately kept vague to offer textbook authors, schools and teachers a certain amount of freedom to interpret the curriculum. The core objectives were never formulated to be a strict guideline at item level. The second dilemma encountered by the curriculum experts pertained to the multiple choice format and the lack of context in the items in the TIMSS Written Test. This raised several questions, such as:

1. How can the multiple choice format be excluded from the judgement, if that format gives a clue to the solution? For example, in item N16 the following is asked:

   Penny had a bag of marbles. She gave one third of them to Rebecca, and then one fourth of the remaining marbles to John. Penny had 24 marbles left in the bag. How many marbles were in the bag to start with?

   A. 36 B. 48 C. 60 D. 96

In this item, a student can first calculate the answer and then check the answer among the alternatives. But it is easier to take the backdoor and try the four alternatives one by one and then skip the wrong answers. Thus, the multiple choice format affects the content of the item, while the experts were asked to ignore it. Thus, the experts were hesitant to make a judgement on item N16, and consensus was not easily reached.

2. Does the lack of daily-life contexts make an item less suitable to the curriculum? For example, item N13 reads:

$$\textit{If } x = 2\textit{, what is the value of } \frac{7x+4}{5x-4}\textit{ ?} \_\_\_\_\_$$

In the Dutch RME-based curriculum, this item would not be stated like this. An item on the substitution of a number into a formula would be set into a context, such as the exemplary item on the prediction of length:

To calculate for any girl her future length as a grown-up, the school doctor uses the following formula:

$$\textit{Length daughter (in cm)} = \frac{\textit{length father (cm)} + \textit{length mother (cm)} - 12}{2} + 3$$

Danielle's father is 1,82 m tall, her mother is 1,68 m.
How tall will Danielle grow according to the formula?

Thus, the mathematical content of item N13 is part of the intended curriculum (substitution of a number into a formula with one variable), but lack of context does not make the item suitable. This made the experts refute the item, not for its mathematical content but for its 'bare-ness'.

The answer to the question whether an item matched with the intended curriculum was not easily found. Only few items in the Written Test resembled the typical RME-based items on which a straightforward answer could be given. Thus, on most items, the considerations had to be weighted with respect to phrasing, content, format, context, and so forth. It made the judgement on some items into a toss-up for the individual expert. However, after some deliberations (by phone), on all mathematics items in the Written Test a consensus was reached. This consensus was needed for the TIMSS Test-Curriculum Matching Analysis (TCMA), in which all country's scores were recalculated based on the

sub-set of items from the Written Test, which were covered by the intended curriculum of one particular country (see section 3.2.2). For TCMA, an absolute yes/no judgement by curriculum experts on items was needed, in order to make two disjoint sets of items, distinguishing between items that matched an intended curriculum and items that did not. Thus, on WT-1995 and WT-1999, the item-curriculum matching index was given on a nominal yes/no scale.
For the Performance Assessment, this clear distinction was not sought because (1) there was no TCMA associated with the Performance Assessment, which needed to create two disjoint sets of items, and (2) the METRIC study aimed at giving room to nuance between 'yes' and 'no'. By rating through a percentage, it was possible to indicate whether an item matched 'more or less' with the intended curriculum, depending on the number of experts indicating that the item matched with the curriculum. This yielded an item-curriculum matching index on a ratio scale.

In the study for the TIMSS Written Test both in 1995 and 1999 the experts' consultation was part of a TCMA, which was carried out in all participating countries (see Beaton et al., 1996; Mullis et al., 2000). For WT-1995, each expert assessed all 150 mathematics items. For WT-2000, each expert assessed all 155 items.
For the Netherlands, a similar instrument as used for the Written Test was developed for the tasks of the Performance Assessment of 1995. This instrument was re-used in 2000. Each expert assessed the separate items across the seven tasks *Dice*, *Calculator*, *Folding*, *Around the Bend*, *Packaging*, *Shadows*, and *Plasticine*. The International TIMSS Study Center had indicated these tasks as 'mathematics' or as combined 'mathematics/science'. In 2000, this set of tasks was supplemented with two more tasks, *Batteries* and *Rubber Band*, because these two tasks contained mathematical activities (combinatorics, extrapolation) as well. Therefore, it was possible that these tasks were covered by the intended mathematics curriculum, too. Thus, for PA-1995, each expert assessed the items across seven tasks and for PA-2000, each expert assessed the items across nine tasks.

Just like with the instrument for the implemented curriculum, in some cases the items were similar and, therefore, grouped together to shorten the questionnaire. For example, in the task *Plasticine* the test items had the sequence: create a lump of 20g - explain how you worked - create a lump of 10g - explain how you worked - create a lump of 15g - explain how you worked – and so forth. In this

case the question on the explanation was grouped into one item for the experts. Consequently, the number of items per task differs between the instruments for the attained curriculum, the implemented curriculum, and the intended curriculum. For all three cases, Table 4.15 gives the number of items per task.

Table 4.15: Number of items per task in the instruments for the intended, implemented and attained curriculum in PA-1995 and PA-2000

Task	*Experts' instrument (in 1995/2000)*	*Teachers' instrument**	*Students' instrument* **(=the test itself)***
Dice	5/5	5	6
Calculator	6/6	6	7
Folding	4/4	2	4
Around the Bend	6/6	6	8
Packaging	3/3	3	3
Shadows	6/6	6	6
Plasticine	3/3	3	8
Batteries	--/4	--	4
Rubber Band	--/6	--	7
Total	33/43	31	53

Note: Dashes indicate the task was not included in the instrument;
* Instruments identical in PA-1995 and PA-2000.

4.4. DATA PROCESSING DESIGN OF THE *METRIC* STUDY

4.4.1 Introduction

The METRIC study gathered data at item level from WT-1995, WT-1999, PA-1995 and PA-2000. Then, the analyses were carried out based on descriptive statistics (percentages, standard errors and correlation coefficients). The current section describes all procedures, which were required to transform the answers from students, teachers and experts into statistics. Students' answers to free response items had to be coded first. This will be explained in section 4.4.2. The judgement by teachers and experts did not require coding as they answered all their questions simply by 'yes' or 'no'. Thereafter, the data were entered into databases. This will be described in section 4.4.3. To verify for the quality of data, several procedures were applied. This stage required procedures for checking reliability of instruments (see section 4.4.4) and procedures to control for the comparability in time between WT-1995 and WT-1999, and between PA-

1995 and PA-2000 (see section 4.4.5). Thereafter, the results were ready to be presented by descriptive statistics. The formulas used are given in section 4.4.6.

4.4.2 Coding students' responses

TIMSS makes a distinction between three categories of items: (1) multiple choice items, (2) short answer items and (3) extended response items. The first category contains questions where students have to choose one from four or five alternative answers. The second category contains questions where students have to provide a short answer, for example a number that results from a calculation. The two categories are one-point items, which are either false or correct. The last category of extended response items contains more complex questions, where students have to write a description or an explanation. These items carry two or three points, so that the score can differentiate between fully correct, partially correct or false. The two categories of short answer items and extended response items together are named *free response items*. Approximately 20% of the items in the Written Test and 100% of the items in the Performance Assessment were in this format. This current section deals with the coding of these items.

Students wrote their answers in the test booklets. Their answers to the free response items were evaluated by coders who were specially trained to use the so-called 'TIMSS scoring rubrics'. This coding system allowed for the identification of common approaches and types of errors in students' responses. The coding system used a two-digit code (10, 11, .., 21, 22, .., 31, 32, etc.). The first digit designates the correctness level of the response (3, 2, 1, or 0 points). Those are the points that honour the quality of the students' responses with full credit, partial credit or no credit at all. The second digit, combined with the first, represents a diagnostic code used to identify specific types of approaches, strategies, or common errors and misconceptions.
For all four tests, coders were trained during a one-day workshop to apply the scoring procedures based on the TIMSS rubrics through a guide (TIMSS Coding Guides for Performance Assessment Population 1 and 2, 1994; TIMSS Scoring Guides for the Mathematics and Science Free-Response Items, 1999). The guide contained the rubrics and explanations of how to apply them, together with examples of student responses for the various rubric categories. Thus, for all four tests, students' answers were interpreted by coders and transformed into a two-digit code. This code was then entered into the database.

For three of the test administrations, WT-1995, WT-1999 and PA-1995, a systematic sub-sample of approximately 10% of the students' responses was coded independently by two different coders. In this way, the inter-coder agreement could display the reliability of coding. This agreement was calculated as the percentage of items on which the two coders agreed with their codes. The agreement on the correctness code (the first digit) is tabulated in Table 4.16. For comparison, the international average range is also given.

Table 4.16: Range of inter-coder agreement per item on the correctness scores in the METRIC study

Sub-study	**D***utch range of exact agreement (min/max)*	*Range of international average agreement (min/max)*
WT-1995	87/100	91/100
WT-1999	85/100	93/100
PA-1995	52/100	67/100
PA-2000	--	--

Note: Dashes indicate that double coding was not applied.

On each of the free response items in both WT-1995 and WT-1999, the two independent Dutch coders agreed well on the correctness of students' answers. On most items their codes were totally identical (100%). On a few items, their codes matched on the majority of students' answers. Both in WT-1995 and WT-1999, the average Dutch percent agreement across items was 99% (not included in the table), which was exactly equal to the international average.

The range of inter-coder agreement for PA-1995 was clearly lower than on WT-1995 and WT-1999, both for the Dutch data as internationally. On one item (from the task *Shadows)*, the coders only agreed on half the number of students' answers. On the other half, they differed in the allocation of points. However, a low inter-coder agreement did not have repercussions. The protocol set no criterion for a sufficient inter-coder agreement. For example, if two coders agreed on less than 70% of students' answers, it could be that one coder made systematic misinterpretations of the coding scheme. This could be repaired by providing re-training. And in the ultimate case of unreliable codes, some items should be removed from analysis.

The exercise of double coding was omitted for PA-2000. A different protocol was developed to control for the consistency of codes. It will be explained in section 4.4.5 in separate sections on reliability and comparability issues. This

additional protocol was developed, because of dissatisfaction with the above described, international protocol, in which two independent coders assessed 10% of students' work. First, unreliable results needed to be sieved out, and second, the data of 1995 and 2000 needed to be well comparable in time. To gain insight in the quality of data, a more precise inspection on reliability and comparability of codes between 1995 and 2000 was devised. Based on that inspection, it was possible to sieve out unwanted effects.

4.4.3 Organising and cleaning data

The data from the experts and the teachers were collected through questionnaires and could readily be entered into a database. This was also the case with the students' data, after the answers to the free response items had been coded.
All data collected within an international context (cf. Table 4.1) could be entered into software that had been prepared by the International TIMSS Study Center in Boston. They provided National Research Centres with manuals and protocols so that all information, both for WT-1995, WT-1999 and PA-1995, was standardised. After the data were entered into a database, a cleaning process followed. This involved several iterative steps and procedures designed to identify, document and correct deviations from the international instruments, file structures and coding schemes. In the end, only the database remained. The raw data were not available to the METRIC study (e.g. no worksheets with students' answers had been archived).

There were no ready-made international procedures for compiling the students' data of PA-2000, as this was a national Dutch study only. But the administration of PA-1995 proved useful. The same manuals, software and cleaning protocols that were used in 1995 could be re-used for PA-2000. As a result of the cleaning process, it was discovered that one of the 27 initially participating schools had manipulated the alphabetical student list. Therefore, the students' scores from this school were removed from the final data set.

The data from the curriculum experts on WT-1995 and WT-1999 were collected as part of the international Test Curriculum Matching Analysis (TCMA) (Beaton et al., 1996; Mullis et al., 2000). The data from the teachers on WT-1995 were collected within the project of the National Option Test (NOT). The data were available to the METRIC study as descriptives at item level.

All remaining data (intended curriculum for PA-1995 and PA-2000; implemented curriculum for WT-1999, PA-1995 and PA-2000) were collected within the METRIC study and, thus, these were available as raw data to trace back anomalies. The data were entered into custom-made software programs. The data were checked for errors through a protocol based on Grzymala-Busse (1991) for sieving out missing data, doubled data, out-of-range data and checking internal and external consistency by comparing school data, teacher data and class data.

4.4.4 Reliability

Reliability of instruments is generally expressed as a coefficient, resulting from a calculation, which checks for internal consistency of the measurement. The methods for calculation of the coefficient are indicated as Cronbach Alpha or KR-20. The calculations become more precise with a larger number of items and a larger number of respondents. In the METRIC study, results of Cronbach Alpha calculations were not available for the instruments for the intended curriculum of WT-1995 and WT-1999 and the implemented curriculum of WT-1995. All other reliability coefficients will be given in this section. The results are presented in Table 4.17. How they were obtained will be described below.

Reliability of instruments used for the Written Test

In the following three paragraphs the calculations of Cronbach Alpha on the Written Test (or lack thereof) will be described. Cronbach Alpha on the experts' data could not be calculated, as the METRIC study did not have the availability over the raw data, but only over the final consensus of the three experts on each item.

The calculation of Cronbach Alpha of the teachers' judgements on WT-1995 was not available, and could not be made due to lack of data from each individual teacher. However, the calculation of Cronbach Alpha of the teachers' data on WT-1999 could be made. It required an intermediate step as the full set of items was split into three groups, so that not all teachers were asked to judge the full set. Thus, there were three groups of teachers who judged their own set of items. Cronbach Alpha was calculated separately for each of the three groups. The average of the three results gave α=0.78.

The reliability of the tests WT-1995 and WT-1999 was reported in the international reports (Beaton et al., 1996; Mullis et al., 2000). It was checked through the KR-20 method. Because not all students had responded to all items, the reliability coefficient was calculated per test booklet. This calculation yielded, for both WT-1995 and WT-1999, a median reliability for the Netherlands of α=0.89.

Table 4.17: Reliability coefficients (Cronbach Alpha or KR-20) of instruments in the METRIC study

Sub-study	*Intended curriculum*	*Implemented curriculum*	**A***ttained curriculum*
WT-1995	--	--	0.89†
WT-1999	--	0.78*	0.89†
PA-1995	0.83	0.89 - 0.93	0.65‡
PA-2000	0.79	0.78 - 0.82	0.68‡

Note: Dashes indicate that data are unavailable;
* Average reliability across three item groups;
† Median reliability across eight test booklets;
‡ Average reliability across seven tasks (5 maths tasks, 2 combined tasks).

Reliability of instruments used for Performance Assessment

In the next three paragraphs the calculations of Cronbach Alpha on the Performance Assessment will be described. Cronbach Alpha of the experts' data for PA-1995 was α=0.83, with three experts and 33 items (from five mathematics tasks and two combined science/mathematics tasks). For PA-2000, the reliability coefficient based on the expert data was α=0.79, with five experts and 43 items (from five mathematics tasks, two combined science/mathematics tasks, and two additional science tasks).

Reliability of teachers' judgement on the Performance Assessment was calculated over the data for OTL-covered and OTL-testing separately. For OTL-covered, the coefficients were as follows: PA-1995 and PA-2000 yielded α=0.89 and α=0.78 respectively. For OTL-testing, the coefficients were as follows: PA-1995 and PA-2000 yielded α=0.93 and α=0.82 respectively.

Finally, for the two test administrations PA-1995 and PA-2000, Cronbach Alpha was calculated. However, not all students had answered all items. In the Performance Assessment students only carried out three, four or five out of twelve tasks. Thus, for each separate task, the reliability coefficients were calculated. After averaging across the seven tasks (five mathematics tasks and two combined science/mathematics tasks), this yielded on PA-1995 and PA-2000 a reliability of α=0.63 and α=0.68 respectively. Although the coefficients were still sufficient (α >0.6), they were lowest of all measurements. Therefore, a breakdown of the calculation was scrutinised. For all tasks, including the five science tasks, the reliability coefficients were reported, both for PA-1995 and PA-2000. See Table 4.18.

The coefficients vary considerably. This was not considered annoying for the following reason. A reliability coefficient is an indicator of internal consistency of items. It means *"individuals at the same level ought to respond the same way to similar items"* (Krathwohl, 1998, p. 436). However, the items within tasks of the Performance Assessment are multi-dimensional and not similar. For example, in the task G1 *Shadows* one item asks for an investigation plan and another for a three dimensional sketch. Therefore, the two items measure different competencies. Students with high reading/writing abilities will score differently from students with high spatial abilities. Consequently, the pattern of scores may seem inconsistent.

Table 4.18: Reliability coefficients of achievement results (Cronbach Alpha) per task in PA-1995 and PA-2000

	Cronbach Alpha	
Task (# items)	*1995*	*2000*
M1 Dice (6)	0.50	0.64
M2 Calculator (7)	0.71	0.68
M3 Folding (4)	0.63	0.76
M4 Around the Bend (8)	0.59	0.62
M4 Packaging (3)	0.61	0.65
S1 Pulse (4)	0.61	0.59
S2 Magnets (2)	0.65	0.36
S3 Batteries (4)	0.54	0.68
S4 Rubber Band (7)	0.58	0.39
S5 Solutions (7)	0.63	0.63
G1 Shadows (6)	0.64	0.61
G2 Plasticine (8)	0.85	0.78

Reliability across tasks of the Performance Assessment was also calculated. This was done by taking all combinations of two tasks and using the scores of students who performed on both tasks. Through the rotation system in which students were assigned to stations, all pairs of tasks were possible combinations. In Table 4.19, the results of the calculations are given in a matrix. For each pair of tasks, the Cronbach Alpha is recorded for the data of 2000. For brevity the tasks are indicated by their code names, which refer to Table 4.18.

In most cases, the paired tasks had $\alpha > 0.6$. This means that the items from the corresponding tasks measured the same features up to a reasonable level. The scores on each item harmonised with the scores on the paired items. However, in

a few cases the answer patterns of students were very different across tasks. For example, task M5 *Packaging* seemed to measure competencies totally different from task S1 *Pulse* (α =0.18), but it measured in a high consistence with task G2 *Balance* (α =0.80). Overall, the tasks S2 *Magnets* and S4 *Rubber Band* showed on average the lowest results, questioning their reliability. These were also the tasks with the lowest within-task reliability (as given in Table 4.18).

Table 4.19: Reliability coefficients of achievement results (Cronbach Alpha) across paired tasks in PA-2000

	M1	*M2*	*M3*	*M4*	*M5*	*S1*	*S2*	*S3*	*S4*	*S5*	**G***1*	**G***2*
M1												
M2	0.63											
M3	0.71	0.74										
M4	0.60	0.48	0.78									
M5	0.65	0.75	0.82	0.67								
S1	0.64	0.72	0.82	0.57	0.18							
S2	0.62	0.59	0.47	0.59	0.61	0.23						
S3	0.50	0.73	0.68	0.76	0.71	0.77	0.48					
S4	0.65	0.72	0.49	0.59	0.49	0.61	0.34	0.45				
S5	0.43	0.81	0.68	0.71	0.66	0.59	0.72	0.64	0.53			
G1	0.66	0.73	0.71	0.72	0.55	0.73	0.51	0.53	0.67	0.77		
G2	0.70	0.53	0.73	0.75	0.80	0.71	0.57	0.81	0.77	0.72	0.68	

Concluding on the reliability of test instruments, the calculations of Cronbach Alpha showed that ten out of twelve tasks had a satisfying reliability: M1 *Dice*, M2 *Calculator*, M3 *Folding*, M4 *Around the Bend*, M4 *Packaging*, S1 *Pulse*, S3 *Batteries*, S5 *Solutions*, G1 *Shadows*, and G2 *Plasticine*. For the two remaining tasks, S2 *Magnets* and S4 *Around the Bend*, the reliability of instruments was lower. It was questionable whether the data based on these tasks were suitable for analysis. Fortunately, they were science tasks that were not essential to the METRIC study. All tasks from the Performance Assessment needed in the METRIC study for the measurement of the attained curriculum showed a satisfying reliability (α >0.6).

4.4.5 Comparability issues

Haertel and Linn (1996) describe categories that need special attention pertaining the comparability of tests, in particular of performance assessments. They state that in performance assessments the following components are difficult to control: (1) the conditions of test administration, and (2) the scoring of responses.

Comparable test conditions

The conditions of test administration in the METRIC study were well scrutinised. The Written Tests were carried out in standardised test circumstances. However, in the Netherlands there is little experience with practical tests. Thus, the Performance Assessments in the METRIC study needed close scrutiny.

The equipment and materials used in 1995 were re-used in 2000 (thermometers, stopwatches, rulers, plasticine, dissolution tablets, glass containers, etc). There were three replacements of equipment. These were:

- for the task *Plasticine,* in 2000 a plastic balance was used instead of a pair of metal scales;
- for the task *Magnets*, in 2000 two alike magnets with different strengths were used instead of a large and strong magnet and a weaker and smaller magnet;
- for the task *Shadows*, in 2000 a torch with a narrowed beam was used instead of a torch with a wider beam.

The adjustments were allowed within the tolerance of the TIMSS Performance Assessment Administration Manual (1994). However, it was possible that students gained time by handling easier equipment (e.g. if the new balance reached its balancing point faster). Students would remain with more time for additional items or for reflection on the task. In that case the adaptations in the equipment had created significant differences in testing conditions. Therefore, a test was needed to explore whether the adaptations in the equipment had created differences. That test will be described after the following section, under the heading 'a comparability test for PA-1995 and PA-2000'.

Comparable codings

The coding of free response items both for the Written Tests and in the Performance Assessment was carried out with care. Coders were trained extensively. They practised in the use of the rubrics through exemplary material from previous testing rounds. During coding, some previously scored papers were returned into the collection, to control for the coder's accuracy. Despite the training, possible incomparabilities between coders could occur. Zuzovksy (1999) has pointed out that experience of coders with coding, subject content and pedagogy can help them to better understand and interpret students' responses. Differences in the experiences can lead to differences in coding patterns. The

difference in experiences did indeed occur in the Netherlands between coders of PA-1995 and PA-2000. The Dutch team of coders in PA-1995 consisted of teacher training students in mathematics and physics with little teaching and coding experience. By contrast, the single coder in PA-2000 was an experienced mathematics teacher who had gained additional coding training in the TIMSS format through a coding job for WT-1999. To test for differences between the groups of coders, a test was needed. It could have been an option to seed a selection of papers from 1995 through the papers of 2000. However, this option was eradicated, as students' work of 1995 had not been archived. This called for a test in which the codings of 1995 and 2000 could be compared.

In coding the responses in 2000, a special problem occurred. In the task *Rubber Band* the rubrics from the coding guide did not cover a strategy that was used by a considerable number of Dutch students. In this task the students had to measure and record the stretching of the rubber band with each washer. The result would yield an irregularly increasing graph (growing with 0-5 mm per washer) with diminishing growth. But approximately 10% of Dutch students did not measure. Instead, they made-up data that grew consistently with exactly 5 mm per washer (see Table 4.20).

Table 4.20: Examples of recordings by two students for the task *Rubber Band* from PA-2000

Number of washers	*Length of rubber band according to Student 104-3*	*Length of rubber band according to Student 8-5*
1	11 cm	11.5 cm
2	11.5 cm	12 cm
3	12 cm	12.2 cm
4	12.5 cm	12.7 cm
5	13 cm	13 cm
6	13.5 cm	13 cm
7	14 cm	12.8 cm
8	14.5 cm	13.1 cm
9	15 cm	13.4 cm
10	15.5 cm	13.4 cm

The graph from the invented data made a perfect straight line (student 104-3), while students with realistic data had an irregular line (student 8-5). There was no appropriate code for the styling strategy, which reduced realism from the onset of

the task. The coding rubrics only covered the number of recordings (at least five) and whether these showed a reasonable trend (increasing length with increasing number of washers). The coding guide did not anticipate that the criteria might be met without actual measurements. Therefore, depending on the interpretation of the coding scheme, a coder could either give full, partial or no credit.
Considering the above, the comparability of the Performance Assessments in 1995 and 2000 required a closer analysis. There were perceptible differences between the two measurements with respect to administration conditions (changes in equipment) and to coding (coders characteristics). It was unknown whether these would have implications on the quality of the data. This will be elaborated in the next paragraphs. For the comparability of the Written Tests of 1995 and 1999, there were no hesitations.

A comparability test for PA-1995 and PA-2000

To sieve out the tasks with a higher comparability, a test was devised in which the codes of students' responses were compared. It was alleged that, in general, Dutch students' performances would change only gradually in time. If abrupt differences in the codes occurred, there was a large probability that this could be ascribed (1) to altered test conditions or (2) to coders' disagreement, or (3) to both. Thus, a test was sought, which compared the students' answers of 1995 and 2000 as represented by the codes. Two examples should clarify this issue of comparability of codes in time. The first example is chosen for the corresponding pattern of codes of 1995 and 2000. The second is chosen for the contrasting pattern of codes of 1995 and 2000.

In Figure 4.2 the correctness scores of the eight items across the task *Around the Bend* are depicted in a stacked bar graph. The correctness scores on the items were 0, 1, 2, or 3. The bars indicate what percentage of students received a certain correctness score. On the left is the graph of the 1995 data, and on the right is the graph of the 2000 data. For example, we can compare the scores on the first item, item 1, which was a 2-point item. 90% of the students in the 1995 sample scored a full credit of 2 points, 8% scored a partial credit of 1 point and a small minority scored nil. On that same item, 93% of the students in the 2000 sample scored 2 points and 7% scored 1 point. The distributions are not very different and the differences can be ascribed to students' accomplishment and sample margins.

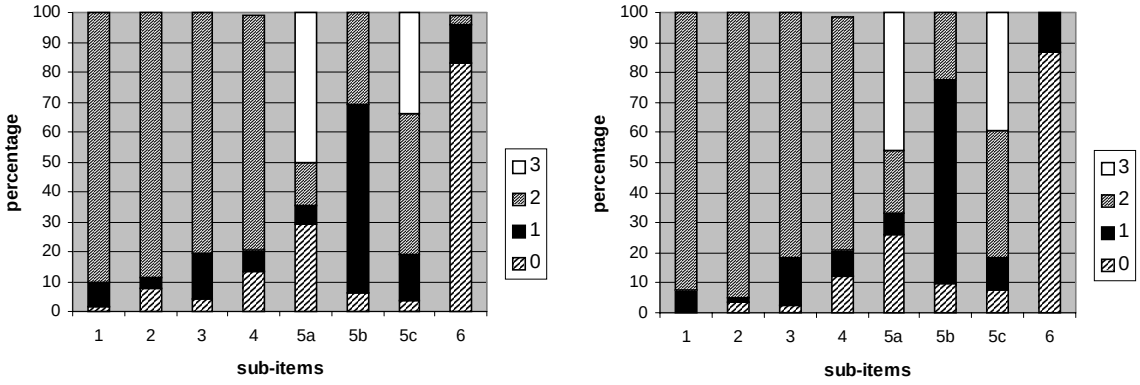


Figure 4.2: Correctness codes on the task *Around the Bend* from PA-1995 and PA-2000

From the stacked bars in Figure 4.2, it can be observed that the score distributions on all items were very similar when comparing 1995 and 2000. Intuitively, the pattern of shaded areas in the graphs seem comparable. This contrasts with the results on another task, the task *Rubber Band*. The correctness scores are shown in a stacked bar graph in Figure 4.3.

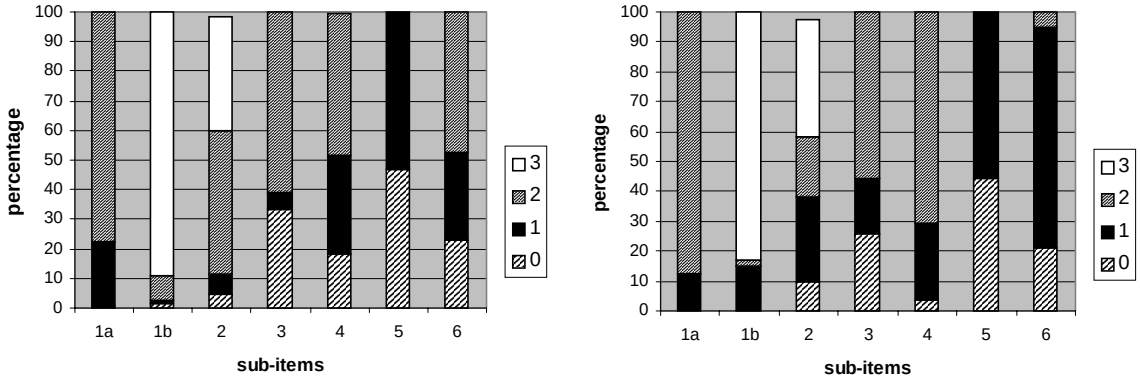


Figure 4.3: Correctness codes on the task *Rubber Band* from PA-1995 and PA-2000

On the sub-items 2 and 6 apparent differences between 1995 and 2000 occurred. For item 2, the grey bar for students in 1995 scoring 2 points seemed to correspond with the combination of 2 and 1 points in 2000. Similarly, for item 6 the large black bar for students in 2000 scoring 1 point seems to correspond with the combination of 2 and 1 points in 1995. For this task, the scores differed remarkably with the scores in 1995 on average being significantly higher. From this, it was concluded that the coders in 1995 were probably more lenient towards those students who did not measure the length of the rubber band, but who had stipulated its growth to be strictly linear. Because of this scoring dissimilarity, it was questionable whether the scores on this task were well comparable between 1995 and 2000.

To answer the question on the comparability of PA-1995 and PA-2000, the χ^2-test for comparing tables was carried out on the cross-tabulation, both for the correctness code (the first digit of the code), and for the diagnostic code (the two digits together). Through this test, large differences in scoring patterns were sieved out. If systematic inconsistencies occurred between the codes of 1995 and 2000, the test would show this. Theoretically, the large differences could have been caused by a tremendous change in students' achievements. However, this probability was considered negligible, based on evidence that students' achievement of any nation only changes gradually (Mullis et al., 2000). If large differences between the codes of 1995 and 2000 occurred, these were most probably caused by (1) different testing circumstances, or (2) differences in coding. The results of the χ^2-tests of each task are recorded in Table 4.21. The outcomes on the χ^2-test are indicated by their significance, which is the probability that the data of 1995 and 2000 have reasonably comparable code distributions. Probabilities $p \leq 0.05$ mean, that the codes of 1995 and 2000 have completely different frequencies.

Table 4.21: Comparability test (χ^2-test) between codes of students' answers in PA-1995 and PA-2000

Task	**N***umber of items*	$p(\chi^2)$ *on correctness scores*	$p(\chi^2)$ *on diagnostic scores*
M1 Dice	6	0.77	0.78
M2 Calculator	7	0.99	0.65
M3 Folding	4	0.53	0.17
M4 Around the Bend	8	1.00	1.00
M4 Packaging	3	0.28	0.18
S1 Pulse	4	0.33	0.34
S2 Magnets	2	0.00*	0.00*
S3 Batteries	4	0.18	0.18
S4 Rubber Band	7	0.00*	0.05*
S5 Solutions	7	0.57	0.48
G1 Shadows	6	0.01*	0.00*
G2 Plasticine	8	0.00*	0.00*

Note: * Significant at 5% level.

The comparison of the data of the tasks *Dice*, *Calculator*, *Folding*, *Around the Bend*, *Packaging*, *Pulse*, *Batteries*, and *Solutions* passed the test. On these tasks, the

probability that the two tests PA-1995 and PA-2000 had yielded comparable data, as expressed both through their correctness code and their diagnostic code, was satisfactory with $p > 0.05$. Four tasks, *Magnets, Rubber Band, Shadows* and *Plasticine* showed lack of comparability, with $p<0.05$. This could have been caused by the differences in test conditions or in coding discrepancies.
The results from the χ^2-test confirmed the intuitive idea about the comparability of data in 1995 and 2000 of the tasks *Around the Bend* and *Rubber Band* as exemplified in the Figures 4.2 and 4.3: the first task passed the test easily and the other did not.

Based on the analysis of quality of data, the following conclusions were drawn. The instruments for the two tasks *Magnets* and *Rubber Band* showed a lower reliability than acceptable. The data of the tasks *Magnets, Rubber Band, Shadows* and *Plasticine* showed a lower comparability than acceptable. Fortunately, the five tasks with a mathematical emphasis were not among the dubious cases. It was therefore advised to eliminate the four tasks *Magnets, Rubber Band, Shadows* and *Plasticine* from data analysis for the attained curriculum. The remaining tasks were *Dice*, *Calculator*, *Folding*, *Around the Bend*, and *Packaging*. This elimination only applied to the data for the attained curriculum. For the implemented and intended curriculum, there had never been questions on the comparability between data from 1995 and 1999/2000, as the teachers and experts were not affected by the change in equipment, and their data did not require coding.

4.4.6 Reporting and comparing data

Presentation of data

In the METRIC study, the research question was to explain students' achievement in light of the intended and implemented curriculum. Therefore, on all available test items, data on students' achievement, teachers' judgement and experts' judgement were gathered. To analyse students' achievement, items were grouped together, based on the judgements by curriculum experts and teachers, respectively. For example, in order to link the attained and the intended curriculum, the items that matched with the Dutch intended curriculum were set aside. The achievement of students on the items was calculated and compared in time, to see whether students' achievement had changed on this item set. In this way, data at the level of the intended and attained curriculum could be compared. Comparison of data from paired samples (the sample of 1995 and the sample of 1999/2000) requires descriptives, such as averages and standard errors. The data

are then compared, using Student's t-test for paired samples. Additionally, the analysis was supported by the calculation of Pearson's correlation coefficients, indicating whether the data sets at the level of the intended, implemented and attained curriculum were aligned. The alpha level of 0.05 was used for all statistical tests.

In the METRIC study, the data were reported in the following ways. Students' achievements on the test items were expressed in *p-values* on each of the mathematics items. The p-value is the percentage of students in the sample who completed the item correctly. Another expression used for the p-value is the 'proportion correct' or the 'percentage correct'.
To compare the achievements of students on the Written Test between 1995 and 1999, the 144 comparable (= identical or cloned) items were used. These items were perceived as the test overlap between WT-1995 and WT-1999. To compare the achievements of students between PA-1995 and PA-2000, the items with reliable and comparable results were used. These items were from the five mathematics tasks.
The presentation of students' achievement through p-values differs from the practice in TIMSS studies where students' achievement is reported using scaling methods, based on Item Response Theory (IRT). The scaling methods make use of plausible values, resulting in an estimated score of each individual student for the whole test and on content areas (numbers, measurement, algebra, etc.), even if the student has not responded to all items. However, for the METRIC study, there was no need for the individual student's score on the whole or part of the test. What was needed, was the achievement of all students to whom a particular test item was administered. This was best reported by p-values and their standard errors. The estimation of standard errors is reported below in a separate section.

For the judgement on the appropriateness of the tests in light of the implemented curriculum, the teachers had been asked to indicate with a 'yes' or a 'no' whether an item could be included into an imaginary test, which covered all content taught. This yielded an indicator for OTL. After aggregation over all teachers, each item had an OTL rate being the percentage of teachers that indicated a 'yes'. The rates were presented together with their standard errors (see below). For WT-1995, 16 items were given an OTL rate (based on data from 91 teachers). For WT-1999, all 155 items in the test received an OTL rate (each based on one third of the 112 teachers). For the METRIC study, only the rates

on the 144 comparable items were considered relevant. For PA-1995 and PA-2000, the OTL rates differentiated between 'content covered' and 'include into a test' (named: OTL-covered and OTL-testing).

For the measurement on WT-1995 and WT-1999 at the level of the intended curriculum, the experts had been asked to indicate with a 'yes' or a 'no' whether an item matched with the intended curriculum. This yielded an item-curriculum matching index on a yes/no scale for all items in WT-1995 and WT-1999.
On PA-1995 and PA-2000, a different strategy was used. With three experts in 1995 and five in 2000, the judgements were aggregated. Thus, each item received an item-curriculum matching index, being the percentage of experts that had indicated 'yes' to the item. For PA-1995, with three experts, the indices would be 0-33-67-100 for respectively 0, 1, 2 or 3 out of 3 experts. For PA-2000, with five experts, the indices would be 0-20-40-60-80-100 for respectively 0, 1, 2, 3, 4 or 5 out of 5 experts. In this way, the item-curriculum matching indices could differentiate between more-or-less matching the intended curriculum, depending on the number of experts who considered the item fit. Considering the number of experts, the indices were not presented with a standard error.

Estimation of standard errors

The p-values, the OTL rates and the item-curriculum matching indices are not precise. These are values that approximate the true values of the whole population. To give an indication of the confidence intervals, the p-values and the OTL rates are stated together with their standard error. The standard error depends on the size of the sample. The larger the sample, the smaller the standard error and the more accurate the p-value (or the OTL rate). Estimation of standard errors of students' achievement (p-values) on each item (whether Written Test or Performance Assessment) is presented by the formula:

$$SE^2 = \frac{P*(1-P)}{N}$$

where N is the number of students that took the item. The same formula applies to the teachers' OTL rates.
However, for the students' achievements on the Written Test, the result of the above formula yields an underestimate. It does not take into account the effects caused by the test design. In the Written Tests of the METRIC study, complete classes were tested, while these tend to be relatively homogeneous, clustering

students' achievement. To obtain a higher quality in the sample, it is generally advised to test a random sample of students from a school (selected from all grade 8 classes) and not complete classes. For research purposes, it is sufficient to select randomly just a small number of students from each class, but for practical purposes, it is often easier to test complete classes (Gonzalez & Foy, 1997; Snijders & Bosker, 1999).

The effect of the test design for clustered students is expressed as a coefficient for the Design Effect (DEff). The coefficient DEff is estimated from the intra-class correlation coefficient ϱ, which indicates the relation between the interclass variance and the between-class variance (Snijders & Bosker, 1999). Gonzalez and Foy (1997) estimate the design effect DEff by comparing results from the jack-knifing repeated replication (JRR) method with results from a simple random sampling design (SRS) through the following formula:

$$DEff = \frac{Var_{jrr}}{Var_{srs}}$$

In this way, Gonzalez and Foy (1997) found DEff=11,15 for the mathematics scores on WT-1995 in grade 8 in the Netherlands. With this coefficient, the sample size can be adapted to an effective sample size, EffN, with the following formula: (Gonzalez & Foy, 1997; Snijders & Bosker, 1999):

$$EffN = \frac{N}{DEff}$$

It means that the effective sample size is smaller than the original sample size, because of the homogeneity of classes. In fact, the students are given a lower weight because of their clustered achievements. With for example DEff = 10, it means that from a class of 30 students, only three randomly selected students are effectively needed for accurate test results. Also, based on DEff, the formula for calculation of the standard error is adjusted as follows:

$$SE_{eff}{}^2 = VAR_{eff}(P) = \frac{P*(1-P)}{EffN} = \frac{P*(1-P)*DEff}{N}$$

Based on this formula, the standard error increases with factor √DEff for items that were taken by all students from a class of size N.

Within the METRIC study, the design effect is not equal for all items. Many items were only taken by a few students from a class, and not by the complete class. For example, in the Performance Assessment, nine students were selected to be tested,

and then each task was taken by only three of them. This created a random selection of three students from each class. Similarly, items in many clusters of the Written Test appeared only in one of the eight test booklets. Thus, the distribution of booklets created a random selection of 3.125 students from each class (one out of eight, with average class size of 25). Therefore, the factor DEff used to estimate the standard errors in the METRIC study were made to differ per item category. The items in the Performance Assessment are perceived as not requiring adjustment (DEff=1). The items in the Written Test were allocated a design effect factor DEff, depending on the number of booklets in which an item appeared. If an item appeared only in one out of the eight booklets, the design effect was smaller than if an item appeared in all test booklets. The scheme for interpolation of DEff is given in Table 4.22. It is based on the relation DEff = 1+ ϱ(b-1), whereby ϱ is the intra-class correlation and b is the number of students in a class who completed the same test booklets. With the coefficients DEff, the standard errors of p-values on items were estimated.

Table 4.22: Interpolation of the design effect factor DEff for items in the achievement tests in the METRIC study

Item category	**WT-8**	*WT-4*	*WT-3*	*WT-2*	*WT-1*	*PA*
DEff	11	5.8	4.1	3.2	1.9	1

Note: WT-8 = items appearing in eight test booklets of WT; WT-4 = items appearing in four test booklets of WT, etc.

The previous section described the estimation of standard errors for single items. When grouping the items, the p-values and the OTL rates were averaged. The associated standard error of the averaged p-value $\hat{p}$ was estimated by the *pooled* standard error, which is estimated as follows:

$$SE_{\hat{p}}^2 = \frac{1}{n}\sum SE_i^2$$

Comparing data

In the METRIC study, it was explored whether students improved on certain items. Therefore, the occurrence of significant differences of means was calculated as follows. To test the difference between two p-values p_1 and p_2 on an item, we made use of the standard errors SE_1, SE_2 and the following formula for the standard error between two means (Gonzalez & Foy, 1997; Freund, 1988):

$$SE_{diff}{}^2 = SE_1^2 + SE_2^2$$

With two measurements in different years (1995 and 1999/2000), the two samples were considered independent. The difference in scores was then tested through a two-tailed t-test, at 95% and at 99% confidence intervals. As support in the testing, in Appendix E two matrices are given. In these, the significant difference between two p-values can be found in the following way:
First, the p-value of the one sample is looked up horizontally. Second, the p-value of the other sample is looked up vertically. By crossing the row and the column, a cross point is found. That cross point will indicate if there is a significant difference between the p-values. If it lies in the shaded areas, there is no significant difference between the two p-values. The comparison of p-values can both be applied to the Written Tests and the Performance Assessment.

To compare data between the intended, implemented and attained curricula, correlation coefficients were calculated. These will be presented together with their significance levels.

In the next chapter, chapter 5, trends in the resulting database of the METRIC study will be described through descriptive statistics. There were 6 measurements in 1995 (two tests, three curricular appearances), which were repeated in 1999/2000. The data of the twelve measurements will be used to answer the research questions in the ensuing chapter 6.

Chapter 

Trend results

~ Cada um sabe onde o sapato aperta. ~

Only the wearer knows where the shoe pinches.
(BRASILIAN PROVERB)

This chapter describes the data gathered in the METRIC. The data were collected through six measurements in 1995, which were repeated in 1999/2000. In the first section, section 5.1, trends in the results at the level of the intended curriculum are presented. In section 5.2, trends in the results at the level of the implemented curriculum are presented. Finally, in section 5.3, trend results at the level of the attained curriculum are presented. These data are needed to answer the research questions in chapter 6.

5.1 INTRODUCTION

The research results of the METRIC study are based on four sub-studies: the international TIMSS Written Test of 1995 and its repeat in 1999 (WT-1995 and WT-1999) and the international TIMSS Performance Assessment of 1995 and its repeat in 2000 (PA-1995 and PA-2000). The data were collected for the three following curricular appearances: the intended curriculum, the implemented curriculum, and the attained curriculum. Therefore, there were twelve measurements. The categories of the data collection are illustrated in Figure 5.1.

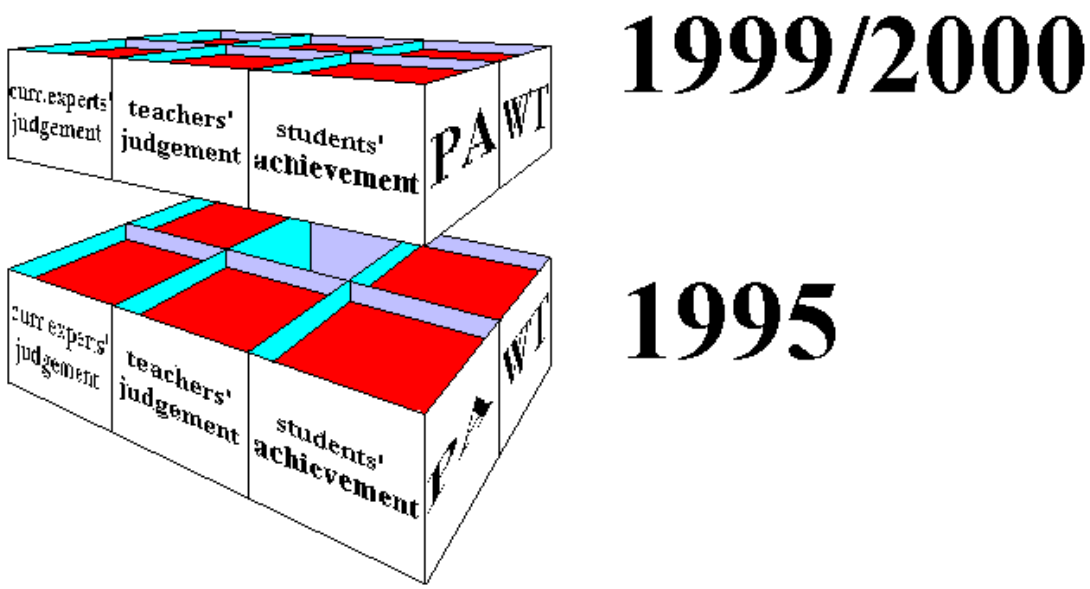


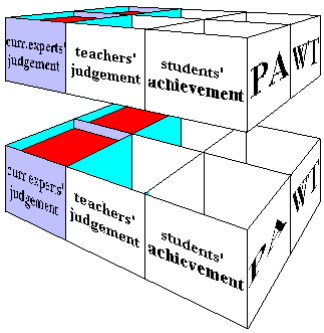
Figure 5.1: Data collection categories in the METRIC study

As explained in the previous chapter, every effort was made to create valid trend comparisons by repeating the sub-studies of 1995 almost identically in 1999/2000. For WT-1995 and WT-1999, data were collected on 144 test items, which were submitted to curriculum experts, teachers and students. For one category, the data set from WT-1995 at the level of the implemented curriculum was limited to 16 items. This could not be complemented anymore within the METRIC study.

For PA-1009 and PA-2000, data collection was successful for all six categories, using the seven tasks *Dice*, *Calculator*, *Folding*, *Around the Bend*, *Packaging, Shadows* and *Plasticine,* although the measurement at the level of the attained curriculum of 1995 and 2000 was limited to the first five mathematical tasks.

In the following sections, this database will be presented in the sequence: intended curriculum, implemented curriculum, attained curriculum. The data collected from each of the curriculum appearances will be separately presented for the Written Test and the Performance Assessment. Thus, trends will be revealed. Based on the trends, the research questions will be answered in Chapter 6.

5.2 THE INTENDED MATHEMATICS CURRICULUM – EXPERTS' APPRAISAL OF THE TWO TESTS



5.2.1 *Trends in Dutch curriculum experts' judgement on the appropriateness of the Written Test*

In the following two sections, the appropriateness of the Written Test and the Performance Assessment will be presented in light of the intended mathematics curriculum. First, results from the Written Tests in 1995 and 1999 will be presented. In section 5.2.2, results from the Performance Assessment will be presented.

To operationalise the intended curriculum, three mathematics curriculum experts judged the appropriateness of the test items in the Written Test. Two out of the three experts of 1995 were re-consulted in 1999. For each item, the experts were asked to indicate whether it matched with the intended curriculum as formulated in the core objectives for junior secondary school (*Kerndoelen Basisvorming*). The experts' judgement was binary: either an item matched or did not match. Therefore, the experts had to reach consensus: if they disagreed on their judgement, it was negotiated (by phone) till an agreement was reached. The judgement of each item yielded an *item-curriculum matching index*, which was expressed on a nominal yes/no scale for the Written Test. For the complete test, the data yielded a *test-curriculum matching index* (see Figure 3.1).

It was expected that a difference between 1995 and 1999 in the judgements of the appropriateness of the test would occur, as the core objectives had been reviewed in 1998. For example, the Pythagorean theorem and the concepts of probability, inter- and extrapolation and congruence had been skipped from the program. The use of computers, calculators and the usefulness of mathematics were further stressed.

The results of the judgement on the appropriateness of the 144 items in the TIMSS Written Test in light of the intended curriculum are presented in Appendix F. In 1995, 99 out of 144 items (69%) matched with the intended mathematics curriculum. In 1999, 102 out of 144 items (71%) matched with the intended mathematics curriculum. The two percentages are very similar.

However, there is a discrepancy between the two results, as they do not point at the same items. The two items sets considered appropriate in 1995 and 1999 were not the same.
The difference in the judgement can be expressed through the correlation coefficient, calculated from the item-curriculum matching indices of 1995 and 1999. The correlation between the experts' judgements of 1995 and 1999 is not meaningful. Because of the binary data, the correlation was calculated through parameter-free methods. Both Kendall's tau and Spearman's rho methods yielded r=0.19 (n=144, p=0.02).

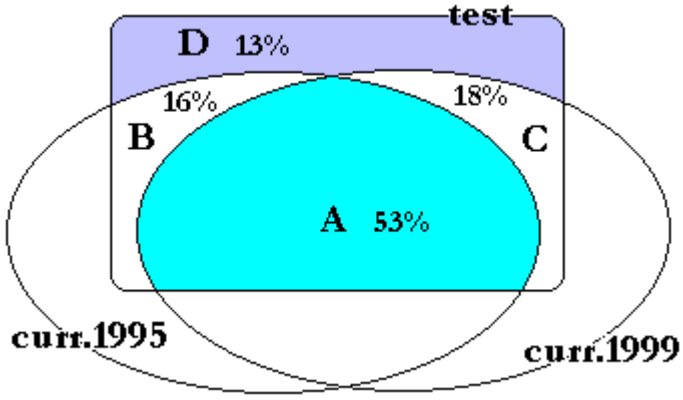


Figure 5.2: Partitioning of the Written Test by the match with the intended curriculum in 1995 and 1999

The item-curriculum matching indices of 1995 and 1999 differ on approximately one third of the items. The complete test can be partitioned according to the judgements of 1995 and 1999. This is illustrated in Figure 5.2. The full set of 144 items is indicated as a rectangle. There are two different test-curriculum matchings, one based on the intended curriculum in 1995 and another based on the intended curriculum in 1999. The figure also illustrates that the curricula cover areas, which are not included by the test.
When looking at the test, there is a subset of items, which is covered by the intended curriculum according to the judgement by the experts in 1995 (parts A and B). This subset of the test differs from the subset, which is covered by the intended curriculum of 1999 (parts A and C). Thus, the matching of the full test with two different intended curricula yields a partitioning into four parts (parts A, B, C and D).

Out of the complete test of 144 items, 76 items were considered to match the intended curriculum of both 1995 and 1999 (part A: 53%). There were 23 items

that matched the intended curriculum in 1995 but not in 1999 (part B: 16%). There were 26 items that did not match the intended curriculum of 1995 (part C: 18%). The remaining 19 items neither matched with the intended curriculum of 1995 nor with that of 1999 (part D: 13%). Table 5.1 lists the four categories.

This analysis of data at item level shows that slightly more than half of the test matched with the intended curriculum of both 1995 and 1999. Together with the part of items that are covered neither in 1995 nor in 1999, there is an agreement between 1995 and 1999 on two-thirds of the items (part A and D: 53%+13%). On one third of the items (part B and C: 16%+18%), the judgement by the curriculum experts changed between 1995 and 1999, either from 'yes' to 'no' or the other way around.

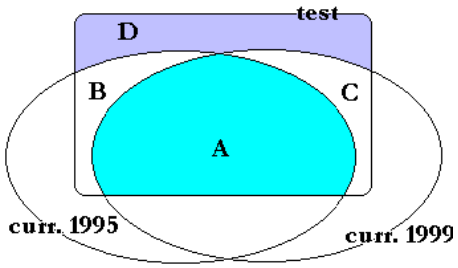


Table 5.1: Percentages of test items matching with the intended curriculum in WT-1995 and WT-1999

Part	*1995 matching intended curr.*	*1999 matching intended curr.*	*# items*	*% of items (n=144)*
A	Yes	Yes	76	53
B	Yes	No	23	16
C	No	Yes	26	18
D	No	No	19	13

Part A contains many items, of which the content has been covered at primary school level, such as applying basic arithmetic algorithms, working with fractions and decimal numbers, reasoning with proportions, measuring, calculating area, and interpreting diagrams and graphs. By contrast, part D contains algebra items on manipulating 'bare' formula (e.g. items K04, Q02 and R09/R10 in Appendix D) and a few items on calculating probabilities.
Bos and Vos (2000) looked at the characteristics of items in the sets B (matching the intended curriculum as measured in 1995 but not in 1999) and C (matching the intended curriculum as measured in 1999 but not in 1995):

Characteristics of set B, in comparison to set C:

- Set B contained more items on fractions and number sense. In 1995, the experts probably still relied on higher skills in mental arithmetic. In 1999, the experts indicated that they had preferred the use of calculators for these items. With calculators being prohibited during administration of the Written Test, the experts judged these items as not matching with the intended curriculum.
- Set B contained more items on two-dimensional geometry. This topic was included into the core objectives of 1993, but was abandoned in the 1998 review of the intended curriculum.

Characteristics of set C, in comparison to set B:

- Set C contained more items on topics that were not part of the core objectives but which could be solved rather straightforwardly by 'common sense'. For example, items on probability, data representation or algebra, which could be solved with proportional reasoning, were included now. The core objectives stress the use of common sense in mathematics, and in 1999 the experts probably trusted that these items matched with the intended curriculum.
- Set C contained more items on algebra. This shift can be ascribed to the minor adaptation in the algebraic notation of multiplication with variables. Notations such as '7a' in WT-1995 had been altered into '7•a' in WT-1999, making them correspond to the current practice in Dutch mathematics education.

Besides the adaptations in the core objectives, there could be other reasons for the differences between 1995 and 1999 in the match of the Written Test with the intended curriculum. These could be:

- The interpretation of the intended curriculum by the experts had changed.
- The judgement on a number of items was complex because of their formal phrasing and the multiple choice format (see chapter 4, section 4.3.5). In these cases the judgement resembled a toss-up. Therefore, in 1995 and 1999, the toss could have ended differently.
- Maybe the instrument used in the measurement was unreliable. This instrument was copied from the TIMSS Test Curriculum Matching Analysis (TCMA), and therefore the reliability was trusted and not questioned. However, in hindsight, this could have been an omission.

Thus, the judgement on the appropriateness of Written Test in 1995 and 1999 in light of the intended curriculum showed (1) consistency on two-thirds of the items and (2) a shift in judgement on one third of the items. In the next section, we will see whether the experts' judgement on the appropriateness of the Performance Assessment did change to the same extent.

5.2.2 *Trends in Dutch curriculum experts' judgement on the appropriateness of the Performance Assessment*

The items in the Performance Assessment were judged by curriculum experts in 1995 and 2000. In both years the same instruments were used, similar to the instrument used for assessing the Written Test. In both measurements, the items across the five mathematics tasks *Dice*, *Calculator*, *Folding*, *Around the Bend* and *Packaging*, and the two combined science/mathematics tasks *Shadows* and *Plasticine* were submitted to the experts. In 2000, two more science tasks (*Batteries* and *Rubber Band*) were submitted because they contained mathematical activities as well. In 1995, three experts responded. In 2000, the same three experts responded, plus two more, making a total of five experts.

In the measurement of the appropriateness of the Performance Assessment in light of the intended curriculum, the data allowed for more detailed proportions than on the Written Test. The data were aggregated over the experts and this yielded an *item-curriculum matching index (per item)*, expressed by a percentage. This presentation in a ratio scale allowed for more moderation than the presentation on a nominal scale, by stating either 'yes' or 'no' (after reaching consensus). The judgement on the appropriateness of Performance Assessment items allowed for compromising between the yes/no dichotomy with no consensus being sought. With three experts in 1995, the judgements on the Performance Assessment were translated into the rates 0-33-67-100 for respectively zero, one, two or three experts indicating that the item matched with the curriculum. Similarly, for 2000, with five exerts, the rates were 0-20-40-60-80-100. The data are presented at item level in Appendix G. In Table 5.2, the item-curriculum matching indices are averaged per task and over the complete test.

There were three tasks with a high average matching index on its items (>75): *Dice*, *Around the Band* (both in 1995 and 2000), and *Rubber Band* (only judged by mathematics experts in 2000). The rates of *Folding, Around the Bend,* and *Plasticine* were high in 1995 (>75) but decreased in 2000. The rates of two tasks remain virtually stable (*Calculator* and *Shadows*).

The two tasks, which were added into the measurement of 2000, *Rubber Band* and *Batteries*, received very different ratings. The task *Batteries* received a very low rate of approval. Just one of the five experts in 2000 rated this task as matching with the intended curriculum. On the other hand, the task *Rubber Band* that had been labelled as a 'science task' by the TIMSS International Study Center, received very high item-curriculum matching indices. According to the judgement by the Dutch mathematics experts, this task was second best to match with the intended curriculum.

Table 5.2: Average item-curriculum matching indices of test items in PA-1995 and PA-2000

	1995 (n=3)		*2000 (n=5)*	
Task (#items)	*Avg item-curr. matching index*	*# items with item-curr. matching index > 50*	*Avg item-curr. matching index*	*# items with item-curr. matching index > 50*
Dice (5)	87	4 out of 5	92	5 out of 5
Calculator (6)	72	6 out of 6	70	5 out of 6
Folding (4)	100	4 out of 4	60	3 out of 4
Around the Bend (6)	100	6 out of 6	83	5 out of 6
Packaging (3)	89	3 out of 3	73	3 out of 3
Shadows (6)	61	3 out of 6	67	6 out of 6
Plasticine (3)	78	3 out of 3	40	1 out of 3
[Batteries (4)]	–	–	[25]	[0 out of 4]
[Rubber Band (6)]	–	–	[87]	[6 out of 6]
Average index (33 items in 7 tasks)	*83*	*29 out of 33 (88%)*	*72*	*28 out of 33 (85%)*

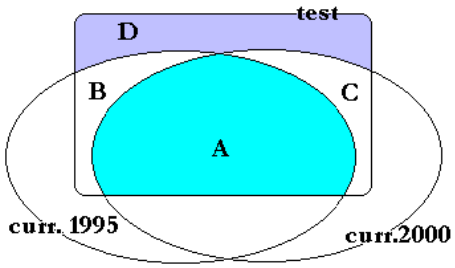
At the bottom of Table 5.2, the curriculum matching indices are averaged over the 33 items in the seven tasks *Dice*, *Calculator*, *Folding*, *Around the Bend*, *Packaging*, *Shadows* and *Plasticine*. Between 1995 and 2000, the average index dropped by 11 percent, from 83 to 72. Thus, it can be concluded that the Performance Assessment received a lower approval of the experts in 2000 than in 1995.

When comparing the curriculum matching indices of 1995 and 2000 at item level, the trend correlation is significant: $r=0.41$ ($n=33$, $p=0.02$). The correlation is not high (Krathwohl, 1998), but it is higher than the trend correlation $r=0.19$ on the item-curriculum matching indices of the Written Test ($n=144$, $p<0.01$). However, the latter were compiled on a dichotomous yes/no scale.

Because of the difference in scales, it is not possible to compare the data on the Performance Assessment straightaway with the data of the Written Test. To make the data on the Performance Assessment comparable to those of the Written Test, the data expressed as percentages of the Performance Assessment were transformed into data on a yes/no scale. The transformation was achieved by taking the *majority judgement*. An item was said to match with the intended curriculum, if a majority of experts had indicated a 'yes'. With odd numbers of experts (three in 1995 and five in 2000), there never occurred a draw at 50-50.
Thus { 0, 20, 33, 40 } → 'not matching' and { 60, 66, 80, 100 } → 'yes, matching'.

For example, the five items across the task *Dice* had item-curriculum matching indices in 1995 of 100-100-100-100-33. These figures indicate that four items (out of five) had a majority approval. The indices of 2000 on the same task were very similar to those of 1995, being 100-100-100-100-60. Thus, on only one out of five items, the index changed apparently. Expressed as a percentage, the sequences differ by 5%. Despite the similarity in the two sequences of indices given above, they leads to a different majority approval: there are four out of five items matching with the intended curriculum in 1995, and five out of five in 2000. This means a difference of 20%. The transformation of the ratings into binary data is included in Table 5.2. The transformation onto a yes/no scale is needed for comparison with the data on the Written Test.
Taking the majority judgement for all items in the Performance Assessment, there were 29 out of 33 test items (85%) covered by the intended curriculum in 1995. By 2000, this figure had hardly changed, with 28 items (88%) covered by the intended curriculum. Thus, the curriculum match of the Performance Assessment in 1995 and 2000, when expressed on a binary scale, seem to be of the same magnitude. And just like with the Written Test, the two sets do not fully overlap.
In the full test there were 33 items. There were 24 items that matched the intended curriculum both in 1995 and 2000 (part A: 73%). There were nine items on which the judgement of 1995 and 2000 differed: four items matched with the intended curriculum in 1995, while they did not in 2000 (part B: 15%). Five items matched in 2000 and did not match in 1995 (part C: 12%). There was not one item from the seven tasks *Dice*, *Calculator*, *Folding*, *Around the Bend*, *Packaging*, *Shadows* and *Plasticine* that matched neither in 1995 nor in 2000 (part D: 0%). The results on the Performance Assessment are tabulated in Table 5.3. The table has the same format as Table 5.1 on the Written Test.

Table 5.3: Percentages of items matching with the intended curriculum in PA-1995 and PA-2000



Part	*1995 matching intended curr.*	*1999 matching intended curr.*	*# items*	*% of items (n=33)*
A	Yes	Yes	24	73
B	Yes	No	5	15
C	No	Yes	4	12
D	No	No	0	0

Comparing Table 5.1 and 5.3 shows that there is a larger proportion of test items in the Performance Assessment matching with the intended curriculum than in the Written Test. 85-88% of the Performance Assessment is covered by the intended curriculum, versus 69-71% of the Written Test. Still, like on the Written Test, on a considerable proportion (25%) of test items in the Performance Assessment the judgement of 1995 and 2000 disagreed. However, this proportion was smaller than in the Written Test, which had a disagreement on 33% of the test items.

Just like on the Written Test, we can look at item characteristics in the parts B and C, which are the parts of the disagreement in judgement between 1995 and 2000. Part B contains the items that were considered as covered by the intended curriculum in 1995 and not in 2000. The five items are spread out over different tasks (*Calculator*, *Folding*, *Around the Bend*, and *Plasticine*). The items can be found at the end of the tasks, being the final item asking for an explanation or a reflection on the task. Generally speaking, these items are 'harder' than the other items. Thus, the majority of the experts in 2000 probably refuted more difficult items, while the majority of the experts in 1995 considered these as covered by the intended curriculum.

Part C contains the items that were considered as covered by the intended curriculum in 2000 and not in 1995. Three of the four items are found in the combined science/mathematics task *Shadows*. That task consists of six items. In

the first three items the students are asked to discover how the size of the shadow changes and to make three measurements of the distances between torch, screen and object. In the last three items, the students are asked to design an investigation to find out the relationship between the distances between the objects. In this sequence the last three items reverse the logical order of doing an investigation (first making a design, then doing the measurements). Thus, the majority of the experts in 1995 refuted the three items on the research design, while they considered the first three items on 'doing the measurements' as covered. The majority of the experts of 2000 did not take this difference into account and considered all items covered.

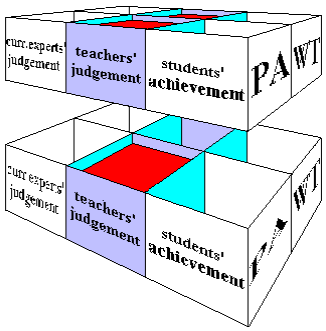
The difference between the parts B and C illustrates that a small majority judgement in the one year (60% or 67%) can be consistent with a large minority judgement in the other year (40% or 33%). With three experts in 1995 and five in 2000, the ratio of 'two out of three' (67%) can become 'two out of five' (40%) without any individual expert changing his/her mind. The extra two experts can then have caused the changing rate. Similarly, a minority such as 'one out of three' (33%) can become without inconsistency a majority such as 'three out of five' (60%). Thus, because of the small numbers of consulted experts, the transformation of the item-matching indices from a ratio scale into a yes/no scale can magnify the differences between the judgements of 1995 and 2000. Similar judgements from 1995 and 2000, such as those on the task *Dice*, can become more diverging. As a result, the judgements of 1995 may appear less consistent with the judgements of 2000 than they were when measured on the ratio scale. This effect is also shown by the correlation coefficients on the two judgements of 1995 and 2000. When measured on the ratio scale, the correlation is $r=0.41$ ($n=33$, $p=0.02$). This correlation coefficient indicates that there is a certain amount of consistency between the judgements of 1995 and 2000. After dichotomisation of the curriculum matching indices into the yes/no scale through the *majority judgement transformation*, the correlation coefficient has become negative and insignificant with $r= -0.18$ (Kendall's tau and Spearman's rho, $n=33$, $p>0.05$). This shows that the operationalisation of the intended curriculum is probably better served by an instrument, which yields rates on a ratio scale that leaves room for compromising between the dichotomy of a yes/no answer.

5.2.3 *Comparison of trends in Dutch curriculum experts' judgement on the appropriateness of both tests*

To operationalise the intended curriculum for the METRIC study, the two tests were submitted to curriculum experts both in 1995 and in 1999/2000. The experts were asked to indicate for each test item whether it matched with the intended curriculum. The instrument was borrowed from the TIMSS Test Curriculum Matching Analysis consisting of a listing of the items together with yes/no boxes to tick. For the Written Test, the answers of the experts were compiled and in the case of disagreement, a consensus was sought. For the Performance Assessment, the answers of the experts were aggregated into item-curriculum matching indices.
The measurements of 1995 and 1999/2000 showed a shift in the judgements of the experts on a number of items. The analysis of item characteristics offer evidence that this shift can be ascribed to the review of the core objectives in between the measurements (see section 5.2.1). Additionally, the shift can be ascribed to the method applied, disallowing for compromising. On a yes/no scale, differences in judgements were enlarged (see section 5.2.2). In particular, the data compiled from the Performance Assessment showed that the judgement on the appropriateness of test items in light of the intended curriculum asked for the need to find middle grounds. The shift between 1995 and 1999/2000 in the judgement by the curriculum experts could additionally have been caused by their changed interpretations of the core objectives, and the complex process of judging multiple choice items while ignoring the multiple choice format (see section 4.3.5).

Comparing the results on the Written Test and the Performance Assessment yields the following observation: the proportion of items that match with the intended curriculum in both years (part A in Figure 5.2) is 53% for the Written Test and 73% for the Performance Assessment. The proportion of items that is covered neither in 1995 nor in 1999/2000 (part D in Figure 5.2) is 13% for the Written Test and 0% for the Performance Assessment. This shows that the Performance Assessment is more closely aligned with the Dutch intended curriculum than the Written Test.

5.3 THE IMPLEMENTED MATHEMATICS CURRICULUM – TEACHERS' JUDGEMENT ON THE APPROPRIATENESS OF THE TESTS



5.3.1 *Trends in Dutch teachers' judgement on the appropriateness of the Written Test*

To operationalise the implemented curriculum, the mathematics teachers of the students sampled judged all items on whether or not the items matched with the content taught before test administration. The instrument used was developed by De Haan (1992) (see section 4.3.4). In this current section, results from the Written Test will be presented. In section 5.3.2, the Performance Assessment will ensue.

For the TIMSS Written Test in 1995, 16 test items were submitted to the teachers as part of the National Option Test (NOT). The 16 items were selected for this national test, because they matched well with the intended mathematics curriculum in 1995. The 16 anchor items were not representative of the complete Written Test. Therefore, in 1999, all mathematics test items in the TIMSS Written Test were submitted to the teachers. The teachers were asked to indicate whether they would include the items into an imaginary test, covering all content taught so far. Their answers were taken as an indicator of OTL (Opportunity to Learn).

Table 5.4 shows the results for the 16 items submitted in 1995 (the texts of these items are all included in Appendix D). The table shows for each item what percentage of mathematics teachers in 1995 answered positively to the OTL question. This percentage is named the OTL rate of the item. The table also contains the OTL rates from 1999. By making use of the identical or cloned items in WT-1999, the teachers' judgement from 1999 can be compared to those from 1995. The first eight items were identical in WT-1995 and WT-1999. The other eight items were released for publication and cloned into new items for WT-1999. For clarification, the names of the cloned items in WT-1999 are indicated separately as some of them changed names (only in clusters L and R).

Table 5.4: OTL trend results on 16 selected items from WT-1995 and WT-1999

Item		**O***TL rate (SE)*	
		1995	*1999*
WT-1995	*WT-1999*	*(n=84)*	*(36≤n≤ 39)*
A02	id.	89 (3)	85 (6)
B08	id.	90 (3)	94 (4)
B10	id.	96 (2)	100 (0)
B11	id.	98 (2)	95 (4)
C05	id.	89 (3)	92 (4)
E05	id.	89 (3)	74 (7)
F12	id.	96 (2)	97 (3)
H07	id.	86 (4)	85 (6)
K06	clone: K06	96 (2)	95 (4)
L08	clone: L09	96 (2)	100 (0)
N15	clone: N15	99 (1)	95 (4)
N19	clone: N19	96 (2)	97 (3)
O09	clone: O09	96 (2)	92 (4)
R07	clone: R08	96 (2)	95 (4)
R10	clone: R11	98 (2)	97 (3)
V02	clone: V02	86 (4)	89 (5)
Average (16 items)		*93 (3)*	*93 (4)*

All items received a high OTL rate, both in 1995 and 1999. In 1999, there were two of the 16 items that received a *full credit* of 100. This means that all responding teachers indicated a 'yes' to these items. On average, in 1995 and 1999 the items had an OTL rate of 93 in both years. There is a significant correlation between the OTL rates of 1995 and 1999 on the 16 items, $r=0.70$ ($n=16$, $p<0.01$). This means that items with a high OTL rate in 1995 received a high OTL rate in 1999 again.

With one exception, all rates were higher than 85. The exception is item E05 that had a high OTL rate in 1995 of 89, which decreased in 1999 to 74. With the small number of respondents, and thus the large standard errors, this decrease is *not* significant ($p=0.06$). The text of item E05 cannot be published, due to the TIMSS item release policy (see section 4.3.2). It is an algebra item, in which several ordered pairs of numbers have to be associated to a formula. The item is abstract and has no real-life context. This could explain why Dutch mathematics teachers decreased their judgement in the period between 1995 and 1999.

The high OTL rates of the 16 items as presented in Table 5.4 are not exemplary for all other items in the TIMSS Written Test. The 16 items are not representative of the complete Written Test. To supplement OTL information in 1999, all items in the test were submitted to the teachers. In Appendix F, the data are reported.

In 1999, the average OTL rate for the 144 items in the TIMSS Written Test was 82 (SE = 6). It means that on average, any of the items was considered suitable for an imaginary test by 82% of the responding mathematics teachers. When looking into detail, it turns out that more than one third of the test items had an OTL rate higher than 90. There were only six items (4%) that received an OTL rate below 50 (these were F08, I01, K04, N18, Q02, and R10 [=R09 in 1995]). Two of the items ask the students to work with probability (F08 and N18), and the other four items ask the students to work with abstract variables. As exemplary items, the text of four of the six items (items F08, I01, Q02 and R10) has been included in Appendix D.

The OTL rates are percentages of teachers in the range between 0 and 100. To get an overview of the OTL rates, ten categories were created. These were labelled 0-10, 10-20, 20-30, and so forth. In this way the items could be grouped into a category according to their OTL rate. Items in the category 0-10 had OTL rates lower than or equal to 10. Items in the category 10-20 had OTL rates higher than 10 and lower than or equal to 20, and so forth. The items in each category were counted and calculated as a percentage of the total item set. These percentages are given in Table 5.5. Two columns were made: one for the frequencies (in percentages) of the 144 items from 1995 and 1999, and another column for the cumulative frequencies.
The data show ample approval of the Written Test by the Dutch mathematics teachers in 1999. On most items, their judgements were very high. The categories with the highest OTL rates have the highest frequencies. Two-thirds of the items (67%) have an OTL rate higher than 80. And 90% of all test items have an OTL rate of 60 or higher. This shows that the TIMSS Written Test matched relatively well with the implemented curriculum in 1999.

Table 5.5: OTL rates on all items in WT-1999 (comparable data for WT-1995 unavailable)

Category for **O***TL rate*	**%** *of math items in Written Test (n=144)*	*Cumulative* **%** *of items * (n=144)*
90 - 100	38	38
80 - 90	29	67
70 - 80	15	82
60 - 70	8	90
50 - 60	6	96
40 - 50	1	97
30 - 40	3	99
20 - 30	0	99
10 - 20	2	100
0 - 10	0	100
Average OTL rate on all items 82 (SE=6)		

Note: * Due to rounding off, the cumulative percentages may seem inconsistent with the percentages.

As a side step, the statistics on teachers' acceptance of the multiple choice format is reported in this paragraph. Teachers were asked to indicate whether they would include an item into an imaginary test, when considering (1) its content and (2) its format (see section 4.3.4). Only their answers on the content of the items were taken as indicator of OTL. However, the consideration on the format yields information on the extent to which teachers would include the item into an imaginary test, while maintaining the multiple choice format. The results of 1999 were as follows: for any item from the Written Test, on average 71% of the teachers indicated: 'yes' to content and 'yes' to format; similarly, on average 11% of the teachers indicated: 'yes' to content and 'no' to format. Therefore, a large majority of teachers did not object to the multiple choice format. This confirms that many teachers considered many items in the TIMSS Written Test as matching with the implemented curriculum.

In order to find out whether the Dutch mathematics teachers' approval of the Written Test was exceptional, Vos and Bos (2001a; 2001b) compared the rates with teachers of other subjects and with teachers in other countries. They

showed that the OTL rates of Dutch mathematics teachers in 1999 on the mathematics items in the Written Test is comparable to the rates of Flemish mathematics teachers on the same items (84) and it is much higher than the approval by Dutch science teachers on the science items in TIMSS-99 (61).
Unfortunately, the OTL data of the Written Test do not allow a comparison in time between data of 1995 and 1999. The number of items (16) on which OTL rates are available in both years is small, and the characteristics of these items do not allow for generalisation, as they are not representative for the whole Written Test.

5.3.2 *Trends in Dutch teachers' judgement on the appropriateness of the Performance Assessment*

To operationalise the implemented curriculum, the mathematics teachers of the participating students judged the test items in the Performance Assessment. Both in 1995 and 2000 all items across the mathematics tasks were submitted to them. The question to the teachers was twofold: whether they had taught the content of an item before test administration, and whether they would include the item into an imaginary test of their own making. Therefore, there will be two types of data: data for 'OTL-covered' (for covered content) and data for 'OTL-testing' (for readiness to include into a test).

Table 5.6 OTL trend results on PA-1995 and PA-2000

	Avg **O***TL-covered rate (SE)*		*Avg* **O***TL-testing rate (SE)*	
Task (#items)	*1995 (n=19)*	*2000 (n=20)*	*1995 (n=19)*	*2000 (n=20)*
Dice (5)	47 (11)	73 (10)	51 (11)	70 (10)
Calculator (6)	43 (11)	68 (10)	56 (11)	82 (9)
Folding (2)	17 (8)	31 (10)	28 (10)	75 (10)*
Around the Bend (6)	35 (10)	68 (10)*	56 (11)	84 (8)*
Packaging (3)	65 (10)	74 (10)	76 (10)	88 (7)
Shadows (6)	30 (10)	33 (10)	47 (11)	67 (11)
Plasticine (3)	23 (9)	40 (11)	26 (10)	70 (11)*
Average over all items (31 items)	*38 (10)*	*58 (10)*	*51 (11)*	*76 (9)*

Note: * Significant difference between data of 1995 and 2000 ($p<0.05$).

Table 5.6 shows the Dutch data of OTL-covered and OTL-testing for 1995 and 2000 on the TIMSS Performance Assessment. In the table the rates are average percentages of teachers who answered positively to the OTL-covered and the OTL-testing question. In 1995, the average OTL rate on an item was 38 for OTL-covered and 51 for OTL-testing. These were low rates, indicating that less than half of the teachers had covered the content tested in the test items. The rates had increased five years later. In 2000, the average OTL rate on an item was 58 for OTL-covered and 76 for OTL-testing. Table 5.6 also presents a breakdown of the OTL-rates for the different tasks of the Performance Assessment.

With the small number of teachers responding, the OTL data have a limited value. Still, the table yields noteworthy results. First, we focus on the columns for OTL-covered rates. In 1995, most tasks have an OTL-covered rate lower than 50 (with the task *Packaging* being the exception). The average rate over the 31 items is 38. There is a notable trend from 1995 to 2000. The OTL-covered rate of the items increases between 1995 and 2000 to 58. On each item, the rates increased by approximately 20 but this increase is only significant on the items in the task *Around the Bend* (two-tailed t-test, $p<0.05$). The tasks *Folding*, *Shadows* and *Plasticine* showed low figures in both 1995 and 2000. The tasks *Dice*, *Calculator*, and *Around the Bend* had low figures in 1995, but not in 2000. Only the task *Packaging* had high figures in both 1995 and 2000.
The low number of respondents result in wide margins of the data. However, more refined methods can test whether the increase of the OTL-covered rates is significant. By means of the sign test, the items are counted on which an increase has taken place. With 31 items, the rates increased on 28 of them. This large number proves that the increase of the OTL-covered rates is significant [$Bin(31,p=½)$, $P(X\geq 28)<0.01$].

The columns on the right of Table 5.6 pertain the question whether teachers would include an item into a test of their own making. This yielded the OTL-testing rates. All OTL-testing rates are higher than the OTL-covered rates. This is reasonable, as teachers can include items, which have not been covered by their teaching, but which connect to it. Thus, they can expect *transfer* from taught content to the items. On the OTL-testing rates, again, the task *Packaging* has the highest figures. Also, there is a notable trend from 1995 to 2000. There is an

average increase from 51 to 76 on the items, and this increase is significant on the items across the tasks *Folding*, *Around the Bend*, and *Plasticine*. In particular, the two tasks *Folding* and *Plasticine* showed very low figures for 1995 (below 30), which increased considerably to above 70 in 2000.
Again, the t-test cannot establish a significant difference on the average OTL-testing rates of 1995 and 2000. By using the sign test, the items are counted on which an increase has taken place. With 31 items, the rates increased on 30 of them. This large number proves that the increase of the OTL-testing rates is significant [Bin(31,p=½), P(X≥30)<0.01].

The correlations at item level of OTL rates on the Performance Assessment of 1995 and 2000 are as follows. For the OTL-covered rates, the correlation is r=0.75 (n=31, p<0.01). For the OTL-testing rates, the correlation is r=0.64 (n=31, p<0.01). The two correlation coefficients confirm that both OTL rates increased consistently on all items. The OTL-covered rates of 1995 increased with approximately 20 and the OTL-testing rates increased with approximately 25. The larger increase of the OTL-testing rates between 1995 and 2000 was also demonstrated by the fact that there was only one task with significant differences between the OTL-covered rates of 1995 and 2000, while there were three tasks with significant differences between the OTL-testing rates. This means that Dutch mathematics teachers changed their minds on the taught content and to a larger extent, they changed their minds on the inclusion of items into an imaginary test. This trend could be caused by an increased confidence in possible transfer of content taught. It can also be that teachers have become more inclined to organise a practical test. However, the increased OTL-testing rate reflects teachers' intentions and it does not necessarily reflect actual classroom practice. As teachers have little experience with these kinds of tests, they might have given imaginary answers.

At the level of the implemented curriculum, the METRIC study had two different data available: OTL-testing and OTL-covered. As a sidestep, there is more to say about the relationship between the two kinds of data. The correlation between OTL-covered data and OTL-testing data in 1995 turns out to be r=0.84 (n=31, p<0.01). This means that the two rates approximated each other fairly. By 2000, the correlation has decreased to r=0.70 (n=31, p<0.01). In Figure 5.3, the correlations between the four data sets are illustrated.

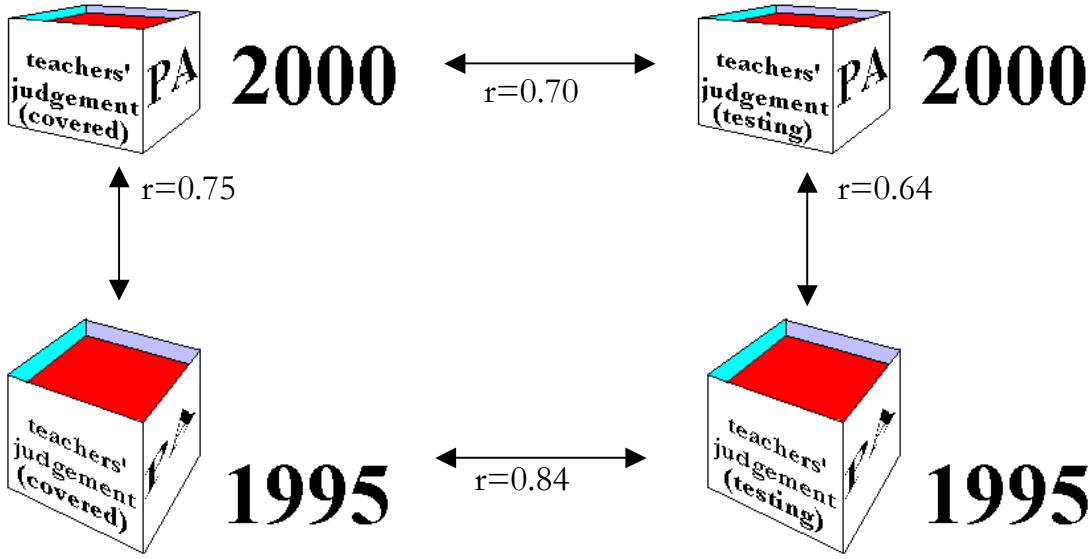


Figure 5.3: Overview of correlations on PA-1995 and PA-2000, at the level of the implemented curriculum (OTL-covered and OTL-testing)

From the correlations, we can deduct that the OTL-covered rates and the OTL-testing rates were more mutually alike in 1995 than in 2000. However, the consistency between the two measurements is still considerable. It probably has decreased because mathematics teachers indicate to an increasing rate that they could include test items into a practical test of their own making, even if the content has not been covered. This could imply that, in the five years time covered by the METRIC study, they have gained more confidence in a possible *transfer* of learnt knowledge and skills to extra-curricular content.
After this short methodological sidestep on OTL data, we return to the review of OTL data. In the next section the OTL data will be compared between the Written Test and the Performance Assessment.

5.3.3 *Comparison of trends in Dutch teachers' judgement*

In this section, we can compare the OTL rates from the Written Test and the Performance Assessment. To measure OTL, the items in both tests were submitted to the teachers. In 1995, only 16 items from the Written Test and all items in the Performance Assessment were submitted. In 1999/2000, all test items in both tests were submitted.
The instrument used differed between the tests. This difference has already occurred in 1995, and was maintained in 1999/2000, in order to keep the

instruments intact over time. For the test items in the Written Test, teachers were asked whether they would included the items into an imaginary test that was meant to cover all content taught in the past (including primary school) until the time of testing. For the Performance Assessment, the questionnaire was slightly adapted because of a presumed lower familiarity with practical tests. Here, teachers were asked whether they had covered the content (OTL-covered) and whether they would include the items into an imaginary test (OTL-testing). For the comparison of Written Test and Performance Assessment we can use both OTL rates of the Performance Assessment.

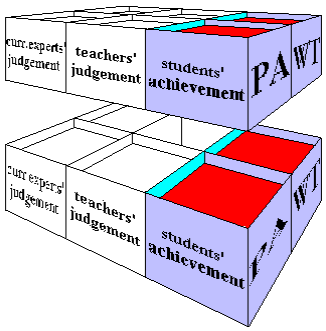
The average OTL rate for the 16 items from the Written Test was equal in 1995 and 1999, being 93. The correlation between the data sets of 1995 and 1999 was high: $r=0.70$ ($n=16$, $p<0.01$) which could have been caused by the connection between the 16 items as they were selected for the National Option Test. However, the measurements for the Performance Assessment in 1995 and 2000 display similar high correlations. The correlation of the OTL-covered rates was $r=0.75$ ($n=31$, $p<0.01$) and the correlation for the OTL-testing rates in 1995 and 2000 was $r=0.64$ ($n=31$, $p<0.01$). It means that there is a considerable consistency in teachers' answers between 1995 and 1999/2000.

The OTL rates on the Written Test and the Performance Assessment further reveal, that the Written Test matches well with the implemented curriculum. The average OTL rate on all items in the Written Test is 82% (SE=6) in 1999. The match of the Performance Assessment with the implemented curriculum is weaker but shows a noteworthy improvement between the measurements of 1995 and 2000. In 1995, the Performance Assessment received lower OTL rates than the Written Test, with average rates 38 (SE=10) and 51 (SE=11) for OTL-covered and OTL-testing respectively. By 2000, the rates had increased to an average OTL-covered rate of 58 (SE=10) and an OTL-testing rate of 76 (SE=9). With the exception of that last figure, the OTL-testing rate of 2000, all OTL data on the Performance Assessment are significantly lower that the OTL rate of the Written Test (two-tailed t-test, $p>0.05$). But in 2000, there are approximately three-quarters of the teachers considering the Performance Assessment items fit for a test of their own making. This figure of 76 on the Performance Assessment approaches the figure on the Written Test where 82 of the teachers had indicated to include an average test item into a test of their own making.

The OTL data on the Performance Assessment also disclose a large discrepancy between the OTL-covered rates and the OTL-testing rates. In 1995, the difference between these is 20, and in 2000 this difference has further increased. This means that there are an increasing number of teachers who indicate that they could include items into a test of their own making, even if the content had not (yet) been covered. This phenomenon cannot be checked for the Written Test, as the differentiation between OTL-covered and OTL-testing was not made, and only one OTL rate was compiled.

With the Written Test matching to a large extent with the implemented curriculum, and the Performance Assessment to a somewhat lesser extent, we will turn in the next section to the attained curriculum. There the students' results on the two tests will be presented.

5.4 THE ATTAINED MATHEMATICS CURRICULUM – DUTCH STUDENTS' ACHIEVEMENT RESULTS



5.4.1 *Trends in Dutch students' achievement results on the Written Test*

In this section, the Dutch students' results on the TIMSS Written Tests of 1995 and 1999 will be presented. In section 5.4.2 the same will be done for the TIMSS Performance Assessment. The students' test results are taken as an operationalisation of the attained curriculum.

The results for the attained curriculum of the TIMSS Written Test were first published in the international TIMSS reports (Beaton et al., 1996; Mullis et al., 2000). In the reports, students' scores were calculated using Item Response Theory (IRT). This theory was used to estimate students' scores on all items even if their test booklet contained only a portion of these. The average scores of students of a country were used as a country's score and re-scaled to an internationally standardised scale with an average of 500 and a standard deviation of 100 (between countries).
The laborious IRT-based method was not suitable for the METRIC study as it did not aim at finding scores over the whole test, but at finding average scores on separate items (p-values). By averaging over sets of items (e.g. the set of items

matching with the intended curriculum), links between curricular appearances could be established. As an indicator of the score on the complete test, the average p-value of all items was calculated for comparison with the IRT-based method from the international reports.
In Table 5.7, the results of both methods are displayed. The scale score for Dutch students for 1995 and 1999 was 529 and 540 respectively. These scores were above the international average of 519 in 1995 and 521 in 1999 (Mullis et al., 2000). The Dutch score seems to have improved in four years time. However, with standard errors of 6.1 and 7.1 respectively, the difference is not significant (two-tailed t-test, $p>0.05$).

Table 5.7: Dutch mathematics achievement results on WT-1995 and WT-1999

	Achievement	
Method of reporting achievement	*1995 (n=1984)*	*1999 (n=2962)*
Dutch scale score in the international TIMSS report (SE)*	529 (6.1)	540 (7.1)
Average p-value on 144 items (SE)	61.5 (3.3)	63.2 (2.7)

Source: * Mullis et al. (2000).

Table 5.7 includes the average p-values (average percentage correct) on 144 items in the Written Tests in 1995 and 1999. These data are based on the database, which was made available by the TIMSS International Study Center (Gonzalez & Miles, 2001). The full list of p-values on the 144 items is given in Appendix F. To investigate whether p-values on an item differ significantly between WT-1995 and WT-1999, the t-test is a suitable instrument. Appendix E is based on the t-test and can serve as a guideline to investigate differences in p-values.
On 144 items in WT-1995 and WT-1999, the average p-values were 61.5 and 63.2 respectively. The difference between the scores is not significant (two-tailed t-test, $p>0.05$), confirming the results from the international report, which were reported in a scale score format. In both representations (whether through a scale score or through average p-values), the minor increase is statistically undetectable using the t-test.

The METRIC study used items in WT-1995 and their clones in WT-1999 to compare data in time. However, there was concern on the well functioning of the paired items from 1995 and 1999 for the trend comparison of students'

achievement. The question was, whether the achievement on the cloned items in WT-1999 was fully comparable to the achievement of the original items in WT-1995. It remained possible that differences in achievement were not caused by students' scoring pattern but by small differences between the clones. Thus, a selection of items was made to study this issue more closely. The 16 anchor items from the National Option Test were selected. In Table 5.8 the students' scores (in p-values) are given for the 16 items. The first eight items were identical in 1995 and 1999 (items A02, B08, B10, B11, C05, E05, F12 and H07). On all these, the p-values are higher in 1999 than in 1995. For one of them (item B10), the difference between the p-values of 1995 and 1999 is significant (two-tailed t-test, $p<0.05$). On all eight items together, the average p-value has increased by 5, which is not significant (two-tailed t-test, $p>0.05$).

Table 5.8: Trends in p-values on 16 selected items from WT-1995 and WT-1999

Item		*Achievement: p-value (SE)*	
WT-1995	***WT-1999***	***1995*** *(n=239 – 1984)*	***1999*** *(n=369 - 2957)*
A02	id.	82 (6)	86 (2)
B08	id.	67 (4)	71 (3)
B10	id.	62 (4)	74 (3)*
B11	id.	80 (3)	82 (2)
C05	id.	57 (4)	60 (3)
E05	id.	66 (4)	74 (3)
F12	id.	69 (3)	75 (3)
H07	id.	81 (2)	82 (2)
Average p-value (8 identical items)		*71 (3)*	*76 (3)*
K06	clone: K06	51 (4)	75 (3)*
L08	clone: L09	69 (3)	82 (2)*
N15	clone: N15	64 (3)	52 (3)**
N19	clone: N19	64 (4)	61 (3)
O09	clone: O09	75 (3)	73 (3)
R07	clone: R08	54 (4)	57 (3)
R10	clone: R11	62 (4)	66 (3)
V02	clone: V02	24 (3)	45 (3)*
Average p-value (8 cloned items)		*58 (4)*	*64 (3)*

Note: * Significant difference with 1999 p-values being higher ($p<0.05$);
** Significant difference with 1999 p-values being lower ($p<0.05$).

The picture for the eight paired items in the bottom-half of the table is fuzzier. On one item the achievement has significantly decreased, and on three items the achievement has significantly increased (two-tailed t-test, $p>0.05$). In these cases, it was possible that the cloned item differed too much from its original, making their level of difficulty different. This resulted then in significant different scores. For example, let us look at the case of the item V02 in WT-1995, which was cloned into item V02 in WT-1999 (see Appendix D). The 1995 item shows two advertisements for rental of office space. The 1999 item shows two advertisements for acquiring a subscription to a journal. In both items the prices have to be compared for a given quantity: 110 square meters of office space, or 24 issues of the journals. This already creates a difference between the two items, as the surface of an office space in square meters is a more difficult concept than the number of issues (a whole number). Also, the amount (110 versus 24) can cause a difference in difficulty level. There is still a third difference between the two items. In the 1995 version, the two given prices have a different time scale (a price per month and a price per year). In the 1999 version, the prices have a different intercept (number of issues that are free). The fourth difference between the two items is the attractiveness of the context: the first version has the adult setting of business people, while the second version has the adolescent setting of a purchase of popular magazines. Consequently, the second item offers a context to which grade 8 students can more easily relate. Therefore, the difference between the p-values of 1995 and 1999 can well be ascribed to the conceptual differences between the clones. Thus, the comparison of the achievement on the cloned pairs between 1995 and 1999 needs care. The average p-values have risen with 6 but this might not always be caused by increased students' knowledge and skills. Therefore, it was decided that the cloned items were not to be useful for trend comparison in achievement in the METRIC study.

At this stage, having learnt that the items really needed to be truly identical for trend comparison of students' achievement, the 48 identical items in the clusters A-H were re-scrutinised. These were the items that were kept secret after 1995 and re-used in 1999. However, not all of them had been left untouched. In the preparations of the 1999 data collection in the Netherlands, small adaptations had been made to a few items (these alterations were made with permission from the TIMSS International Study Center). Two adaptations were made to connect

with the prevailing practice in Dutch mathematics education at grade 8 level. The algebraic notation for multiplying with a variable, such as '7a', was changed into '7•a'. Similarly, the word' congruent' was replaced by the phrase 'having the same size and shape'. The adaptations could affect the difficulty of the items and thus hinder the comparison of achievement at item level. The adaptations had been applied to seven items in the clusters A-H (items A05, B12, C03, D10, E02, G06, H10). Thus, *there remained 41 items that were exactly identical in 1995 and 1999*. Their number was still considered large enough to give a sound basis for analysis, although the topics covered did no longer include congruency or the multiplication of variables. However, there was still a large variety of topics covered, ranging from handling numbers (including fractions and decimal numbers), reading graphs, interpreting perspective drawings, estimating the size of angles, or dealing with the probability of events. All items had the multiple choice format.
When comparing the achievements in 1995 and 1999 at item level on the 41 identical items from the Written Test in 1995 and 1999, there turned out to be a very high correlation of $r=0.97$ ($n=41$, $p<0.01$) between the scores of 1995 and 1999. This correlation meant that items with a high p-value in 1995 had again a high p-value in 1999, and items with a low p-value had again a low p-value in 1999. The average p-value on the 41 identical items in 1995 was 72 (SE= 3) and in 1999 the average p-value was 75 (SE= 3). The averages do not differ significantly (two-tailed t-test, $p>0.05$).

A closer look at item level revealed that the increase in p-values between 1995 and 1999 occurred on almost all items, but only on two items the increase was significant. The two items are item B10 on the comparison of decimal numbers (see Appendix D) and item G05 on fractions (the item cannot be published). The increase on these items could indicate that Dutch students had improved their skills in handling numbers. When handling a 95% probability margin, it is acceptable that two out of 41 items (5%) have a significant increase in p-value.
The general statistical method to compare the data of two samples is the two-tailed t-test. This test did not reveal a significant difference between the achievements on the 41 items in 1995 and 1999. However, the slight improvement seemed to happen over the full range of items. Therefore, a more refined method was used to detect differences between the scores from 1995 and 1999. This was achieved by using the non-parametric sign test for paired samples on the items in the Written Test that were exactly equal in 1995 and 1999. These

were the 41 identical items. On four of these, the achievement had decreased, on one item it had remained equal, and on the 35 remaining items the achievement had increased. The sign test revealed that the increased achievement was small but detectable [Bin(41,p=½), P(X≥35)<0.01].
This finding was striking. The slight, though significant increase in students' achievement is an indicator of the performance of the Dutch educational system. The 41 items did not reflect the full diversity of items in the whole test, as they all had the multiple choice format and some topics were not covered (e.g. multiplication with variables, congruence). However, the 41 items contained a broad spectrum of topics, ranging from handling fractions, ordering decimal numbers, calculating probabilities, finding symmetry to interpreting graphs. They were selected on the criterion of equivalence between 1995 and 1999. Thus, the significant increase in students' achievement was observed on a broad representation of the Written Test.

From the previous findings, the following was concluded: between 1995 and 1999, Dutch students showed a slight progress on 41 identical items in the TIMSS Written Test. Although the increase in their scores was undetectable when measuring by means of Student's t-test, the use of the sign test proved that the small increase was significant. This conclusion did not contradict the analysis as reported in the international reports (cf. Table 5.7), as the sign test used in the METRIC study allowed for a more refined approach.
The analysis also showed, that the use of clones was not suitable for a trend analysis of achievement data. In a number of cases the slight differences between the paired items of 1995 and 1999 lead to large differences in students' achievement. This analysis of the functionality of the clones reduced the METRIC database.

As a result of the above finding, the functionality of the cloned items for analysing trends at the level of the intended and implemented curriculum was checked. For the latter, the OTL trend data in section 5.3.1 (e.g. Table 5.4) showed that the judgements on paired items from WT-1995 and WT-1999 were very similar and differences could not be ascribed to whether the paired items were identical or cloned. The item, on which the OTL rate changed most (item E05 with OTL rate 89 in 1995, and 74 in 1999) was identical in both years.
At the level of the intended curriculum, trends in the judgements by the curriculum experts on the appropriateness of items did not show a relation to

cloning. It was related to the adaptation of the notation for multiplying variables ('7a' in WT-1995 and '7•a' in WT-1999). This applied to six algebra items in WT-1995 (B12, D10, G06, L17, O07, P11), on which the dot-notation was applied in WT-1999, resulting in an altered judgement by the experts. The items did not match with the intended curriculum in 1995, while their clones in WT-1999, with the adapted notation, matched with the intended curriculum in 1999. The proportion of 6 out of 144 items (4%), on which the experts' judgements of 1995 and 1999 could not well be paired, was considered small. The 6 items could not account for the shift in experts' judgement between 1995 and 1999, which occurred on 33% of the items (see section 5.2.1). Consequently, the pairing of items from WT-1995 and WT-1999 through the use of comparable items (identical and cloned) remained a valid base for the trend analysis of teachers' and experts' judgement.

5.4.2 *Trends in the Dutch students' achievement results on the Performance Assessment*

In this section, the students' results on the TIMSS Performance Assessment in 1995 and 2000 will be presented. This test was not altered in-between the years and exactly the same items were submitted in both 1995 and 2000. The results on the TIMSS Performance Assessment are presented per task, and averaged over the complete set of items in the test.

According to the International TIMSS Study Center, there were twelve tasks in the Performance Assessment: five mathematics tasks, five science tasks and two combined science/mathematics tasks. In the international report, for the international comparison of students' mathematics achievement, six tasks were used: the five mathematics tasks, plus one of the two combined tasks (Harmon et al., 1997). In the previous chapter, it was explained that both combined tasks did not yield comparable and reliable data (see section 4.4.5). Therefore, in the METRIC study, the trend analysis of students' achievement between 1995 and 2000 was based on the five mathematical tasks *Dice*, *Calculator*, *Folding*, *Around the Bend*, and *Packaging*, altogether consisting of 28 items. The resulting average p-values over the items are given in Table 5.9. In 1995, the average percentage correct over the 28 items in the five tasks was 67 (SE=3). In 2000 this was 68 (SE=5). These figures do not differ significantly (two-tailed t-test, $p>0.05$).

Additionally, for each task separately, the average p-values of the items were calculated. Detailed data for the separate items across each task are given in Appendix G.

The data for the Performance Assessment were compiled through a smaller sample size than in the Written Test. This makes the precision margins of scores wider than in the Written Test. To compare the p-values between items, Appendix E can serve as a guideline.

Table 5.9 presents for each task the achievement results, expressed as average p-value for its items in 1995 and 2000. The results show that both in 1995 and 2000, the tasks *Calculator* and *Packaging* were the more challenging tasks, which resulted in lower average p-values than for the other tasks. The scores for 2000 on the five mathematics tasks showed a striking similarity to the scores for 1995. On any task, the average p-value differed less than six percent between 1995 and 2000. There was no significant difference between these (two-tailed t-test, $p<0.01$).

Table 5.9: Dutch mathematics achievement results on PA-1995 and PA-2000

	Achievement avg p-value (SE)	
Task (#items)	***1995*** *(n=437)*	***2000*** *(n=234)*
Dice (6)	77 (3)	74 (4)
Calculator (7)	62 (4)	60 (5)
Folding (4)	73 (4)	77 (5)
Around the bend (8)	68 (3)	70 (4)
Packaging (3)	52 (4)	58 (5)
Average p-value on 28 items	*67 (3)*	*68 (5)*

When correlating the p-values of 1995 and 2000 on the Performance Assessment at item level, we found $r=0.96$ ($n=28$, $p<0.01$). This correlation is high (Krathwohl, 1998), meaning that the p-values in 1995 and 2000 followed the same pattern on the items. This correlation had a similar magnitude as the correlation of $r=0.97$, found between the p-values of 1995 and 1999 on the 41 identical items from the Written Test.

With the minimal increase between the two averages, it was tested by the Sign test whether this was significant. The same had been done for the Written Test. With 28 items in the Performance Assessment, there was an increase on 14 items, stability on one item and a decrease on 13 items. This gave evidence that there was no significant difference in the achievement of 1995 and 2000

[Bin(28,p=½), P(X≥14)>0.05]. Thus, there was a null-trend in the achievement of Dutch students on the TIMSS Performance Assessment.

5.4.3 *Comparison of trends in Dutch students' achievement results*

In this section the achievements of Dutch students between 1995 and 1999/2000 on the TIMSS Written Test and the TIMSS Performance Assessment are compared. The two tests revealed a high level of consistency for the attained curriculum. On the Written Test, the Dutch students' results increased slightly on the 41 items that were exactly identical in WT-1995 and WT-1999. The increase could not be perceived as significant when tested through the t-test, but it showed to be significant when tested through the sign test for two paired samples. On the Performance Assessment, Dutch students' results did not increase significantly between 1995 and 2000.

For both tests, the results of 1995 correlated highly with the results of 1999/2000. Trend correlation coefficients were r=0.97 for the Written Test and r=0.96 for the Performance Assessment. These correlation coefficients were much higher than the trend correlation coefficients of the OTL data (for the implemented curriculum). These were r=0.70 on 16 items from the Written Test, and on all items in the Performance Assessment r=0.75 on 'OTL-covered' data and r=0.64 on the 'OTL-testing' data. The trend correlation of the item-curriculum matching indices (for the intended curriculum) were even lower, with r=0.19 on the Written Test and r=0.41 on the Performance Assessment.

In Tables 5.10a and 5.10b, the trend results, as described in this current chapter, are summarised. Besides the results at the level of the attained curriculum, the descriptive statistics at the other two curricular appearances are given in Table 5.10a. Trend correlation coefficients between data of 1995 and 1999/2000 are added in Table 5.10b. The coefficients are highest at the level of the attained curriculum, showing the high stability of students' achievement. The coefficients are lower for the OTL rates, and lowest for the item-curriculum matching indices.

Table 5.10a: Descriptive statistics of trend results in the METRIC study

Sub-study	*Intended curriculum Test-curriculum matching index*	*Implemented curriculum Avg OTL rate*	*Attained curriculum Achievement Avg p-value (SE)*
WT-1995	69 (nom.) -- (rat.)	--	72 (3)
WT-1999	71 (nom.) -- (rat.)	82	75 (3)
PA-1995	88 (nom.) 83 (rat.)	38 (cov.) 51 (tst.)	67 (3)
PA-2000	85 (nom.) 72 (rat.)	58 (cov.) 76 (tst.)	68 (3)

Note: Dashes indicate data are unavailable;
nom.=on a nominal scale; rat.=on a ratio scale; cov.=OTL-covered; tst.=OTL-testing.

Table 5.10b: Trend correlations in the METRIC study

Paired sub-studies	*Intended curriculum Correlation between item-curriculum matching indices*	*Implemented curriculum Correlation between OTL rates*	*Attained curriculum Correlation between p-values*
WT-1995 – WT-1999	0.19	--	0.97
PA-1995 – PA-2000	0.41	0.75 (cov.) 0.64 (tst.)	0.96

Note: Dashes indicate data are unavailable;
cov.=OTL-covered; tst.=OTL-testing.

The analysis of students' achievement results revealed that the METRIC database for the attained curriculum had to be reduced from 144 to 41 items from the Written Test. The small differences between items in 1995 and their clones in 1999 possibly induced differences in achievement. But the full set of 144 items was kept intact for the analysis of data from the intended and implemented curriculum. These data proved to be useful for the METRIC study to answer the research questions, which were posed in chapter 1. In the next chapter, the data will be used to search for replies to these questions.

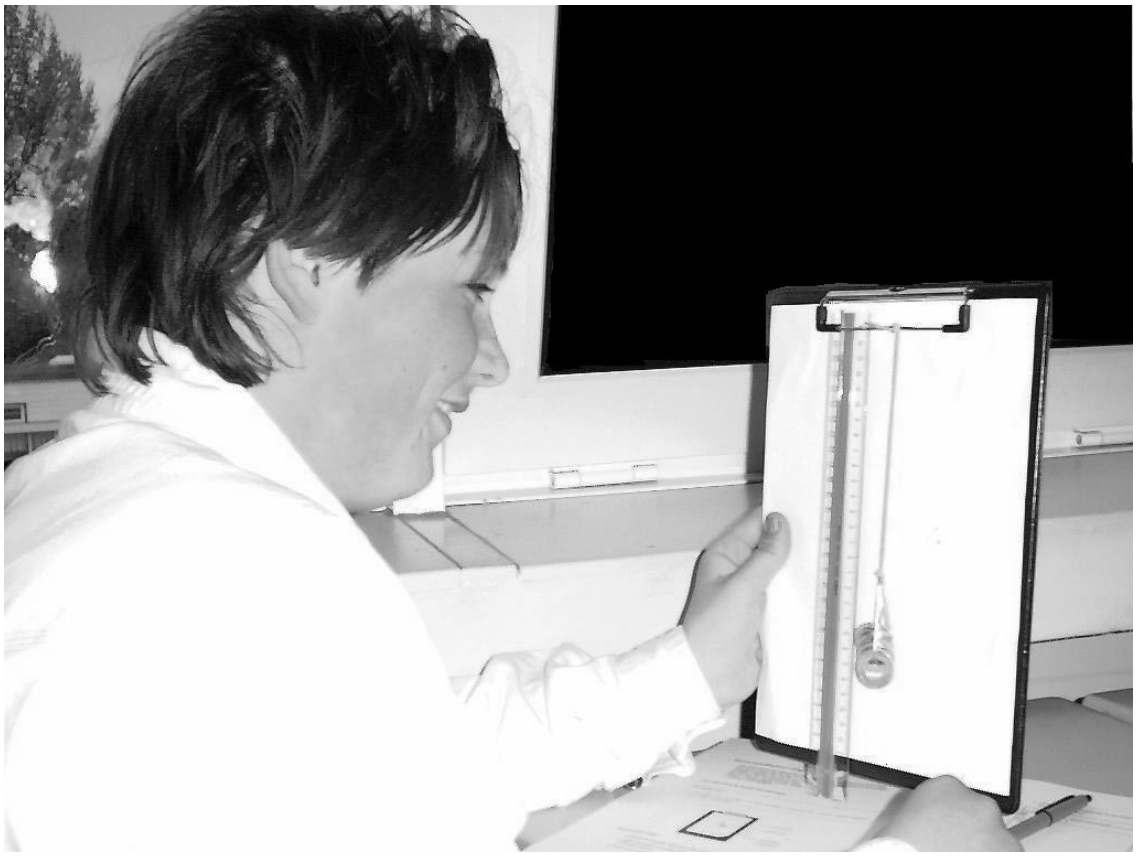


Figure 5.4: Student working on the task *Rubber Band* from the TIMSS Performance Assessment

Chapter 

Research results

~ Es fällt kein Meister vom Himmel.~

Masters don't come out of the blue.
*(*GERMAN PROVERB*)*

This chapter presents answers to the research questions. Section 6.1 introduces the research questions and summarises how the previous chapters have prepared for finding answers. In the following section (section 6.2), the first question, pertaining the inter-test achievement discrepancy is answered. In the subsequent section (section 6.3), the second question is answered (as far as possible). In the last section, the conclusions of the METRIC study are drawn.

6.1 INTRODUCTION

The METRIC study was initiated to find answers explaining the discrepancies, which were observed after the TIMSS Written Test and the TIMSS Performance Assessment were administered in the Netherlands in 1995. The first discrepancy pertained the achievement results: on the Written Test, Dutch students' achievement score differed significantly above the international average score, while on the Performance Assessment, their score was near the international average (see section 1.2.2). This discrepancy was termed the inter-test achievement discrepancy. To explain this discrepancy, the two tests were replicated in order to investigate whether the achievement results had been well measured and would re-occur. Therefore, the Written Test was repeated in 1999, and the Performance Assessment was repeated in 2000.

An additional reason for replicating the studies after a few years was the curriculum reform, which was legislated in 1993. The newly introduced RME-based mathematics curriculum for junior secondary schools focused more on learning mathematics that would be useful and authentic to students (see section 2.3). The first cohort learning through the new curriculum contained the students tested in TIMSS in 1995. It was assumed that the introduction of the new curriculum created insecurity among teachers, which could have an effect on students' achievement results, especially on those from the Performance Assessment. The context of the recent curriculum reform asked for a replication of the two tests after a few years, giving teachers ample time to adjust to the new curriculum.

The judgement of the two tests in light of the new, intended curriculum lead to the observation of another discrepancy: the intra-curricular discrepancy. Curriculum experts had been asked to judge the tests on their appropriateness in light of the intended curriculum. The judgements contrasted with students' achievement. In light of the new RME-based curriculum, the Written Test was not considered very appropriate to test for students' achievement, but students' achievement was relatively high. On the other hand, the Performance Assessment aligned better with the intended curriculum, but students' achievement was relatively lower than on the Written Test. Thus, the intra-curricular discrepancy was observed on both tests, but in a reversed arrangement.

The METRIC study aimed at compiling data on the appropriateness of the tests, as a context for explaining students' achievements in 1995. Therefore, not only judgements from curriculum experts were gathered, but additionally from the mathematics teachers of the tested students as well. The instruments have been described in the sections 4.3.4 and 4.3.5. The data of these measurements were obtained from available data from 1995, and repeat measurements were integrated into the replication of the achievement tests in 1999/2000.

The previous chapter presented the trend results of all data gathered. These were needed, to position the discrepancies. The current chapter deals with the discrepancies.

6.2 THE INTER-TEST ACHIEVEMENT DISCREPANCY

This section focuses on the first research question the METRIC study, which was based on the finding that Dutch students' achievement seemed relatively better in the international comparison of the Written Test than of the

Performance Assessment. This discrepancy was named the inter-test achievement discrepancy. The first research question was:

> *To what extent does a repeat of the TIMSS Written Test in 1999 and a repeat of the TIMSS Performance Assessment in 2000, result in an inter-test achievement discrepancy similar to 1995?*

The data set for this research question pertained to all data, which were gathered at the level of the attained curriculum, as illustrated in Figure 6.1.

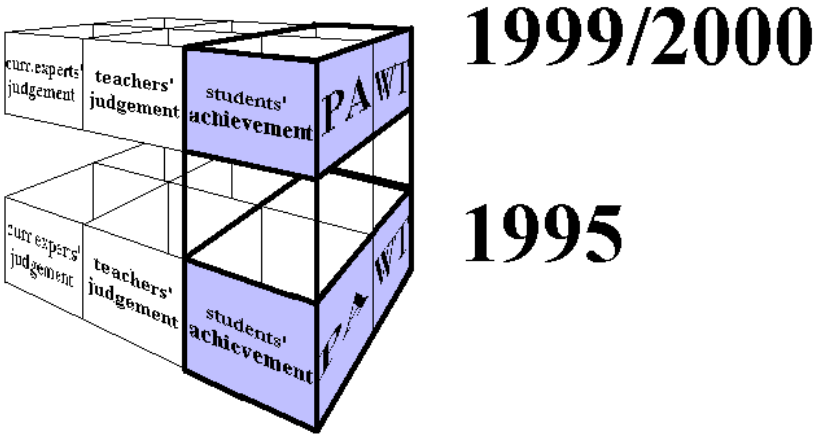


Figure 6.1: *Data set for the first research question*

In the sections 5.4.1 and 5.4.2, it has already been described how the results of the attained curriculum as measured through both the TIMSS Written *Test and the TIMSS Performance Assessment showed stability over time. Dutch students' achievement improved slightly on the* Written *Test (significantly when measured by the sign test) between 1995 and 1999.* However, *their achievement did not change on the Performance Assessment.*
The research question, as quoted above, refers to the inter-test achievement discrepancy. This is the discrepancy, which was observed in 1995 between the international results of the Dutch students on the two tests. The achievements on the Written *Test were perceived as being higher than on the Performance Assessment.* When *comparing internationally, Dutch students reached just below the top-scoring countries on the* Written *Test while they were near the international average on the Performance Assessment. In Table 1.2 in chapter 1 the Dutch score and ranking was shown on both tests*: *in the column of the* Written *Test the* Netherlands *ranked much higher than in the column of the Performance Assessment.*

Both tests were repeated in order to find out whether the achievement on the Performance Assessment would improve. With *stable results over time, it looks as if this inter-test achievement discrepancy remained intact.* However, *in 1995 the results on the TIMSS Performance Assessment were described as disappointing because of the median ranking in the international league table (Bos et al., 2001). This observation focused on the position in the ranking of the countries. This position of countries was based on the scores on five mathematics tasks plus the combined mathematics/science task Plasticine (*Harmon *et al., 1997).* However, *when looking at the country's scores, instead of looking at the rankings, the situation for the Dutch students was less dramatic. The Dutch students' score was not significantly different from the scores of a large group of other countries.*

In Figure 6.2, the data of Table 1.2 are shown again, but visualised in two scatter diagrams. In both diagrams, the horizontal axis holds the country's results in 1995 on the Written Test. *The vertical holds the country's results in 1995 on the Performance Assessment. In the diagram on the left, the results are expressed as rankings. A diagonal line has been added, indicating an equal position on both tests. Countries above the diagonal ranked higher on the Performance Assessment than on the* Written Test; *countries below the line ranked higher on the* Written Test *than on the Performance Assessment. In the diagram on the right, the results are expressed as average scores. Two vast lines were added to indicate the international average score on either test.*

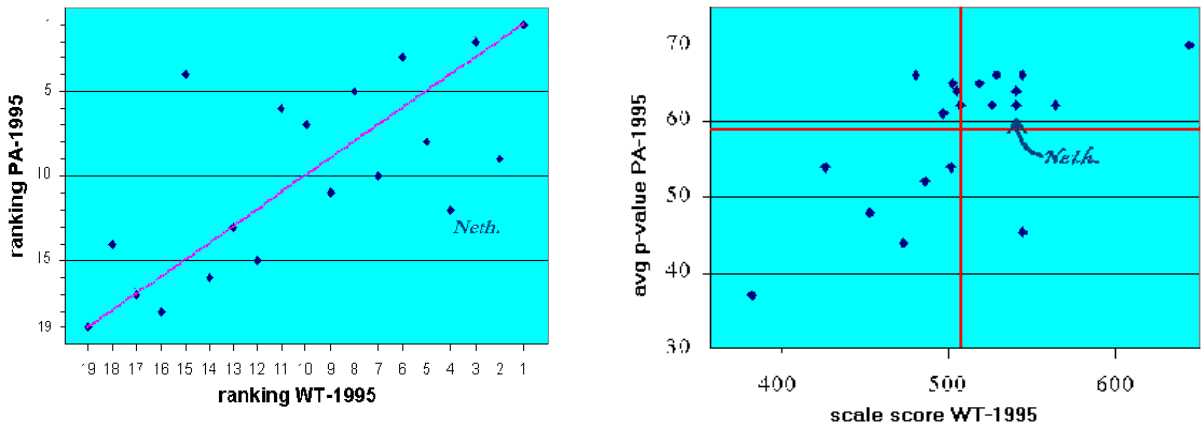


Figure 6.2: Scatter diagrams of mathematics achievement results of countries on WT *1995 and PA1995*
(on the left: according to ranking; on the right: according to score)

Each point in Figure 6.2 indicates a country. The identity of a country can be looked up from the co-ordinates in Table 1.2. The position of the Netherlands *among the 19 countries is marked as* **Neth.** *Its position in the left diagram is 'far' from the diagonal, and seems to be exceptional.* However, *the position of the* Netherlands *in the right diagram turns out to be among a group of other countries, whose position is above the international average score on both tests. Thus, by looking at the scores of countries, instead of the rankings, the picture of the inter-test discrepancy between the Dutch students' scores on the* Written Test *and the Performance Assessment was nuanced.*

Additionally, the scores on the Performance Assessment needed a re-calculation, as the results on some of the task had been dubious. In section 4.4.5, it was explained that the results of the task Plasticine were dubious. It was one of the tasks, of which the students' results in 1995 and 2000 were considered unreliable or incomparable. Therefore, the results of these tasks were skipped from the analysis of students' achievement.
When recalculating the scores based on the remaining five mathematics tasks in the Performance Assessment (Dice, Calculator, Folding, Around the Bend and Packaging), the 'disappointing' test results changed totally. Particularly the results of the Plasticine task had been pulling down the overall Dutch mathematics results. Omission of this task raised the Dutch score, and as a consequence, its position in the international comparison. Table 6.1 shows the achievement results of Dutch students in 1995 internationally, when comparing between all countries that participated in both the Written Test *and the Performance Assessment for grade 8. The first column 'TIMSS* Written Test *1995' shows the mathematics scores on the* Written Test, *based on the international scale score as given in Beaton et al. (1996). The second column 'Performance Assessment 1995 (international report)' shows the mathematics results on the Performance Assessment, expressed as average p-values, and based on the six tasks (calculated through the average percentage correct) as was reported in* Harmon et al. *(1997). The third column shows the same results, but now based on the five mathematics tasks.*

Table 6.1: Mathematics achievement results of countries participating in both WT 1995 and PA1995 (before and after re-calculation)

TIMSS Written Test 1995		Performance Assessment 1995 (international report)		Performance Assessment 1995 (five mathematics tasks)	
	Country	Scale points*	Country	Avg p-value	Country	Avg p-value**
1	Singapore	643	1 Singapore	70	1 Singapore	71
2	Czech Rep	564	2 Australia	66	2 Romania	67
3	Switzerland	545	3 Romania	66	3 England	65
4	Netherlands	541	4 Switzerland	66	4 Netherlands	65
5	Slovenia	541	5 Norway	65	5 Norway	65
6	Australia	530	6 Sweden	65	6 Australia	65
7	Canada	527	7 Slovenia	64	7 Switzerland	64
8	Sweden	519	8 England	64	8 Slovenia	64
	Intl average	509	9 Czech Rep	62	9 Sweden	63
9	New Zealand	508	10 Canada	62	10 New Zealand	61
10	England	506	11 New Zealand	62	11 Scotland	61
11	Norway	503	12 Netherlands	62	12 Canada	61
12	USA	502	13 Scotland	61	13 Czech Rep	61
13	Scotland	498	Intl average	59	Intl average	59
14	Spain	487	14 USA	54	14 USA	54
15	Romania	482	15 Iran	54	15 Spain	54
16	Cyprus	474	16 Spain	52	16 Portugal	50
17	Portugal	454	17 Portugal	48	17 Iran	48
18	Iran	428	18 Cyprus	44	18 Cyprus	42
19	Colombia	385	19 Colombia	37	19 Colombia	36

Note: * The figures are based on Beaton et al. (1996). The difference with the scores in Table 5.12 is caused by the re-calculation of scores for trend comparison (Mullis et al., 2000);

** The difference with the results in Table 5.14 is caused by the difference in calculation method. Table 6.1 has average p-values across tasks, which was the method used in the international TIMSS report (Harmon et al., 1997), while Table 5.14 has average p-values across items.

When comparing the two columns for the Performance Assessment, the positions of most countries are not considerably different. Most countries stay approximately at the same position, rising or falling just one, two or three positions. The change in ranking is largest for the Netherlands, *with a leap of 8 positions upwards.*

Not the rankings, but the scores tell the story of the tests. Rankings can create apparent differences between countries, while their scores do not show significant differences. Therefore, to eliminate misunderstanding emerging from

the rankings, the International TIMSS Study Center issues tables for multiple comparison in their international reports (Beaton et al., 1996; Mullis et al., 2000). The comparison tables take into account the inaccuracy of the scores. On the Written Test of 1995, there were many countries that scored at a comparable level as the Netherlands. Based on these tables, the Dutch score on the Written Test in 1995 did neither differ significantly from the score of the Czech Republic (on ranking 2) nor from Sweden (ranking 8). Therefore, in the left column of Table 6.1, there are six countries with a comparable score as the Netherlands: the Czech Republic, Switzerland, Slovenia, Australia, Canada, and Sweden. The Netherlands is just one country in the large group of countries that scored above the international average and below the top-performing countries.

The same applies to the Performance Assessment. With the smaller samples, the precision margins of scores are wider. Therefore, the difference in scores between countries is even more difficult to make. The international TIMSS report on the Performance Assessment did not undertake the audacious enterprise to create a table for multiple comparisons between countries, as the data were not considered reliable enough.
Harmon et al. (1997) reports standard deviations of magnitude ±2 for country's scores on the Performance Assessment. Thus, by rule of thumb, the scores of countries with rankings 2 to 13 have approximately a comparable score as the Netherlands. The Netherlands is just one of the group of countries that scored above the international average but below top-performing Singapore. It can, therefore, be concluded that the international position of the Netherlands on the Written Test in 1995 was of comparable magnitude as the position on the Performance Assessment. This nuanced the inter-test discrepancy to a large extent.

For 2000, it is not possible to compare the Dutch scores on the Performance Assessment with other countries, as the Netherlands was the only country that repeated this sub-study. However, a relative international position can be extrapolated by reasoning:

1. *The repeat of the TIMSS Written Test has shown that in four years time, the average students' achievement of most nations only shows a marginal change (Mullis et al., 2000). Assuming that the same applies to the Performance Assessment with a five-year time span, it means that between 1995 and 2000 the average students' achievement of most nations on the TIMSS Performance Assessment will only show a minimal change.*

2. *In five years time, between 1995 and 2000, the scores of Dutch students on the Performance Assessment did not change significantly (see section 5.4.2).*

Considering the above, it can be extrapolated that if *the TIMSS Performance Assessment had been carried out in a similar international setting as in 1995, the Dutch relative score would be comparable to the score of 1995. Then, the Netherlands would rank among the countries above the international average, but below the top-scoring countries.*

This leads to the following conclusions pertaining to the first research question.

1. *Dutch students' achievement on the TIMSS Written Test improved slightly (significantly when measured by the sign test, see section 5.4.1) between 1995 and 1999. When comparing internationally, the average score of Dutch students on the TIMSS Written Test remained stable between 1995 and 1999.*
2. *Dutch students' achievement on the TIMSS Performance Assessment remained stable between 1995 and 2000. As a result, the international comparative score of Dutch students on the TIMSS Performance Assessment would probably have remained stable.*
3. *The perception of an inter-test achievement discrepancy in 1995 was influenced more by the country rankings and less by the country's scores. A closer look at the scores revealed that the Dutch scores in 1995 on the Written Test couldn't be perceived as "better" (as stated in Bos et al., 2001, p. 90) than the scores on the Performance Assessment. Considering measurement margins and reliability of students' results, both scores have a comparable international level.*

It can therefore be concluded, that if ever *there was an inter-test achievement discrepancy, it remained the same from 1995 to 1999/2000.*

6.3 RELATING STUDENTS' ACHIEVEMENT TO THE INTENDED CURRICULUM

6.3.1 Introduction

The second research question of the METRIC study pertained the intra-curricular discrepancies, which were the discrepancies between students' achievement results and judgements by curriculum experts and mathematics teachers on the appropriateness of the tests in light of the intended and

implemented curriculum, respectively. The second research question was formulated as follows:

> *To what extent are students' results on both tests at both periods in time aligned with (a) the appropriateness of the tests in light of the intended curriculum, (b) the appropriateness of the tests in light of the implemented curriculum and (c) possible discrepancies between these?*

In fact, the second research question consists of three sub-questions. First, the achievement of students as an operationalisation of the attained curriculum will be related to the intended curriculum. Thereafter, the achievement will be related to the implemented curriculum. Finally, the attained curriculum will be related to a possible discrepancy between the intended and the implemented curriculum. For all three sub-questions, the results on the Written Test *and the Performance Assessment will be dealt with in separate sections, before preliminary conclusions can be drawn.*
In all sections, the achievement of students will be analysed in light of the intended or implemented curriculum. This will be done is several ways: (1) by grouping the items according to their match with the intended or implemented curriculum and then look at the p-values of students on the disjoint item sets; (2) by correlating the p-values to the item-curriculum matching indices or to the OTL rates. In the end, final conclusions will be drawn, based on all findings described before.

6.3.2 Relating trends in achievement on the Written Test to the intended curriculum

In this section, Dutch students' achievement on the Written Tests *of 1995 and 1999 will be dealt with, in light of the intended curriculum.*
In section 5.2.1, it was observed that in 1995 and 1999 a considerable number of items in the Written Test *(approximately 70%) matched with the Dutch intended mathematics curriculum. It was also observed, that the judgements by the curriculum experts of 1995 and 1999 yielded a shift on approximately one third of the test items. To measure a trend, the 41 identical items from the* Written Test *will be used. First, the achievement of students will be separated according to both measurements for the intended curriculum. In Tables 6.2a and 6.2b the achievement of 1995 and 1999 on the* Written Test *are tabulated again (cf. Table 5.1). A distinction is made by grouping the items according to whether they*

match with the intended curriculum as measured in 1995 and 1999. There are four item groups based on the intended curricula of 1995 and 1999 (the grouping of items has been presented in section 5.2.1):

- *items that match with the intended curriculum of both 1995 and 1999 (part A),*
- *items that match with the intended curriculum of 1995, but not with that of 1999 (part B),*
- *items that do not match with the intended curriculum of 1995, but that match in 1999 (part C), and*
- *items that match neither with the intended curriculum of 1995 nor with that of 1999 (part D).*

In Table 6.2a, the average p-values of items is given for the full set of 41 items, and additionally for each of the four parts. The average p-value is taken as the indicator of students' achievement. Dutch students' achievement did not change significantly between 1995 and 1999 (two-tailed t-test, p>0.05). However, *with the parameter-free sign test for paired samples, the small improvement on the full set of 41 items, and the small improvement on part A were significant* [*Bin(41,p=½), P(X≥35)<0.01*; *Bin(28, p=½), P(X≥25)<0.01*]. *The sign test did not detect significant differences between the scores of 1995 and 1999 on the parts B, C and D.*

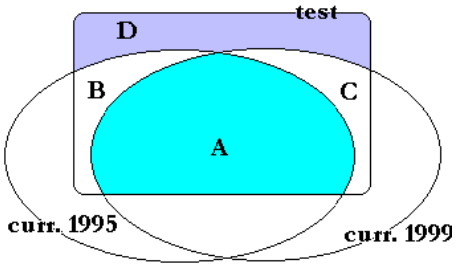


Table 6.2a: *Achievement results on 41 identical items from* WT-*1995 and* WT-*1999, related to the intended curriculum*

Group of items	Achievement Avg p-value (SE) WT-1995 *(n=239-1984)*	WT-1999 *(n=369-2957)*
Identical items in WT-*1995 and* WT-*1999 (n=41)*	*72* (3)	*75* (3)*
A. Items matching with both intended-95 and intended-99 (n=28)	*72* (3)	*76* (3)*
B. Items matching with intended-95, but not matching intended-99 (n=4)	*61* (3)	*64* (4)
C. Items not matching with intended-95, but matching intended-99 (n=7)	*76* (3)	*80* (3)
D. Items matching with neither intended-95 nor intended-99 (n=2)	*72* (3)	*76* (3)

Note: * Significant difference between 1995 and 1999 (sign test, p<0.01).

A noteworthy fact that emerged from Table 6.2a was that the items in part B seemed to be the most 'difficult'. Both in 1995 and 1999, the average p-value was lowest on this part. The highest scores were achieved on the items in part C. Thus, it looked as if the curriculum experts in 1995 had selected 'more difficult' items, while the curriculum experts in 1999 had selected 'easier' items.
This particular observation was possibly caused by the fact that the 41 items were sieved out of the test, after deleting the cloned items. This selection of 41 items contained only multiple choice items, and no items on algebra or congruency. They were not a representative selection of the complete Written Test. *Therefore, a similar table as Table 6.2a was generated, to check for the distribution of p-values on the full item set of 144 items (cloned and identical). Thus, Table 6.2b has the same format as Table 6.2a. The average p-values in this table were calculated over the identical and the cloned items together. As a result, this table is not suitable for comparison of achievements between 1995 and 1999, because differences in the achievements can be caused by the differences between the clones. The table just serves the purpose of checking whether the parts B and C contained items that could be competed correctly by different numbers of students.*

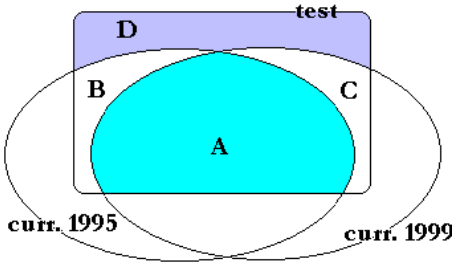


Table 6.2b: *Achievement results on 144 items in* WT-*1995 and* WT-*1999 (identical and cloned items), related to the intended curriculum*

Group of items	Achievement Avg p-value (SE) WT-1995 *(n=239-1984)*	WT-1999 *(n=369-2957)*
Overlapping items in WT-*1995 and* WT-*1999 (n=144)*	*62* (3)	*63* (3)
A. Items matching with both intended-95 and intended-99 (n=76)	*63* (3)	*66* (3)
B. Items matching with intended-95, but not matching intended-99 (n=23)	*60* (3)	*58* (4)
C. Items not matching with intended-95, but matching intended-99 (n=26)	*66* (3)	*68* (3)
D. Items matching with neither intended-95 nor intended-99 (n=19)	*52* (4)	*51* (4)

Table 6.2b shows that the items in part D were the most 'difficult'. Both in 1995 and 1999, the average students' achievement is lowest on these 19 items. These are the items that were considered as not matching with the intended mathematics curriculum, neither in 1995 nor in 1999. It is therefore noteworthy, that still more than 50% of the Dutch students were able to complete these correctly. This notable achievement on these particular 19 items explains why there is only a small difference in achievement between items that match and those that do not match with the intended curriculum. As a result, Dutch students' achievement on the complete Written Test *does not diverge completely from their achievement on only the items that matched with their curriculum. This observation was also made in the TIMSS Test Curriculum Matching Analysis (Beaton, 1998; Beaton et al., 1996).*

H*owever, it remains to be noted that the instrument for the judgement by the experts defined the intended curriculum as the curriculum for more than 50% of the students. For the* N*etherlands, this was operationalised as the mavo/havo-level. Consequently, the items considered as not matching with this operationalisation of the intended curriculum could well have been included in an intended curriculum of higher ability tracks.* U*nfortunately, it is beyond the scope of the METRIC study to find out whether the average of 52% of the students in 1999 scoring on the items in part D (see Table 6.2b) were indeed in higher ability tracks than mavo/havo.*

Other possible explanations for the average p-value of above 50 on items not matching with the intended curriculum could be: (1) students had gained knowledge and skills needed for these items in other subjects than mathematics, or out of school, (2) students were able to make a transfer from their existing knowledge and skills to complete the unfamiliar items, or (3) teachers had taught the content of the items, although this was not required.

Additionally, Table 6.2b yields information on the p-values for part B and C. The items in part C seem to be 'easiest' as this part has the largest average p-values. This confirms the findings from Table 6.2a. Moreover, just like in Table 6.2a, the items in part B yield lower average p-values (both in 1995 and 1999) than the items in part C. Thus, the items matching with the intended curriculum in 1999 (part A and C together) can be completed by a larger number of students than the items matching with the intended curriculum in 1995 (part A and B together).

The above analyses were based on Table 6.2b, which was not suitable for a trend analysis of achievement. The use of clones was not considered suitable, as small differences in items could yield large differences in achievement results. Therefore, Table 6.2a was generated based on 41 identical items. The table yields the following observations:

1. *(comparing the columns) The achievement of students increased slightly but significantly (using the sign test) between 1995 and 1999 on all 41 identical items from the* Written Test. *The achievement of students increased slightly but significantly (using the sign test) between 1995 and 1999 on the items that matched with the intended curriculum both in 1995 and 1999 (part A).*
2. *(comparing the rows) The achievement of students (both in 1995 and 1999) is lower on the items that match with the intended curriculum in 1995 and not in 1999 (part B). The achievement of students (both in 1995 and 1999) is higher on the items that match with the intended curriculum in 1999 and not in 1995 (part C).*

As a result, the average p-value in 1995 on the items matching with the intended curriculum in 1995 (part A and B together) is 71, and on the complementary set of items (part C and D) the average p-value is 75. This looks very contradictory, as students are expected to do 'better' on the items that match with their intended curriculum. However, *the contradiction is largely caused by the choice of items in part* B.
The above-observed contradiction does not hold for 1999. The average p-value in 1999 on the items matching with the intended curriculum in 1999 (part A and C together) is 77, and on the complementary set of items (part B and D), the average p-value is 68.

From the above, a preliminary conclusion can be drawn. It is a conclusion that needs caution, as it is only based on the 41 identical items from the Written Test *in 1995 and 1999. In four years time, students' achievement has increased, particularly on items matching with the intended curriculum of 1999. This increase is partly due to the slight but significant increase in students' achievement (their increased score on part A), but it is also due to the changing judgement on the appropriateness of the test by the curriculum experts (the difference between part B and C). Therefore, students' achievement of 1999 is more aligned with the intended curriculum of 1999 than it is the case for 1995.*

A second way of linking the attained curriculum to the intended curriculum is by calculating a correlation index between item-curriculum matching indices and p-values. In 1995, this correlation is r=0.04 (Kendall's tau & Spearman's rho, n=144, p=0.62), which is not significant. By 1999, this correlation has changed to a significant r=0.23 (Kendall's tau & Spearman's rho, n=144, p<0.01).
In Figure 6.1, two trend correlation coefficients are added: one is between experts' judgement of 1995 and 1999, the other is between students' achievement of 1995 and 1999. Between the data from the four measurements, only three correlations are significant. Therefore, no arrow is drawn between the experts' judgement of 1995 and students' achievement of 1995.

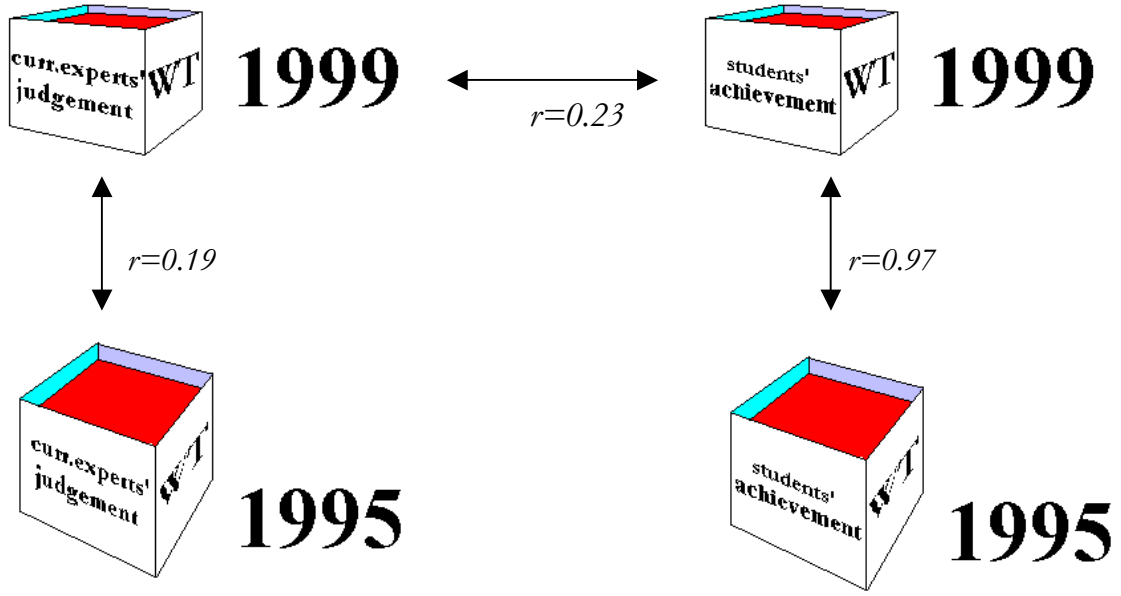


Figure 6.3: Overview of correlations on the Written Test, 1995-1999, *at the level of the intended and attained curriculum*

The picture arising from the correlation coefficients confirms the preliminary conclusions from a different perspective: the high correlation of r=0.97 (n=41, p<0.01) between students' achievements of 1995 and 1999 means that items with a high p-value in 1995 had again a high p-value in 1999. Similarly, items with a low p-value in 1995 had again a low p-value in 1999.
The lack of a significant correlation between the p-values in 1995 and the judgements by the experts in that same year mean that there was no coherence

between these. Students performed well on some items that were not considered to match with the intended curriculum, and conversely, students did not perform well on items that were considered fit.
The correlation between the curriculum experts' judgements of 1995 and 1999 was r=0.19 (see chapter 5, section 5.2.1; both through Kendall's tau and Spearman's rho method, n=144, p=0.02). It means that the two judgements were consistent, although the judgement changed on a considerable number of items (in fact, on one third of the items, see section 5.2.1). The change in the judgement by the experts between 1995 and 1999 resulted in a changing correlation with the p-values. While *in 1995 the correlation was absent, in 1999 it became r=0.23. This means that, generally, in 1999 students did well on items that were considered to match with the intended curriculum, and students did less well on items that did not match with the intended curriculum. The correlation is not high, which means that on a considerable number of items, the judgements and the achievements were contradictory.* However, *the correlation, which had been totally absent in 1995, is not repeated. As a result, in 1999 students' achievement and curriculum experts' judgement on the appropriateness of the* Written Test *have become more aligned in the time span of four years.*

6.3.3 Relating trends in achievement on the Performance Assessment to the intended curriculum

In this section the attained curriculum is linked to the intended curriculum with regard to the TIMSS Performance Assessment of 1995 and 2000. This will first be carried out at task level. For comparison with the data from the Written Test, *additionally, the exercise will be carried out to partition the test into parts, which are either covered or not covered by the intended curricula.* However, *this is based on the transformation of data from a ratio scale into a yes/no scale. The availability of data on a ratio scale offers firm ground to, finally, link intended and attained curriculum through the calculation of correlation coefficient.*

First, trends in average item-curriculum matching indices and students' achievement are presented per task from the Performance Assessment. See Table 6.3.

Table 6.3: *Trends in average item-curriculum matching indices (average percentage of experts) and achievement results on PA1995 and PA2000*

Task	Average item-curr. matching index		Achievement Avg. p-value (SE)	
(#items in experts' instrument/ #items in test)	1995 *(n=3)*	2000 *(n=5)*	1995 *(n=437)*	2000 *(n=234)*
Dice (5/6)	*87*	*92*	*77* (3)	*74* (4)
Calculator (6/7)	*72*	*70*	*62* (4)	*60* (5)
Folding (4/4)	*100*	*60*	*73* (4)	*77* (5)
Around the Bend (6/8)	*100*	*83*	*68* (3)	*70* (4)
Packaging (3/3)	*89*	*73*	*52* (4)	*58* (5)
Shadows (6/6)	*61*	*67*	--	--
Plasticine (3/8)	*78*	*40*	--	--

Note: Dashes indicate the achievement results were deemed unreliable.

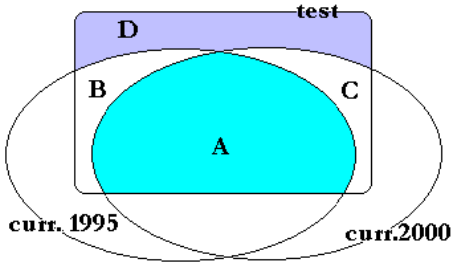
The results in Table 6.3 do not show a clear pattern between intended and attained curriculum. There is a task with a high average item-curriculum matching index (Dice), which has high p-values. But there is also a task with a high average item-curriculum matching index (Around the Bend), which has average p-values. Similarly, there is a task with a lower curriculum matching index (Calculator), which has lower p-values. But there is also a task with lower curriculum matching index (Folding, in particular the rate in 2000), which has above average p-values. This does not give a clear picture.

Another method of linking intended curriculum to attained curriculum has been tried with the Written Test *already. This method consisted of partitioning the test into four parts (A, B, C and D), depending on the judgement by the curriculum experts.* However, *the partitioning of the Performance Assessment is somewhat different from that of the* Written Test. *In the Performance Assessment, part D (which is not covered in 1995 and 2000) is empty. The parts B and C had nine items, but only four of these had reliable p-values. As described in chapter 4, the p-values on some tasks were discarded because of comparability issues (see section 4.4.6). Table 6.4 has the same format as Table 6.2a.*
In 1995, the average p-value on any item of the TIMSS Performance Assessment was 67. This figure is not significantly different from the average p-value of 68 in

2000. When looking at the items that match with the intended curriculum as measured in 1995 and 2000 (part A), the achievement is higher than on the full set of items, with average p-values of 74 in 1995 and 75 in 2000. This means that part A contains the items from the test that many students could complete well. It means that the curriculum experts considered those items to match well with the intended curriculum that could be completed by many students. The items that were not considered relevant to the intended curriculum were obviously 'difficult', as many students could not score on these.

The average p-values on parts B and C have been incorporated into Table 6.4 for the sake of completeness and comparability with Table 6.2a. However, the small number of items in these parts does not allow for generalisation. As a result, in this case it is not possible to make a link between the intended curriculum and the attained curriculum through partitioning as the partitioning is not successful: there are too few items in the Performance Assessment that do not match with the intended curriculum and have reliable data simultaneously.

Table 6.4: Achievement results on PA1995 and PA-2000, related to the intended curriculum



Group of items	Achievement Avg p-value (SE) PA-1995 *(n=437)*	PA-2000 *(n=243)*
Overlapping items in PA1995 and PA2000 (n=28)	*67* (3)	*68* (5)
A. Items matching with both intended-1995 and intended-2000 (n=24)	*74* (3)	*75* (4)
B. Items matching with intended-1995 and not-matching with intended-2000 (n=3)	*31* (3)	*28* (4)
C. Items not matching with intended-1995 but matching with intended-2000 (n=1)	*23* (3)	*32* (5)
D. Items matching with neither intended-1995 nor intended-2000 (n=0)	--	--

Note: Dashes indicate there were no items to obtain achievement results from.

The second way of linking the attained and intended curriculum is to calculate correlation coefficients between p-values and item-curriculum matching indices. As the latter were available on a ratio scale, these can be plotted against the p-values in a scatter diagram. See Figure 6.4.

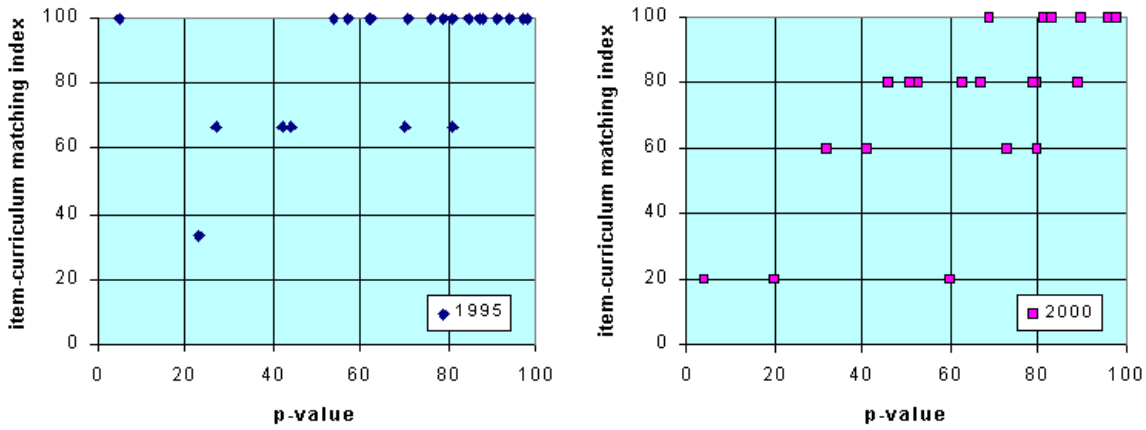


Figure 6.4: Item-curriculum matching indices and achievement results (inp-values) of PA-1995 and PA-2000

In the two scatter plots, each dot represents an item from the TIMSS Performance Assessment. In 1995, there were only three experts, and therefore the values of the item-curriculum matching indices can only be 0-33-37-100. In 2000, there were five experts, and therefore the values can be 0-20-40-80-100.

The scatter diagrams tell that in 1995, the experts judged the items as more appropriate in light of the intended curriculum than in 2000. Also, a larger number of experts included more 'difficult' items than in 2000. As a result, the correlation in the diagram for 1995 is lower with $r=0.55$ ($n=24$, $p<0.01$) than in the diagram for 2000 with $r=0.77$ ($n=24$, $p<0.01$). This difference is partly due to the larger number of experts in 2000, which caused a higher precision in the experts' judgement ratings. However, this higher correlation is also caused by the fact that in 2000, the experts did include fewer items with low p-values than in 1995. The correlations between experts' judgement and students' achievement is illustrated in Figure 6.5.

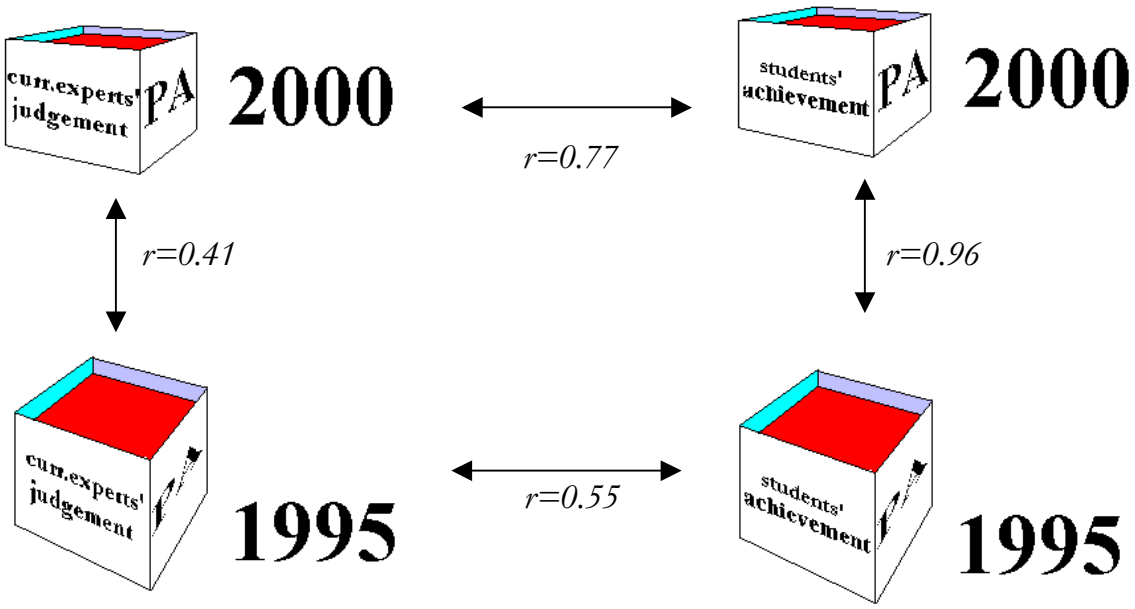


Figure 6.5: Overview of correlations on PA1995 and PA2000, at the level of the intended and attained curriculum

Based on the Performance Assessment, the following observations can be made:

1. *According to the experts' judgement, the tasks in the Performance Assessment match well with the intentions of the RME-based curriculum, although the rate of approval had dwindled in five years.*
2. *Students' achievement on the Performance Assessment displays a striking steadiness.*
3. *The experts' judgement has changed between 1995 and 2000 in favour of items with higher p-values.*

As a result, in 2000 students' achievement on the Performance Assessment is more aligned with the judgement by the experts than in 2000. This alignment is only due to the difference between the experts' judgement of 1995 and 2000, and not due to a change in students' achievements.

6.3.4 Conclusions

In the two previous sections (6.3.2 and 6.3.3) the attained curriculum has been linked to the intended curriculum, through the TIMSS Written Test and the TIMSS Performance Assessment respectively. Below are three general observations:

1. *In 1999/2000, students' achievement is slightly better than in 1995. The difference is significant on the Written Test (by means of the sign test), but it is not significant on the Performance Assessment.*

2. *Between 1995 and 1999/2000, the judgement on the appropriateness of the tests by the curriculum experts changed. Their judgement on the appropriateness of items in the light of the intended curriculum shifted towards a higher preference for items with a higher p-value.*
3. *The correlations at item level between the curriculum experts' judgement and the p-values increased considerably between 1995 and 1999/2000. On the* Written Test *the correlations were lower than on the Performance Assessment, and this could be caused by the dichotomous yes/no scale in which the judgement was expressed. Therefore, a non-parametric method for calculating correlations was used. Either way, the correlations increased considerably.*

It can, therefore, be concluded that in 1999/2000 the students' achievement in 1999/2000 and the judgement on the appropriateness of the items in light of the intended curriculum are better aligned than before. This cannot obviously be attributed to changes in students' achievement. However, *it can largely be attributed to the changed judgement on the appropriateness of the tests by the curriculum experts. This answers the first part of the second research question.*

6.4 RELATING STUDENTS' ACHIEVEMENT TO THE IMPLEMENTED CURRICULUM

6.4.1 Introduction

In the previous sections, the attained curriculum was linked to the intended curriculum. In the next sections, the attained curriculum will be linked to the implemented curriculum. Again, this will be done for the Written Test *and, next, for the Performance Assessment.*

6.4.2 Relating trends in achievement on the Written Test to the implemented curriculum

The implemented curriculum was operationalised through the judgement on the appropriateness of the test items by the teachers. The data on items represented the number of teachers, who indicated that they would include the item into an imaginary test covering all content taught until the time of testing. The terminology used was 'OTL rate' (students had an 'Opportunity To Learn'). For the measurement of OTL of the Written Test, *we do not have availability over the teachers' judgement on the full set of items. As said, in 1995, the OTL rates*

were only measured for the sixteen anchor items (see Appendix F). For 1999, we have OTL rates for all items.
First, the OTL rates on the sixteen anchor items will be presented as a trend in Table 6.5. This table combines the Tables 5.4 and 5.8.

As discussed before, the p-values are incomparable between the cloned items. Nevertheless, *we still can compare the OTL rates with the p-values of the same year.* With *little variation of the OTL rates, the correlation between the OTL rates and the p-values are low and insignificant. The correlation of 1995 yields an insignificant coefficient of r=0.17 (n=16, not significant, p=0.54). This means that no relation can be discerned between what teachers indicate as content that can be included into a test covering all content taught, and what was learnt. For 1999, the correlation between OTL rates and p-values on the above 16 items is even worse. The correlation is negative and insignificant, r=-0.13 (n=16, not significant, p=0.65).*

Table 6.5: *Trends in OTL results and achievement results on 16 selected items from* WT-*1995 and* WT-*1999*

Item name		OTL rate (SE)		Achievement p-value (SE)	
		1995	1999	1995	1999
WT-1995	WT-1999	*(n=84)*	*(36≤n≤ 39)*	*(n=239-1984)*	*(n=369-2957)*
A02	*id.*	*89* (3)	*85* (6)	*82* (6)	*86* (2)
B08	*id.*	*90* (3)	*94* (4)	*67* (4)	*71* (3)
B10	*id.*	*96* (2)	*100* (0)	*62* (4)	*74* (3)
B11	*id.*	*98* (2)	*95* (4)	*80* (3)	*82* (2)
C05	*id.*	*89* (3)	*92* (4)	*57* (4)	*60* (3)
E05	*id.*	*89* (3)	*74* (7)	*66* (4)	*74* (3)
F12	*id.*	*96* (2)	*97* (3)	*69* (3)	*75* (3)
H*07*	*id.*	*86* (4)	*85* (6)	*81* (2)	*82* (2)
K06	*clone*: *K06*	*96* (2)	*95* (4)	*51* (4)	*75* (3)
L08	*clone*: *L09*	*96* (2)	*100* (0)	*69* (3)	*82* (2)
N*15*	*clone*: N*15*	*99* (1)	*95* (4)	*64* (3)	*52* (3)
N*19*	*clone*: N*19*	*96* (2)	*97* (3)	*64* (4)	*61* (3)
O09	*clone*: *O09*	*96* (2)	*92* (4)	*75* (3)	*73* (3)
R07	*clone*: *R08*	*96* (2)	*95* (4)	*54* (4)	*57* (3)
R10	*clone*: *R11*	*98* (2)	*97* (3)	*62* (4)	*66* (3)
V*02*	*clone*: V*02*	*86* (4)	*89* (5)	*24* (3)	*45* (3)

Consequently, based on these sixteen items, we can hardly conclude that students' achievement is related to OTL rates. The lack of correlation could be caused by the characteristics of the sixteen items: they were selected for the National Option Test in 1995 on the criterion of matching with the intended curriculum. This characteristic can have caused the OTL rates to become high, resulting in a ceiling effect. Therefore, the OTL rates do not differentiate.

Due to the lack of more OTL data from the Written Test in 1995, any further trend analysis could not be based on the Written Test. This aspect needed to be covered by data from the Performance Assessment. The data from the Written Test are still valuable to compare the judgements by the teachers with students' achievement at one stage. For 1999 data are available on 144 items in the Written Test. The data can be used to link the implemented curriculum to the attained curriculum. In chapter 5, at the level of the implemented curriculum, the items were grouped into categories according to their OTL rates (see section 5.3.1, Table 5.5). Items were grouped into categories by OTL rates 0-10, 10-20, and so forth, which indicated the percentages of teachers, who would include the item into an imaginary test.
Below, the same grouping is used again. Each OTL rate category contains items, which were considered fit by approximately a comparable number of mathematics teachers. The average p-values on the items were then calculated. The result is presented in Table 6.6.

Table 6.6: Achievement results on items in WT-1999, grouped per OTL rate

OTL rate category (# items)	Achievement Avg p-value (SE) *(n=369-2957)*
≤ *50* (6)	*47* (3)
50 – 60 (8)	*45* (3)
60 – 70 (12)	*54* (3)
70 – 80 (22)	*60* (3)
80 – 90 (41)	*62* (3)
90 – 100 (55)	*72* (3)
Average (144 items)	*63 (3)*

The items with the lower OTL rates (below 50 and 50-60) have the lowest p-values, all being below 50. For higher OTL rates, the p-values increase together with the OTL rates. Items, which have OTL rates above 90, can be correctly

completed by many Dutch students as demonstrated by the average p-value of 72. This figure is higher than the average p-value over all 144 items. This implies that items with a high match with the implemented curriculum (OTL>90), yield on average more satisfying achievement results than items with lower OTL rates.

The above observation is confirmed by the calculation of the correlation coefficient between the OTL data and the p-values of 1999. The correlation is first illustrated in Figure 6.6 through a scatter diagram, whereby each dot represents one of the 144 items in the Written Test. *On the horizontal axis, the p-value is indicated. On the vertical axis the OTL rate is indicated.*
Also included in the Figure 6.6 are the sixteen items on which the OTL measurement of 1995 was based. These items are situated 'high' in the scatter, as in 1995 none of these items had an OTL rate lower than 86.

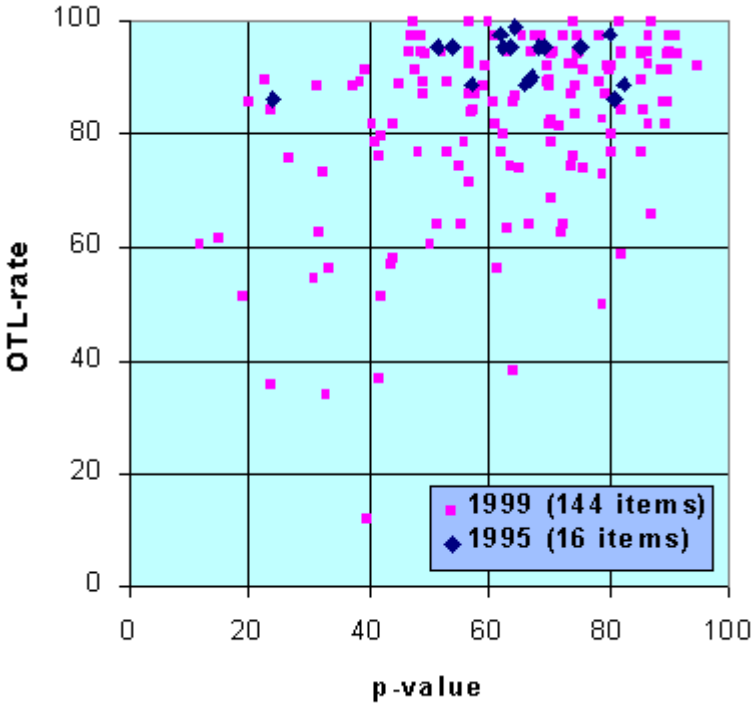


Figure 6.6: Scatter diagram of OTL rates and achievement results (p-values) of 144 mathematics items in WT-*1999 (and sixteen items from* WT-*1995)*

The scatter diagram for all 144 items in 1999 seems a blur. There are items in almost all corners. This means that there are items with a low OTL rate that have high p-values, or the other way around: high OTL rate and low p-values. It looks as if there is a concentration of items in the three upper right blocks where the OTL rates are >80 and the p-values are >40. Also, there seems to be a tendency

towards the upper triangle: most items are situated above or to the right of the (0,0)-(100,100)-diagonal. This means that there are more items with an OTL rate that is higher than the p-value, and fewer items for which the p-value is higher than the OTL rate. The correlation on the 144 items for 1999 is significant, r=0.41 (n=144, p<0.01).

We have no trend data from 1995 to compare these with. However, it is possible to compare the correlation of OTL rates and the p-values in other ways. Originally, De Haan (1992) developed the OTL instrument and she found a correlation with the p-values of r=0.5 (n=51, p<0.01). This can be perceived as a rough maximum, as she found her results under well-controlled circumstances. For example, she used a test, which was specially developed for the content to be taught and it was based on the textbooks the teachers were using, and she had homogeneous classes (all in the mavo/havo track). Thus, the correlation of r=0.41 which was established through the TIMSS Written Test, approaches De Haan's value of r=0.5.

In Figure 6.7, the found correlation coefficient of r=0.41 between OTL rates and p-values for WT-1999 is combined with the trend correlation coefficients, which were already reported in chapter 5 (section 5.4.1). Because of the missing OTL data from WT-1995, the picture is void. It has been included for comparison with Figures 6.3 (comparison of experts' judgement and students' achievement for WT01995 and WT-1999) and 6.9 (forthcoming: comparison of teachers' judgement and students' achievement for PA-1995 and PA-2000).

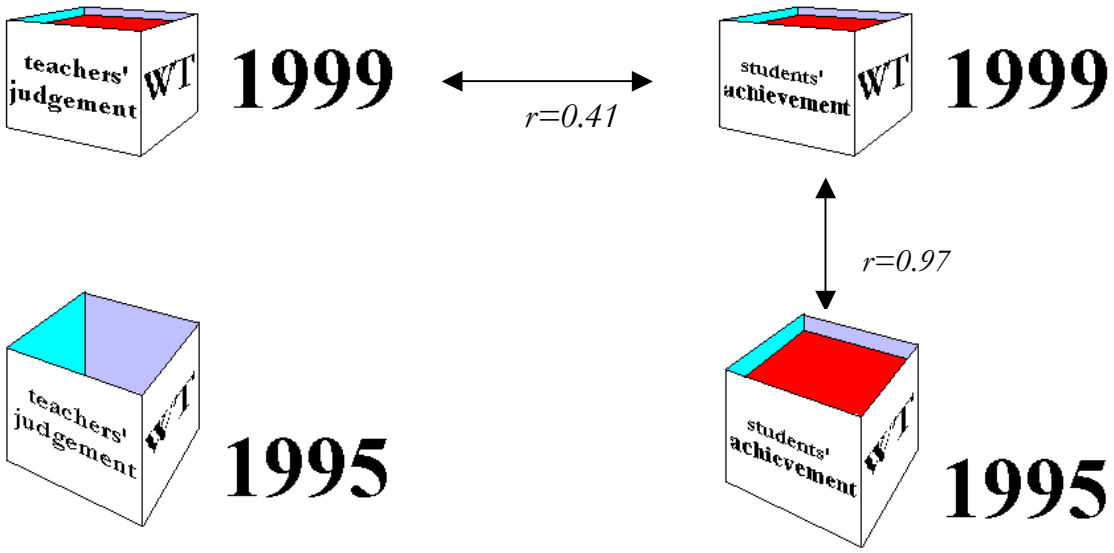


Figure 6.7: Overview of correlations on WT-1995 and WT-1999, at the level of the implemented and attained curriculum

Lack of OTL data from WT-1995 prohibits the analysis of a trend. We can only conclude that the alignment of implemented and attained curriculum is measurable in 1999 (there is a positive correlation, r=0.41), but it cannot be established whether it has increased or not. Maybe, the next section, on the Performance Assessment, will give a more solid foundation for further conclusions.

6.4.3 Relating trends in achievement on the Performance Assessment to the implemented curriculum

In this section, we continue to link the implemented curriculum to the attained curriculum, but here the analysis in based on data gathered through the Performance Assessment. This data collection was not hindered by incompleteness of data. The implemented curriculum was operationalised through the judgement on the appropriateness of the test items by the teachers. For the Performance Assessment, the 'Opportunity To Learn' question to the teachers was asked two-fold: (1) whether the content of the item had been covered before test administration, and (2) whether the teachers would include the test item into a (practical) test of their own making. Thus, the data gathered at the level of the implemented curriculum were named OTL-covered and OTL-testing. These OTL data were available for both 1995 and 2000 and can be compared to students' achievement on the items.
In Table 6.7, all data are reported per task. The p-values of the two tasks Shadows and Plasticine were not comparable between 1995 and 2000 and, therefore, omitted from the analysis. The table is a combination of tables that were already presented in chapter 5: Tables 5.6 and 5.9.

From the table, we can see that the task with the lowest average p-values, the task Packaging, has highest OTL rates. This means that many teachers indicated that they had covered the content of this task and *that they would include the Packaging items into a test of their own making, but that only half of the students were able to complete this task correctly. By contrast, the task with the lowest OTL rates, Folding, has high p-values. This means that only few teachers indicated that they had covered the content of the task Folding, but nevertheless, more than three-quarter of the students was able to complete this task correctly.*

Table 6.7: Trends in OTL rates (average percentage of teachers) and p-values on PA-1995 and PA-2000

Task	Average OTL-covered rates (SE)		Average OTL-testing rates (SE)		Achievement Avg p-value (SE)	
(#items in OTL/test)	1995 *(n=19)*	2000 *(n=20)*	1995 *(n=19)*	2000 *(n=20)*	1995 *(n=437)*	2000 *(n=234)*
Dice (5/6)	*47* (11)	*73* (10)	*51* (11)	*70* (10)	*77* (3)	*74* (4)
Calculator (6/7)	*43* (11)	*68* (10)	*56* (11)	*82* (9)	*62* (4)	*60* (5)
Folding (2/4)	*17* (8)	*31* (10)	*28* (10)	*75* (10)*	*73* (4)	*77* (5)
A the Bend (6/8)	*35* (10)	*68* (10)*	*56* (11)	*84* (8)*	*68* (3)	*70* (4)
Packaging (3/3)	*65* (10)	*74* (10)	*76* (10)	*88* (7)	*52* (4)	*58* (5)
Shadows (3/6)	*30* (10)	*33* (10)	*47* (11)	*67* (11)	--	--
Plasticine (3/8)	*23* (9)	*40* (11)	*26* (10)	*70* (11)*	--	--

Note: Dashes indicate the achievement results were deemed unreliable;
* Significant difference between data of 1995 and 2000 ($p<0.05$).

The data at task level do not generate a clear picture. But the data can also be analysed at item level. With the OTL rates being available on a ratio scale, they can be plotted against the p-values in a scatter diagram. See Figure 6.8.
The resulting correlation coefficients are as follows:

- *between OTL-covered rates and p-values in 1995*: *r=0.26 (n=22, not significant with p=0.25),*
- *between OTL-covered rates and p-values in 2000*: *r=0.44 (n=22, significant, with p=0.04),*
- *between OTL-testing rates and p-values in 1995*: *r=0.38 (n=22, not significant with p=0.08), and*
- *between OTL-testing rates and p-values in 2000*: *r=0.24 (n=22, not significant with p=0.28).*

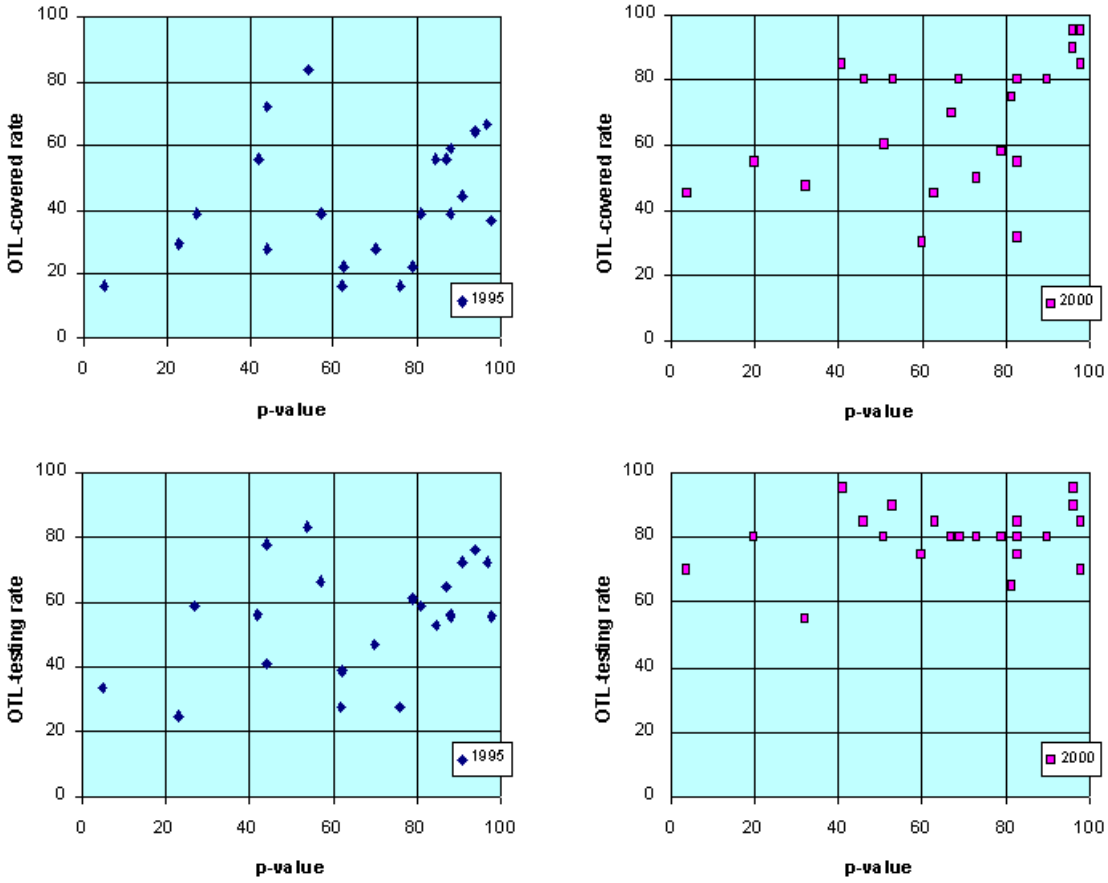


Figure 6.8: Scatter diagram of OTL rates (OTL-covered and OTL-testing) and achievement results (p-values) of PA-1995 and PA-2000

It turns out that only the correlation between the OTL-covered rates and the p-values are significant in 2000. This means that in 2000, the more items were indicated by the teachers as covered, the more students were able to complete these correctly. The coefficient of r=0.44 is similar to the correlation coefficient of r=0.41 which was established through the Written Test, *and it also approaches De* Haan's *value of r=0.5.*

The correlation between OTL-covered rates and p-values was not present in 1995, and it was not present either for the OTL-testing rates. This results in Figure 6.9.

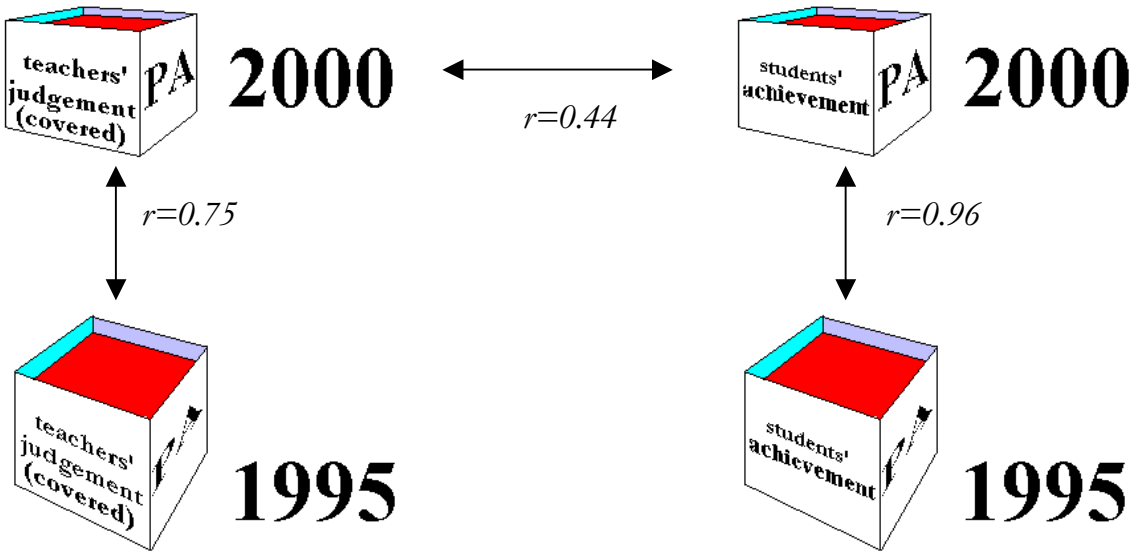


Figure 6.9: Overview of correlations on PA-1995 and PA-2000, at the level of the implemented curriculum (OTL covered) and attained curriculum

The results can be interpreted as follows: from 1995 to 2000, an increasing number of mathematics teachers indicated that they had covered the content of items in the Performance Assessment (see section 5.3.2). In 1995, on average 38% of the teachers indicated that items were covered, and in 2000, this percentage had significantly increased to 58%. In that same time span, students' achievement did not change significantly, meaning that the increased content coverage cannot be related to improved students' achievement. From Figure 6.9, we can see that the intra-curricular (horizontal) correlation is absent in 1995. Thus, students were able to perform well on items that the teachers indicated as not covered and, conversely, students did not do well on items that the teachers indicated as covered. By 2000, there is a correlation of r=0.44 between the teachers' judgement and students' achievement. This means that the increased OTL-covered rates became more aligned with students' stable achievement during the five years that the METRIC study covered.

Figure 6.9 is derived from the data on the Performance Assessment. The figure can be compared with Figure 6.7, which is based on data from the Written Test. The correlations between the students' achievement in time (between 1995 and 1999/2000) and the correlation between teachers' judgements and students' achievement in 1999/2000 shows a striking similarity in magnitude.

Additionally to the OTL-covered rates, the METRIC study had gained OTL-testing rates. The OTL-testing rates increased significantly between 1995 and 2000. As a result, a ceiling effect can have occurred (see the fourth scatter diagram in Figure 6.8), as almost all dots have an OTL-testing rate > 60, which results in a lower differentiation. The OTL-testing rates did not show any relationship with students' achievements, neither in 1995 nor in 2000. In both years, the correlation is insignificant. A possible explanation may be found in the fact that teachers were asked to think of an imaginary practical mathematics test, while they have little experience with these kinds of tests. Thus, teachers might have given imaginary answers. However, to explain the lack of fit between the OTL-testing rates and the p-values would require further study, which is beyond the scope of the METRIC study. It would require for example, interviews with teachers to interpret these findings.

6.4.4 Conclusions

In the two previous sections (6.4.2 and 6.4.3), the attained curriculum has been linked to the implemented curriculum, for the TIMSS Written Test and the TIMSS Performance Assessment respectively. Below are four general observations:

1. *In 1999/2000, Dutch students' achievement has slightly improved when compared to their achievement in 1995. The difference is significant on the Written Test (by means of the sign test), but it is not on the Performance Assessment.*
2. *Because of lack of sufficient OTL data on the Written Test, trends in the judgement on the appropriateness of this test in light of the implemented curriculum could not be analysed. However, between 1995 and 2000, teachers' judgement on the appropriateness of the Performance Assessment changed considerably. Their assessment of shifted towards (a) a significantly higher coverage of the content tested, and (b) a significantly higher tendency to include test items from the Performance Assessment into a test of their own making.*
3. *The correlations at item level between teachers' judgement and the p-values were absent in 1995 (for the Written Test, this was due to absence of data; for the Performance Assessment this was an insignificant correlation). The correlation was present in 1999/2000 with r=0.41 on the Written Test and r=0.44 on the Performance Assessment (between OTL-covered rates and p-values). These coefficients are slightly lower than the coefficient of r=0.5*

found by De Haan (1992). The correlation between OTL-testing rates and p-values was absent, both in 1995 and 2000, probably hindered by teachers' judgement on the inclusion of items into an imaginary practical test, while they have little experience with these.

It can, therefore, be concluded that in 1999/2000, students' achievement and the judgement on the appropriateness of the items in light of the implemented curriculum are quite well aligned. It is difficult to observe a trend. When only based on the correlations between OTL-covered data and students' achievement data from the Performance Assessment, the alignment has increased between 1995 and 2000. This is a noteworthy result in itself, but the second part of the second research question cannot completely be answered with the available data.

6.5 RELATING STUDENTS' ACHIEVEMENT TO THE DISCREPANCY BETWEEN INTENDED AND IMPLEMENTED CURRICULUM

6.5.1 Introduction

The second research question links Dutch students' achievement on the two TIMSS tests to the intended and implemented curriculum. In the two previous sections (section 6.3 and 6.4), the two have been dealt with separately. In the current section, the intended and implemented curriculum will be mutually compared. This is needed for the last part of the second research question, where the point was raised how to relate the students' achievements on the two tests to a possible discrepancy between the judgements at the level of the intended and implemented curriculum. This implies a description of the discrepancies between these.

6.5.2 Relating trends in achievement on the Written Test to the discrepancy between intended and implemented curriculum

The items in the Written Test were submitted to curriculum experts and teachers, in order to assess the items in light of both the intended and implemented curriculum respectively. This happened in 1995 and in 1999 for all 144 items, except for the measurement in 1995 at the level of the implemented curriculum when only 16 items were submitted which all were considered relevant to the Dutch RME-based intended curriculum.

The OTL rates on the 16 items were very high, being 93 in both 1995 and 1999 (see chapter 5, section 5.3.1). This meant that on average 93% of the Dutch mathematics teachers indicated that they considered the items fit for an imaginary test. Considering the fact that these items were specially selected for the National Option Test on the criterion that they matched with certain areas of the intended curriculum, this means that those areas of the intended curriculum were largely implemented both in 1995 and 1999.

In 1999, all items in the Written Test were submitted to the teachers. As quite a number of the items were not considered relevant to the intended curriculum, it was possible to assess whether teachers considered these as suitable for an imaginary test, covering all content taught before test administration.

In Table 6.8, the average OTL rates are presented, when separating the test items according to their judgement by the curriculum experts. For comparison with other tables, the format of Tables 6.2a, 6.2b and 6.4 is maintained. Thus, the test items are again separated into parts A, B, C and D. The column for the OTL rates of 1995 is largely empty because of lack of data. Only the average OTL rate for part A is included, although this figure was calculated using limited data.

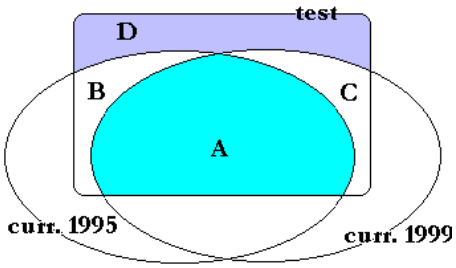


Table 6.8: Average OTL rates on 144 items in WT-1999 (identical and cloned items), related to the intended curriculum

Group of items	Average OTL rate (SE) WT-1995 *(n=84)*	WT-1999 *(n=36-39)*
Overlapping items in WT-1995 and WT-1999 (n=144)	*n/a*	*82* (6)
A. Items matching with both intended-95 and intended-99 (n=14 for 1995; n=76 for 1999)	[*93* (3)]	*89* (5)
B. Items matching with intended-95, but not matching with intended-99 (n=23)	*n/a*	*80* (6)
C. Items not matching with intended-95, but matching with intended-99 (n=26)	*n/a*	*79* (6)
D. Items matching with neither intended-95 nor intended-99 (n=19)	*n/a*	*61* (7)

In 1999, the average OTL rates are highest on the 76 items from the Written Test *that match with the intended curriculum both in 1995 and 1999 (part A). Here, the average OTL rate of 89 testifies of an alignment of intended and implemented curriculum: a vast majority of teachers indicate that they consider these items that match with the intended curriculum suitable for an imaginary test. The items largely test for content taught at primary schools, such as applying basic arithmetic algorithms, working with fractions and decimal numbers, reasoning with proportions, measuring, calculating area, and interpreting diagrams and graphs. The lowest OTL rates of items in part A (<65) occurred on three algebra items (E02, I01 and P15/P09).*

On the parts B and C, the average OTL rates in 1999 are slightly lower. These parts contain 49 (=23+26) items that either matched with the intended curriculum in 1995 and not with the intended curriculum in 1999 (part B) or the other way around (part C). Still, the average OTL rate of 79 and 80 are quite high. The average OTL rates in 1999 are lowest on the 19 items that match neither with the intended curriculum in 1995 nor with that in 1999 (part D). These were largely algebra items on manipulating 'bare' formula and a few items on calculating probabilities (cf. section 5.2.1). The average OTL rate of 60 indicates that still a majority of the mathematics teachers consider these items suitable for an imaginary test. The OTL rate of 60 is an average over 19 items, ranging from OTL rate 12 (item R10 on the formulation of axioms for real numbers, see appendix D) to OTL rate 100 (item O04, on rounding off a decimal number to the nearest hundredth).

The high OTL rates on all four parts indicate that Dutch mathematics teachers at grade 8 level prefer to include items that match with the intended curriculum into a test of their own making, but they do not decline items that do not match with the intended curriculum. Considering the fact that 46% of the participating teachers had higher ability classes (havo/vwo; see section 4.2.2, Table 4.6), this means that there must be a considerable number of teachers with lower ability track classes, who would include items into a test, although these items are not relevant to the intended curriculum. This could mean that many mathematics teachers cover more content than is actually intended. It could explain why large number of students completed these items correctly, as shown by an average p-value of more than 50 on the items in part D (see Table 6.2b). As a consequence, the p-values do differ little between items that match with the intended curriculum and items that match not.

A second way of linking the implemented curriculum to the intended curriculum is by calculating a correlation index between the item-curriculum matching indices and the OTL rates. In 1995, this correlation could not be calculated because of lack of data. For 1999, this correlation is significant r=0.44 (Kendall's tau & Spearman's rho, n=144, p<0.01). This means that there is a relation between the two, with the higher OTL rates for the items that are covered by the intended curriculum. With lack of OTL data for 1995 the full picture remains vague. It is illustrated in Figure 6.10 and will need to be supplemented with results on the Performance Assessment.

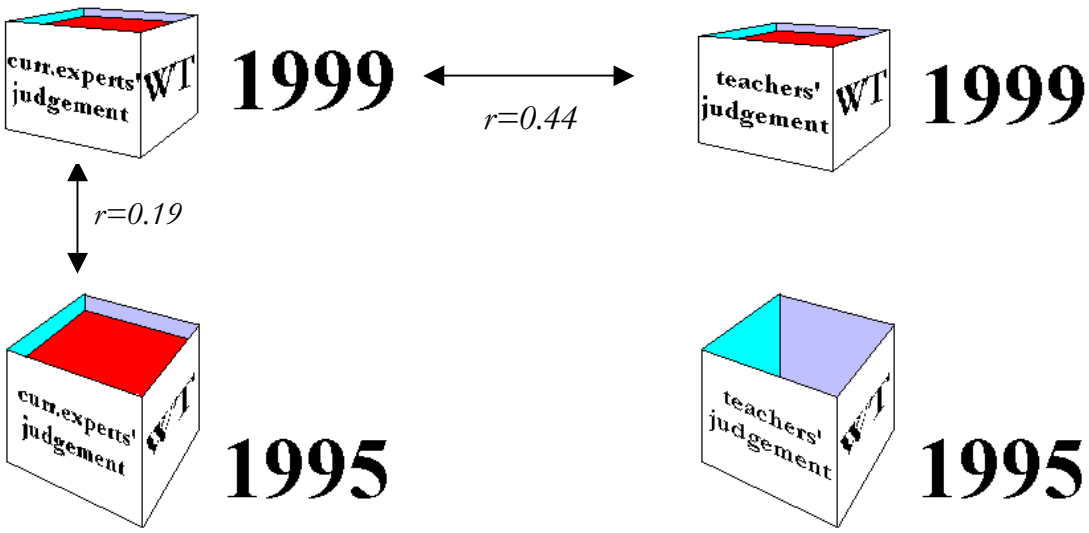


Figure 6.10: Overview of correlations on WT-1995 and WT-1999, at the level of the intended and implemented curriculum

The last part of the second research question pertained the link of students' achievement to a possible discrepancy between the intended and implemented curriculum. At present, we can already make the following observations, based on prior observations described in previous sections, and based on the data presented in this current section:

(from section 6.3.2)

1. *Students' achievement increased slightly but significantly between 1995 and 1999 on all 41 identical items from the Written Test. The achievement of students increased slightly but significantly between 1995 and 1999 on the items that matched with the intended curriculum both in 1995 and 1999 (part A).*

2. *Students' achievement (both in 1995 and 1999) is lower on the items that matched with the intended curriculum in 1995 and not in 1999 (part B). The achievement of students (both in 1995 and 1999) is higher on the items that matched with the intended curriculum in 1999 and not in 1995 (part C).*

(from section 6.4.2)

3. *In 1999, students' achievement was quite well aligned with the OTL rates (r=0.41). Items with high OTL rates were generally also the items with high p-values, and conversely, items with lower OTL rates had lower p-values. Because of lack of data, no trend could be established.*

(from the current section)

4. *In 1999, the OTL rates are quite well aligned with the item-curriculum matching indices, which indicate the match with the* intended *curriculum (r=0.44). The OTL rates are higher on items that match with the intended curriculum, and the OTL rates are lower on the items that do not match with the intended curriculum.*
5. *The OTL rates on items that do not match with the intended curriculum are on average above 60. This means, that a majority of the teachers consider most items in the* Written Test *to match with the content taught.*

From the above observations, it can be concluded, that in 1999, many teachers consider the items in the Written Test *as suitable for an imaginary test, covering all content taught. Items that match with the intended curriculum receive higher OTL rates, and items that do not match with the intended curriculum receive somewhat lower rates, although these are still high (average OTL rate >60). This result at the level of the implemented curriculum is reflected at the level of the attained curriculum. The test items in the* Written Test *were on average completed correctly by many students. Items that match with the intended curriculum received had higher p-values, and items that do not match with the intended curriculum had somewhat lower p-values, which is still high (average p-value >60).*

Based on data from the Written Test, *it is not possible to establish a trend. To supplement the above observations, we will now turn to the Performance Assessment.*

6.5.3 Relating trends in achievement on the Performance Assessment to the discrepancy between intended and implemented curriculum

The items in the Performance Assessment were submitted to curriculum experts and teachers, in order to assess them in light of the intended and implemented curriculum respectively. This happened both in 1995 and 2000 for all items. The first yielded data on the appropriateness of the items in light of the intended curriculum, expressed as an average item-curriculum matching index. The second yielded data on OTL. The latter was asked in two ways: whether the content of items was covered (OTL-covered) and whether the items could be included into a test of their own making (OTL-testing). In Table 6.9 the trend data are reported per task.

Table 6.9: Trends in item-curriculum matching indices and OTL rates from PA1995 and PA-2000

Task	Avg item-curr. matching index		Avg OTL-covered rates (SE)		Avg OTL-testing rates (SE)	
(#items in experts' instrument/in teachers' instrument)	1995 *(n=3)*	2000 *(n=5)*	1995 *(n=19)*	2000 *(n=20)*	1995 *(n=19)*	2000 *(n=20)*
Dice (5/5)	*87*	*92*	*47* (11)	*73* (10)	*51* (11)	*70* (10)
Calculator (6/6)	*72*	*70*	*43* (11)	*68* (10)	*56* (11)	*82* (9)
Folding (4/2)	*100*	*60*	*17* (8)	*31* (10)	*28* (10)	*75* (10)*
A. the Bend (6/6)	*100*	*83*	*35* (10)	*68* (10)*	*56* (11)	*84* (8)*
Packaging (3/3)	*89*	*73*	*65* (10)	*74* (10)	*76* (10)	*88* (7)
Shadows (3/3)	*61*	*67*	*30* (10)	*33* (10)	*47* (11)	*67* (11)
Plasticine (3/3)	*78*	*40*	*23* (9)	*40* (11)	*26* (10)	*70* (11)*
Average (33/31 items across 7 tasks)	*83*	*72*	*38 (10)*	*58 (10)*	*51 (11)*	*76 (9)*

Note: * Significant difference between data of 1995 and 2000 ($p<0.05$).

The results in Table 6.9 do not show a clear pattern between item-curriculum matching indices and OTL rates. There is a task with a high average item-curriculum matching index (Dice), which has higher OTL rates. But there is also a task with a high average item-curriculum matching index (Folding), which has low OTL rates, especially in 1995. Similarly, there is a task with a lower average item-curriculum matching index (Shadows), which has low OTL rates. But there is also a task with a lower average item-curriculum matching index (Calculator), which has above average OTL rates. This does not give a clear picture, just like in previous sections the Tables 6.3 and 6.7 did not.

A second method of analysing the data is to calculate correlation coefficients at item level. With *the item-curriculum matching indices and the OTL rates both being available on a ratio scale, they can be plotted in scatter diagrams. See Figure 6.11.*

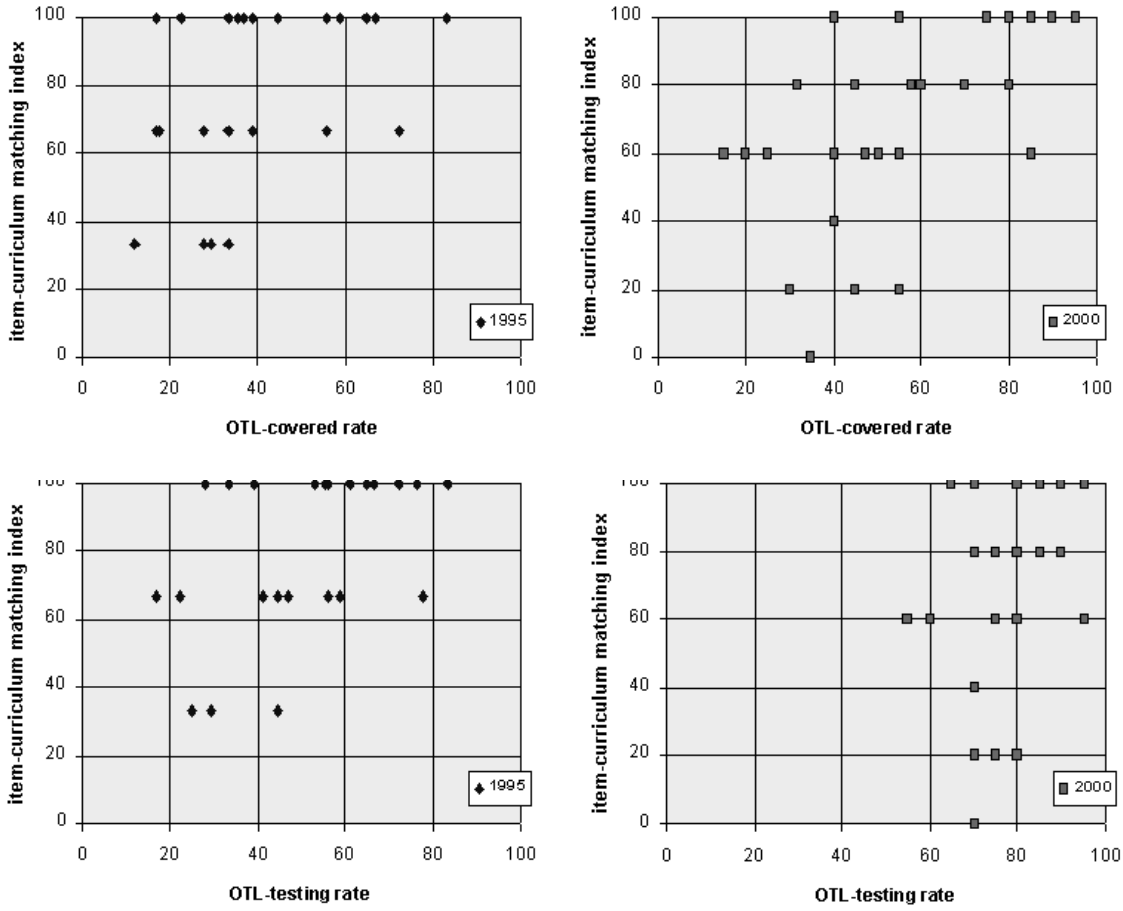


Figure 6.11: *Scatter diagram of OTL rates (OTL-covered and OTL-testing) and item-curriculum matching indices of PA1995 and PA2000*

The resulting correlation coefficients are as follows:

- *between OTL (covered) and item-curriculum matching indices in 1995*: *r=0.28 (n=31, not significant with p=0.12),*
- *between OTL (covered) and item-curriculum matching indices in 2000*: *r=0.61 (n=31, significant, with p<0.01),*
- *between OTL (testing) and item-curriculum matching indices in 1995*: *r=0.39 (n=31, significant with p=0.03), and*
- *between OTL (testing) and item-curriculum matching indices in 2000*: *r=0.32 (n=31, not significant with p=0.09).*

It follows from the calculations, that there is a noteworthy difference between the data of 1995 and 2000. In 1995, the item-curriculum match as rated by the experts is not related to OTL-covered. This means that in 1995 those items, which the experts indicated as being covered by the intended curriculum, did not receive higher OTL-covered rates by the teachers. And conversely, in 1995 those items, which the experts indicated as not *being covered by the intended curriculum, did* not *receive lower OTL-covered rates from the teachers. By 2000, the OTL-covered rates have changed in such a way that they highly relate to the item-curriculum match. In 2000, those items, which the experts indicated as being covered by the intended curriculum, did obviously receive higher OTL-covered rates by the teachers. And conversely, in 1995, those items, which the experts indicated as not being covered by the intended curriculum, did receive lower OTL-covered rates from the teachers. This shows that within five years the judgement at the level of the intended curriculum became better aligned with the judgement at the level of the implemented curriculum.*

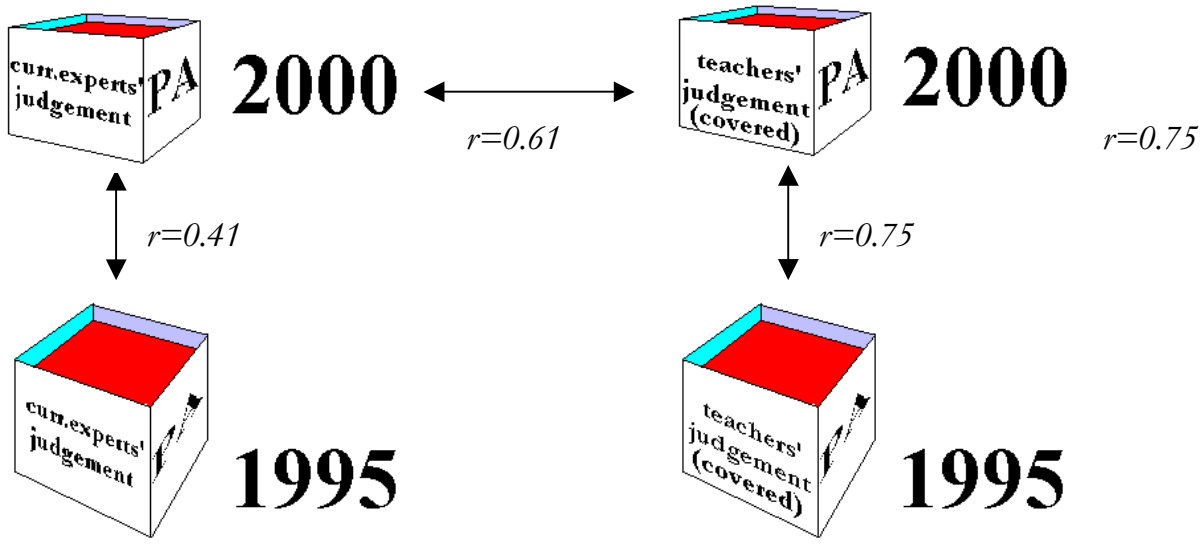


Figure 6.12a: Overview of correlations on PA1995 and PA2000, at the level of the intended and implemented curriculum (OTL-covered)

On the other hand, the inclination of teachers to include test items from the Performance Assessment into a test of their own making, has increased to such an extent, that the dots in the fourth scatter diagram (Figure 6.11) are all on the right hand side (OTL-testing > 60), creating a ceiling effect, which hinders the differentiation between the item-curriculum matching indices.

The correlation between the OTL-covered rates and item-curriculum matching indices is significant in 2000. The correlation was significant in 1995. This results in Figure 6.12a. By contrast, the correlation between the OTL-testing rates and item-curriculum match is significant in 1995. This correlation was not significant in 2000. This results in Figure 6.12b.

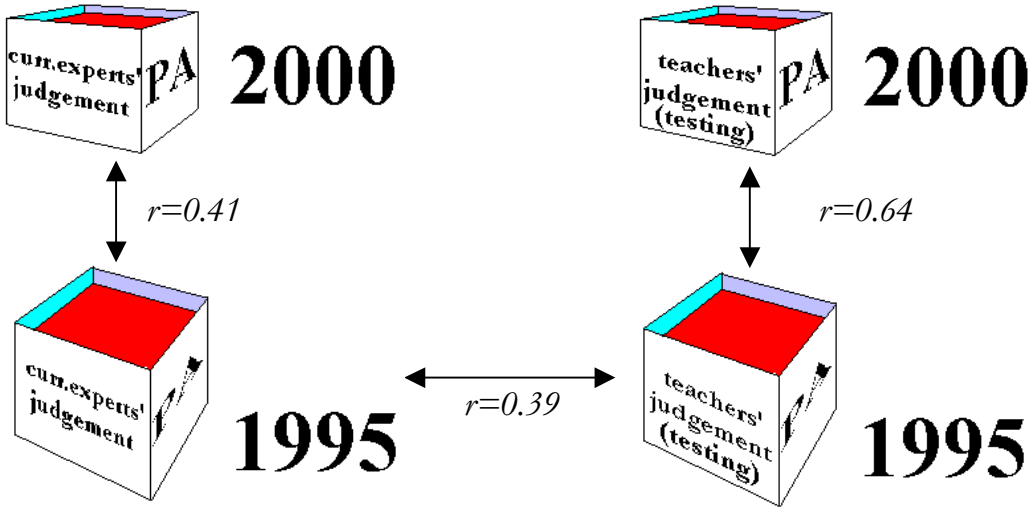


Figure 6.12b: Overview of correlations on PA1995 and PA2000, at the level of the intended and implemented curriculum (OTL-testing)

The OTL-testing rates are an indicator of teachers' intentions to include items from the Performance Assessment into a test of their own making. In 1995, these intentions are aligned with the intended curriculum, while in 2000, the OTL-testing rates have increased to such an extent, that they do not correlate any more with the intended curriculum coverage rates. By 2000, the intended curriculum is better aligned with the OTL-covered rates (r=0.61).

In this current section, data from the intended and implemented were compared in search of answers to the second research question, which pertained linking students' achievement to the intended and implemented curriculum. Based on data from the Performance Assessment, we can make the following observations, based on prior observations described in previous sections, and based on the data presented in this current section:
(from section 6.3.3)

1. *Students' achievement showed stability over time. It did not change significantly between 1995 and 2000 on the Performance Assessment.*

2. *Between 1995 and 2000, the average item-curriculum matching indices (indicating the match with the* intended *curriculum) on the Performance Assessment decreased from 83 to 72. In 1995, this practical test received higher approval than the* Written Test*, but the high approval has dwindled since.*
3. *Students' achievement (both in 1995 and 2000) is higher on the items that match with the intended curriculum in 1995 and in 1999 (part A). The achievement of students (both in 1995 and 1999) is lower on the items that did not match with the intended curriculum, either in 1995 or in 2000 (part B and part C). The correlation between item-curriculum matching indices and p-values increased from r=0.55 in 1995 to r=0.77 in 2000.*

(from section 6.4.3)

4. *The OTL rates on the Performance Assessment changed significantly between 1995 and 2000. Both the OTL-covered rates and the OTL-testing rated testified of a changing attitude among Dutch mathematics teachers.*
5. *The OTL rates on the Performance Assessment were lower than the OTL rates on the* Written Test*. Thus, the* Written Test *aligned better with the implemented curriculum than the Performance Assessment. This observation may well reflect current Dutch assessment practice in mathematics classrooms.*
6. *In 1995, students' achievement showed little coherence with the OTL rates (either OTL-covered or OTL-testing), resulting in an insignificant correlation coefficient. In 2000, the correlation between OTL-covered and the p-values was significant (r=0.44). Items with high OTL-covered rates were generally also the items with high p-values, and conversely, items with lower OTL-covered rates had lower p-values. In 2000, the high OTL-testing rates probably caused a ceiling effect.*

(from the current section)

7. *In 1995, the OTL-covered rates showed no correlation with the item-curriculum match, but the OTL-testing rates, which reflect teachers' intentions, did (r=0.39).*
8. *In 2000, the OTL-covered rates are well aligned with the item-curriculum matching indices (indicators of the match with the* intended *curriculum), which is expressed through a reasonable high correlation coefficient (r=0.61). The OTL-covered rates are obviously higher on items that match with the intended curriculum, and the OTL-covered rates are low (<50) on the items that do not match with the curriculum. Again, in 2000, the high OTL-testing rates probably caused a ceiling effect.*

Based on the above observations, the following conclusion can be drawn. In 1995, there was a discrepancy between teachers' and experts judgement on the appropriateness of the Performance Assessment. This discrepancy has decreased by 2000. Nevertheless, *the decreased item-curriculum matching indices, and the increased OTL rates did not affect students' achievement.*

6.5.4 Conclusions

The second research question pertained the link of data gained at the levels of the intended, implemented and attained curriculum, in order to find trends, which could explain students' achievement. There were data based on the Written Test, *and data based on the Performance Assessment.*
The two tests were different in kind. The Written Test *contained many short items, mainly of the multiple choice format, asking for isolated knowledge and skills. Each item was considered to take a maximum of three minutes to complete. By contrast, the Performance Assessment asked students to carry out a practical investigation and stay concentrated on that one task for 15 or 30 minutes. As a result of the differences between the tests, the data yielded complementary observations.*
The analyses were carried out by grouping the items according to the judgement on their appropriateness in light of the intended and implemented curriculum. The average achievement was then calculated on the different sub-sets of items. Additionally, correlation coefficients were calculated. These are summarised in Table 6.10.

Table 6.10: Intra-curricular correlation coefficients in the METRIC study

Sub-study	Between intended and attained curriculum	Between implemented and attained curriculum	Between intended and implemented curriculum
WT-*1995*	* (nom.)	--	--
WT-*1999*	*0.23* (nom.)	*0.41*	*0.44*
PA-1995	*0.55* (rat.)	* (cov.) * (tst.)	* (cov.) *0.39* (tst.)
PA-2000	*0.77* (rat.)	*0.44* (cov.) * (tst.)	*0.61* (cov.) * (tst.)

Note: Dashes indicate data are unavailable; nom.=on a nominal scale; rat.=on a ratio scale; cov.=OTL-covered; tst.=OTL-testing;
* Insignificant correlation.

First, the intended and attained curriculum were linked. The analysis was carried out by separating the items according to the judgement by the experts. For the Written Test, *the judgements by the experts on the appropriateness of items had made a considerable change, while students' achievement had increased slightly, but significantly. It was found that in 1995 the experts had preferred 'harder' items than in 1999/2000. Thus, it was concluded that the judgement at the level of the intended curriculum had made concessions, in the sense that a larger number of students were able to complete the items that were considered to match with the intended curriculum. For the Performance Assessment, the judgement by the experts had changed, in the sense that their enthusiasm for this test dwindled overall. On this test, the same phenomenon as for the* Written Test *was observed: in 2000, the 'harder' items yielded lower item-curriculum matching indices than in 1995. These were the particular items at the end of a task (the more explanatory or reflective items). The increased correlation coefficients confirmed that the intended and the attained curriculum had become more aligned, mainly due to concessions at the level of the intended curriculum.*

Second, the implemented and attained curriculum were linked. The data at the level of the implemented curriculum were different between the Written Test *and the Performance Assessment. The OTL rates were generally high for the* Written Test, *displaying that the content of most of the test items matched with the content covered in class. This contrasted with the OTL rates on the Performance Assessment, which were lower, although they increased significantly between 1995 and 2000.*

Linking the data with students' achievement was carried out, among others, by calculating correlation coefficients at item level between the judgement data by teachers and the scores of the students (OTL rates and p-values). Neither *for the* Written Test *nor for the Performance Assessment, the 1995 data gave a significant correlation coefficient, although for the* Written Test *this was due to lack of data. For the Performance Assessment, the lack of a correlation meant that students were able to perform well on items that the teachers indicated as inappropriate, or the other way around. By 1999/2000, this had changed. The correlation coefficients of r=0.41 on the* Written Test *and r=0.44 on the Performance Assessment (for 'OTL-covered') respectively showed that, generally, students were able to score well on items which the teachers had indicated as matching with the implemented curriculum. Comparing the coefficients to the coefficient, which was established by De* Haan *(1992): r=0.5, this indicated that in 1999/2000, the attained and the implemented curriculum were quite well aligned.*

Finally, the intended and implemented curriculum were linked. Here, *again, the items were separated according to the judgement by the experts. Additionally, correlation coefficients at item level were calculated between the judgement data of the experts and of the teachers (test-curriculum matching indices and OTL rates). For the* Written Test, *no trend could be observed, because of lack of data, but on the Performance Assessment, there was an increased alignment, with insignificant coefficients in 1995, which became significant by 1999/2000. In 1999/2000, the coefficients were r=0.44 on the* Written Test *and r=0.61 on the Performance Assessment.*

Several aspects hindered the answering of the second research question. First, insufficient OTL data on the Written Test *in 1995 disabled several comparisons. Second, the OTL-testing data on the Performance Assessment increased to a large extent in 2000, witnessing of a changing opinion of mathematics teachers on practical tests, but also causing a possible ceiling effect in 2000. The picture arising from the remaining data shows the following*:

1. *The data at the level of the intended curriculum show a considerable change in the judgement of the appropriateness of the two tests, with a declining preference towards the Performance Assessment. In 2000, the curriculum experts pointed to a larger extent at 'easier' items.*
2. *The data at the level of the implemented curriculum show a preference towards the* Written Test, *but the appropriateness of the Performance Assessment in light of the implemented curriculum has increased.*
3. *The data at the attained curriculum have remained stable, only showing minimal change.*
4. *Within the period of four/five years, the data from the intended and attained curriculum show an increased alignment (see section 6.3.4). Without establishing a trend, the data from the implemented curriculum show a clear alignment with the other data in 1999/2000 (see section 6.4.4 and this current section).*

All data were gathered against the background of a newly introduced, RME-*based curriculum. In 1995, the students tested in the METRIC study were from the first cohort of students learning through the new intended curriculum. From the perspective of the* RME-*based curriculum, the curriculum experts judged the Performance Assessment to match better with the intended curriculum than the* Written Test. *This was obviously not the case for the implemented curriculum*: *a large number of teachers indicated that they had not covered the content of*

items from the Performance Assessment. This had changed by 2000, with increased OTL-covered rates and even more increased OTL-testing rates. At the same time, in 2000 many teachers indicated that items from the Written Test *were suitable for inclusion into a test covering all content taught until test administration. This is possibly related to the slight, but significant improved achievement on the* Written Test. *On the other hand, teachers' changing attitude towards the Performance Assessment cannot be related to any change in students' achievement on this test.*
This answers the second research question, as far as possible.

Figure 6.11: Student working on the task Packaging from the TIMSS Performance Assessment

Chapter 7

Conclusion

~ Nunya, adidoe, asi metunee o.~

Wisdom is like a baobab tree; not one individual can embrace it.
(GHANAIAN PROVERB)

This last chapter concludes the METRIC study. Section 7.1 summarises the study, by recapitulating on its questions, its methods and its findings. The subsequent section discusses how the METRIC study can be positioned in the research field, methodologically and scientifically. In the last section, recommendations are given on further policies in mathematics education and for further research.

7.1 SUMMARY

The story of the METRIC study started in 1995, when the TIMSS-95 study was conducted. TIMSS-95 was a large-scale international comparative study in science and mathematics education. Students from many countries were given an identical test, apart from differences in the language used. By comparing the average achievements of students between countries, stakeholders were enabled to analyse the performance of their educational system from an international perspective. The Netherlands was one out of 41 countries participating in TIMSS-95.

Within TIMSS-95, two different tests, combining the subjects of mathematics and science, were administered among Dutch grade 8 students. One test was a

standardised paper-and-pencil test, which consisted of many multiple choice questions and a few free response questions. This test was named the Written Test. The other test brought students into a laboratory environment asking them to conduct empirical investigations. This test was named the Performance Assessment. In 1996, the international comparative results were published (Beaton, et al., 1996; Harmon et al., 1997). Stakeholders raised their eyebrows over Dutch students' mathematics achievement results. The results were as follows: on the mathematics items of the Written Test, the Dutch students scored significantly above the international average score, while on the mathematics tasks of the Performance Assessment, the Dutch students scored near the international average score. This was considered a discrepancy between the achievement results on the two tests (see chapter 1).
The difference between results on the two tests was considered even more striking, as Dutch mathematics curriculum experts had raised their concern about the mathematics items in the Written Test (De Lange, 1997a, 1997b). These items were not considered suitable to test for knowledge and skills, which students learnt in Dutch mathematics classrooms. Curriculum experts objected to the large number of multiple choice items and against the 'bare-ness' of the items. The mathematics items in the Written Test contrasted with items in Dutch mathematics education, which integrated mathematics and context. The treatise, which had helped to shape the Dutch intended mathematics curriculum, is known as Realistic Mathematics Education (RME). Backgrounds of RME were given in chapter 2. The RME-based curriculum at grade 8 level was expected to match better with the Performance Assessment than with the Written Test. Thus, it came as a surprise that students' achievement results were contradicting these expectations. This was considered a discrepancy between curriculum experts' views and students' achievement.

The discrepancies resulted in the so-called METRIC study. It aimed at finding explanations for the discrepancies by replicating both tests in 1999/2000. First, there was the discrepancy between the results on the two tests, which was termed the *inter-test achievement discrepancy.* The first research question focused on it. It was formulated as follows:

> *To what extent does a repeat of the TIMSS Written Test in 1999 and a repeat of the TIMSS Performance Assessment in 2000, result in an inter-test achievement discrepancy similar to 1995?*

Second, there was the discrepancy between the intended curriculum and the attained curriculum, which was termed the *intra-curricular discrepancy*. The discrepancy occurred on both tests: the Written Test was not considered to match well with the intended curriculum, while students' achievement was above the international average. At the same time, the Performance Assessment was considered to match well with the intended curriculum, while students' achievement was near the international average. To study the intra-curricular discrepancies, a conceptual framework for curricular appearances was used, as developed by IEA and adapted for TIMSS (Robitaille et al., 1993). This framework consists of three levels: the macro, the meso and the micro level. At each level, the curriculum has a different appearance. These appearances were distinguished as follows:

- the *intended* curriculum is found at system level; it consists of what society requires students to master,
- the *implemented* curriculum is found at school and classroom level; it consists of the content as it is interpreted by the teachers and taught to the students, and
- the *attained* curriculum is found at student level; it consists of the knowledge and skills that students have acquired indeed.

The METRIC study used the three appearances of the mathematics curriculum at grade 8 in the Netherlands to operationalise its second research questions on the intra-curricular discrepancies. The question was formulated as follows:

> *To what extent are students' results on both tests at both periods in time aligned with (a) the appropriateness of the tests in light of the intended curriculum, (b) the appropriateness of the tests in light of the implemented curriculum and (c) possible discrepancies between these?*

The study derived its name form the conceptual framework. The six letters in the acronym METRIC stood for a study in **M**athematics **E**ducation, investigating **T**rends in, and **R**elations between the **I**ntended, implemented and attained **C**urriculum.

After a study on the history of the RME-based curriculum (chapter 2) and a study of literature on how the three curricular appearances can be described and analysed (chapter 3), methods were developed to tackle the research questions.

First, the two tests were replicated to analyse, whether the *inter-test achievement discrepancy* re-occurred. Second, data on the appropriateness of the tests in light of the intended curriculum were gathered. To analyse the appropriateness of the two tests, curriculum experts were invited to indicate for each test item whether it matched with the intended curriculum. This judgement was an operationalisation of the intended curriculum at macro level, based on instruments used for the TIMSS Test Curriculum Matching Analysis (TCMA) (Beaton, 1998; Beaton et al., 1996). By combining students' achievement and experts' judgement on each test item, discrepancies between the attained curriculum and the appropriateness of the items in light of the intended curriculum were captured.
Third, at the level of the implemented curriculum, mathematics teachers were asked to indicate for each item whether the content matched with the content taught before test administration. This judgement yielded an indication, on whether students had received an opportunity to learn (OTL) the knowledge and skills, which were required to master the items. This was an operationalisation of the implemented curriculum, based on a valid and reliable instrument developed by De Haan (1992). By combining students' achievement and teachers' judgement, discrepancies the attained curriculum and the appropriateness of the tests in light of the implemented attained curriculum were captured. Similarly, by combining teachers' judgements and experts' judgement on all test items, discrepancies between the appropriateness of the items in light of the intended and the implemented curriculum were captured.

The instruments for measuring students' achievement consisted of the Written Test and the Performance Assessment. Both tests had been carefully developed in a process, in which subject matter specialists from all participating countries were consulted. Besides free response items, the Written Test contained many multiple choice questions, which enabled automated and uniform scoring, ensuring reliable and comparable results between countries. The robustness of the Written Test was high, as TCMA showed. When a selection of items was made, and achievement was re-calculated based on the selected items only, the relative achievement results between countries did not change significantly. The items in the Written Test covered a wide range of topics, from primary school level arithmetic, to transformation geometry and algebra on substituting values into functions (Beaton et al, 1996; Mullis et al., 2000). Most countries participating in TIMSS had indicated that the Written Test matched well with

their intended curriculum. When comparing the matching of the Written Test with the intended curricula of nations, it turned out that the Netherlands differed from the international mainstream (see chapter 1, section 1.1.2).
The second TIMSS test was the Performance Assessment. With little experience in conducting such innovative practical tests on a large scale, the Performance Assessment was considered less reliable than the Written Test. The reliability was affected by the complexity of coding students' answers on empirical investigations (Haertel & Linn 1996; Linn, & Baker, 1996; Zuzovsky, 1999). The replication of the Performance Assessment for the METRIC study confirmed this observation (see chapter 4, sections 4.4.4 and 4.4.5). On the other hand, the validity of practical tests like the Performance Assessment was considered high (e.g. Birenbaum & Dochy, 1996; Burton, 1996; Clarke, 1996; Darling-Hammond & Ancess 1996; Hermans, 1992; Kind, 1997; Niss, 1993; Wolf, 1994; Zuzovsky & Harmon, 1999).

The METRIC study gathered data through the above named instruments. It drew on the data, which were already compiled in 1995, and on the international data collection of the Written Test in 1999. In 2000, the METRIC study replicated the Performance Assessment to complete the research design, yielding four sub-studies: WT-1995, PA-1995, WT-1999, and PA-2000. With two tests and three curricular appearances, there were six measurements in 1995 and six measurements in 1999/2000. Each measurement yielded data at item level (144 items in the Written Test, 31 items in the Performance Assessment). The judgements in light of the intended curriculum yielded item-curriculum matching indices, the judgements in light of the implemented curriculum yielded OTL rates, and students' achievement yielded p-values. In chapter 4, the research design of the twelve measurements was described.
The METRIC study was initiated after the 1995 measurements had been carried out. To ensure comparability in time, the instruments were kept intact for replication, guided by the motto *if you want to measure change, do not change the measure*. Nevertheless, the comparison between the two tests, and between the three curricular appearances was a complex exercise:

- at the level of the intended curriculum, the data on the appropriateness of items from the Performance Assessment were available on a ratio scale (as percentages) and these nuanced the experts' judgement. In this way, the judgement could be expressed as 'more or less' matching with the intended curriculum. The data on the appropriateness of items from the Written Test

were only available in a binary yes/no scale. The two different scales hindered the inter-test comparison at the level of the intended curriculum, unless one of the data sets was transformed onto the scale of the other (see section 4.3.4).

- at the level of the implemented curriculum, the OTL data were gathered for all test items in WT-1999, PA-1995 and PA-2000 (see section 4.3.4). For WT-1995, the OTL data were only available on 16 items (10% of the test). The 16 items were not a random selection of the whole test, as they had been selected on their match with the intended curriculum (as part of the National Option Test, NOT, see section 1.2.2 and 3.1.4). This omission could not be compensated anymore within the METRIC study.
- at the level of the implemented curriculum, the OTL data on the two tests were compiled using a different format. It was caused by the different contexts, in which the data had been compiled in 1995. The OTL data for WT-1995 were compiled as part of NOT, asking the teachers 'would you include the item into a test, which tests for all content taught before test administration?'. The OTL question was phrased differently in the teacher' instrument for PA-1995, where the data were compiled as a national option within the international data collection. Here, teachers were asked two different questions: 'was the content of the item covered?' and 'would you include the item into a practical test?' For the sake of trend comparison, the instruments were kept intact for the replicated studies, although this hampered the inter-test comparison at the level of the implemented curriculum. As a result, the Written Test yielded an OTL rate for each item, while the Performance Assessment yielded two different rates for each item: the OTL-covered rate and the OTL-testing rate. The OTL-covered rate indicated the number of teachers who had covered the content of an item before test administration, while the OTL-testing rate indicated the number of teachers who alleged that they could include the test items into a practical test of their own making (see section 4.3.4).
- At the level of the attained curriculum, the METRIC study closely scrutinised the trend comparability of students' achievement between 1995 and 1999/2000. On the Written Test, only 41 items had been left untouched between 1995 and 1999. All other items had undergone slight adaptations (in their notations, in their numbers or in their contexts). On a number of items, these small adaptations resulted in significant different student scores (see section 5.4.1). The items of the Performance Assessment were identical in 1995 and in 2000. However, a number of these items showed unreliable

achievement results, because (1) the test circumstances had slightly changed, and (2) the coding of students' answers was not fully consistent between 1995 and 2000 (see section 4.4.5). The unreliable achievement results were discarded from the analysis. This left the METRIC study with 41 items from the Written Test and 28 items across five tasks from the Performance Assessment for trend comparison at the level of the attained curriculum.

The large database of the METRIC study was described in chapter 5, where trends in the data were presented. These were as follows:

At the level of the intended curriculum

1. The judgement on the appropriateness of the Written Test in light of the intended curriculum showed that 69% of the test items matched with the intended curriculum in 1995, and 71% in 1999. Thus, the overall test-curriculum matching rate remained virtually equal, but the two proportions overlapped on only 53% of the items. The judgement had changed on 33% of the test items between 1995 and 1999 (see section 5.2.1).
2. The judgement on the appropriateness of the Performance Assessment in light of the intended curriculum was measured at item level on a ratio scale. In 1995, the average item-curriculum matching index was 83, while in 2000 it was 72. To compare the appropriateness between the two tests, these results were not suitable, because the scales differed in nature. Therefore, the data were transformed onto a nominal scale, which indicated that 88% of the test items in 1995, and 85% of the test items in 2000 matched with the intended curriculum (see section 5.2.2).
3. In 1995, Dutch curriculum experts considered the Performance Assessment to be better aligned with the intended mathematics curriculum than the Written Test. The judgement on the overall appropriateness of the Written Test did not change substantially between 1995 and 1999. According to the experts' judgement, the appropriateness of the Performance Assessment decreased between 1995 and 2000 (see section 5.2.3).

At the level of the implemented curriculum

4. In 1999, the judgement on the appropriateness of the Written Test in light of the implemented curriculum showed a good match. The content of items had been considered suitable for an imaginary test by, on average, 82% of the teachers. This represented a vast majority of the teachers. Only six items in

the Written Test had an OTL-rate below 50. No trend could be established, because for 1995, there were only data available for 16 items (with an average OTL rate of 93). The judgements on this subset of items were very consistent with the 1999 judgements on these same items (again, with an average OTL rate of 93). However, it was too far fetching to extrapolate from this consistency in teachers' judgements that the 1995 judgement would also have shown a good match with the implemented curriculum (see section 5.3.1).

5. The judgement on the appropriateness of the Performance Assessment in light of the implemented curriculum was low in 1995. The OTL-covered rate was 38. The figure rose to 58 by 2000. Both average rates were obviously lower than the OTL rate on the Written Test of 1999 (82).
   The average OTL-testing rates of items from the Performance Assessment increased from 51 in 1995 to 76 in 2000. However, the increased OTL-testing rate reflects teachers' intentions on organising a practical test and it does not necessarily reflect actual classroom practice (see section 5.3.2).
6. Teachers' judgement revealed that the Written Test matched better with the implemented curriculum than the Performance Assessment (see section 5.3.3).

<u>At the level of the attained curriculum</u>

7. Between 1995 and 1999, students' average achievement on the Written Test increased slightly, but not significantly, when measured through the t-test. However, the increase occurred on almost all items, and by using the sign test, the increase proved to be significant. This was an indicator of slightly improved students' knowledge and skills (see section 5.4.1).
8. Students' average achievement on the Performance Assessment did not show any significant change between 1995 and 2000 (see section 5.4.2).

In Tables 7.1, and 7.2, the trend results of the METRIC study, as described in chapter 5, are summarised. The descriptive statistics at all curricular appearances are given in Table 7.1. Trend correlation coefficients between the data of 1995 and 1999/2000 are added in Table 7.2. The coefficients are highest at the level of the attained curriculum, showing the high stability of students' achievement. The coefficients are lower for the OTL rates, and lowest for the item-curriculum matching indices.

Table 7.1: Descriptive statistics of results in the METRIC study

Sub-study	**Intended curriculum** *Test-curriculum matching index*	**Implemented curriculum** *Avg OTL rate*	**Attained curriculum** *Achievement Avg p-value (SE)*
WT-1995	69 (nom.) -- (rat.)	--	72 (3)
WT-1999	71 (nom.) -- (rat.)	82	75 (3)
PA-1995	88 (nom.) 83 (rat.)	38 (cov.) 51 (tst.)	67 (3)
PA-2000	85 (nom.) 72 (rat.)	58 (cov.) 76 (tst.)	68 (3)

Note: Dashes indicate data are unavailable;
nom.=on a nominal scale; rat.=on a ratio scale; cov.=OTL-covered; tst.=.OTL-testing.

Table 7.2: Trend correlations in the METRIC study

Paired sub-studies	**Intended curriculum** *Correlation between item-curriculum matching indices*	**Implemented curriculum** *Correlation between OTL rates*	**Attained curriculum** *Correlation between p-values*
WT-1995 – WT-1999	0.19	--	0.97
PA-1995 – PA-2000	0.41	0.75 (cov.) 0.64 (tst.)	0.96

Note: Dashes indicate data are unavailable;
cov.=OTL-covered; tst.=OTL-testing.

Table 7.3: Intra-curricular correlation coefficients in the METRIC study

Sub-study	**Between intended and attained curriculum**	**Between implemented and attained curriculum**	**Between intended and implemented curriculum**
WT-1995	* (nom.)	--	--
WT-1999	0.23 (nom.)	0.41	0.44
PA-1995	0.55 (rat.)	* (cov.) * (tst.)	* (cov.) 0.39 (tst.)
PA-2000	0.77 (rat.)	0.44 (cov.) * (tst.)	0.61 (cov.) * (tst.)

Note: Dashes indicate data are unavailable;
nom.=on a nominal scale; rat.=on a ratio scale; cov.=OTL-covered; tst.=OTL-testing;
* Insignificant correlation.

After the presentation of the METRIC database in chapter 5, the research questions were examined in chapter 6. The first question pertained the *inter-test achievement discrepancy*. This question was based on the discrepancy observed between students' international achievement results on WT-1995 and PA-1995. This discrepancy faded when observing the scores more in-depth, including the standard errors, which indicated the probability margins of the results. When considering these margins of precision, Dutch students' scores were comparable to the students' scores of a large group of countries, both on WT-1995 and PA-1995. Additionally, the replication of the Performance Assessment turned out to be useful for finding evidence on the reliability of instruments of PA-1995. Test circumstances had been difficult to control and the coding of free-response answers had been a complex exercise. After discarding the unreliable test items from analysis, the discrepancy turned out to be minimal. Therefore, the *inter-test achievement discrepancy* of 1995 was reduced to a non-existing phenomenon (see section 6.2). On both WT-1995 and PA-1995, Dutch students' achievement in mathematics was above the international average, just like it had already been in an earlier IEA study, the Second International Mathematics Study, conducted in 1982 (cf. Section 3.1.4).
The replication of both WT-1995 and PA-1995 showed that Dutch students' achievements in mathematics had only changed minimally. The score on the Written Test had increased slightly, but significantly (by means of the sign test, see section 5.4.1), while the score on the Performance Assessment had not changed significantly. Thus, the ocean liner had changed its course to a minor degree. Whether the vessel was heading in the intended direction, was another issue.

The second research question of the METRIC study linked the data gained at the levels of the intended, implemented and attained curriculum. By searching for intra-curricular relations, possible trends in students' achievement could be explained. The intended and attained curriculum were linked by separating the items according to the judgement by experts. There were items that matched with the intended curriculum in 1995, and other items that matched with the intended curriculum in 1999/2000. The judgements on the appropriateness of both tests in light of the intended curriculum had changed for 25%-33% of the items between 1995 and 1999/2000. The changes reflected the adaptations of the intended curriculum between the two test administrations (less mental arithmetic, more calculator usage), and the adaptations in the notation of multiplication with variables ('7a' in WT-1995 had been altered into '7•a' in WT-1999). Other reasons

for the change, given in chapter 5, were: confusion in judging the items (e.g. ignoring the format, while this could not be ignored), and a shift in interpretations of the intended curriculum (see section 5.2.1). The possibility of a low reliability of the instrument was also stated, in the case of the instruments on judging the Written Test. However, the parallel instrument for measuring the judgement on the appropriateness of the Performance Assessment had an acceptable reliability (α=0.83 and α=0.79, see section 4.4.4), and the results from that measurement also showed the shift in experts' judgement.
Moreover, the METRIC study discovered that students' achievement on the items that matched the intended curriculum in 1995 was lower than students' achievement on the items that matched the intended curriculum in 1999/2000. This was caused by the fact that, in 1995, the experts had preferred 'harder' items than in 1999/2000, while they had no access to students' achievement results. It was concluded that, so to say, the intended curriculum had made concessions towards the attained curriculum. It meant that, in 1999/2000, more students were able to complete the items that matched with the intended curriculum, not because students' achievements had changed, but because the judgement on the appropriateness of items had changed. This finding was observed both on the Written Test and on the Performance Assessment, independently (see sections 6.3.2 and 6.3.3). The correlation coefficients between experts' judgement and students' achievement confirmed the increased alignment (see Table 7.3).

For the link between the implemented and attained curriculum, correlation coefficients at item level were calculated between teachers' judgement data and students' achievement data (OTL rates and p-values). The calculation could not be carried out for WT-1995 because of lack of data. For PA-1995, the available data yielded no significant correlation coefficient. This meant that there was no clear relation between teachers' judgement and students achievement. Students were able to perform well on items that the teachers indicated as inappropriate, or the other way around. By 1999/2000, the situation had changed. The OTL data collection was completed for all items in the Written Test. Combining the OTL data with achievement data resulted in a correlation coefficient of r=0.41 for WT-1999. When comparing to other research results, a coefficient of r=0.41 was considered satisfactory (see section 6.4.2). On PA-2000, both OTL judgements by the teachers (OTL-covered and OTL-testing) had increased overall since 1995, and the combination with the achievement results yielded r=0.44 for OTL-covered. However, no significant correlation for OTL-testing.

The latter could be ascribed to mathematics teachers' unfamiliarity with practical tests; their increased judgements on including items into an imaginary practical test was noteworthy in itself, but it might have yielded imaginary answers (see section 6.4.3). As a result, a trend on the link between implemented and attained curriculum could only be based on the correlation of OTL-covered data and achievement results from the Performance Assessment. Despite the limitation, the data showed that, in 1999/2000, the attained and the implemented curriculum were better aligned than in 1995 (see sections 6.4.2 and 6.4.3).

For the link between intended and implemented curriculum, again, the items were separated according to the judgement by the experts. Additionally, correlation coefficients at item level were calculated between experts' and teachers' judgement data (test-curriculum matching indices and OTL rates). Again, the calculations could not be made for WT-1995 because of lack of data. The data from the Performance Assessment showed an increased alignment, with insignificant coefficients in 1995, indicating that there was no clear relation between the data. In 1999/2000, the correlation had become significant r=0.61 for the Performance Assessment (for OTL-covered). The alignment was confirmed by the correlation coefficient from WT-1999, which was r=0.44. This meant, that an increasing number of teachers indicated that the intended curriculum indeed matched with implemented curriculum. Contrariwise, it could also mean that the experts were to an increasing rate indicating that the implemented curriculum was the content that was indeed intended to be covered (6.5.2 and 6.5.3).
The data indicate that between 1995 and 1999/2000:

1. The appropriateness of the Written Test in light of the intended curriculum changed considerably, but with only approximately 70% of the test items matching with the intended curriculum, the appropriateness was relatively low when compared internationally. The appropriateness of the Performance Assessment in light of the intended curriculum decreased. However, the Performance Assessment remained more appropriate than the Written Test.
2. In 1999, the appropriateness of the Written Test in light of the implemented curriculum was satisfactory (insufficient data available for 1995); the appropriateness of the Performance Assessment in light of the implemented curriculum was low in 1995, but had increased in 2000.
3. Students' achievement remained highly stable, only showing minimal change.
4. Within a timeframe of four/five years, the data from the three curricular appearances show an increased alignment between all three levels.

Hence, it appeared that in 1995, the ocean liner followed a course, which deviated from the course steered, and from the course due. By 1999/2000, the vessel had changed its course to a minor degree; while the course steered and the course due had been altered in such a way, that the three aligned better.

Interpreting the findings

The data of the METRIC study and the resulting analyses should be interpreted against the background of the curriculum reform, which was legislated in 1993. The new mathematics curriculum had been designed by an enthusiastic project group, W12-16. Mathematics was to be integrated into contexts, focusing on its usefulness. Multiple choice items were renounced. The objections of the curriculum developers against the many multiple choice and 'bare' mathematics items in the TIMSS Written Test were profound, as the items reminded of the abandoned curriculum. The ardour of the W12-16 project group spread out high expectations, for example towards the Performance Assessment. Although there was no assessment practice similar to this test, the experts expected it to match with the design of the intended curriculum. Although the intended curriculum was obviously formulated in core objectives, it was not yet fully known to what extent it could be implemented or was attainable.

In 1995, mathematics teachers did not share the enthusiasm. They had been informed on the new curriculum, but its intentions had not yet settled, for example on how the practical usefulness of mathematics was to be implemented in classrooms. The judgement on the appropriateness of the Performance Assessment indicated that on average only 38% of the teachers had covered the content tested in the items, and only half the teachers (51%) indicated that they could include items from the Performance Assessment into a practical test of their own making. In 1995, probably many teachers were still engaged in a process of discovering and appraising the intended curriculum, while offering their students a mix of traditional and new content, without setting priorities.

Four years later, the curriculum experts had become more aware of the boundaries and hurdles of the initial core objectives. These objectives were reviewed, and probably, the enthusiasm of the W12-16 project group had dwindled as well. The experts changed their judgement on the appropriateness of items in the Written Test. Their changed judgement pointed at 'easier' items. Their approval of the Performance Assessment decreased, in particular on items, which concluded the tasks (the more reflective tasks).

In 1999, many teachers indicated the items from the Written Test suitable for a

test of their own making covering all content taught until test administration (the average OTL rate was 82). This could mean that the content was covered indeed, or that the teachers estimated that their students were 'ready' for them. Despite the multiple choice format of many items, and their 'bare' nature (not integrating context and mathematics), only six out of 144 items in the Written Test had an OTL-rate below 50. This could mean, that many Dutch teachers still offered their students items similar to those from the Written Test, and thus, reflecting the abandoned curriculum.

A trailing implementation of the new curriculum, and the adherence of teachers to more traditional assessment practices could explain the slight but significant improvement of students' achievement on the Written Test between 1995 and 1999. The prudence of teachers could also explain the lack of improvement of students' achievement on the Performance Assessment. Despite the prudence, in 1999, teachers' judgement had become better aligned with the experts' judgement ($r=0.44$ for WT-1999 and $r=0.61$ for PA-2000, OTL-covered). This is another noteworthy result of the METRIC study, as it indicates that the intended curriculum aligned better with the implemented curriculum. This could have resulted in the slight, but significant improvement of students' achievement on the particular items that matched with the intended curriculum (see section 6.3.2). At the same time, the mathematics teachers increased their appreciation of the Performance Assessment considerably (OTL-testing). This result pointed at a possible change in teachers' attitude. It could be another indication of teachers' increased adherence to the new, intended curriculum and of alignment of intended and implemented curriculum. This is noteworthy. However, the changing attitudes did not lead to any significant improvement in students' achievement on the practical test (see section 6.3.3).

From an international perspective, it appears that the achievement results show that the Dutch educational system yields satisfying results. However, it would require further study to establish whether the achievement results were satisfactory by national standards. This would require further investigation of the nature of the two tests: which items were more elementary or more challenging in nature, to what extent were mathematical activities required, and to what extent was the Dutch intended curriculum for junior secondary schools covered? Thus, the Dutch ocean liner seems to have a captain who keeps a relatively correct course, when compared to the course of other nations' vessels. The course became closer to what the authorities' intended, but this does not mean its heading was correct by national standards.

7.2 REFLECTION

7.2.1 The effect of small differences

The METRIC study has shown that small differences can have large effects. This observation was made on several occasions. This effect was perceived at the level of the attained curriculum. The small adaptations in the equipment of the Performance Assessment induced different testing circumstances, which resulted in different achievements. The difference between an easy to use, plastic balance and delicate, metal scales was allowed according to the TIMSS protocols. However, the first reached its point of balance faster and this gave students more time to complete the remainder of their tasks. Thus, a small difference in equipment could cause a large difference in scores.
Similarly, the small adaptations through the 'cloning' of test items (making small adaptations, e.g. by changing the numbers) made some items harder and others easier. In one example from the Written Test, the item was changed into a more attractive context (V02 - a boy selecting a subscription to a journal instead of a businessman selecting office space to rent). This small adjustment almost doubled the number of Dutch students being able to complete the item correctly.

Small differences having large effects was also noticed in the METRIC study in the reports of average students' scores of nations. This was observed when the results of the tests were presented through league tables in which the participating countries were ranked in their order of score (e.g. Bos & Vos, 2000; Beaton et al., 1996; Harmon et al., 1997; Kuiper et al., 1997, 1999; Mullis et al., 2000). Insignificant differences between countries may seem enlarged, when the results are expressed on the ordinal scale of rankings. The rankings are popular among journalists and politicians, making these league tables go a long way. However, the underlying probability intervals become marginalized. In fact, the METRIC study was initiated, partly, because of this indiscretion. The observed discrepancy between Dutch students' achievements on the two different TIMSS-95 tests had capitalised on the rankings in the league table. That observation was deficient for two reasons: it was disproportioned by the rankings, and it was affected by the results on the task *Plasticine* from the Performance Assessment, which consisted of unreliable data and pulled down the Dutch score. The METRIC study established that the Dutch students' results on the Performance Assessment were of the same magnitude as the results on the Written Test, when comparing internationally, while ignoring the rankings and focusing on scores.

The observation of small differences having large effects also occurred at the level of the intended and the implemented curriculum. Teachers and experts had to judge items on a nominal yes/no scale. There was no option to differentiate between the two ends, forcing the respondents to make long deliberations and find small differences to make an absolute choice. The exercise was even more complex as the judgement had to interpolate between the ability tracks of students and to ignore the format of the items (see section 4.3.5).
The absence of any compromise was repeated when aggregating the experts' judgements. For the Written Test, the aggregation of experts' judgement was carried out by asking them to reach consensus, the result of which was, again, given on the binary yes/no scale. This method yielded absolute answers in cases where a middle course was needed. Therefore, the rigid method was not applied for the judgements by experts on the Performance Assessment, or for the teachers' judgement on either test. In these cases, the judgements were aggregated into a percentage on a ratio scale. The percentage indicated the number of teachers or experts who judged the item as 'fit'. If many teachers or experts had given their 'yes', the rate would be high; if many teachers or experts had given their 'no', the rate would be low. This gave room to nuance the judgement on the appropriateness of items. The need for a middle course occurred, in particular, on items that were formulated in an unfamiliar way. For example, the mathematical content of an item could be considered relevant, but the presentation of an item could differ from practice (e.g. the multiple choice format, or the lack of context). Expressed on a binary yes/no scale, the judgements on the appropriateness of the Written Test in light of the intended curriculum produced very diverging results between 1995 and 1999 ($r=0.19$). The judgements on the appropriateness of the Performance Assessment, which were aggregated on a ratio scale, showed more consistency between 1995 and 2000 ($r=0.41$). The same judgements on the Performance Assessment, when transformed onto a binary yes/no scale, seemed inconsistent. Thus, the METRIC study showed that the divergence of the judgement on the appropriateness of the Written Test between 1995 and 1999 could have been caused by the method of aggregation. This showed, that the scales on which the judgement were represented, did matter indeed (see section 5.2.2 and 5.2.3). It remains an open question whether the consistency would further increase if, for example, both the teachers and the experts were asked to judge the items on a four-point scale, instead of on the binary yes/no scale.

7.2.2 The effect of large-scale curriculum reforms

The METRIC study aimed at finding answers to an observed discrepancy between the 1995 achievement results of Dutch grade 8 students on the TIMSS Written Test and the TIMSS Performance Assessment. The achievement of students was associated with the coinciding curriculum reform for junior secondary schools. At that level, the traditional intended mathematics curriculum of 1968 was radically reformed into an RME-based curriculum in 1993. Some topics were abolished or postponed (e.g. sets, Euclidean geometry, parabolas) and many topics were reshaped in an approach, which embedded mathematical activities into a meaningful context. The students tested in 1995 in TIMSS were selected from the first cohort for whom the new RME-based curriculum was compulsory.

After the TIMSS results were published in 1997, the 'underachievement' of Dutch students on the Performance Assessment (when comparing the Dutch rankings of the Performance Assessment and the Written Test) was explained by pointing at a possible delay in the implementation of the new curriculum. It was alleged (see e.g. Kuiper et al., 1997) that teachers were still inexperienced with the new curriculum, which would prepare students to do well on the practical test. After a few years, teachers would be more experienced with training their students on inquiry-based tests and thus the Dutch students' would perform better in the international comparison. As a result, a replication of the Performance Assessment was called for, and the METRIC study originated. However, as the METRIC study has shown, the achievement of Dutch students has remained stable, not in the least on the Performance Assessment. The METRIC study has also shown that teachers showed a growing inclination towards the use of practical tests. The changing attitude is noteworthy but has not paid off (yet) in increased students' scores. It could therefore be alleged, that the curriculum implementation is taking many more years than the four/five years that the METRIC study covered.

One of the reasons given for delayed curriculum implementation, is the lack of exemplary curriculum materials (Van den Akker, 1988, 1998; Kuiper 1993), and the continuous sustenance of teachers' inspiration (Kuipers, 1999). However, the exciting curriculum materials produced by the curriculum designers of W12-16 are no longer available (Goddijn & Kindt, 2001). The initial ideas on assessment as expressed by Dekker (1993), Van Dormolen (1992) and Kok et al. (1992) were diluted to standard paper-and-pencil tests, with mostly short answer questions

(see e.g. Appendix A). Therefore, there is obviously need for valuable additional assessment materials to support continuous initiatives to better implement the intended curriculum. The TIMSS Performance Assessment can serve as such. The judgement by the curriculum experts testifies of its appropriateness in light of the intended curriculum. Therefore, it can serve to illustrate the intentions of the curriculum to teachers.

The METRIC study observed that the curriculum experts showed a considerable shift in their judgement on the appropriateness of the test items. Items that were considered covered in 1995, were considered not covered in 1999/2000, or the other way around. The shift in judgements is probably a reflection of the review of the legislated core objectives in 1998. However, it is also possible that changing interpretations of the curriculum by the experts caused the shift. The core objectives, as formulated in 1993, intentionally left room for interpretations (ten Hove & Van der Zwaart, 1993; Van der Zwaart & Boertien, 1998). Thus, in 1995, the outlines of the intended curriculum were probably still dim. It still had to find its final shape. Teachers had to accommodate it to their own situation. They had to choose new teaching materials, develop new teaching strategies and to find new beliefs (Fullan, 1991). They could find their information in the core objectives, in additional explanatory texts, in the experimental national tests and exams as set by the National Institute for Educational Measurement (Cito), and in commercial textbooks based on the new curriculum. As reported in Kuiper et al. (1997) in 1995, still 50% of the teachers used the textbooks based on the abandoned curriculum. Therefore, it can be alleged that the intended and implemented curriculum for grade 8 in the Netherlands were truly in transition, each in their own way.

The intended and implemented curriculum, being in transition, did not substantially affect students' achievement on the two TIMSS tests. Dutch students' achievement in the international comparison remained constant, continuing the high score they attained decades earlier in 1982 (see section 3.1.4). Students' achievement stayed right on course, just slightly changing. To explain this stability, the attained curriculum needs to be seen in perspective. Several authors have expressed, that the process of curriculum implementation is disorderly (e.g. Van den Akker, 1998; Fullan, 1991; Kuiper 1993; Kuiper et al., 2001).

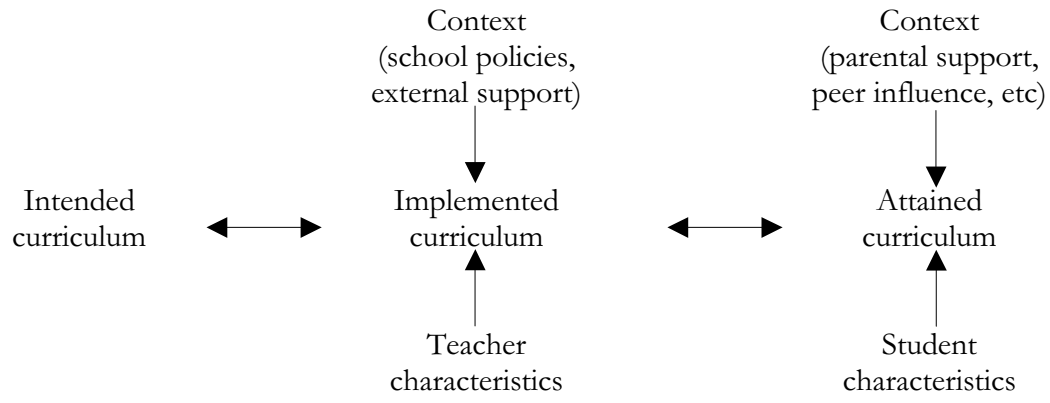


Figure 7.1: Conceptual framework for curriculum implementation
Source: Kuiper et al. (2001), adapted from Van den Akker (1998).

The conceptual framework for curriculum implementation in Figure 7.1. illustrates that many variables affect an attained curriculum. The three curricular appearances are central in the figure, but there are strong variables added. The implemented curriculum has its own context (e.g. within-school variables, such as school policies, and external support), and depends directly on teachers' characteristics (e.g. training, experience with prior reforms, beliefs). Similarly, the attained curriculum has its own context (e.g. parental support, peer influence) and depends directly on students' characteristics (e.g. attitudes, beliefs). In this framework, the intended curriculum is a variable on one end, competing with many other strong variables. This illustrates that an intended curriculum is not the only variable affecting the implemented curriculum. Similarly, an implemented curriculum is not the only variable affecting students' achievement. With other variables remaining unaltered, there is a sound base for a stable attained curriculum. Hence, it requires careful planning, and deliberate, continuous action to really change the course of an ocean liner. Or, to paraphrase a metaphor in Van den Akker (1998): the choppy breakers at the surface cause only ripples at the bottom of the sea.

7.2.3 Educational research and mathematics education research

The METRIC study originated from IEA's TIMSS study. The key-role players in that study earned their merits in educational research, in particular in psychometric and methodological areas, drawing on experience and research in large-scale surveys. From the onset, subject matter specialists were consulted. But to date, TIMSS has not yet lead to a fruitful co-operation between mathematics education researchers and educational researchers. Some mathematics educators

have expressed scepticism towards TIMSS, among others, because they doubted the relevance of large-scale international research and because of the lack of innovative mathematics educators involved (see Keitel, 2000). The dichotomy between educational specialists and mathematics education researchers is caused by the difference in research focus. Generally, educational researchers focus on more comprehensive, often policy-relevant, research questions, not bound to any specific subject domain. Mathematics education researchers aim at understanding how students learn mathematics and how mathematics education can best be changed or improved. Between the two fields, TIMSS tries to run with the hare and hunt with the hounds, without falling between two stools.

TIMSS has offered mathematics education researchers an abundant amount of data. All data are readily available, offering mathematics researchers a wide range of research opportunities. For example, the two-digit code for free response items is an instrument to study students' strategies and misconceptions, but to date, no mathematics education researcher has taken up this challenge. Thus, in future, TIMSS could and should lead to more fruitful co-operation between the two sectors of research.

In the METRIC study, the dichotomy between the two research fields was also felt. The METRIC study used the refined methodology from large-scale educational research, which is not predominant in research on mathematics education. This approach lead for example to the use of large samples enabling generalisation to a national level, and it lead to an emphasis on the validity and reliability of instruments. Mathematics educators have little experience with the first, and the latter is often overlooked in mathematics education research. For example, mathematics educators can refute the use of multiple choice items because they believe they offer little space, for example, for creativity. But on the other hand, this format allows for wide content coverage and facilitates efficient and reliable international comparative studies. Without multiple choice items, these studies would become too complex and too laborious to carry out successfully.

However, the general, not subject-bound background of TIMSS also weakened the operationalisation of the research questions in the METRIC study. For example, for an educational researcher, the test-curriculum match can be a clear dichotomous concept: either a test item is covered by a curriculum or it is not. In this perception test items are plain instrumental units for measuring educational variables. For a mathematics educator, a test item is something totally different. From the perspective of the subject matter specialist, items are multi-

dimensional, with aspects of language, culture, history, instruction, misconceptions and, of course, the format that gives the respondent room for replying to the question posed. In the test-curriculum matching judgement, curriculum experts were asked to ignore the multiple choice format of an item, and just assess its content. This may seem trivial to an educational researcher, while it is a complex exercise to a specialist in mathematics education. Within the METRIC study, there was little room to investigate this aspect, but a qualitative appraisal of items by curriculum experts could well yield a wealth of information on the compositional aspects of the test items (e.g. with respect to format, culture, language, etc.).
This difference in the perception of items also occurred at the level of the implemented curriculum. Teachers had to assess a considerable number of items, of which the content probably had been covered at primary schools. Although it would be practical for secondary school teachers to know what content is taught at primary schools, in practice, there is a considerable gap between these two levels of schooling. Many secondary school teachers do not know exactly what content is taught at primary schools. It is therefore questionable whether the answers of the teachers were completely valid. Additionally, a number of items were difficult to assess, because they originated from a different educational culture. Thus, the poor familiarity of teachers with some items can have hindered the validity of teachers' answers in the METRIC study.

From the perspective of an educational researcher, the OTL instrument of De Haan (1992) had proven its validity and reliability. It was carefully developed in a study, in which the content tested had been taught in the immediate period prior to the test. Thus, the items were conceptualised within the same educational culture, as they originated from the textbooks used. As a result, teachers were familiar with the items, their phrasing and their format. When asked, whether the teachers would include the items into an imaginary test, the familiarity with the items made it easier to identify whether the item had been covered or not. This familiarity aspect was not fully present in the METRIC study, as items were submitted to teachers that looked partly familiar and partly 'strange', with for example items that covered recognisable content, but phrased in an unfamiliar way (e.g. as 'bare' mathematics). It can therefore be questioned whether the validity and reliability of the instrument in the METRIC study were still the same as in the original version of De Haan.

Another example of the dichotomy between the fields of educational research and the field of mathematics educational research, which had consequences for the METRIC study, was the use of the mathematics curriculum framework of Robitaille et al. (1993). This framework can be criticised for its lack of fit with the Dutch RME-based curriculum (cf. Section 3.1.2). Based on the TIMSS framework, test items were designed, resulting in a test, which *"reflects much more the eighth-grade focus on arithmetic in the United States than the eighth-grade focus elsewhere"* (Schmidt, et al., 2001, p. 2.2). Many TIMSS researchers, not being familiar with the subject matter, took the framework for granted, including the allocation of items to topics from the framework. Thus, the task *Rubber Band* once having been indicated as 'science' was from then on ignored as an instrument to measure mathematical skills. The same applied to some items from the Written Test. For example item L09 was labelled as a physics item, because it was said to deal with the conversion of energy. Item O14 was labelled as an earth science item, because it dealt with the size and distance of planets. They read like this:

Item L04 (TIMSS-95, released after 1995)
Machine A and Machine B are each used to clear a field. The table shows how large an area each cleared in 1 hour and how much gasoline each used.

	Area of field cleared in 1 hour	Gasoline used in 1 hour
Machine A	2 hectares	3/4 liter
Machine B	1 hectare	1/2 liter

Which machine is more efficient in converting the energy in gasoline to work? Explain your answer.

Item O14 (TIMSS-95, released after 1995)
The Sun is bigger than the Moon, but they appear to be about the same size when you look at them from the Earth. Why is this?

Both items match well with the Dutch RME-based curriculum (at approximately grade 6 level). They offer a context, in which mathematical activities are needed to find an answer. The first item requires proportional reasoning on the gasoline spent; the second requires proportional reasoning from a spatial perspective. In both cases, the proportional reasoning becomes substantive because of the context. After reshaping the context into a (mental) model, and deducting an

answer, the answer has a meaning in real-life. Few other items in the Written Test show what distinguishes the RME treatise from traditional mathematics. It is therefore illustrative that these exemplary items were not labelled as 'mathematics' in TIMSS, but were traced among the science items. Consequently, the international science achievement results include the report on skills learnt in mathematics classes, and the international mathematics results do exclude the report on applied mathematical activities. In hindsight, it would have been interesting to include Dutch students' results on these items into the national reports on mathematics achievement.

From the Dutch perspective, the challenge remains to improve and to innovate test items in TIMSS, including the development and application of a salient curriculum framework. The exemplary items in Appendix A may serve as such. The METRIC study considered the existing framework used by TIMSS unsuitable for reporting Dutch students' achievement (see section 3.1.2). Additionally, the allocation of TIMSS items to topics and cognitive activities may need improvement. From an RME point of view, it is unclear how and why some labels were dispatched to certain items. For example, item V02 (on the comparison of two advertisements for the rent of an office) was labelled 'data representation' while this item does not ask the students to visualise their data. This item deals with linear relations between variables (the price depends on the area rented) and the comparison of incremental rates. The item could thus also be labelled as algebra because it contains the comparison of two linear relations. It seems that the absence of abstract variables (x and y) and the phrasing within a context, made the TIMSS consultants decide to label it under 'data representation'.

Due to the poor fit of the TIMSS mathematics curriculum framework, the METRIC study did not use the item categories provided by TIMSS nor any other categorisation. It was beyond the scope of the METRIC study to develop, test and apply an RME-based curriculum framework, although an initiative was taken (see section 3.1.1). An RME-based framework is visualised in Figure 7.2. This framework proves to be useful to check whether the four mathematising activities (modelling, reformulating, interpreting, reflecting) are well covered in a test (this will be further explained in the forthcoming section 7.3.4). The framework can be useful, not only in TIMSS, but also in test development by the Dutch National Institute for Educational Measurement (Cito).

Content	Vertical mathematising		Horizontal mathematising	
	Reflecting	Reformulating	Modelling	Interpreting
Arithmetic, measuring and estimating				
Algebra				
Geometry				
Data processing and statistics				

Figure 7.2: Framework for an RME-based curriculum

7.2.4 The innovative nature of the TIMSS Performance Assessment

The METRIC study observed an underestimation of the validity of the TIMSS Performance Assessment, in particular for mathematics education. In 1995, to the international TIMSS researchers, the Performance Assessment was just another test with a questionable reliability. In none of the international publications, any pride over this innovative study for mathematics education shines through. Only the Dutch researchers honoured the validity of this test, and pleaded for its repeat (Bos et al., 2001; Kuiper et al., 1999). This resulted in the METRIC study, and in the replication of the Performance Assessment in the Netherlands. It is unfortunate that no other country accepted the invitation to join in, thus leaving the METRIC study with the premiere to replicate a practical test and thus assess reliability pertaining the comparability of testing circumstances and the consistency of codings. The replication resulted in methodological findings and experience with these kinds of alternative achievement tests.

Originally, the Performance Assessment was developed from a science perspective, allowing to test for empirical skills in physics, biology, chemistry, and so forth. A few mathematics tasks were added because, in TIMSS, all tests cater for both mathematics and science (Zuzovsky & Harmon, 1999). Mathematics could not be left out from the Performance Assessment. This lead to the development of a range of hands-on mathematics tasks. Besides the usual hands-on mathematics tasks on space and shape, the items in the Performance Assessment also related to topics such as numbers and probability. Within the METRIC study, one of the science tasks, *Rubber Band* was also identified as having strong mathematical aspects, as the measurements of the stretching rubber band had to be plotted, described and extrapolated. This task contained a 'decreasingly increasing' relationship between two variables, which is a pre-

calculus concept that is part of the intended RME-based curriculum in the Netherlands.

The mathematical hands-on tasks from the Performance Assessment offer examples for both an operationalisation and an attractive enrichment of the RME-based classroom. Traditional mathematics educators associate manipulatives generally with 'fun mathematics', as opposed to 'real mathematics' (Moyer, 2002). Through the combination of manipulatives with assessment, the contrast between 'fun' and 'real' dissolves. Examinable hands-on tasks do fit well into many mathematics curricula. For example, in the Netherlands, 5% of instructional time is to be spent on *integrated mathematical activities* (Geïntegreerde Wiskundige Activiteiten - GWA). According to teacher guides, in GWA, there is room for students to do investigative project work (Van Dormolen, 1993). However, textbooks have limited their paragraphs on GWA to thematic reproductive items. For example, the thematic chapter can be on the statistics of the Blue Whale, consisting of items in which their size, breeding habits and possible extinction have to be extracted from graphs. The context can be attractive to many students, but the activities associated with these themes do not ask students to gather data and model these. The textbook authors already gathered and modelled the data. As a result, GWA is skipped by many teachers, or is considered as rehearsal of prior content. Some authors (e.g. Inspectie voor het Onderwijs, 1999b; Vink, 2001) noted that GWA was not settling as a modelling activity, because of lack of teaching ideas and materials for GWA, and because there is no incentive for teachers to organise GWA.

There are few modelling activities for students in Dutch classrooms, although these activities are part of their intended curriculum. This finding was done when using the RME-based framework of Figure 7.2. Textbooks and exemplary test items mostly offer ready-made tables and graphs, as part of the texts describing the context. The exemplary RME-based items in Appendix A illustrate this. When mapping the items into the RME-based framework, it turns out that the three items focus more on interpreting and reformulating, than on modelling and reflecting. The first item in Appendix A, on the differences in time zones, provides students with the table, instead of asking students to compile this table from investigations. The second item, on the different views of the candleholder, supplies students with the scale drawing of the situation, instead of asking students to make a sketch of the object, when viewed from above. The third item, on the prediction of growth, provides students with the formula, instead of asking students to generate a formula themselves. In all situations, students have

to start with the interpretation of the provided mathematical models. The situation, whereby students are offered ready-made models is an anti-didactical inversion (cf. Freudenthal, 1973) of the sequence of mathematising activities (see section 2.2.1). This repeated omission of modelling activities can give students the impression that formula, tables, graphs and other mathematical representations emerge out of thin air, and not from modelling activities.

Because the format of tests and exams has remained paper-and-pencil based, and time-restricted, Dutch mathematics teachers are not compelled to detach themselves from the textbooks, and to undertake project work (Kleijne, 1999). The main changes with regard to assessment after 1993 were the denouncement of multiple choice items and the linking of mathematical activities to a context. The new assessment practice offers little space for students' empirical investigations and for mathematising activities, such as modelling and reflecting. Because of this omission, teachers do not feel compelled to offer their students training opportunities for these skills. As a result, Dutch students to not develop all mathematising skills to the same extent, with the focus remaining on interpreting and reformulating. However, to fix this omission, the TIMSS Performance Assessment offers exemplary items for the mathematising activities such as modelling and reflecting. If similar test items are included into compulsory tests, assessment practice could become more balanced.

7.3 RECOMMENDATIONS

7.3.1 For Dutch educational policy and practice

The METRIC study used the distinction between the intended, implemented and attained curriculum. This distinction lead to the observation that Dutch students' mathematical knowledge and skills are satisfactory in the international comparison, when measured either by the TIMSS Written Test or by the TIMSS Performance Assessment. It basically means that, within international comparative studies such as TIMSS, the Dutch educational system has relatively good merits at the level of the attained curriculum. The METRIC study has also shown that the judgement on the appropriateness of the tests in light of the intended curriculum and the implemented curriculum changed. However, independent of this judgement, students' achievement did only change minimally within the time period of four/five years.

The METRIC study did not investigate to what extent the tests covered the intended curriculum. It is possible that the intended curriculum embraces more than what was tested. Therefore, the METRIC study was not able to draw conclusions on whether the Dutch educational system has good merits in light of its own standards (cf. Kuiper, et al., 2002). Therefore, Dutch authorities could consider to ask for research-based data, giving insight into the relevance of the TIMSS tests to their intended curriculum.

In the second place, the METRIC study established that there is room for improvement. As explained in section 7.2.4, the current mathematics assessment practice is not balanced, giving less focus on modelling and reflecting than on interpreting and reformulating. Curriculum experts judged one of the tests used in the METRIC study, the Performance Assessment, to match well with the intended curriculum. The Performance Assessment offers students opportunities to model their data found in self-conducted investigations. Nevertheless, tests of this practical nature are not included into the current Dutch mathematics assessment practice. The Performance Assessment can serve as exemplary material, as advocated by Van den Akker (1998), to improve the implementation of the intended curriculum.

As established in the METRIC study, mathematics teachers have indicated that they could include practical test items, such as those used in the Performance Assessment, into tests of their own making. This intention needs to be supported by all means in order to materialise. Therefore, Dutch authorities might consider to review the Performance Assessment, and to ask for the development of similar hands-on, examinable mathematics tasks. Additionally, they need to develop policies to motivate and support teachers by all means to use the tasks.

The METRIC study focused on Dutch mathematics education only. Of course, stakeholders in other countries can learn from the results as well. Thus, the above recommendations can also be taken at heart in other nations.

7.3.2 For further research

In this section, recommendations for further research will be given. These pertain the international comparative studies in mathematics education, in particular TIMSS, and the emerging Dutch mathematics education research field.

TIMSS

The METRIC study drew on data, which were accumulated through two tests, the TIMSS Written Test and the TIMSS Performance Assessment. Both tests have been developed in a particular educational culture and at a particular stage in time. However, educational cultures differ between countries and they change over time. The adaptation of tests to the circumstances in each country at a particular period in time is difficult to combine with international comparative research. Additionally, considering the motto *if you want to measure change, do not change the measure*, comparative studies such as TIMSS need to keep a large portion of items intact to establish trends. Therefore, the motto induces conservatism and hinders the TIMSS ocean liner from changing its course.

For comparison between countries and over time, identical instruments are required. Therefore, in TIMSS, the comparability over time is endorsed by keeping a secured set of items unpublished and use them again at a later stage. In this way, one section of the test is kept unaltered for calibration, and another section of the test leaves room for innovation. Cloning many items, as was done between 1995 and 1999, is not innovating. By incorporating a growing number of more 'modern' items and abandoning traditional items, the test can be adapted over time to new topics or approaches. Therefore, TIMSS might consider innovating their policy of replacing items, adapting the test to the most recent developments in mathematics education. In this way, TIMSS can also serve to spread exemplary, innovative materials to many nations.

Additionally, the grouping of items into clusters and different test booklets offers room for adaptation of the tests to different educational cultures. Presently, the tests are 'too easy' for students in some countries, while many students from 'lower achieving countries' cannot solve any of the items correctly. The TIMSS tests could be split into two parts: (1) an international section of items contained in all test booklets of all participating countries for international calibration, and (2) a national section with items that suit the local situation (or a regional situation). In this way, the large amount of items in the Written Test that Dutch curriculum experts consider as 'primary school items' and 'traditional algebra drill' can be reduced for Dutch students. These items can be maintained for students from countries where these items are considered relevant. Similarly, RME-based items can be included without frustrating students from other countries where items on, for example, word-formulas and visual geometry are virtually absent. By adapting part of the tests to the intended curriculum of a country, this could further enhance research into the effects of a particular intended curriculum. In

the METRIC study, this was obviously not the case: the TIMSS test items covered a wide range of topics, but few items could be traced down as being truly RME-based. Therefore, the performance of the new RME-based curriculum could not be measured well, as it was unclear to what extent it was covered by the test. Thus, by offering a national section in the test, the comparability between countries can be kept intact while the tests gain validity at a national level. In this way, TIMSS can become more pluralistic.

The METRIC study showed, that an item-based OTL instrument can well supplement the TIMSS instruments for gathering data at the level of the implemented curriculum. TIMSS might consider including such an item-based OTL instrument into their international instruments. The instrument used in the METRIC study, asked teachers to indicate whether they would include the items into a test of their own making, covering all content taught until test administration. The instrument was item-based instead of topic-based, because topics are not easily defined, especially when the terminology describing a range of associated activities is omitted ('linear equations' in stead of 'applying/creating/organising/solving linear equations'). For WT-1999, all items were submitted to teachers, splitting the large item set into sub-sets. Thus, all items yielded an OTL rate. However, the development of alternative instruments to gain information at the level of the implemented curriculum can be considered. For example, teachers can be asked to release the assessment papers, which they used recently for testing their students. Or teachers can be asked to appraise a set of test items from TIMSS and indicate an order of preference for inclusion into a test of their own making. Alternatively, teachers can be asked to assess items, and indicate how they would edit them to make them suitable for a test for their students.

Supplementing the visions of others (e.g. Mullis, et al., 2001; Schmidt, et al., 2001), the METRIC study questioned the TIMSS mathematics curriculum framework used for development of the TIMSS tests (see section 3.1.2). As Table 1.1 testifies, the Written Test based on the framework matched with the intended curricula of many countries, but it disregarded the intended curricula of other countries, such as the Netherlands. From a Dutch perspective, the test was not *equally unfair*. Additionally, the framework did not function well for the allocation of test items to topic areas. One of the weak points in the existing framework, noticed in the METRIC study, was the lack of verbs in the

definitions of mathematics topic areas. TIMSS might consider further developing and refining their mathematics curriculum framework, and TIMSS might consider allowing items testing for higher order skills to fit into several cells of the test grid. An additional option could be, to ask participating nations to categorise the test items according to their own judgement. This could yield additional information on the nature of the intended curricula of participating nations.

The METRIC study repeated the TIMSS Performance Assessment. The replication proved valuable, as it established new insight into validity and reliability. TIMSS might reconsider the validity of their Performance Assessment, and anticipate on its use in international comparative studies, supplementing the Written Test. The METRIC study helped to gain experience pertaining the comparability of testing circumstances and the consistency of codings. The replication resulted in methodological findings, which can help future performance-based studies.

Research in Dutch mathematics education

In the past decades, mathematics education research has expanded in the Netherlands. The research mainly pertains the development of innovative instructional materials, to a large extent at the level of the intended curriculum and to a lesser extent at the level of the implemented curriculum. However, formative and summative evaluations to measure the effects of these materials are scarce. At the level of junior secondary school, there is the following example on the instructional approach to operations with negative numbers. In many textbooks, this topic is an endeavour. To tackle this, the authors of the textbook *Moderne Wiskunde* have developed the metaphor of 'the Witch' (Van den Born, et al., 1999). The Witch cooks soup in a big bowl and throws special blocks into it to change the temperature. Throwing 'plus-blocks' heats the soup, throwing 'minus-blocks' cools the soup. Besides adding the blocks, the witch can also remove them. By removing 'plus-blocks' the temperature falls. What happens if minus-blocks are removed from the bowl? Through this metaphor, students can 'see' that the temperature will rise. This illustrates the positive effect of two combined negatives (removal and minus-blocks). This 'Witch' metaphor is famous nation-wide, but has never been described in a research context, nor has it ever been subject to a comparative study with other instructional approaches to negative numbers. Thus, although many Dutch mathematics educators have

an opinion about 'the Witch', it has never been systematically evaluated. There are only anecdotal descriptions but no evidential data on the 'Witch'. Consequently, a potentially successful instructional approach is not shared internationally either. Therefore, Dutch mathematics educators might consider to establish strong, research-based evidence on the effectiveness of established teaching strategies in mathematics education, in particular on the effectiveness of integrating mathematics and context. To what extent do contexts enhance mathematics learning and are there any contexts, which hinder mathematics learning?

The lack of data on mathematics education in the Netherlands hindered the METRIC study. To find a source on a Dutch interclass-correlation coefficient for mathematics (see chapter 4, section 4.4.6), the METRIC study consulted a foreign source. This was caused by the shortage of research on between-school and within-school effectiveness in Dutch secondary mathematics education.

The METRIC study leaves a number of dangling threads for further research. For example, the achievement data of 1995 on the Performance Assessment were administered among a subset of the students who were tested through the Written Test. This arrangement still waits for a secondary analysis on the between-test correlation at student level.

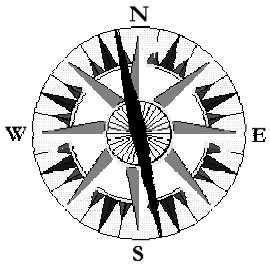
As said in section 7.3.1, the Performance Assessment calls for further development, to innovate mathematics assessment practices in Dutch junior secondary schools. Hands-on tasks offer students tangible materials as context, which require less text than the current descriptions in RME-based items. The open-ended tasks ask students to think about the *how* and *why* of their activities, which connects to the heuristic approach of RME-based mathematics, as advocated by Van Streun (2001). Additionally, hands-on tasks can assist Dutch teachers to better interpret the intended curriculum and offer their students more time for projects, group work and address topics such as data processing and computer usage. This could overcome some of the shortcomings as were observed by the Inspectorate for Education (Inspectie van het Onderwijs, 1999b).

The METRIC study gathered OTL data at the level of the implemented mathematics curriculum. The data available invite for further study at class level. When one teacher indicated that he/she had covered a certain item, how did his/her students then perform? The scope of the METRIC study did not allow going into this depth, but any researcher taking up the challenge will have a premiere.

Moreover, to date, at this level very little research has been carried out on what teachers believe, do, or are able to. Classroom observations are scarce, anecdotal and unsystematic. Descriptions on teachers' attitudes, motivation, ideas, ideals and strategies are only found implicitly in the journals for mathematics teachers. There is still much research to be carried out, while this is badly needed to support the improvement of instruction and curriculum implementation in mathematics education in the Netherlands. The TIMSS Video Study gathered an abundance of materials by shooting videos in 80 randomly selected grade 8 mathematics classes in the Netherlands. The materials will soon be released. Many research questions can be studied, using this unique material. For example, to what extent to teachers make use of the contexts in which mathematics items are integrated? Or: when watching the videos, can criteria for a 'good' mathematics lesson be formulated?

Considering the fact that Dutch mathematics education research is still in an early stage, many open questions remain to be answered. In light of that urgent necessity, the METRIC study was only a pilot boat.

References [1]



Achtergronden van het nieuwe leerplan Wiskunde 12-16. Band 1 en 2 (1992) [**Backgrounds to the new curriculum** W12-16]. Utrecht: **Freudenthal Instituut, Rijksuniversiteit** Utrecht/Enschede: **SLO.**

Adams, R.J., & Gonzalez **e.J.** (1996). **The TIMSS test design. In M.O. Martin** & **E.L. Kelly** (**Eds.**), *Third International Mathematics and Science Study Technical Report. Volume I: Design and development* (**pp. 3-1 – 3-36**). **Boston, MA**: **Boston College.**

Akker, J.J.H. van den (1988). *Ontwerp en implementatie van natuuronderwijs* [**Design and implementation of science education**]. **Amsterdam**/Lisse: **Swets** & Zeitlinger.

Akker, J.J.H. van den (1998). *De uitbeelding van het curriculum* [**Representing the curriculum**] [**Inaugural lecture**]. **Enschede**: Universiteit **Twente.**

Akker, J. van den, & **Terwel**, J. (2001). **Tien opgaven voor het curriculum in de bèta-vakken – stolbeschouwing** [**Ten exercises for the science and mathematics curricula – final considerations**]. *Tijdschrift voor Didaktiek der β-wetenschappen, 18*(1), 95-99.

Algebragroep W12-16 (1990). **En de variabelen, hoe staat het daarmee**? [**And the variables, how about them**?]. *Nieuwe Wiskrant, 10*(1), **12-19.**

Alkin, M.C. (**Ed.**). (1992). *Encyclopedia of educatinal research. Volume 1*. New York: **Macmillan.**

Aukema-Schepel, A. (1991). Van **de bestuurstafel** [**From the management's table**]. *Euclides 67*(**3**), 94-95.

Aukema, A., & Jansen H. (1992). **Twee ontwikkelaars geven weerwoord** [**Two developers retort**]. *Euclides 67*(**7**), 194-199.

Baroody, A.J., & Ginsburg, H.**P.** (1986). **The relationship between initial meaningful and mechanical knowledge of arithmetic. In** J. Hiebert (**Ed.**), *Conceptual and procedural knowledge: the case of mathematics* (**pp. 75-112**). London: Lawrence **Erlbaum.**

Baxter, G.**P.**, & **Shavelson**, J.**R.** (1994). **Science performance assessments**: **Benchmarks and surrogates.** *International Journal of Educational Research, 21*(**3**), 279-298.

Beaton, A. (1998). **Comparing cross-national student performance on TIMSS using different items.** *International Journal of Educational Research, 29*(6), 529-542.

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, MA**: **Boston College.**

---

[1] Dutch surnames that are preceded by prepositions or articles such as 'De', 'Ten' or 'Van' are ordered alphabetically on the first letter of the surname in this list of references. For example, 'De Haan' is to be found under H, and 'Van den Akker' under A.

Bergh, H. **van den** (**1**988). *Examens geëxamineerd* [**Examining exams**]. **The** Hague: **Instituut voor** Onderzoek **van het** Onderwijs.

Berwald, **I.** (1988). **Buitenlandse** rekenmethoden [**Foreign methods in arithmetic**]. *Nieuwe Wiskrant, 8*(**1**), **19-22.**

Birenbaum, **M.**, & **Dochy**, **F.J.R.C.** (**1**996). *Alternatives in assessment of achievements, learning processes, and prior knowledge.* **Dordrecht**: Kluwer.

Bishop, **A.J.**, **FitzSimons**, G.**E.**, & **Seah**, W.**T.** (**2**001). **Do teachers implement their intended values in mathematics classrooms? In M. van den** Heuvel-Panhuizen (**Ed.**), *Proceedings of the 25th Conference of the International Group for the Psychology of Mathematics Education, July 12-17, 2001* (**V**olume **2**, **pp. 2-1**69 – **2-17**6). Utrecht: **Freudenthal Institute.**

Blij, **F. van der**, & **Treffers**, **A.** (**1**985). *Werkdocumenten basisvorming in het onderwijs, WB 7 Rekenen/wiskunde* [**work documents core curriculum**, W**B 7**, **mathematics**]. **The** Hague: W**etenschappelijke Raad voor het Regeringsbeleid.**

Bloom, **B.S.** (**1**956). *Taxonomy of educational objectives; the classification of educational goals.* London: Longman.

Boaler, J. (**1**99**7**). *Experiencing school mathematics: teaching styles, sex and setting.* **Buckingham**: Open University **Press.**

Boon, **P.** (**2**001). **Eindexamens vbo/mavo C/D**, **eerste tijdvak 2**001 [**National exams vbo/mavo C/D**, **first period 2**001]. *Euclides, 77*(**1**), **4-7.**

Borg, W.**R.** (**1**979). **Teacher coverage of academic content and pupil achievement.** *Journal for Educational Psychology, 71,* 6**3**5-6**4**5.

Born, W**. van den**, **Breugel**, **I. van**, **Dijkstra**, J., Gelderblom, G., Goemans, H., Horst, **A. van der**, Kok, **D.**, Koning, **A.**, **Ramaker**, W., **Roode**, G. **de**, **Streun**, **A. van**, Udding, H., Verkooijen, W., & Visser, W. (**1**999). *Moderne Wiskunde - 7e editie, 2a mavo havo*. Groningen: Wolters Noordhoff.

Bos, K.**Tj.** (**2**002). *Benefits and limitations of large-scale international comparative achievement studies – the case of IEA's TIMSS study*. **Doctoral dissertation**, **Universty of Twente**, **Enschede.**

Bos, K.**Tj.**, & Kuiper, W.**A.J.M.** (**1**998). **Praktische vaardigheden internationaal vergeleken** [**Practical skills internationally compared**]. *NVOX, 23*(**7**), **3**66-**372.**

Bos, K.**Tj.**, Kuiper, W.**A.J.M.**, & **Plomp**, **Tj.** (1999). **Student performance and curricular appropriateness in the** Netherlands. *Studies in Educational Evaluation, 25*, **2**69-**27**6.

Bos, K.**Tj.**, Kuiper, W.**A.J.M.**, & **Plomp**, **Tj.** (2001). **TIMSS results of Dutch grade** 8 **students in international perspective**: **performance assessment and written test.** *Studies in Educational Evaluation, 27*, **7**9-9**4.**

Bos, K.**Tj**, & Vos, **F.P.** (**2**000). *Nederland in TIMSS-1999, exacte vakken in leerjaar 2 van het voortgezet onderwijs* [**The** Netherlands **in TIMSS-**99, **mathematics and science at grade** 8 **level**]. **Enschede**: Universiteit Twente.

Broekman, H.G.**B.**, **Spijkerboer**, L.**C.**, & **Terlingen**, J.**J.M.** (**1**991). *Algoritmen en heuristieken in contextrijk reken-wiskundeonderwijs* [**Algorthms and heuristics in mathematics education with rich contexts**]. Utrecht: OW&O**C/Ri**jksuniversiteit Utrecht.

Brown, M. (2000). **Does research make a contribution to teaching and learning in school mathematics? Reflections on an article from Diane Ravitch. In T.** Nakahara & **M.** Koyama (**Eds.**), *Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education, July 23-27, 2000* (**Volume 1, pp. I-**80 – **I-83**). Hiroshima: Hiroshima University.

Burstein, L. (**Ed.**). (**1993**). *The IEA study of mathematics III: Student growth and classroom processes*. **Tarrytown**, NY: **Pergamon Press.**

Burton, L. (1996). **Assessment of mathematics**: What is the agenda? **In M. Birenbaum** & **F.J.R.C. Dochy** (**Eds.**), *Alternatives in assessment of achievements, learning processes, and prior knowledge* (**pp. 31-62**). **Dordrecht**: Kluwer.

Burghes, D. (1999). *IPMA: International Project on Mathematics Attainment.* **Exeter**: **Centre for Innovation in Mathematics Teaching.**

Carlson, J.L., & Ostrosky, **A.L.** (1992). **Item sequence and student performance on multiple-choice exams**: **Further evidence.** *Journal of Economic Education, 23*(**3**), **232-35.**

Carroll, J**.B.** (1963). **A model of school learning.** *Teachers College Records, 64,* **723-733.**

Clarke, D. (1996). **Assessment. In A.J. Bishop**, K. **Clement**, **C.** Keitel, J. Kilpatrick, & **C.** Laborde (**Eds**), *International handbook of mathematics education* (**pp. 327-370**). **Dordrecht**: Kluwer.

Clarke, **D.J.** (**in press**). *Perspectives on practice and meaning in mathematics and science classrooms.* **Dordrecht**: Kluwer.

Cobb, P., & **Bauersfeld**, H. (**Eds.**). (1995). *The emergence of mathematical meaning: Interaction in classroom cultures*. **Hillsdale**, NJ: **Lawrence Erlbaum.**

Comber L**.C.**, & Keeves, J**.P.** (**1973**). *Science education in nineteen countries: An empirical study.* **Stockholm**: **Almquist** & **Wicksell**/New York: **John Wiley.**

Commissie Ontwikkeling Wiskundeonderwijs (1992). *Trajectenboek wiskunde 12-16* [**Book for learning trajectories mathematics 12-16**]. Utrecht: **Freudenthal Instituut**/ **Enschede**: SLO.

Cremers-van Wees, L**.M.C.M.**, **A**kkermans, L**.M.**W., & **Brandsma**, H**.P.** (1999). *Ontwikkelingen in de BAVO in de periode1990-1998: Modernisering en harmonisering in de basisvorming* [**Developments in the core curriculum in** 1990-1998: **Modernisation and harmonisation in the core curriculum**]. **Enschede**: University of Twente.

Cronbach, L**.J.** (1982). *Designing evaluations of educational and social programs*. **San Francisco**: Jossey-Bass.

Darling-Hammond, L., & **Ancess**, J. (1996). **Authentic assessment and school development. In J.B. Baron** & **D.P.** Wolf (**Eds.**), *Performance-based student assessment: challenge and possibilities* (**pp.** 52-83). **Chicago**: **The** University **of Chicago Press.**

Dekker R. (1991). *Wiskunde leren in kleine heterogene groepen* [**Learning mathematics in small heterogeneous groups**]. **De** Lier: **Academisch Boeken Centrum.**

Dekker, T. (1993). **De basisvorming getoetst** [**The core-curriculum tested**]. *Nieuwe Wiskrant, 13*(**2**), 5-9.

Doolaard, S., Creemers-van Wees, L.M.C.M., & Bosker, R.J. (1999). *Basisvorming in 1996; Beschrijving en vergelijking met de periode van invoering* [**Junior secondary education in 1996; Description and comparison with the period of introduction**]. **Enschede**: Universiteit **Twente.**

Doran R.L., & Tamir, P. (Eds.). (1992). **An international assessment of science practical skills.** *Studies in Educational Evaluation, 18(1)*, **1-102.**

Dormolen, J. van (1974). *Didaktiek van de wiskunde* [**Instructional methods in mathematics**]. Utrecht: Oosthoek.

Dormolen, J. van (1993). *Wiskunde werklokaal- het gebruik van materialen en instrumenten bij het leren van wiskunde* [**Mathematics laboratory – the use of materials and equipment for learning mathematics**]. Utrecht: **Algemeen Pedagogisch Studiecentrum.**

Dormolen, J. van, & Zwaneveld, B. (1992). *Heruitgave van drie brochures: Vaardigheden, Handelen om te begrijpen, Instappen en toepassen* [**New edition of three brochures: Skills, Acting to understand, Start and apply**]. Leusden: Nederlandse Vereniging van Wiskundeleraren.

Dormolen, J. van (1999). **Reaktie op het inspectierapport** [**Reaction on the report of the Inspectorate for Education**]. *Nieuwe Wiskrant 19*(2), 51.

Dunham, J. (1990). *Journey through genius – the great theorems of mathematics*. New York: Wiley.

Eggen, Th.J.H.M., & Sanders, P.F. (Eds.). (1993). *Psychometrie in de praktijk* [**Psychometrics in practice**]. **Arnhem: Cito.**

Ehrenfest-Afanassjewa, T. (1923). *Übungensammlung zu einer geometrische Propädeuse* [**Set of exercises for an introduction course in geometry**]. **The** Hague: **Martinus** Nijhoff.

Eisner, E.W. (1979). *The educational imagination: On the design and evaluation of school programs*. New York: **MacMillan.**

Ernest, P. (1988). **The impact of beliefs on the teaching of mathematics. In P. Ernest (Ed.)**, *Mathematics teaching: The state of the art* (**pp. 249-254**). London: **Falmer Press.**

Est, W.T. van (1993). Hans **Freudenthal** (**17 September** 1905 **- 13** October 1990). *Educational Studies in Mathematics, 25*, 59-69.

Fehr, H.F. (1961). **New Thinking in School Mathematics. Paris**: OEEC.

Fey-den Boer, A. (1999). **Schoolboeken en studieresultaten** [**Schoolbooks and study results**]. *Euclides, 75(3)*, **101-103.**

Frary, R.B. (1985). **Multiple-choice versus free-response: a simulation study.** *Journal of Educational Measurement, 22*, **21-31.**

Freeman, D.J., Belli, G.M., Porter, A.C., Floden, R.E., Schmidt, W.H., & Schwille, J.R. (1983). **The influence of different styles of textbook use on instructional validity of standardised tests.** *Journal of Educational Measurement, 20*(**3**), 259-270.

Freudenthal, H. (1969). **Further training of mathematics teachers in the Netherlands.** *Educational Studies in Mathematics, 1*, **184-492.**

Freudenthal, H. (1973). *Mathematics as an educational task*. **Dordrecht: Reidel.**

Freudenthal, H. (1975). **Pupils' achievements internationally compared – the IEA.** *Educational Studies in Mathematics, 6*, **127-1**86.

Freudenthal, H. (**1**991). *Revisiting mathematics education – China lectures*. **Dordrecht**: Kluwer.

Freund, J.**E.** (1988). *Modern elementary statistics (7th edition)*. **Englewood Cliffs**, NJ: **Prentice** Hall **International.**

Fullan, **M.**G. (**1**991). *The* New *meaning of educational change* (**2**nd **ed.**). New York: **Teachers College Press.**

Gaans, W.**M.**J.**M. van** (**1**991). *Vakwerk in de basisvorming – wiskunde* [**Professional/subject** work **in basic secondary education – mathematics**]. **Arnhem**: Lambo.

Garden, **R.A.**, & Orpwood, G. (1996). **Development of the TIMSS achievement tests. In M.**O. **Martin** & **D.**L. Kelly (**Eds.**), *Third International Mathematics and Science Study, Technical report, Volume I: Design and development* (**pp. 2-1 – 2-1**9). **Boston**, **MA**: **Boston College.**

Garden, **R.A.** (1999). **Development of TIMSS performance assessment tasks.** *Studies in Educational Evaluation, 25*, **217-241.**

Glatthorn, **A.A.** (**1**996). *Performance Assessment and standards-based curricula: The achievement cycle.* Larchmont, NY: **Eye on Education.**

Goddijn, **A.**, & Kindt, **M.** (2001). Knelpunten en toekomstmogelijkheden voor de wiskunde in het VO [**Bottlenecks and perspectives for secondary school mathematics**]. *Tijdschrift voor Didaktiek der β-wetenschappen, 18*(**1**), 6**7-9**4.

Goodlad, J.**I.**, Klein, **M.F.**, & **Tye**, K.**A.** (1973). **The domains of curriculum and their stduy. In** J.**I.** Goodlad (**Ed.**). *Curriculum inquiry: The study of curriculum practice (pp. 43-76).* New York: **McGraw-**Hill.

Goodlad, J.**I.**, & **Richter** Jr., **M.**N. (1966). *The development of a conceptual system for dealing with problems of curriculum and instruction*. Los **Angeles**, **CA**: UCLA.

Gonzales, **E.**J., & **Foy. P.** (1997). **Estimation of sampling variability, design effects, and effective sample sizes. In M.**O. **Martin** & **D.**L. Kelly (**Eds.**), *Third International Mathematics and Science Study, Technical report, Volume II: Implementation and analysis* (**pp.** 8**1-**100). **Boston**, **MA**: **Boston College.**

Gonzales, **E.**J., & **Miles**, J.**A.** (**Eds.**). (2001). *TIMSS 1999 User guide for the international database.* **Boston**, **MA**: **Boston College.**

Gravemeijer, K.**P.E.** (1994). *Developing realistic mathematics education*. Utrecht: **CD** β **Press.**

Gray, **E.**, & **Tall**, **D.** (1994). **Duality, ambiguity, and flexibility: a 'proceptual' view of simple arithmetic.** *Journal for Research in Mathematics Education, 25*(**2**), 116-140.

Grzymala-Busse, J.**R.** (1991). *Managing uncertainty in expert systems.* **Boston**: Kluwer.

Groen, W. (2000). Honderd jaar leerplanwijzigingen -vernieuwingsbewegingen met en zonder gevolg [Hundred years of curriculumreforms - innovation movements with and without impact]. **In F.** Goffree, **M. van** Hoorn, & **B.** Zwaneveld (**Eds.**), *Honderd jaar wiskundeonderwijs - een jubileumboek* [Hundred years of mathematics education, a jubilee book] (**pp. 223 – 23**8). Leusden: Nederlandse Vereniging van Wiskundeleraren.

Gulmans, J., Loon, L.J. van, & Pelgrum, W.J. (1981). *Doelstellingenonderzoek middelbaar beroepsonderwijs; Deel I: opzet en uitvoering.* Enschede: Technische Hogeschool Twente.

Haan, D.M. de (1992). *Measuring test-curriculum overlap*. Doctoral dissertation, Universiteit Twente, Enschede.

Haan, J. de, Meij, J. van der, & Pakkert, M. (1997). *Curriculaire geschiktheid van de TIMSS praktische vaardigheidstoets* [Curricular appropriateness of the TIMSS performance assessment]. Unpublished report, Enschede, University of Twente.

Haertel, E.H., & Linn, R.L. (1996). Comparability. In G.W. Phillips (Ed.). *Technical issues in large-scale performance assessment.* Washington: NCES.

Harmon, M., Smith, T.A., Martin, M.O., Kelly, D.L., Beaton, A.E., Mullis, I.V.S., Gonzales, E.J., & Orpwood, G. (1997). *Performance assessment in IEA's Third International Mathematics and Science Study*. Boston, MA: Boston College.

Hardy, R.A. (1984). Measuring instructional validity: a report of an instructional validity study for the Alabama high school graduation examination. *Journal of Educatonal Measurement, 20*(3), 291-301.

Hawker, D., & Ollerton, M. (1999). National tests in mathematics: Two perspectives. *Mathematics Teaching, 168*, 16-25.

Hermans, P.H.L. (Ed.). (1992). Denken en doen: het toetsen van praktische vaardigheden. Arnhem: Cito.

Heuvel-Panhuizen, M. van den (1996). *Assessment and realistic mathematics education*. Utrecht: CD β Press.

Hiebert, J. (1999). Relationships between research and the NCTM Standards. *Journal for Research in Mathematics Education, 30,* 3-19.

Hiele, P.M. van (1973). *Begrip en inzicht* [Understanding and insight]. Purmerend: Muusses.

Hodson, E. (1984). The effect of changes in item sequence on student performance in a multiple-choice chemistry test. *Journal of Research in Science Teaching, 21*(5), 489-495.

Hodson, E. (1987). How important is question sequence? *Education in Chemistry, 24*, 11-22.

Hoeben, W.Th.J.G. (1993). Evaluatie van onderwijsbeleid [Evaluation of educational policies]. In W.J. Nijhof, H.A.M. Franssen, W.Th.J.G. Hoeben, & R.G.M. Wolbert (Eds.), *Handboek curiculum - modellen, theorien, technologien*. Lisse: Swets & Zeitlinger.

Hoogland, K. (1992). Eindoordeel? Beginoordeel [Final judgement? Starting judgement]. *Euclides 67*(9), 289-292.

Hoogland, K (1998). Redactioneel [Editorial]. *Euclides 74*(2), 38.

Hove, J. ten, & Zwaart, P. van der (1993). *Bouwstenen voor de basisvorming – leerplan wiskunde* [Bricks for the core curriculum – curriculum mathematics]. Groningen: Wolters Noordhoff/Enschede: SLO.

Howell, K.W., & Nolet, V. (2000). *Curriculum-based evaluation: teaching and decision making*. Belmont: Wadsworth/Thomson Learning.

Huck, S.W., & Bowers, N.E. (1972). Item difficulty level and sequence effects in multiple-choice achievement tests. *Journal of Educational Measurement, 9*(2), 105-111.

Husén, T., & Bloom, B.S. (Ed.). (1967). *International studies of achievement in mathematics: A comparison of twelve countries. Volumes I and II*. Stockholm: Almquist & Wicksell/New York: John Wiley.

Inspectie van het Onderwijs (1998). *Onderwijsverslag over het jaar 1998*. The Hague: SDU.

Inspectie van het Onderwijs (1999a). *Werk aan de basis. Evaluatie van de basisvorming na vijf jaar. Algemeen rapport*. Utrecht: Inspectie van het Onderwijs.

Inspectie van het Onderwijs (1999b). *Wiskunde in de basisvorming. Evaluatie van de eerste vijf jaar*. Utrecht: Inspectie van het Onderwijs.

Kawanaka, T., Stigler, J.W., & Hiebert, J. (1999). Studying mathematics classrooms in Germany, Japan and the United States: Lessons from TIMSS Videotape Study. In G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (pp. 86-103). London: Falmer.

Keitel, C. (2000). *Cultural diversity, internationalization and globalization: Challenges or perils for mathemativs education.* Keynote address at the 8th annual meeting of the Southern African Association for Research in Mathematics and Science Education (SAARMSE 2000). Port Elisabeth, South Africa, 19-22 January.

Kilpatrick, J. (1993). The chain and the arrow. In: M. Niss (Ed.). *Investigations into assessment in mathematics education. An ICMI-study* (pp. 31-46). Dordrecht: Kluwer.

Kind, P.M. (1999). Performance Assessment in science, what are we measuring? *Studies in Educational Evaluation, 25*(3), 179-194.

Kindt, M. (2000). De erfenis van al-Khwarizmi – over veranderingen in de schoolalgebra [The legacy of Al-Khwaizmi – about change in school algebra]. In F. Goffree, M. van Hoorn, & B. Zwaneveld (Eds.), *Honderd jaar wiskundeonderwijs - een jubileumboek* [Hundred years of mathematics education, a jubilee book] (pp. 57-71). Leusden: Nederlandse Vereniging van Wiskundeleraren.

Kleijne, W. (1999). Wiskunde in de basisvorming. Evaluatie van de eerste vijf jaar [Mathematics in the core curriculum. Evaluation of the first five years]. *Euclides, 75*(3), 75-80.

Kline, M. (1953). *Mathematisc in western culture*. Oxford: University Press.

Kok, D., Meeder, M., Wijers, M., & Dormolen, J. van (1992). *Wiskunde 12-16, een boek voor docenten* [Mathematics W12-16, a book for teachers]. Utrecht: Freudenthal Institute/Enschede: SLO.

Kool, M. (1999). *Die conste vanden getale - een studie over Nederlandstalige rekenboeken uit de vijftiende en zestiende eeuw, met een glossarium van rekenkundige termen* [The art of numbers – a study on Dutch mathematics textbooks from the 15th and 16th century]. Hilversum: Verloren.

Krathwohl, D.R. (1998). *Methods of educational and social science research; an integrated approach.* New York: Addison Wesley.

Kuhlemeier, H., Kleintjes, F., & Bergh, H. van den (2001). Effect van toetsvorm en vraagtype op de moeilijkheid van de afsluitingstoetsen basisvorming; een toepassing van multiniveau analyse met random kruisclassificatie [Effect of test type and item format on the difficulty of the final tests in the core curriculum; an application of multilevel analysis with random cross-classification]. *Pedagogische Studiën, 78*(3), 197-211.

Kuiper, W.A.J.M. (1993). *Curriculumvernieuwing en lespraktijk* [Curriculum reform and teaching practice]. Doctoral dissertation, Universiteit Twente, Enschede.

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: Universiteit Twente.

Kuiper, W.A.J.M., Bos, K.Tj., & Plomp, Tj. (1999). Mathematics achievement in the Netherlands and appropriateness of the TIMSS mathematics test. *Educational Research and Evaluation, 5*(2), 85-104.

Kuiper, W.A.J.M., Bos, K.Tj., & Plomp, Tj. (2000). The TIMSS national option test. *Studies in Educational Evaluation, 26*, 43-60.

Kuiper, W., Boersma K., & Akker, J. van den (2001). Discrepanties in onderzoeksresultaten omtrent de kwaliteit van de exacte vakken in de basisvorming [Discrepancies in research results on the quality of science and mathematics education in the core curriculum]. *Tijdschrift voor Didaktiek der β-wetenschappen, 18*(2), 140-162.

Kuipers, W. (1999). Kijken met je handen [Looking with your hands]. *Nieuwe Wiskrant, 19*(2), 49-50.

Kuyper, H. (1999). Wiskunde in VOCL [Mathematics in VOCL]. *Nieuwe Wiskrant, 19*(2), 25-29.

Lagerwerf, B. (1994). *Wiskundeonderwijs in de basisvorming* [Mathematics education in junior secondary schools]. Groningen: Wolters Noordhoff.

Lapointe, A.E., Mead, N.A., & Askew, J.M. (1992). *Learning mathematics*. New Jersey, NJ: IAEP & ETS.

Lange, J. de (1987). *Mathematics, insight and meaning*. Doctoral dissertation, IOWO, Utrecht.

Lange, J. de (1992). Nieuwe curricula 12-16: De basis gevormd [New curricula 12-16 - the basis created]. *Euclides 67*(9), 259-262.

Lange, J. de (1992). Assessment: No change without problems. In M. Stephens & J. Izard (Eds.). *Reshaping assessment practices: Assessment in the mathematical sciences under challenge. Proceedings from the First National Conference on Assessment in the Mathematical Sciences, Geelong, Victoria, 20-24 November 1991* (pp. 46-76). Camberwell, Australia: ACER.

Lange, J. de (1997a). *Learning from TIMSS: Looking through the TIMSS mirror from a teaching angle.* Paper prepared for a National Research Council Symposium on the results of the Third International Mathematics and Science Study. Washington DC, 3-4 February.

Lange, J. de (1997b). De betekenis van TIMSS voor het Nederlandse wiskundeonderwijs [The meaning of TIMSS for Dutch mathematics education]. *Nieuwe Wiskrant, 17*(2), 22-25.

Leinhardt, G. (1983). Overlap: Testing whether it is taught. In G.F. Madeus (Ed.), *The courts, validity and minimum competency testing* (pp. 153-170). The Hague: Kluwer Nijhoff.

Linden, W.J. van der (1998). A discussion of some methodological issues in international assessments. *International Journal of Educational Research, 29*(6), 569-577.

Linn, R.L., & Burton, E. (1994). Performance-based assessment: Implications of task specificity. *Educational Measurement: Issues and Practice, 13(1)*, 5-15.

Linn, R.L., & Baker, E.L. (1996). Can performance-based assessments be psychometrically sound? In J.B. Baron & D.P. Wolf (Eds.), *Performance-based student assessment: Challenge and possibilities* (pp. 84-103). Chicago: The University of Chicago Press.

Luyn, J. van (1998). *Basisvorming: De basis van het studiehuis* [Basic secondary education: The basis for the "studyhouse"]. The Hague: PMVO.

Luyten, H. (2000). Wiskunde in Nederland en Vlaanderen - wat vinden (en vonden) de leerlingen ervan? [Mathematics in the Netherlands and Flandres – what did the students think?]. *Pedagogische Studiën, 77*(4), 206-221.

Maanen, J.A. van (1987). *Facets of seventeenth century mathematics in the Netherlands*. Zwolle: De Boer.

Madaus, G.F. (1983). *The courts, validity and minimum competency testing.* The Hague: Kluwer Nijhoff.

Marsh, C., & Willis, G. (1995). *Curriculum – alternative approaches, ongoing issues*. Englewood Cliffs, NJ: Merril/Prentice Hall.

Martin, M.O. (1996). Third International Mathematics and Science Study: An overview. In M.O. Martin & D.L. Kelly (Eds.), *Third International Mathematics and Science Study, Technical report, Volume I: Design and development* (pp. 1-1 – 1-19). Boston, MA: Boston College.

Mehrens, W.A., & Phillips, S.E. (1987). Sensitivity of item difficulties to curricular validity. *Journal of Educational Measurement, 24*(4), 357-370.

Merrill, M.D. (1983). Component display theory. In Ch.M. Reigeluth (Ed.), *Instructional design and models* (pp. 279-333). Hillsdale: Erlbaum.

Miller, J.P., & Seller, W. (1985). *Curriculum, perspectives and practice*. New York: Longman.

Ministerie van OC&W (1998). *Kerndoelen basisvorming, 1998-2003, relaties in beeld* [Attainment targets core curriculum, 1998-2003, relations in the picture]. Zoetermeer: Ministerie van Onderwijs, Cultuur en Wetenschappen.

Moor, E.W.A. de (1999). *Van vormleer naar realistische meetkunde* [From theory of forms to realistic geometry]. Utrecht: CD β Press.

Moore, D.S., & McCabe G.P. (1993). *Statistiek in de praktijk [Statistics in practice].* Schoonhoven: Academic Service.

Moyer, P. (2002). Are we having fun yet? How teachers use manipulatives to teach mathematics. *Educational Studies in Mathematics, 47*, 175-197.

Mulder, H.B.G.W.J. (1996). *Concept herziene kerndoelen basisvorming: Verantwoordingsrapport voortgezet onderwijs* [Concept reviewed attainment targets core curriculum: Accountability report secondary education]. Enschede: SLO.

Mullis, I.V.A., Martin, M.O., Gonzalez, 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 the Third International Mathematics and Science Study at the Eighth Grade.* **Boston, MA: Boston College.**

Mullis, I.V.A., Martin, M.O., Smith, T.A., Garden, R.A., Gregory, K.D., Gonzalez, E.J., Chrostowski, S.J., & O'Connor, K.M. (2001). *TIMSS Assessment Frameworks and Specifications 2003.* **Boston, MA: Boston College.**

Niss, M. (**Ed.**). (1993). *Investigations into assessment in mathematics education, an ICMI Study.* **Dordrecht: Kluwer.**

OECD (2000). *Measuring student knowledge and skills. The Pisa 2000 assessment of reading, mathematical and scientific literacy.* **Geneva: OECD.**

Pelgrum, W.J., Eggen, Th.J.H.M., & Plomp, Tj. (1983a). *Tweede Wiskunde Project, beschrijving van uitkomsten* [**Second Mathematics Project, description of outcomes**]. **Enschede: Technische Hogeschool Twente.**

Pelgrum, W.J., Eggen, Th.J.H.M., & Plomp, Tj. (1983b). *Tweede Wiskunde Project, analyses van uitkomsten: Leerstofaanbod en resultaten* [**Second Mathematics Project, analyses of outcomes; Content offer and results**]. **Enschede: Technische Hogeschool Twente.**

Pelgrum, W.J., Eggen, Th.J.H.M., & Plomp, Tj. (1986). *The implemented and attained mathematics curriculum - a comparison of eighteen countries*. **Enschede: Technische Hogeschool Twente.**

Pelgrum, W.J. (1990). *Educational assessment: monitoring, evaluation and the curriculum*. **De Lier: Academisch Boekencentrum.**

Pepin, B. (2001). *Equivalence in cross-national comparisons: Developing an understanding of culturally salient concepts through qualitative research.* **Paper presented at ECER 2001, Lille, France, 5-8 September.**

Perrenet, J.C. (1995). *Leren probleemoplossen in het wiskundeondewijs: samen of alleen. Onderzoek van wiskunde leren bij 12- tot 16-jarigen* [**Learning problem solving in mathematics education: together or alone. Research on learning mathematics by 12-16 year olds**]. **Amsterdam: Universiteit van Amsterdam.**

Peschar, J.L. (1988). *Evaluatie van de basisvorming. Kader voor het uitvoeringsplan* [**Evaluation of the core curriculum. Framework for the executive plan**]. **The Hague: SVO.**

Piaget, J. (1952). *The child's conception of number*. **London: Routledge & Kegan Paul** (**translated from French**).

Piper, K. (1979). *Curriculum style and social learning. ACER Monograph No. 4.* **Hawthorn, Victoria: ACER.**

Pimm, D. (1987). *Speaking mathematically; Communication in mathematics classroom*. **New York: Routledge & Kegan Paul.**

Reeuwijk, M. van (1992). **The standards applied: Teaching data visualization**. *Mathematics Teacher, 85*(7), 513-518.

Resnick, L.B., & Resnick, D.P. (1989). *Assessing the thinking curriculum: New tools for educational reform*. **Pitsburgh, PA: Learning Research and Development Center, University of Pittsburgh and Carnegie Mellon University.**

Robitaille, D.F., & **Garden, R.A.** (**Eds.**). (1989). *The IEA Study of Mathematics II: Contexts and outcomes of school mathematics*. Oxford: Pergamon Press.

Robitaille, D.F., & **Garden, R.A.** (**Eds.**). (1996). *Research questions and study design. TIMSS Monograph nr. 2*. Vancouver: Pacific Educational Press.

Robitaille, D.F., **Schmidt**, W.H., **Raizen, S.**, **McKnight, C.**, **Britton E.**, & **Nicol, C.** (1993). *Curriculum frameworks for mathematics and science. Timss Monograph nr. 1*. Vancouver: Pacific Educational Press.

Roelofs, E.C. (1996). **Toepassingsgerichtheid en authentiek leren. In** N.A.J. Lagerweij, G. Kanselaar, J.L. van der Linden, **E.C. Roelofs**, O. **Treep**, J.C. Voogt, L.J.A. Vriens, & L. van Wessum, *Rapport 2: Basisvorming op de voet gevolgd - de invoering* [**Report 2**: **Basis education followed closely**: **the introduction**]. Utrecht: Universiteit Utrecht.

Roelofs, E.C., **Franssen**, H.A.M., Houtveen **A.A.M.**, & Lagerweij, N.A.J. (1999). **Een dieptestudie naar authentiek leren in de basisvorming. Docentgedrag, methodengebruik en leerlingpercepties** [**An indepth study for authentic learning in junior secondary schools. Teacher attitude, text book usage and student perceptions**]. *Pedagogische Studiën, 76(***4**), 258-272.

Roelofs, E., Vermeulen, **CJ.**, & Houtveen, **A.A.M.** (1998). *Basisvorming op weg, onderzoek naar de meningen van docenten over de realisatie van de basisvorming [The core curriculum for junior secondary schools under way: research for teachers' opinions about its realisation]*. Utrecht: ISOR/Onderwijsresearch.

Sawada, T. (1999). **The Japanese perspective on TIMSS.** *Zentralblatt für Didaktik der Mathematik, 99*(6), 170-174.

Schalkwijk, L.Th.J.M. (1998). *Onderzoekend wiskunde leren*. **Doctoral dissertation**, Katholieke Universiteit Nijmegen, Nijmegen.

Schama, S. (1991). *The embarrassment of riches, an interpretation of Dutch culture in the Golden Age*. Londen: Fontana Press.

Schmidt, W.H., **Jorde, D.**, **Cogan**, L.S., **Barrier, E.**, Gonzalo, **I.**, **Moser**, U., **Shimizu**, Y., **Sawada, T.**, Valverde, G., **McKnight, C.**, **Prawat, R.**, Wiley, **D.E.**, **Raizen, S.**, **Britton, E.D.**, & Wolfe, **R.**G. (1996). *Characterising pedagogical flow.* **Dordrecht**: Kluwer.

Schmidt, W.H., Valverde, G., **McKnight, C.**, Houand, **R.T.**, & Wiley, **D.E.** (1997). *Many visions, many aims - Volume I: A cross-national investigation of curricular intentions in school mathematics.* **Dordrecht**: Kluwer.

Schmidt, W.H., & **Cogan**, L. (1996). **Development of the TIMSS context questionnaires. In M.**O. **Martin** & **D.**L. Kelly (**Eds.**), *Third International Mathematics and Science Study, Technical report, Volume I: Design and development* (**pp.** 5-1 – 5-22). **Boston**, **MA**: **Boston College.**

Schoemaker, G. (1989). **Kolom 7** [**Column 7**]. *Euclides 64*(6), 159.

Schoenfeld, A.H. (2000). **Purposes and methods of research in mathematics education.** *Notices of the American Mathematical Society, 47*, 641-649.

Schuring, H. (2000). **Examens door de jaren heen** [**Exams throughout the years**]. **In F. Goffree, M. van Hoorn**, & **B.** Zwaneveld (**Eds.**), *Honderd jaar wiskundeonderwijs - een jubileumboek* [**Hundred years of mathematics education, a jubilee book**] (**pp.** 139 – 148). Leusden: Nederlandse Vereniging van Wiskundeleraren.

Shavelson, R.J. (1994). **Guest editor's preface.** *International Journal of Educational Research, 21*(3), 235-238.

Shavelson, R.J., McDonnell, L.M., & Oakes, J. (**Eds.**). (1989). *Indicators for monitoring mathematics and science education, a sourcebook*. **Santa Monica, CA**: **RAND.**

Shavelson, R.J., Baxter, G.P., & **Pine**, J. (1992). **Performance assessment**: **Political rhetoric and measurement reality.** *Educational Researcher, 21*(4), **22-27.**

Shavelson, R.J., Baxter, G.**P.**, & Xiaohong, G. (1993). **Sampling variability of performance assessments.** *Journal of Educational Measurement, 30*(3), 215-232.

Skemp, **R.R.** (1986). *The psychology of learning mathematics*. **London**: **Penguin** Hammondsworth.

Sluijter, C., Kleintjes, **F.G.M., Schalk**, H.H., **Roosmalen**, W. **van**, Hermans, **P.**H.L., & **Bogaerts, C.A.M.J.** (1996). *De constructie van beoordelingsschalen bij afsluitingstoetsen voor de basisvorming* [**The construction of scoring scales with final tests for the core curriculum**]. **Arnhem**: **Cito.**

Smid, H.J. (2000). **Wiskundeonderwijs op bijna vergeten scholen - wiskunde op de** (m)**ulo** [**Mathematics education at almost forgotten schools**]. **In F.** Goffree, **M. van** Hoorn, & **B.** Zwaneveld (**Eds.**), *Honderd jaar wiskundeonderwijs - een jubileumboek* [**Hundred years of mathematics education, a jubilee book**] (**pp. 121 – 138**). Leusden: Nederlandse **Vereniging van** Wiskundeleraren.

Snijders, T.A.B., & **Bosker, R.J.** (1999). *Multilevel analysis – an introduction to basic and advanced multilevel modelling*. **London**: **SAGE Publications.**

Sosniak, L.**A., Ethington, C.A.**, & Varelas, **M.** (1994). **The myth of progressive and traditional orientations**: **teaching mathematics without a coherent point of view. In I.** Westbury, **C.A. Ethington**, L.**A. Sosniak**, & **D.P. Baker** (**Eds.**), *In search of more effective mathematics education; examining data from the IEA second international mathematics study* (**pp.** 95-**112**). **Norwood**, NJ: **Albex.**

Streun, A. van (1985). Heuristisch **wiskunde-onderwijs** [**Heuristic mathematis education**]. **Doctoral dissertation, Rijksuniversiteit** Groningen, Groningen.

Streun, A. van (1985). **Slagen onderwijsexperimenten altijd**? [**Are educational experiments always succesful**?]. *Nieuwe Wiskrant, 5*(2), 15-17.

Streun, A. van (1990). Wishful thinking **en nieuwe leerplannen** [Wishful thinking **and new curricula**]. *Euclides, 65*(1), 186-188.

Streun, A. van (2001). *Het denken bevorderen* [**Enhance thinking**] [**Inaugural lecture**]. Groningen, Netherlands: **Faculteit der** Wiskunde **en** Natuurwetenschappen, **Rijksuniversiteit** Groningen.

Stroomberg, H.**P.** (1973). *Verslag van een empirisch onderzoek van onderwijsdoelstellingen. Deel I: Rekenen* [**Report of an empirical research for educational objectives**]. **Amsterdam**: **Research Instituut voor de Toegepaste Psychologie.**

Szendrei, J. (1996). **Concrete materials in the classroom. In A.J. Bishop**, K. **Clement**, **C.** Keitel, J. **Kilpatrick**, & **C.** Laborde (**Eds.**), *International handbook of mathematics education* (**pp. 411-434**). **Dordrecht**: Kluwer.

TIMSS performance assessment administration manual for the main survey (1994). **Boston, MA**: **Boston College.**

TIMSS coding guide for performance assessment Populations 1 and 2 (1994). **Boston, MA**: **Boston College.**

TIMSS scoring guides for the mathematics and science free-response items (1999). **Boston, MA**: **Boston College.**

TIMSS test-curriculum matching analysis (1995)**. Boston, MA**: **Boston College.**

TIMSS-R survey operations manual (1998)**. Boston, MA**: **Boston College.**

TIMSS-R test-curriculum matching analysis (1998)**. Boston, MA**: **Boston College.**

Thompson, D.R., & **Senk, S.**L. (2001)**. The effects of curriculum on achievement in second-year algebra: The example of the university of Chicago School Mathematics Project.** *Journal for Research in Mathematics Education, 32*(**2**), 58-84**.**

Travers K.J., & **Westbury, I.** (1989)**.** *The IEA study of mathematics I: Analysis of mathematics curricula***.** Ox**ford**: **Pergamon Press.**

Treffers, A. (1987)**.** *Three dimensions. A model of goal and theory descriptions in mathematics instruction - the Wiskobas Project***. Dordrecht**: **Reidel.**

Treffers, A. (1993)**.** Wiskobas en Freudenthal**, realistic mathematics education.** *Educational Studies in Mathematics, 25*, 89-108**.**

Verhage, H., & Wijers, **M.** (1991)**.** Uitkomsten enquete regionale bijeenkomsten [Outcomes survey regional meetings]**.** *Euclides 66*(9), 264-268**.**

Verhoef, N., & **Bos, D.** (1992)**.** Wiskundewerklokaal en de lerarenopleiding te Zwolle [Mathematics working room and teacher training in Zwolle]**.** *Nieuwe Wiskrant 12*(**3**), 5-10**.**

Vink, **A.** (2001)**. Praktische opdrachten in het vmbo.** *Euclides 76*(8), 308-315**.**

Vink, **A.**, Zegers, **A.**, & **Ballering F.** (1993)**. Reactie** [reaction]**.** *Euclides 68*(5), 149-152**.**

Vos, **F.P.**, Kuiper, W.**A.J.M.**, & **Bos**, K.**Tj.** (2000)**.** *Predecessor items and students' achievement***. Paper presented at ECER** 2000, **Edinburg, Scotland**, 20-23 **September.**

Vos, **F.P.**, & **Bos**, K.**Tj.** (2001a)**. Trends** (1995 – 1999) **in performances of Dutch grade** 8 **students in TIMSS against the background of the realistic mathematics curriculum. In M. van den** Heuvel-**Panhuizen** (**Ed.**), *Proceedings of the 25th Conference of the International Group for the Psychology of Mathematics Education.* U**trecht**, N**etherlands**: **Freudenthal Institute.**

Vos, **F.P.**, & **Bos**, K.**Tj.** (2001b)**.** *Comparing three curricular levels of TIMSS-95 and TIMSS-99 mathematics results in the Netherlands with Belgian (Flemish) data.* **Paper presented at ECER** 2001, Lille, **France**, 5-8 **September.**

Werf, **M.P.C. van der**, Lubbers **M.**J., & Kuyper, H. (1999)**.** *Onderwijsresultaten van VOCL'89 en VOCL '93 leerlingen* [**Education results of students in** VO**C**L'89 **and** VO**C**L'9**3**]**.** Groningen: **Rijksuniversiteit** Groningen**.**

Wiggins, G. (1989)**. A true test**: **Towards more authentic and equitable assessment.** *Phi Delta Kappan, 76*(9), 703-713**.**

Wijers, **M.** (2000)**. Algebra, een praktijkprobleem** [**Algebra, a practical problem**]**.** *Nieuwe Wiskrant 19*(**3**), 36-41**.**

Wijers, M., & Kemme, S. (2000). Welke algebra heb je nodig in 4 vwo? [Which algebra is needed in grade 10?]. *Nieuwe Wiskrant 20*(1), 22-25.

Wijnstra, J.M. (Ed.). (1990). *Verantwoording van de rekenpeiling medio en einde basisonderwijs 1987* [Account of the mathematics assessment in primary education 1987]. Arnhem: Instituut voor Toetsontwikkeling.

Wolf, R.M. (1994). Performance assessment in IEA studies. *International Journal of Educational Research, 21*(3), 239-245.

Wolf, R.W. (1998). Validity issues in international assessments. *International Journal of Educational Research 29*, 491-501.

Wood, R. (1991). *Assessment and testing*. Cambridge: Cambridge University Press.

Zuzovsky, R., & Harmon, M. (Eds). (1999). TIMSS performance assessment. *Studies in Educational Evaluation, 25*(3), 269-276.

Zuzovsky, R. (1999). Problematic aspects of the scoring of the TIMSS practical performance assessment: Some examples. *Studies in Educational Evaluation, 25*(3), 315-323.

Zwaart, P. van der, & Boertien, H. (1998). *Naar een schoolbrede aanpak van de basisvorming; Een handreiking bij de herziene kerndoelen wiskunde* [Towards a school-wide approach for the core curriculum; A helping hand for the reviewed core objectives]. Enschede: SLO.

Zwaneveld, B. (1999). *Kennisgrafen in het wiskundeonderwijs* [Concept mapping in mathematics education]. Maastricht: Shaker.

Samenvatting

Als een oceaanstomer die van koers verandert

Het wiskundecurriculum in de tweede klas van
het voortgezet onderwijs in Nederland, 1995-2000

~ Een schip op het strand is een baken op zee. ~
A ship on the beach is a beacon at sea. (learning from other people's mistakes)
(DUTCH PROVERB)

AANLEIDING VOOR DE STUDIE

In 1995 nam Nederland deel aan een grootschalig internationaal vergelijkend onderzoek naar de leeropbrengsten in het onderwijs van de exacte vakken. Dit project heette de Third International Mathematics and Science Study (TIMSS), uitgevoerd onder auspiciën van de International Association for the Evaluation of Educational Achievement (IEA). De prestaties van leerlingen werden gemeten middels identieke toetsen in alle deelnemende landen, met dien verstande dat de toets naar de landstaal was vertaald.
Een van de populaties waarin het onderzoek werd uitgevoerd betrof leerjaar 2 van het voortgezet onderwijs (leerlingen van gemiddeld 14 jaar). Voor deze populatie werden twee toetsen gebruikt om de prestaties te meten. Enerzijds was er een schriftelijke toets (de TIMSS Written Test, hierna afgekort als WT-1995) en anderzijds was er een praktische vaardigheidstoets (de TIMSS Performance Assessment, hierna afgekort als PA-1995).

De scores van de Nederlandse leerlingen op WT-1995 in de internationale vergelijking waren goed te noemen. Nederlandse leerlingen behoorden wat betreft wiskunde tot de mondiale subtop. Hun gemiddelde score lag beduidend hoger dan het internationale gemiddelde. Dit was opmerkelijk omdat leerplanexperts voor het vak wiskunde oordeelden dat deze internationale schriftelijke toets slechts in beperkte mate aansloot op het beoogd curriculum (de Kerndoelen voor de Basisvorming). De leerplanexperts gaven aan dat slechts 69% van de wiskundige toetsitems paste bij het beoogd curriculum. Enkele getoetste onderwerpen zoals letterrekenen en congruentie waren nog nauwelijks in het lesprogramma aan bod geweest. De wiskunde-experts maakten bezwaar tegen de kale redactiesommen. Bovendien werd, met 75% meerkeuze items, de vorm van de testitems minder geschikt geacht voor het toetsen van vaardigheden waaraan in de basisvorming belang wordt gehecht.

In 1995 werd in aanvulling op de schriftelijke toets ook een praktische vaardigheidstoets afgenomen in 19 landen waaronder Nederland. Belangrijk argument voor het complementaire karakter van deze tweede toets was dat met een schriftelijke toets vooral 'kennis' en in mindere mate 'vaardigheden' onderzocht kunnen worden. Deze praktische vaardigheidstoets (PA-1995) was een grensverleggende toets die bestond uit wiskunde- en sciencetaken en werd afgenomen in een practicumomgeving. Leerlingen werden voorzien van manipulatieve materialen (klei, magneten, vouwblaadjes, plakband, enz.) en eenvoudige meetapparatuur (thermometer, liniaal, weegschaal, enz.). Zij werden getoetst met opdrachten als: het opzetten en uitvoeren van experimenten, het doen en beschrijven van waarnemingen, het rekenen met een zakrekenmachine, het zoeken naar regelmaat en het noteren en interpreteren van meetgegevens.
Geconsulteerde curriculumexperts oordeelden in 1995 dat deze internationale praktische toets, gegeven de gerichtheid op het meten van toepassingsgerichte vaardigheden, goed aansloot bij de Kerndoelen.
De scores van de Nederlandse leerlingen op PA-1995 onderscheidden zich echter nauwelijks van het internationale gemiddelde. Dit contrasteerde opvallend genoeg zowel met de prestaties op de schriftelijke toets (waarop wél een hoge internationale ranking werd bereikt) alsook met gegevens over de grote mate van geschiktheid van de toets in het licht van het beoogde curriculum voor de basisvorming.

De vraag rees of deze discrepanties konden worden toegeschreven aan de aandacht die in de basisvorming tot dan toe was besteed aan toepassingsgerichte vaardigheden in de exacte vakken. Wellicht volgde het onderzoeksmoment (1995) te kort op de formele invoering van de basisvorming (1993) en bleef de uitvoering van het nieuwe leerplan nog achter. Deze veronderstelling werd ook geschraagd door het lerarenoordeel over PA-1995. In het kader van onderzoek naar het uitgevoerd curriculum werd aan de wiskundeleraren van de getoetste leerlingen de taken uit de toets voorgelegd. Zij oordeelden over het algemeen afwijzend op deze taken in tegenstelling tot de curriculumexperts. Dit kon erop duiden dat niet alle getoetste inhouden en vaardigheden ook daadwerkelijk in de les aan de orde waren geweest.

Gezien de hierboven beschreven onverwachte resultaten werd het een maatschappelijk en wetenschappelijk relevante vraag geacht of herhaalde afname van beide toetsen zou resulteren in prestaties vergelijkbaar met die uit 1995. Daarnaast lag de vraag voor de hand of vakexperts en docenten tot een soortgelijk oordeel zouden komen over de geschiktheid van toetsen als in 1995 het geval was. Aldus werd besloten om aan te haken bij de internationale herhaling van de schriftelijke toets (WT-1999) en, bij gebrek aan internationale belangstelling, de praktische toets alleen in Nederland te herhalen (PA-2000). De herhaalde afname van beide toetsen zou trendgegevens kunnen verschaffen over de waargenomen discrepanties.
De volgende onderzoeksvragen stonden centraal.

- In hoeverre resulteert afname van PA-2000 in combinatie met de herhaalde afname van de schriftelijke toets WT-1999 in dezelfde verschillen in prestaties op PA-1995 en WT-1995?
- In hoeverre zijn de leerlingprestaties op beide toetsen terug te voeren op de geschiktheid van deze toets in het licht van de Kerndoelen voor de exacte vakken in de basisvorming (beoogd curriculum) en het feitelijk gegeven onderwijs in die vakken (uitgevoerd curriculum), en op eventuele discrepanties hiertussen?

CONCEPTUEEL KADER EN CONTEXT VAN DE STUDIE

De studie werd uitgevoerd tegen de achtergrond van een conceptueel kader waarin drie verschijningsvormen van curricula voor de exacte vakken worden onderscheiden: het beoogde, het uitgevoerde en het gerealiseerde curriculum.

Het beoogde curriculum verwijst naar Kerndoelen, examenprogramma en methoden. Het uitgevoerde curriculum verwijst naar het feitelijke onderwijsaanbod op school- en vooral klassenniveau. Het gerealiseerde curriculum verwijst naar de opbrengst van het onderwijs in de exacte vakken in termen van verworven kennis, vaardigheden en houdingen.

Met de schriftelijke en de praktische TIMSS-toetsen worden kennis en vaardigheden getoetst. De praktische toets is binnen TIMSS ontwikkeld vanuit een visie waarin natuurwetenschappelijk onderwijs zich baseert op een samenhang tussen processen en inhoud (procedurele en declaratieve kennis). Centraal staan open opdrachten om verklaringen voor onderzochte verschijnselen te geven. Een practicum is daarmee niet langer een ondersteunende illustratie van concepten (als instap vóór of als demonstratie ter rechtvaardiging van de theorie), maar een zelfstandige, toetsbare onderwijsactiviteit. Ervaringen met praktische tests zijn van recente datum en de betrouwbaarheid in een internationaal vergelijkende context is vooralsnog enigszins problematisch. Het coderen van de leerlingantwoorden op de open vragen is een complex proces. Door het gebruik van een meerderheid aan meerkeuzevragen doet dit probleem zich in veel geringere mate voor bij de schriftelijke toets. Hierdoor kunnen de toetsitems aan grote hoeveelheden leerlingen voorgelegd worden en kunnen de antwoorden geautomatiseerd verwerkt worden. Dit maakt de toets veel betrouwbaarder, zowel voor vergelijking in de tijd als tussen landen.

Internationale onderwijsinnovaties waarin leerlingen een plaats krijgen als betrokkenen die hun kennis construeren vanuit contextrijke ervaringen, hebben in TIMSS geleid tot de ontwikkeling van de praktische vaardigheidstoets, in het wiskundeonderwijs tot de ontwikkeling van het *realistisch reken/wiskunde-onderwijs* en in Nederland tot de basisvorming. In de laatste staan de begrippen toepassing, vaardigheid en samenhang (TVS) centraal. Met deze trefwoorden werd in 1993 een verplicht leerplan voor alle leerlingen in de onderbouw van het Nederlands voortgezet onderwijs ingevoerd. Naast invoering van nieuwe vakken werden leerplannen van bestaande vakken veranderd. Wat betreft wiskunde werd aangesloten bij eerdere leerplanvernieuwingen in het primair onderwijs (Wiskobas) en de bovenbouw van het secundair onderwijs (Hewet en Hawex). Een enthousiaste ontwikkelwerkgroep, W12-16, ontwikkelde een nieuw wiskundeleerplan met als leidraad dat het geleerde niet alleen in het latere leven

zinvol moest zijn maar ook op het moment van leren. Wiskundige begrippen werden geïntegreerd binnen contexten. Abstractere onderwerpen als letteralgebra en transformatiemeetkunde werden uitgesteld en vervangen door aantrekkelijkere onderwerpen als *grafiekentaal* en *kijkmeetkunde*. Daarmee werd aangesloten op de wens van vermaatschappelijking van de leerinhouden waarmee deze sterker dan voorheen uitgaan van levensechte, voor de leerlingen herkenbare situaties. Tevens streefde men naar een verwetenschappelijking van het onderwijs waarin een belangrijke plaats is ingeruimd voor strategisch leren denken en problemen leren oplossen. De basisvorming moest naast een vernieuwde lesinhoud samengaan met een didactiek van authentiek leren, die echter in de eerste jaren na de invoering nog slechts mondjesmaat uit de verf kwam. Met name het aanbieden van activerend, onderzoeksgericht werk aan de leerlingen schoot tekort. Daarmee sluit de TIMSS praktische vaardigheidstoets niet alleen goed aan bij het beoogde leerplan van de basisvorming, maar voorziet het als exemplarische onderwijsactiviteit tevens in een leemte van het uitgevoerde curriculum. De praktische vaardigheidstoets is geschikt voor gebruik naast de bestaande methodes.

OPZET EN UITVOERING VAN HET ONDERZOEK

De studie baseerde zich op de reeds uitgevoerde studies WT-1995 en PA-1995. Er werd aangehaakt bij de replicatiestudie van de schriftelijke toets in internationaal verband (WT-1999). Door tevens de praktische toets te herhalen werd het toetsdesign gecompleteerd. Vanwege financiële beperkingen werd de replicatiestudie van de praktische toets een jaar uitgesteld (PA-2000). Instrumentarium, steekproefopzet en methoden van dataverzameling en -verwerking van de replicatiestudies volgden met het oog op vergelijkbaarheid alle procedures voor de dataverzameling van WT-1995 en PA-1995.
Voor elk van de vier deelstudies WT-1995, PA-1995, WT-1999 en PA-2000 werden drie soorten data verzameld: een toetsafname bij leerlingen ter operationalisatie van het gerealiseerde curriculum; het bevragen van leraren op de geschiktheid van de betreffende toets in het licht van het uitgevoerde curriculum; het bevragen van curriculumexperts op de geschiktheid van de toets in het licht van het beoogde curriculum. Met deze consultatie werd de operationalisatie van het conceptueel kader gecompleteerd. Aldus ontstonden er twaalf metingen waarop de studie zich baseerde.

Ter operationalisatie van het beoogd curriculum werden een aantal curriculumexperts geraadpleegd (drie voor WT-1995, WT-1999 en PA-1995; vijf voor PA-2000). De betrokken personen waren werkzaam bij de landelijke Pedagogische Centra, het Freudenthal Instituut, lerarenopleidingen en het CITO. Aan de experts zijn alle toetsitems voorgelegd en hen is gevraagd of deze passen bij het beoogd curriculum als omschreven in de Kerndoelen voor het vak wiskunde in de basisvorming. De verstrekte antwoorden leverden per toetsitem een *item-curriculum matching index* op. Voor de schriftelijke toetsen waren deze alleen beschikbaar op een nominale ja/nee-schaal. Voor de praktische toets waren deze beschikbaar op een ratio schaal waardoor de resultaten meer genuanceerd konden worden. Dit laatste bleek nuttig omdat er bij diverse toetsitems ambivalentie ontstond over de geschiktheid in het licht van de Kerndoelen, bijvoorbeeld door de 'kaalheid' en de meerkeuzevraagvorm. In deze gevallen kon de ambivalentie uitgedrukt worden als percentage van experts die het item geschikt achtten. Het nadeel van deze aanpak was dat de resultaten met betrekking tot het geschiktheidsoordeel van de twee toetsten alleen konden worden vergeleken indien de resultaten op de ene toets werden getransformeerd naar de schaal waarop de resultaten van de andere toets waren uitgedrukt.

Ter operationalisatie van het uitgevoerd curriculum werd, tegelijk met de toets, een vragenlijst aan de wiskundeleraren van de betreffende toetsklassen voorgelegd. Deze vragenlijsten bevatten alle toetsitems en over ieder item werd een geschiktheidoordeel gevraagd. Aan de leraren werd binnen WT-1995 en WT-1999 gevraagd te beoordelen of zij de items uit de toets geschikt vonden voor opname in een eigengemaakt proefwerk over alle tot dan behandelde stof (opportunity to learn, OTL). Laatstgenoemde maat is gebaseerd op onderzoek van De Haan (1992). De meting middels het OTL instrument diende een indicatie te geven voor het gegeven onderwijs en of de leerlingen de gelegenheid hadden gekregen de betreffende kennis en vaardigheden aan te leren. Naast een meting van het gegeven onderwijs (content coverage) kon het oordeel ook aangeven dat de leraar opname van het betreffende item mogelijk achtte, omdat zijn/haar leerlingen voldoende aansluitende bagage hadden meegekregen. Het percentage leraren dat een positief antwoord gaf werd aangeduid als de *OTL rate* van een item.

Helaas was in 1995 slechts een beperkt aantal toetsitems uit de schriftelijke toets aan de leraren voorgelegd. Deze opgaven maakten deel uit van de Nationale

OptieToets (NOT), die was gebaseerd op de Kerndoelen. Naast 20 nieuw ontwikkelde items behelsde deze toets 16 items uit WT-1995. De NOT toetste daarmee het beoogde leerplan, maar tevens konden middels de 16 overlappende *ankeropgaven* de prestaties tussen NOT en WT-1995 vergeleken worden. Zoals gezegd, deze 16 ankeropgaven waren aan de leraren ter beoordeling voorgelegd, en door hun selectiegeschiedenis waren zij niet representatief voor de overige toetsitems uit WT-1995. Voor WT-1999 werden wel alle toetsitems aan de leraren voorgelegd.
Het leraren-instrument verschilde tussen de praktische en de schriftelijke toetsen. Omdat aangenomen werd dat leraren minder ervaring hebben met het organiseren van praktische toetsen werden de vragen binnen PA-1995 en PA-2000 uitgesplitst. Er werd gevraagd of (1) de stof van de items onderwezen was en (2) de items geschikt werden geacht voor opname in een praktische toets over de tot dan toe behandelde stof c.q. vaardigheden. Dit leverde per item een *OTL-covered rate* en een *OTL-testing rate* op.

Ter operationalisatie van het uitgevoerd curriculum werden de twee toetsen aan een representatieve steekproef van Nederlandse leerlingen voorgelegd. De twee schriftelijke toetsen WT-1995 en WT-1999 kwamen overeen qua aard en moeilijkheidsgraad. Een deel van de opgaven uit WT-1995 was geheim gehouden en ongewijzigd opnieuw opgenomen in WT-1999. Het resterende voor publicatie vrijgegeven deel uit WT-1995 was *gekloond* (gelijke inhoud, maar met andere getallen) voor WT-1999. Na verificatie van vergelijkbaarheid bleken er 41 items bruikbaar voor de trendvergelijking van de prestaties en 144 items bruikbaar voor de overige vergelijkingen.
De items uit PA-1995 en PA-2000 waren identiek waarbij de science-taken waren gehandhaafd zodat de taakinteractie-effecten gelijk zouden blijven Dit zijn effecten waarbij leerlingen bij de ene taak ervaringen opdoen die hen kunnen helpen bij een volgende taak. Voor de analyse van de prestaties waren alleen de wiskundetaken *Dobbelstenen, Rekenmachine, Vouwen en knippen, De bocht om, Inpakken* en de gecombineerde science/wiskunde taken *Schaduwen* en *Klei* van belang.
Na controle van de metingen bleek dat de toetsomstandigheden en het coderen van de leerlingantwoorden niet in alle gevallen vergelijkbare resultaten hadden opgeleverd. Daarom werden de resultaten van de taken *Schaduwen* en *Klei* niet meegenomen in de analyse. De vijf resterende taken omvatten 31 items.

TREND RESULTATEN

Het paarsgewijs vergelijken van de resultaten uit de zes metingen van 1995 (drie voor WT-1995 en drie voor PA-1995) met de resultaten van de zes metingen uit 1999/2000 leverde de volgende observaties op:

Op het niveau van het beoogd curriculum:

- De geschiktheid van WT-1995 en WT-1999 in het licht van de kerndoelen bleef vrijwel constant: in 1995 werden 69% van de opgaven, en in 1999 werden 71% van de opgaven als geschikt aangemerkt door de leerplanexperts. Dit oordeel stemde slechts overeen op 53% van de opgaven. Op 33% van de opgaven wisselde het oordeel tussen 1995 en 1999. Redenen hiervoor konden zijn: de wijzigingen in de wiskundekerndoelen in 1998, wijziging in de interpretatie van de kerndoelen en de methode van aggregatie (de resultaten waren 'absoluut'; een 'twijfelgeval' kon in de ene meting een ander resultaat opleveren dan in de andere meting).
- De geschiktheid van PA-1995 en PA-2000 in het licht van de kerndoelen was hoger dan voor de schriftelijke toets. De gemiddelde *item-curriculum matching index* daalde van 83 naar 72.

Op het niveau van het uitgevoerd curriculum:

- Bij gebrek aan voldoende OTL-data over WT-1995 kon er geen trend worden waargenomen. De OTL-meting van WT-1999 geeft aan dat deze toets als geschikt beschouwd kan worden in het licht van het uitgevoerd curriculum. Een item werd door gemiddeld 82% van de wiskundeleraren aangemerkt als geschikt voor opname in een proefwerk over alle tot op dat moment behandelde leerstof. Slechts 6 van de 144 items had een *OTL rate* lager dan 50.
- De praktische toets bleek minder goed aan te sluiten bij het uitgevoerd curriculum dan de schriftelijke toets. De gemiddelde *OTL-covered rate* was 38 voor PA-1995 en steeg tot 58 voor PA-2000. De gemiddelde *OTL-testing rate* was 51 voor PA-1995 en steeg spectaculair naar 76 voor PA-2000. Daarmee was deze test nog steeds minder geschikt dan de schriftelijke toets (in het licht van het uitgevoerd curriculum), maar de toegenomen waarden duidden op een attitudeverandering onder leraren.

Op het niveau van het bereikte curriculum:

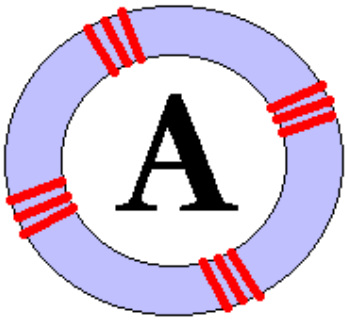
- Op de schriftelijke toets stegen de scores licht op de 41 identieke items uit WT-1995 en WT-1999. Deze toename kon als klein doch significant worden aangemerkt.
- Op de praktische toets waren geen verschillen waar te nemen tussen de leerling resultaten van PA-1995 en PA-2000.

ONDERZOEKSRESULTATEN

De eerste onderzoeksvraag had betrekking op de leerlingresultaten van WT-1995 en PA-1995. In de internationale vergelijking waren de Nederlandse leerlingen redelijk 'hoog' geëindigd (significant boven het internationale gemiddelde) in tegenstelling tot het resultaat op de praktische toets.
De replicatiestudies leverden twee nieuwe inzichten op. Enerzijds bleek dat de bovenbeschreven tegenstelling deels berustte op het *ranglijstdenken* waarin de olympische volgorde van landen in de internationale tabellen de gedachten domineerde. Indien naar de scores werd gekeken, met inachtneming van de statistische onzekerheidsmarges, dan bleken beide toetsresultaten minder divergent dan de ranglijst deed vermoeden. Anderzijds bleek dat de resultaten van de taak *Klei* uit PA-1995 onbetrouwbaar waren geweest en de score van de Nederlandse leerlingen enigszins omlaag trokken. Na verwijdering van deze taak en herberekening van de scores van de deelnemende landen bleek de score op de praktische toets goed overeen te komen met die op de schriftelijke toets. De verschillen in prestaties op de twee toetsen uit 1995 bleken daarmee niet te bestaan. De leerlingprestaties in 1995 waren op beide toetsen in de internationale vergelijking 'goed' te noemen (een gemiddelde score boven het internationale gemiddelde), zoals ze dat ook al waren geweest in 1982 op eerder internationaal vergelijkend onderzoek.

De tweede onderzoeksvraag relateerde de leerlingprestaties op beide toetsen aan de geschiktheidsoordelen in het licht van het beoogd en het uitgevoerd curriculum.
Zoals reeds gemeld was het geschiktheidsoordeel over de toetsitems in het licht van de Kerndoelen enigszins verschoven. Door deze oordelen aan de leerlingprestaties te koppelen werd ontdekt dat de experts in 1995 meer 'moeilijke' opgaven geschikt geacht hadden (opgaven die door een kleiner aantal leerlingen correct kon worden voltooid) dan in 1999/2000. Hieruit volgde dat het beoogde en het bereikte curriculum in 1999/2000 meer op elkaar afgestemd waren dan in 1995.
De analyse van de relatie tussen leerlingprestaties en het geschiktheidsoordeel in het licht van het uitgevoerde curriculum werd enigszins bemoeilijkt door tekort aan data over WT-1995 en doordat de *OTL-testing rates* mogelijk vervuild werden omdat wiskundeleraren weinig ervaring hebben met het samenstellen van praktische toetsen. Desondanks kon worden geconstateerd dat het uitgevoerde

en het bereikte curriculum in 2000 meer op elkaar afgestemd waren dan in 1995. Tevens bleek dat de geschiktheidsoordelen verstrekt door de leerplanexperts en de wiskundeleraren in 1999/2000 beter correleerden dan in 1995.
Met deze analyses bleek dat de internationaal relatief goede leerlingprestaties op de schriftelijke toets niet goed terug te voeren waren op de mindere geschiktheid van deze toets in het licht van de kerndoelen, maar wel op de hoge mate van geschiktheid van deze toets in het licht van het uitgevoerd curriculum. Toch namen de prestaties van de leerlingen met name toe op de opgaven die door de experts als geschikt waren aangemerkt. Dit kan worden toegeschreven aan een toegenomen afstemming tussen beoogd en uitgevoerd curriculum.
De internationaal relatief goede leerlingprestaties op de praktische toets konden niet worden teruggevoerd op de mindere geschiktheid van deze toets in het licht van het uitgevoerd curriculum. Wel is het verheugend om een toegenomen belangstelling onder wiskundeleraren voor deze alternatieve toetsmogelijkheid waar te nemen. Indien deze attitudeverschuiving benut en ondersteund wordt kan dit leiden tot een verrijking van het wiskundeonderwijs in de basisvorming.

Appendix 

Exemplary, RME-based items for grade 8

Tijdsverschil

Wanneer je vanuit Nederland met iemand in het buitenland wilt telefoneren, moet je rekening houden met het tijdsverschil.
In de tabel hieronder staat het tijdsverschil tussen Nederland en een aantal andere landen.

Land	*Tijdsverschil vanuit Nederland*
België	0 uur
Chili	- 5 uur
Griekenland	+ 1 uur
Marokko	- 1 uur
Nieuw-Zeeland	+ 12 uur
Suriname	- 4 uur
Zweden	0 uur

1. Een zakenman in Rotterdam wil 's morgens om 10 uur een bedrijf in Chili bellen.
   Leg uit waarom het niet verstandig is om op die tijd op te bellen.

   Antwoord:

2. In Nederland is het 14.30 uur.
   Hoe laat is het op dat moment in Nieuw-Zeeland?

   Antwoord:

3. Wat is het tijdsverschil tussen Marokko en Griekenland?
   Geef een toelichting.

   Antwoord:

Kandelaar

Petra heeft foto's gemaakt van een kandelaar met vier kaarsen erin.
Alle kaarsen zijn tegelijk aangestoken.
Op foto A zie je dat de kandelaar vier kaarsen heeft.

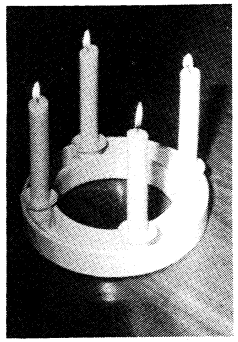


foto A

Foto B is van bovenaf genomen.

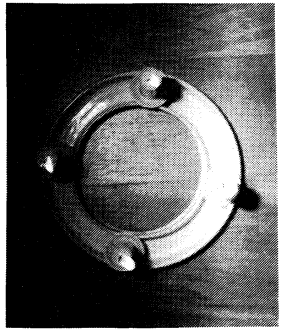
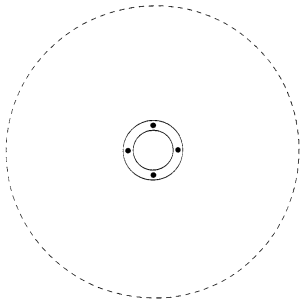


foto B

Hieronder is een bovenaanzicht getekend van de kandelaar met de kaarsen erin. In het midden zie je de kandelaar met de kaarsen. De stippellijn geeft de plaatsen aan van waaruit foto's worden gemaakt.

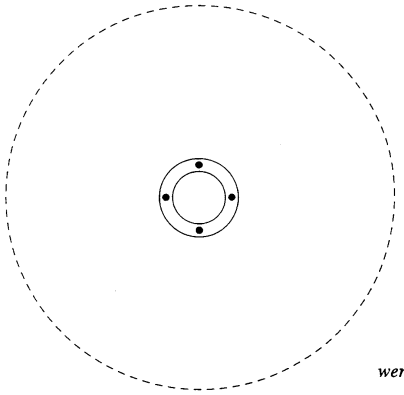


bovenaanzicht

In de volgende opgaven is het bovenaanzicht vergroot als werktekening.

Je kunt een foto maken waarbij je precies drie kaarsen ziet.

4. Geef in de werktekening hieronder één plaats op de stippellijn aan van waaruit die foto genomen kan zijn. Uit je tekening moet duidelijk zijn hoe je die plaats gevonden hebt.



werktekening

Hieronder staat foto C.

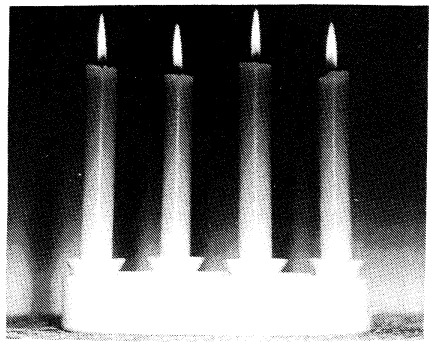
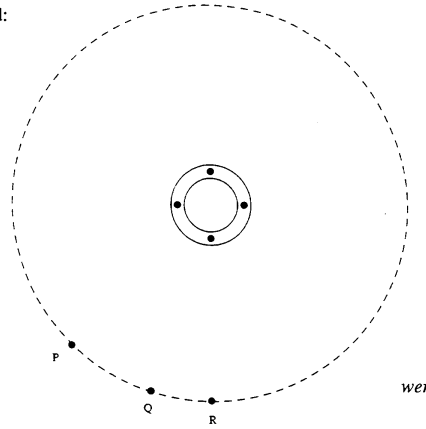


foto C

Hieronder zie je de werktekening nog een keer.

5. Vanaf welke plaats op de stippellijn kan foto C genomen zijn, plaats P, Q of R?

   Antwoord:



werktekening

6. Geef in de werktekening bij vraag 5 nóg twee plaatsen aan van waaruit foto C genomen kan zijn. Noem die plaatsen S en T. Uit je tekening moet duidelijk zijn hoe je de plaatsen van S en T gevonden hebt.

Lengte

Om te berekenen hoe lang een meisje ongeveer zal worden, gebruikt een schoolarts de volgende formule:

$$\textit{lengte dochter}\ (\text{in cm}) = \frac{\textit{lengte vader}\ (\text{cm}) + \textit{lengte moeder}\ (\text{cm}) - 12}{2} + 3$$

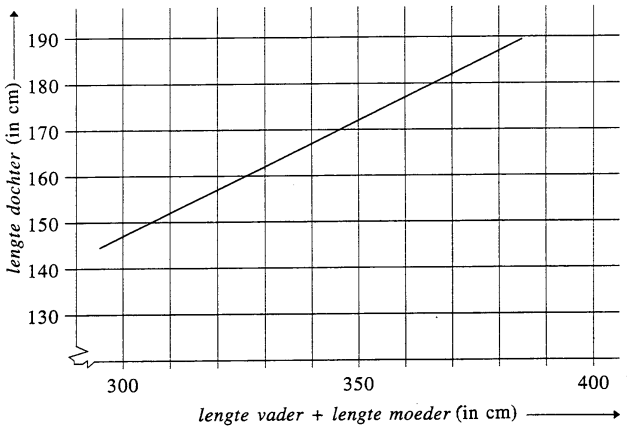
7. De vader van Daniëlle is 1,82 m lang, haar moeder is 1,68 m.
   Hoe lang zal Daniëlle volgens deze formule worden?

   Antwoord:

8. Is het mogelijk dat - volgens de formule - een dochter groter wordt dan haar vader?
   Licht je antwoord toe. Dat mag ook met een voorbeeld.

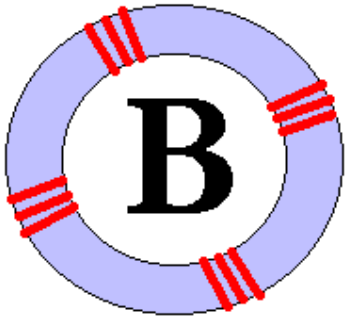
   Antwoord:

Hieronder is de grafiek getekend die hoort bij de formule.



Wanneer de dochter kleiner blijft dan 1,57 m, is ze "nogal klein".
Wanneer de dochter groter wordt dan 1,80 m, is ze "nogal groot".

9. Geef in de tekening hierboven op de grafiek aan welk gedeelte hoort bij "nogal klein" en welk gedeelte hoort bij "nogal groot".
   Gebruik een kleurpotlood en zet erbij "nogal klein" en "nogal groot".

Appendix 

Core objectives of the intended mathematics curriculum for junior secondary schools, 1998-2003

Numbered items refer to general objectives.
Non-numbered items refer to specific mathematical objectives.

Content	Vertical mathematising		Horizontal mathematising	
	Reflecting	Reformulating	Modelling	Interpreting
Arithmetic, measuring and estimating	Develop and learn to apply calculating skills (2.4). Develop reasoning skills to tackle problems (3.2 and 3.3). Work systematically (3.4 and 3.5). Process personal experiences and take a position based on arguments (3.6 and 3.7). Communicate (4). Reflect on the learning process (5). Reflect on the future (6).	Understand relations between proportions, fractions and decimal numbers and use arithmetic models to perform calculations. Work with current measures for length, area, volume, time, angles and money. Select appropriate instrument for operations (calculator, mental arithmetic or calculate). Use calculator adequately for conversion of fractions, percentage, roots, and exponents. Estimate.	Relate concrete phenomena to mathematics (1.2). Work with proportions and scale. In meaningful situations order, add and subtract negative numbers.	Relate mathematics to concrete phenomena (1.2). Work with proportions and scale.
Algebraic relations	Develop reasoning skills to tackle problems (3.2 and 3.3). Work systematically (3.4 and 3.5). Process personal experiences and take a position based on arguments (3.6 and 3.7). Communicate (4). Reflect on the learning process (5). Reflect on the future (6).	Transform phrase, table, (word-) formula and graph into another representation. With a given graph, determine whether the relationship is constant, increasing, decreasing or periodical. Determine when two relationships have the same outcome, on which interval one is larger than the other. Determine regularity in number patterns and tables, characterise and extend these. Use computers for working with relationships	Relate concrete phenomena to mathematics (1.2). Transform simple relations between two variables from reality into phrase, table, (word-) formula and graph. Process changes in relations into phrase, table, (word-) formula and graph.	Relate mathematics to concrete phenomena (1.2). Transform phrase, table, (word-) formula and graph into simple relations between two variables from reality. Read off, compare and interpret relationships. Recognise characteristics and interpret these. Draw conclusions for the connected context from specific points, course and shape of a graph.
Geometry	Develop reasoning skills to tackle problems (3.2 and 3.3). Work systematically (3.4 and 3.5). Process personal experiences and take a position based on arguments (3.6 and 3.7). Communicate (4). Reflect on the learning process (5). Reflect on the future (6).	Use concepts like parallel, perpendicular, direction. Describe regularities and properties of geometrical objects. Estimate, measure and calculate angles, lengths, areas and volumes. Use instruments.	Relate concrete phenomena to mathematics (1.2). Make views, nets and designs. Describe, imagine and represent 3-d situations through 2-d pictures. Use scale.	Relate mathematics to concrete phenomena (1.2). Interpret 2-d pictures of 3-d situations.
Data processing and statistics	Gather, describe and arrange data (2.3). Develop reasoning skills to tackle problems (3.2 and 3.3). Work systematically (3.4 and 3.5). Process personal experiences and take a position based on arguments (3.6 and 3.7). Communicate (4). Reflect on the learning process (5). Reflect on the future (6).	Manipulate data from tables, graphs, and diagrams. Use computers.	Relate concrete phenomena to mathematics (1.2). Visualise data for problems from reality, assess whether these give an adequate picture.	Relate mathematics to concrete phenomena (1.2). Read off, and interpret statistical representations. Extrapolate from models with respect to expectations for future events and developments.

Source: Ministerie van OC&W (1998)

Appendix C

Exemplary mathematics tasks from the TIMSS Performance Assessment

DOBBELSTENEN

Bij deze opstelling heb je nodig:

Een dobbelsteen
Een beker

Lees nauwkeurig ALLE aanwijzingen.

Je opdracht:

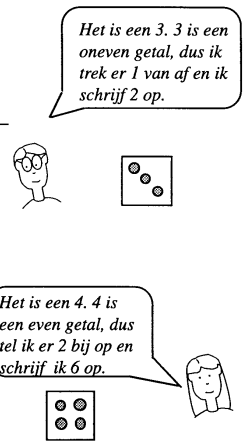
Onderzoek wat er gebeurt als we een regel gebruiken om de getallen te veranderen die je krijgt als je met een dobbelsteen gooit.

De regel om de getallen te veranderen is:

- Als je een ONEVEN getal gooit, trek je er 1 van af. Schrijf het getal op dat je dan krijgt.
- Als je een EVEN getal gooit, tel je er 2 bij op. Schrijf het getal op dat je dan krijgt.

1. In de tabel hieronder zijn twee nieuwe getallen al voor je ingevuld. Gebruik de regel om uit te zoeken wat de andere nieuwe getallen zijn. Noteer die in de tabel.

Getal op de dobbelsteen	Nieuwe getallen
1	
2	
3	2
4	6
5	
6	



2. Wat valt je op aan de getallen die je in de tabel hebt genoteerd?

3. Gooi 30 keer met de dobbelsteen. Verander bij iedere worp het gegooide getal volgens de regel. Noteer in de tabellen hieronder na iedere worp welk getal je hebt gegooid en wat het nieuwe getal is.

Getal op de dobbelsteen	Nieuwe getal

Getal op de dobbelsteen	Nieuwe getal

Sla de pagina om.

4. Kijk nog eens naar de tabel die je bij vraag 3 hebt ingevuld. Hoe vaak heb je de volgende getallen genoteerd in de kolom 'Nieuwe getal'?

Nieuwe getal	Aantal keren genoteerd
0	
1	
2	
3	
4	
5	
6	
7	
8	

5a. Welk nieuwe getal heb je het vaakst genoteerd?

5b. Waarom is het op deze manier gegaan?

LEG ALLES WEER TERUG ZOALS JE HET AANGETROFFEN HEBT.

OOK IEMAND ANDERS KAN DE OPSTELLING DAN WEER GEBRUIKEN.

REKENMACHINE

Bij deze opstelling heb je nodig:

Een rekenmachine

Je opdracht:

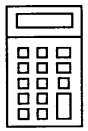
Onderzoek met behulp van een rekenmachine de regelmaat in enkele getallen en probeer vervolgens de ontbrekende getallen te vinden.

Lees deze aanwijzingen voordat je antwoord geeft op de vragen.

Als je de rekenmachine gebruikt:

- Zorg er voor dat je op de goede toetsen drukt.
- Zorg er voor dat je het venster zorgvuldig afleest.

1. Reken met behulp van de rekenmachine de volgende vermenigvuldigingen uit.



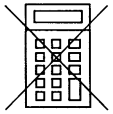
34×34 = _____

334×334 = _____

3334×3334 = _____

2. Wat valt je op aan de vermenigvuldigingen en aan de regelmaat in de uitkomsten?

3. Kijk nog eens goed naar de regelmaat in de uitkomsten en noteer wat, volgens jou, de uitkomst is van de volgende vermenigvuldiging. Je mag hierbij GEEN rekenmachine gebruiken.



$33334 \times 33334 =$ _____

4. Noteer wat ,volgens jou, de uitkomst is van de vermenigvuldiging hieronder. Je mag hierbij GEEN rekenmachine gebruiken.

$3333334 \times 3333334 =$ _____

5. Hoe ben je aan de antwoorden op vraag 3 en 4 gekomen?

6. Tjeerd zegt tegen Sandra dat hij met behulp van een rekenmachine twee gehele getallen met elkaar vermenigvuldigd heeft en dat het antwoord 455 was, maar hij is de getallen vergeten. Hij kan zich er nog twee dingen van herinneren:
   - Beide getallen bestonden uit twee cijfers.
   - Beide getallen waren kleiner dan 50.

Sla de bladzijde om.

Sandra probeert verscheidene getallen. Ze begint met 6 x 64 met de rekenmachine uit te rekenen. Maar Tjeerd zegt: "Ik kan je tenminste drie redenen geven waarom dat niet de getallen zijn die ik heb gebruikt". Welke redenen zijn dat?

a.

b.

c.

Na een tijdje over het probleem te hebben nagedacht, probeert Sandra het nog een paar keer en vindt de twee getallen.

- Probeer nu zelf de getallen te vinden die Sandra heeft gevonden.

Je mag dat doen op de manier(en) die jij het beste vindt. Schrijf elke manier op die je probeert.

VOUWEN EN KNIPPEN

Bij deze opstelling heb je nodig:

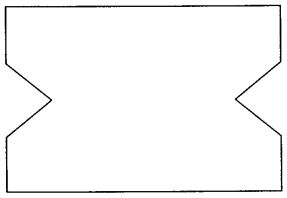
9 velletjes papier
Een schaar
Een envelop

Je opdracht:

Vouw en knip velletjes papier op zo'n manier dat ze overeenkomen met gegeven vormen. Voor iedere vorm mag je het papier zo vaak vouwen als je wilt, maar je mag maar EEN KEER RECHT knippen.

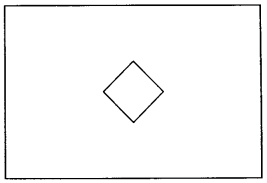
1. Kijk eens naar vorm 1 hieronder. Vouw een vel papier zo vaak als nodig is en knip EEN KEER RECHT zodat, wanneer je het vel papier open vouwt, het dezelfde VORM heeft als vorm nummer 1. Je velletje papier en de weggeknipte stukjes hoeven niet even GROOT te zijn als in de tekening is aangegeven.. Als het niet gelukt is, mag je het nog een keer proberen met een ander stukje papier. Je mag het in totaal DRIE keer proberen.
   - Schrijf nummer 1 op ieder velletje papier dat je voor deze opdracht gebruikt hebt.
   - Schrijf je voornaam op ieder velletje papier.

Vorm 1



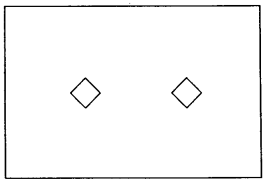
2. Doe hetzelfde voor vorm 2. Denk eraan dat je maar EEN KEER RECHT mag knippen. Je mag het in totaal DRIE keer proberen.
   - Schrijf nummer 2 op ieder velletje papier dat je voor deze opdracht gebruikt hebt.
   - Schrijf je voornaam op ieder velletje papier.

Vorm 2



3. Doe hetzelfde voor vorm 3. Denk eraan dat je maar EEN KEER RECHT mag knippen. Je mag het in totaal DRIE keer proberen.
   - Schrijf nummer 3 op ieder velletje papier dat je voor deze opdracht gebruikt hebt.
   - Schrijf je voornaam op ieder velletje papier.

Vorm 3



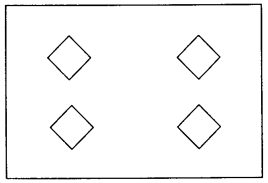
Sla de bladzijde om.

4. Hieronder is vorm 4 getekend. In plaats van vorm 4 te knippen of te vouwen, vragen we van je NA TE DENKEN over hoe je deze vorm zou kunnen maken door te vouwen en één keer recht te knippen. Dus GEEN PAPIER VOUWEN OF KNIPPEN BIJ DEZE VRAAG.

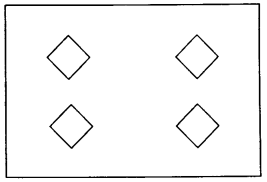
   Teken in plaats daarvan in de tekening hieronder de LIJNEN die je zou zien op een velletje papier dat gevouwen en geknipt was.

   Hieronder staan twee tekeningen van vorm 4 voor het geval dat je niet tevreden bent met je eerste poging en je het nog eens wilt proberen. Denk eraan, teken alleen lijnen om te laten zien waar het vel papier gevouwen zou moeten worden.

Vorm 4



Vorm 4



STOP AL JE VELLETJES PAPIER IN JE ENVELOP, OOK JE MISLUKTE POGINGEN EN ZET JE NAAM OP DE ENVELOP.

GOOI ALLE RESTJES PAPIER WEG.

DE BOCHT OM

Bij deze opstelling heb je nodig:

Twee kleine kartonnen rechthoeken (A en B) die meubelstukken voorstellen
Ruitjespapier. Hiermee kun je rechthoeken maken die andere meubelstukken voorstellen
Een schaar
Een liniaal van 30 centimeter lang
Een plastic zak en stickers
Paperclips
Een kartonnen model dat een gang in een huis voorstelt

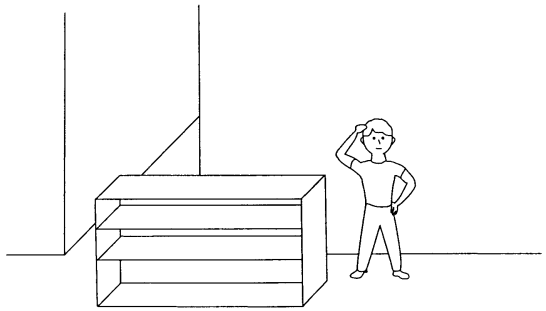
Je opdracht:

Ga na welke meubelstukken van een bepaalde afmeting door de bocht in een gang kunnen.

Lees dit voordat je antwoord geeft op de vragen:

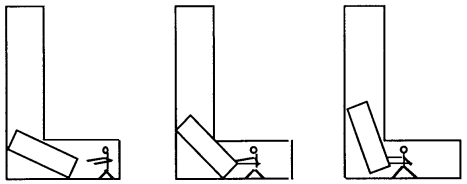
Ray gaat verhuizen. Als je zijn nieuwe huis via de voordeur binnen gaat, kom je in een gang. De belangrijkste kamers liggen aan het einde van deze gang. In de gang zit een bocht.

Welke meubelstukken kunnen, wat betreft de afmetingen, door deze bocht in de gang?



Ray wil een paar grote meubelstukken de bocht om zien te krijgen met de goede kant boven. Hij wil de meubelstukken niet op hun zijkant draaien. Hij gebruikt de kartonnen modellen van de gang en de meubelstukken om uit te zoeken welke meubelstukken de bocht om kunnen.

Hieronder zie je een paar tekeningen (niet op schaal) die laten zien wat er zou kunnen gebeuren.



Je hebt twee kleine kartonnen rechthoekjes die meubelstukken voorstellen en een kartonnen model van de gang in Ray's nieuwe huis. Rechthoekjes en model zijn op schaal. *Schaal: 4 centimeter staat voor 1 meter.*

1. Meet de lengte en de breedte van de twee modellen van meubelstukken in centimeters.

   A is _____ cm lang en _____ cm breed.

   B is _____ cm lang en _____ cm breed.

2. Wat is de lengte en de breedte van de twee meubelstukken in meters?

   A is _____ m lang en _____ m breed.

   B is _____ m lang en _____ m breed.

3. Hieronder staan enkele meubelstukken:

éénpersoonsbed	salontafel	3-persoonsbank	leunstoel	wieg
tweepersoonsbed	eettafel	2-persoonsbank	porseleinkast	

   Afgaande op hun grootte:

   Welk meubelstuk is waarschijnlijk A?_____

   Welk meubelstuk is waarschijnlijk B?_____

4. Welk meubelstuk of welke meubelstukken (A of B of allebei) kunnen de bocht om in de gang van Ray's nieuwe huis en welke niet?

Sla de bladzijde om.

5. Gebruik het grafiekpapier om andere modellen van meubelstukken te maken volgens de maten zoals opgegeven in de tabel hieronder. Alle maten zijn gegeven in meters.

   Geef in de tweede kolom van de tabel aan wat het meubelstuk zou kunnen zijn.

   Zoek in de derde kolom uit of het meubelstuk de hoek om kan en controleer het goede antwoord.

	Maten van meubelstuk		Welk meubelstuk	Kan het de bocht om?		
	Lengte (m)	Breedte (m)	zou dit kunnen zijn?	Ja, makkelijk	Ja, net	Nee
C	0,5	0,5				
D	1,5	0,5				
E	2	0,5				
F	1	1				
G	1,5	1				
H	2	1				

6. Of een meubelstuk al dan niet de hoek om kan, hangt af van zijn lengte en breedte. Bekijk de resultaten die je hebt voor alle meubelstukken A, B, C, D, E, F, G en H.
   - Probeer een regel te vinden om uit de lengte en breedte van een meubelstuk af te leiden of een meubelstuk al dan niet de bocht om kan.

STOP DE MEUBELSTUKKEN DIE JE HEBT GEMAAKT IN DE PLASTIC ZAK EN SCHRIJF JE NAAM OP DE STICKER.

MAAK DE ZAK AAN DEZE BLADZIJDE VAST MET EEN PAPERCLIP.

LAAT DE MODELLEN A EN B BIJ DE OPSTELLING ACHTER.

INPAKKEN

Bij deze opstelling heb je nodig::

4 pingpong-ballen in een vierkant doosje
Dubbelzijdig plakband om ervoor te zorgen dat de ballen niet gaan rollen
Wit papier om een doosje met de ballen op te tekenen
Een passer
Een liniaal van 30 centimeter lang
Twee stukken karton om de ballen mee op te meten
Een schaar
Plakband

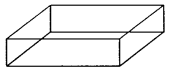
Je opdracht:

Ontwerp een aantal doosjes waar de 4 pingpong-ballen precies in passen.

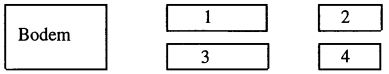
Lees dit voordat je antwoord geeft op de vragen:

Het volgende laat zien wat we bedoelen met de uitslag van een doosje.

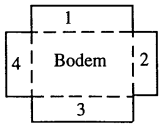
Dit doosje heeft een bodem en 4 zijkanten.



De bodem en de zijkanten kunnen ieder apart uitgeknipt worden:

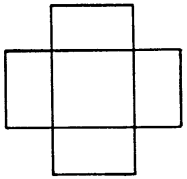


Je kunt de bodem en de zijkanten ook uit één stuk knippen en daarna langs de stippellijnen vouwen, zoals in de tekening hieronder:

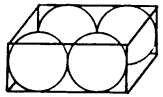


Dit is de uitslag van een doosje.

Dit is de vorm van de uitslag van het doosje waar de 4 pingpong-ballen in zitten. De uitslag is niet op schaal getekend, maar als dat wel zo was, zou je de zijkanten omhoog kunnen vouwen en het doosje kunnen maken.



De vier pingpong-ballen passen precies in het doosje dat je hebt gekregen (zie de tekening hieronder).



Het is mogelijk meer doosjes te maken waar de ballen precies in passen, maar die anders van vorm zijn.

. Gebruik de ballen om 3 andere doosjes te bedenken waar de 4 ballen precies in passen. Maak een tekening van ieder doosje met de 4 ballen erin.

Sla de bladzijde om.

2. Maak nu een tekening van de uitslag van ieder doosje.

3. Kies EEN van de doosjes die je getekend hebt. Neem het witte velletje papier. Teken op dit papier de uitslag van het doosje dat je hebt gekozen. Teken de uitslag op ware grootte. Als je het doosje echt zou maken zouden de 4 ballen er precies in moeten passen.

MAAK DE UITSLAG AAN DEZE BLADZIJDE VAST MET EEN PAPERCLIP.

LAAT ALLE ANDERE DINGEN ACHTER ZOALS JE ZE AANGETROFFEN HEBT.

SCHADUWEN

Bij deze opstelling heb je nodig:

Een zaklamp op een verhoging (dit noemen we 'de lamp')
Een stuk karton van 5 cm^2 op een standaard
Een stuk wit papier (van 30 x 40 cm) om een schaduw van de kaart op te maken (dit noemen we 'het scherm')
Een papieren meetlint van een meter lang
Een liniaal van 30 centimeter lang

Lees nauwkeurig <u>ALLE</u> aanwijzingen.

Als het stukje karton tussen de lamp en het scherm staat, komt er een schaduw van het stukje karton op het scherm.

Je opdracht:

Onderzoek hoe de grootte van de schaduw verandert als je het stukje karton beweegt.

Dit moet je doen:

- Houd het stukje karton stil en beweeg de lamp naar het stukje karton toe en van het stukje karton af.

1. Wat gebeurt er met de grootte van de schaduw?

2. Waarom is de schaduw altijd groter dan het stukje karton? Je mag bij je antwoord een tekening/diagram maken.

3. Zoek nu tenminste drie plaatsen waar je de lamp en de standaard met het stukje karton kunt neerzetten om een schaduw te maken die twee keer zo breed is als het stukje karton. Meet de afstand van het stukje karton tot het scherm en van de lamp tot het stukje karton voor elk van de drie plaatsen. Schrijf alle gemeten afstanden op.

Je gaat nu via een proef een algemene regel proberen te vinden voor hoever de lamp en het stukje karton van het scherm af moeten staan om de schaduw te maken die twee keer zo breed is als het stukje karton.

Zorg ervoor dat je:

- weet wat je gaat meten
- weet hoe je je meetresultaten duidelijk en eenvoudig weer zult geven.
- conclusies uit je meetresultaten kunt trekken.

4. Beschrijf hoe je de proef hebt uitgevoerd. Bedenk dat je bij de beschrijving een tekening zou kunnen maken.

5. Geef je meetresultaten zo duidelijk mogelijk weer.

6. Welke algemene conclusie kun je uit deze resultaten trekken? Probeer een regel te bedenken die zegt wanneer de schaduw altijd twee keer zo breed is als het stukje karton.

LEG JE MATERIAAL WEER TERUG ZOALS JE HET BIJ HET BEGIN VAN DE OPDRACHT HEBT AANGETROFFEN.

OOK IEMAND ANDERS KAN DEZE OPSTELLING DAN WEER GEBRUIKEN.

ELASTIEKJE

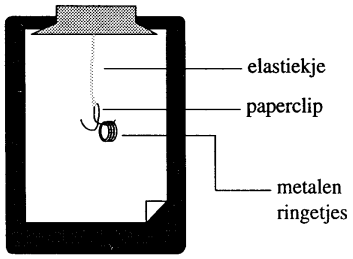
Bij deze opstelling heb je nodig:

Een klembord (zie tekening) met daaraan vastgeklemd een elastiekje
Een grote paperclip die aan het uiteinde van het elastiekje is bevestigd
Metalen ringetjes om aan de grote paperclip te hangen
Een liniaal van 30 centimeter lang
Enkele velletjes papier zonder lijntjes
2 velletjes grafiekpapier

Lees nauwkeurig ALLE aanwijzingen.

Je opdracht:

Onderzoek hoe de lengte van het elastiekje verandert als er steeds meer ringetjes aan worden gehangen.



Dit moet je doen:

- Hang de metalen ringetjes één voor één aan de paperclip.
- Meet de lengte van het elastiekje iedere keer nadat je er een nieuw ringetje aan hebt gehangen.
- Noteer je meetresultaten in de tabel.

1. Noteer je meetresultaten in de tabel hieronder. Denk eraan om boven iedere kolom een koptekst te zetten.

2. Maak een grafiek van je resultaten op het grafiekpapier. Je mag een lijn- of een staafdiagram maken.

BEANTWOORD DE VRAGEN 3 TOT EN MET 6 DOOR GEBRUIK TE MAKEN VAN JE TABEL EN JE GRAFIEK.

3. Als er 2 ringetjes aan de paperclip hangen en er worden er nog 3 bij gehangen, hoeveel langer wordt het elastiekje dan?

   Het elastiekje wordt dan _____ centimeter langer.

4. Beschrijf hoe de lengte van het elastiekje veranderde toen er steeds meer ringetjes aan werden gehangen.

Sla de bladzijde om.

5. Hoe lang denk je dat het elastiekje wordt als je er nog 2 ringetjes meer aan zou kunnen hangen dan je in werkelijkheid hebt?

   Ik denk dat het elastiekje in totaal _____ centimeter lang zal worden.

6. Waarom denk je dat?

LEG ALLES WEER TERUG ZOALS JE HET AANGETROFFEN HEBT.

OOK IEMAND ANDERS KAN DE OPSTELLING DAN WEER GEBRUIKEN.

KLEI

Bij deze opstelling heb je nodig:

Klei
Een balans
Een plastic zak
Een gewicht van 20 gram en een gewicht van 50 gram
Kleine gekleurde ronde stickers

Lees nauwkeurig <u>ALLE</u> aanwijzingen.

Je opdracht:

Gebruik de balans om zo zorgvuldig mogelijk enkele stukjes klei af te wegen. Geef vervolgens aan hoe je die stukjes gemaakt hebt.

Voordat je aan je opdracht begint:

ZORG DAT DE SCHALEN VAN DE BALANS IN EVENWICHT ZIJN ALS ZE LEEG ZIJN. ALS DAT NIET ZO IS, STEEK DAN JE VINGER OP EN VERTEL HET AAN DE TOETSLEIDER.

1a. Gebruik de balans om een stukje klei te maken dat 20 gram weegt.

- Wanneer je het stukje van 20 gram hebt gemaakt, schrijf dan '20 gram' op een sticker en plak die sticker op het stukje klei. Stop het stukje klei in de plastic zak.

1b. Schrijf op hoe je het stukje van 20 gram hebt gemaakt.

2a. Gebruik de balans om een stukje klei te maken dat 10 gram weegt.

- Wanneer je het stukje van 10 gram hebt gemaakt, schrijf dan '10 gram' op een sticker en plak die sticker op het stukje klei. Stop het stukje in de plastic zak bij het stukje van 20 gram.

2b. Schrijf op hoe je het stukje van 10 gram hebt gemaakt.

3a. Gebruik de balans om een stukje klei te maken dat 15 gram weegt.

- Wanneer je het stukje van 15 gram hebt gemaakt, schrijf dan '15 gram' op een sticker en plak die sticker op het stukje klei. Stop het stukje in de plastic zak bij de andere twee stukjes klei.

3b. Schrijf op hoe je het stukje van 15 gram hebt gemaakt.

4a. Gebruik de balans om een stukje klei te maken dat 35 gram weegt.

- Wanneer je het stukje van 35 gram hebt gemaakt, schrijf dan '35 gram' op een sticker en plak die op het stukje klei. Stop het stukje klei in de plastic zak bij de andere stukjes klei.

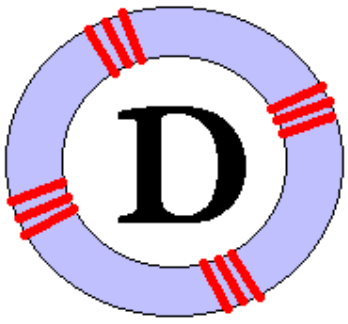
4b. Schrijf op hoe je het stukje van 35 gram hebt gemaakt.

LEVER DE ZAK MET DE AFGEWOGEN STUKJES KLEI IN BIJ DE TOETSLEIDER.

ZORG ERVOOR DAT JE NAAM OP DE ZAK STAAT.

LAAT DE REST ACHTER ZOALS JE HET BIJ HET BEGIN VAN DE OPDRACHT HEBT AANGETROFFEN.

Appendix



Exemplary items from the TIMSS Written Test, 1995-1999

This appendix contains a selection of mathematics items in the Written Test (WT-1995 and WT-1999). For each item, it is indicated: (1) in which test it was used, and (2) when it was released.

The items were used in the Netherlands and thus, they are stated in Dutch. The items are indicated by an alphabetical character and a number (e.g. item 'B08'). The character indicates a cluster of items (e.g. cluster 'B'). The items were rotated as clusters over the eight test booklets.

The selection includes the 16 *anchor items* from the National Option Test used in 1995 (Kuiper, Bos & Plomp, 1997, 1999). The anchor items were selected for matching well with the intended curriculum. They are indicated by an asterisk (*).

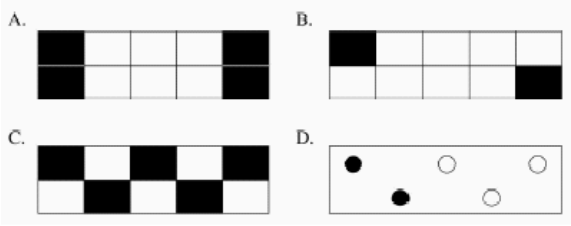
To illustrate the similarity between clones from WT-1995 and WT-1999, these are paired in the list below.

Some of the items are not fully revealed, as the TIMSS Item Release Policy does not allow their publication. These are described by a rough description *in italics* (as published in the TIMSS-99 Almanacs).

A02* (identical in WT-1995 and WT-1999; not yet released).
Objects balanced on scale. Format: multiple choice. Topic: algebra.

B08* (identical in WT-1995 and WT-1999; released after 1999).
Als er in 100 g van een bepaald voedingsmiddel 300 calorieën zitten, hoeveel calorieën zitten er dan in een portie van 30 g van ditzelfde voedingsmiddel?

A. 90 B. 100 C. 900 D. 1000 E. 9000

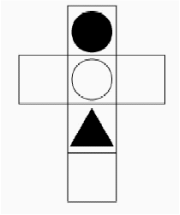


B09 (identical in WT-1995 and WT-1999; released after 1999).

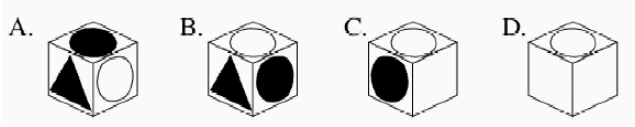
Welke figuur geeft aan dat $\frac{2}{5}$ gelijk is aan $\frac{4}{10}$?

B10* (identical in WT-1995 and WT-1999; released after 1999).
Welke van de onderstaande getallen is het kleinst?

A. 0,625 B. 0,25 C. 0,375 D. 0,5 E. 0,125



B11* (identical in WT-1995 and WT-1999; released after 1999).



Met de uitslag wordt een kubus gevouwen. Zoek uit welke kubus je dan krijgt.

B12 (identical in WT-1995 and WT-1999; released after 1999).
n is een getal. Als je *n* vermenigvuldigt met 7 en er dan 6 bij optelt, is de uitkomst 41. Welke vergelijking hoort hierbij?

A. 7n+6 = 41 B. 7n-6 = 41 C. 7n x 6 = 41 D. 7(n+6) = 41

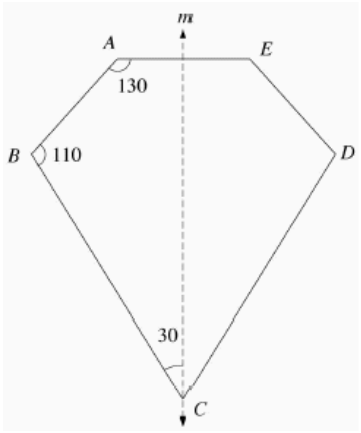
With permission from ISC, in 1999 the alternatives of item B12 were altered into:

A. 7•n+6 = 41 B. 7•n-6 = 41 C. 7•n x 6 = 41 D. 7•(n+6) = 41

C05* (identical in WT-1995 and WT-1999; not yet released).
Number of matchsticks continuing pattern. Format: multiple choice. Topic: algebra.

D07 (identical in WT-1995 and WT-1999; released after 1999).
Lijn *m* is een symmetrie-as van figuur ABCD. De grootte van hoek BCD is

A. 30°
B. 50°
C. 60°
D. 70°
E. 110°

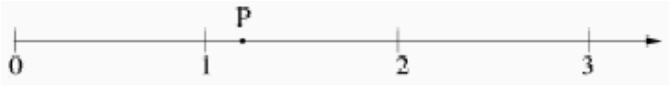


D08 (identical in WT-1995 and WT-1999; released after 1999).
7 en 13 verhouden zich als *x* en 52. Hoe groot is *x*?

A. 7 B. 13 C. 28 D. 364

D11 (identical in WT-1995 and WT-1999; released after 1999).
In welke eenheid kun je het gewicht (massa) van een ei het beste uitdrukken?

A. centimeter B. millimeter C. gram D. kilogram



D12 (identical in WT-1995 and WT-1999; released after 1999).
Wat is de beste schatting voor het getal bij punt P op de getallenlijn?

A. 1,2 B. 1,2 C. 1,4 D. 1,5

E05* (identical in WT-1995 and WT-1999; not yet released).
Sets of ordered pairs of numbers. Format: multiple choice. Topic: algebra.

F07 (identical in WT-1995 and WT-1999; released after 1999).
Een hardloopster rent 3000 meter in precies 8 minuten. Wat is haar gemiddelde snelheid, in meters per seconde?

A. 3,75 B. 6,25 C. 16,0 D. 37,5 E. 62,5

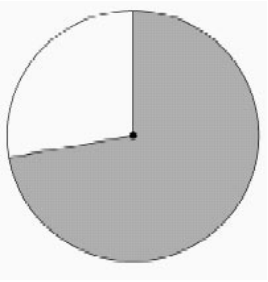
F08 (identical in WT-1995 and WT-1999; released after 1999).
Als een geldstuk wordt opgegooid is de kans dat kop boven komt gelijk aan ½. In vier worpen achter elkaar komt het geldstuk telkens met kop boven. Wat gebeurt er waarschijnlijk als de munt een vijfde keer opgegooid wordt?

A. Het is waarschijnlijker dat munt boven komt in plaats van kop.
B. Het is waarschijnlijker dat kop boven komt in plaats van munt.
C. Het is even waarschijnlijker dat kop of munt boven komt.
D. Er is meer informatie nodig om deze vraag te beantwoorden.

F09 (identical in WT-1995 and WT-1999; released after 1999).
Welk van de volgende getallen ligt tussen 0,07 en 0,08?

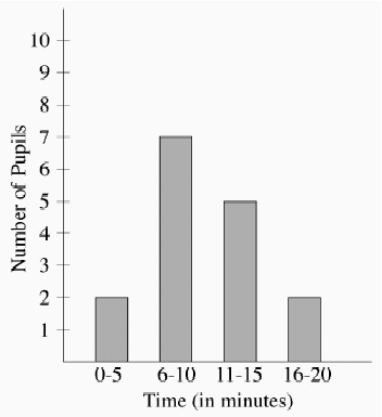
A. 0,00075 B. 0,0075 C. 0,075 D. 0,75

F12* (identical in WT-1995 and WT-1999; released after 1999). Welk deel van de cirkel is grijs gemaakt?



A. Tussen 0 en ¼ B. Tussen ¼ en ½
C. Tussen ½ en ¾ D. Tussen ¾ en 1

H07* (identical in WT-1995 and WT-1999; released after 1999).
Aan een aantal leerlingen is gevraagd hoe lang ze fietsen van school naar huis. De antwoorden staan in de grafiek hiernaast. Hoeveel leerlingen zijn **langer** dan 10 minuten onderweg?



A. 2
B. 5
C. 7
D. 8
E. 15

H10 (identical in WT-1995 and WT-1999; released after 1999).
De tabel laat het verband zien tussen *x* en *y*.

x	2	3	4	5
y	7	10	13	16

Welke formule hoort hierbij?

A. $y = x + 5$ B. $y = x - 5$ C. $y = \frac{1}{3}(x - 1)$ D. $y = 3x + 1$

With permission from ISC, in 1999 the alternatives of item H10 were altered into:

A. $y = x + 5$ B. $y = x - 5$ C. $y = \frac{1}{3} \bullet (x - 1)$ D. $y = 3 \bullet x + 1$

J11 (used in WT-1995; released after 1995).
Een vierhoek **MOET** een parallellogram zijn als aanwezig zijn:

A. een paar aangrenzende gelijke zijdes
B. een paar parallelle zijdes
C. een diagonaal als as van symmetrie
D. twee aangrenzende gelijke hoeken
E. twee paar parallelle zijdes

J11 (used in WT-1999; released after 1999).
Welk antwoord is **NIET** voor alle rechthoeken waar?

A. De tegenoverliggende zijdes zijn evenwijdig.
B. De tegenoverliggende zijdes zijn gelijk.
C. Alle hoeken zijn rechte hoeken.
D. De diagonalen zijn gelijk.
E. De diagonalen staan loodrecht op elkaar.

J12 (used in WT-1995; released after 1995).

Reken uit: $\frac{8}{45} : \frac{4}{15} =$

Antwoord: _____

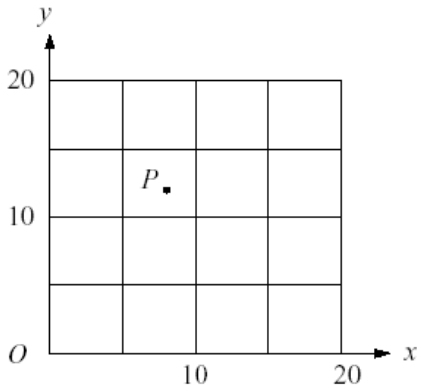
J12 (used in WT-1995; released after 1995).

Reken uit: $\frac{6}{55} : \frac{3}{25} =$

Antwoord: _____

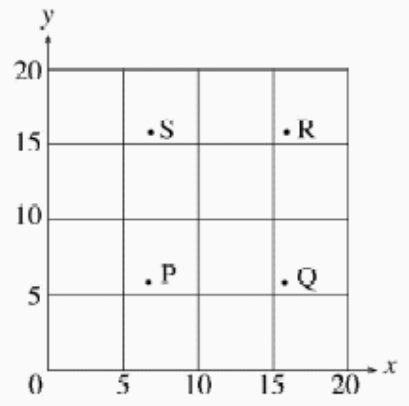
J14 (used in WT-1995; released after 1995).
Welke van de volgende coördinaten zouden van punt P kunnen zijn?

A. (8,12)
B. (8,8)
C. (12,8)
D. (12,12)



J14 (used in WT-1999; released after 1999).
Welk punt in de grafiek zou de coördinaten (7,16) kunnen hebben?

A. Punt P
B. Punt Q
C. Punt R
D. Punt S



K04 (used in WT-1995; released after 1995).

$\frac{x}{2} < 7$ betekent hetzelfde als:

A. $x < \frac{7}{2}$ B. x<5 C. x<14 D. x>5 E. x>14

K04 (used in WT-1999, not released).
Inequality equivalent to x/3>8. Format: multiple choice. Topic: algebra.

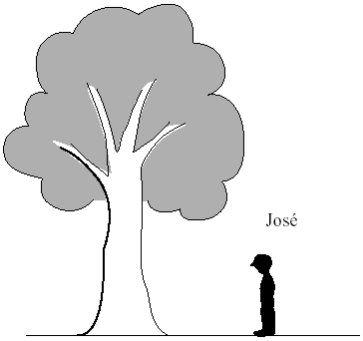
K06* (used in WT-1995; released after 1995).
Vorig jaar bedroeg het aantal leerlingen op het Thorbecke College 1172.
Dit jaar is het aantal leerlingen ongeveer 15 procent hoger dan het vorig jaar.
Hoeveel leerlingen heeft het Thorbecke College dit jaar ongeveer?

A. 1800 B. 1600 C. 1500 D. 1400 E. 1200

K06 (used in WT-1999, not released).
Approximate tons of fertilizer sold. Format: multiple choice. Topic: numbers.

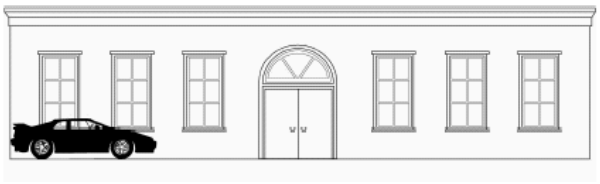
L08* (used in WT-1995, released).
José is 1,5 meter lang. Hoe hoog is de boom ongeveer?

A. 4 m B. 6 m C. 8 m D. 10 m

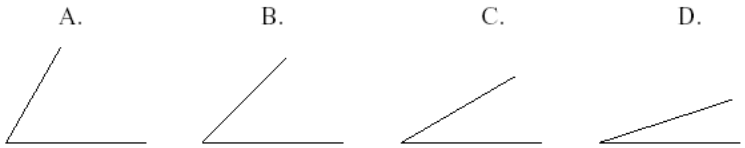


L09 (used in WT-1999, released).
De auto is 3,5 meter lang. Hoe lang is het gebouw ongeveer?

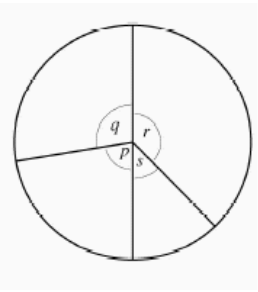
A. 4 m B. 6 m
C. 8 m D. 10 m



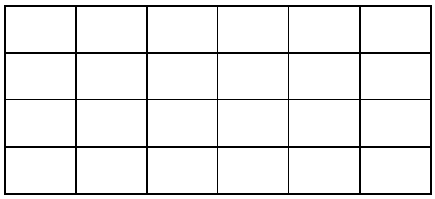
N15* (used in WT-1995, released after 1995).
In welke figuur is een hoek van ongeveer 30° getekend?



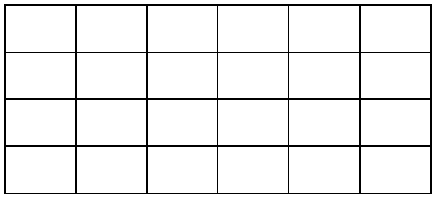
N15 (used in WT-1999, released after 1999).
Welke hoek in de figuur is ongeveer 45°?



N19* (used in WT-1995, released after 1995).
Kleur ⅝ deel van de hele figuur.



N19 (used in WT-1995, released after 1999).
Kleur ⅜ deel van de hele figuur.



O09* (used in WT-1995, released after 1995).
Peter traint elke dag door 5 km hard te lopen. De baan waarop hij loopt is ¼ km lang. Hoeveel rondjes loopt hij elke dag op de baan?

Antwoord:_____

O09 (used in WT-1999, not released).
Scoops of flour needed to fill bag. Format: short answer question. Topic: numbers.

Q02 (used in WT-1995; released after 1995).
Reken uit $\frac{2x}{9} - \frac{x}{9} =$

A. $\frac{1}{9}$ B. 2 C. x D. $\frac{x}{9}$ E. $\frac{x}{81}$

Q02 (used in WT-1999, not released).
Subtract fractons involving x. Format: multiple-choice. Topic: algebra.

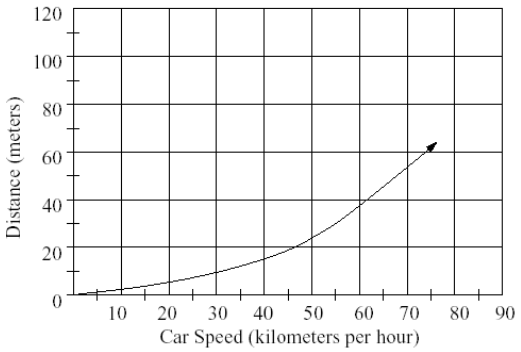
R07* (used in WT-1995, released after 1995).
Een stapel van 200 gelijke vellen papier is 2,5 cm dik. Hoe dik is één vel papier?

A. 0,008 cm B. 0,0125 cm C. 0,05 cm D. 0,08 cm

R08 (used in WT-1999, released after 1999).
500 zoutkristallen wegen samen 6,5 g. Wat is het gemiddelde gewicht van één zoutkristal?

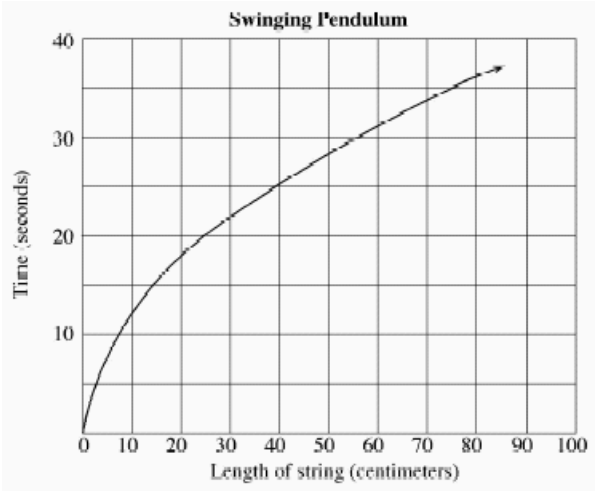
A. 0,0078 g B. 0,013 g C. 0,0325 g D. 0,078 g

R08 (used in WT-1995, released after 1995).
De grafiek laat de afstand zien die een gemiddelde auto nog aflegt als op de remmen wordt getrapt.
Een auto rijdt 80 km/uur. Hoever rijdt deze auto nog door als er op de remmen is getrapt?



A. 60 m B. 70 m C. 85 m D. 100 m

R09 (used in WT-1999, released after 1999).
De grafiek laat de tijd zien die een slinger nodig heeft om 20 keer heen en weer te gaan bij verschillende lengtes van het touwtje dat aan de slinger vast zit.
De lengte van één van de touwtjes is 90 cm. Hoeveel tijd kost het de slinger dan ongeveer om 20 keer heen en weer te gaan?



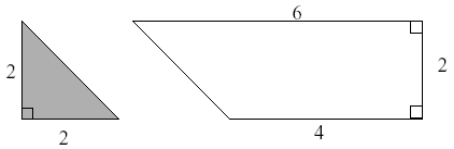
A. 35 seconden B. 38 seconden C. 42 seconden D. 45 seconden

R09 (used in WT-1995, released after 1995).
Gegeven is dat *a*, *b*, en *c* verschillende reële getallen zijn. Welke bewering is FOUT?

A. $(a + b) + c = a + (b + c)$
B. $ab = ba$
C. $a + b = b + a$
D. $(ab)c = a(bc)$
E. $a - b = b - a$

R10 (used in WT-1999, released after 1999).
Gegeven is dat *a*, *b*, en *c* verschillende reële getallen zijn. Welke bewering is waar?

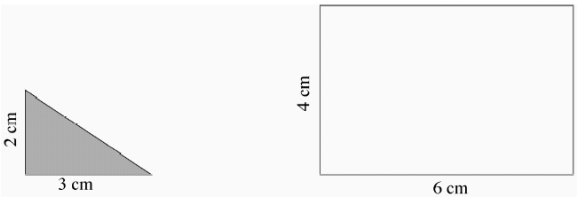
A. $a - b = a - b$
B. $a \bullet (b + c) = b \bullet (c - a)$
C. $b - c = c - b$
D. $a \bullet b = b \bullet a$
E. $a \bullet b - c = a \bullet c - b$



R10* (used in WT-1995, released after 1995).
Hiernaast is een grijsgemaakte driehoek getekend en daarnaast een trapezium. Hoeveel van deze driehoeken passen precies in het trapezium?

A. Drie B. Vier C. Vijf D. Zes

R11 (used in WT-1999, released after 1999).
Hieronder is een grijsgemaakte rechthoekige driehoek getekend en daarnaast een rechthoek. Hoeveel van deze grijsgemaakte rechthoekige driehoeken zijn er nodig om de oppervlakte van de rechthoek precies te bedekken?



A. Vier B. Zes C. Acht D. Tien

R12 (used in WT-1995, released after 1995).

Reken uit: 6000
<u>2369 -</u>

A. 4369 B. 3742 C. 3631 D. 3531

R13 (used in WT-1999, released after 1999).

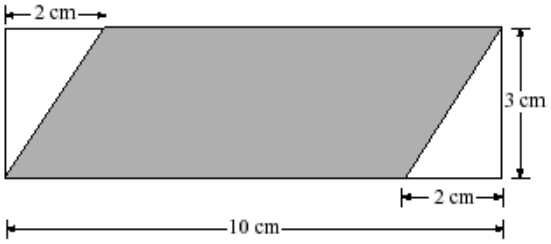
Reken uit: 7003
<u>4078 -</u>

A. 2035 B. 2925 C. 3005 D. 3925

V04 (used in WT-1995, released after 1995).

In de figuur zie je een grijsgemaakte parallellogram binnen een rechthoek.

Hoe groot is de oppervlakte van het grijze parallellogram? Antwoord: _____



U03 (used in WT-1999, not released).

Area of triangle inside a square. Format: short answer question. Topic: measurement.

V02* (used in WT-1995, released after 1995).
Deze twee advertenties stonden in de krant van een land waar de *zed* de munteenheid is.

GEBOUW A
Kantoorruimte te huur
Als u kantoorruimte tussen 85 - 95 vierkante meter huurt, dan kost u dat 475 *zeds* per maand.
Als u tussen 100 – 120 vierkante meter huurt, dan kost u dat 800 *zeds* per maand.

GEBOUW B
Kantoorruimte te huur
Als u kantoorruimte tussen 35 - 260 vierkante meter huurt, dan kost u dat 90 *zeds* per vierkante meter per jaar.

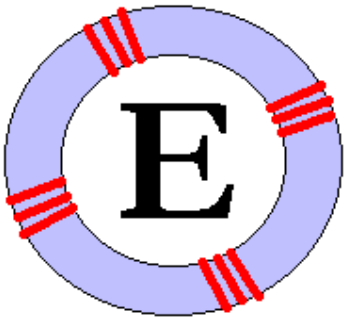
Een bedrijf in dat land wil een kantoor van 110 vierkante meter voor een jaar huren. In welk gebouw, A of B, zouden ze kantoorruimte moeten huren om het goedkoopst uit te zijn? Laat zien hoe je aan je antwoord komt.

V02 (used in WT-1999, released after 1999).
Chris wil zich voor 24 nummers op een tijdschrift abonneren. Hij leest de volgende advertenties voor twee tijdschriften. *Ceds* is de munteenheid in het land waar Chris woont.

Computer Tijdschrift
24 nummers
Eerste vier nummers
GRATIS
De rest
voor 3 *ceds* per nummer.

Basketball Tijdschrift
24 nummers
Eerste zes nummers
GRATIS
De rest
voor 3.5 *ceds* per nummer.

Voor welk tijdschrift zijn 24 nummers het goedkoopst? Hoeveel goedkoper? Laat zien hoe je aan je antwoord komt.

Appendix 

Tables for comparison of p-values of paired samples

Table for the significant difference for two p-values p_1 from PA-1995 and p_2 from PA-2000

The numbers indicate $P(p_1 = p_2)$

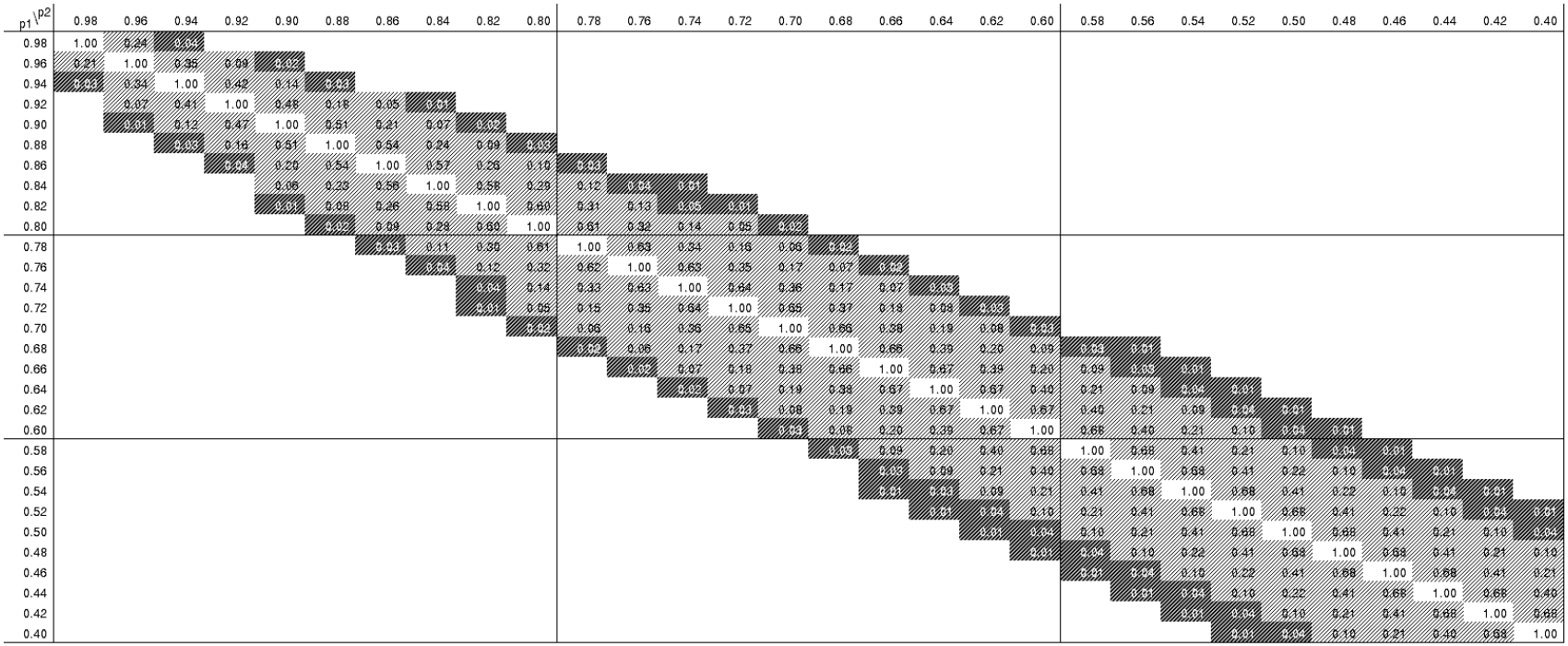
$p_1 \backslash p_2$	0.98	0.96	0.94	0.92	0.90	0.88	0.86	0.84	0.82	0.80	0.78	0.76	0.74	0.72	0.70	0.68	0.66	0.64	0.62	0.60	0.58	0.56	0.54	0.52	0.50	0.48	0.46	0.44	0.42	0.40
0.98	1.00	0.42	0.17	0.07	0.03																									
0.96	0.38	1.00	0.52	0.25	0.11	0.05	0.02																							
0.94	0.11	0.50	1.00	0.58	0.31	0.15	0.07	0.03	0.01																					
0.92	0.03	0.21	0.57	1.00	0.62	0.35	0.19	0.09	0.04	0.02																				
0.90		0.07	0.28	0.61	1.00	0.65	0.39	0.22	0.11	0.05	0.02	0.01																		
0.88		0.02	0.12	0.33	0.65	1.00	0.68	0.42	0.24	0.13	0.06	0.03	0.01																	
0.86			0.04	0.15	0.37	0.67	1.00	0.69	0.44	0.26	0.15	0.08	0.04	0.02																
0.84			0.01	0.06	0.19	0.40	0.69	1.00	0.71	0.46	0.28	0.16	0.09	0.04	0.02															
0.82				0.02	0.09	0.22	0.43	0.70	1.00	0.72	0.48	0.30	0.18	0.10	0.05	0.02	0.01													
0.80					0.04	0.11	0.24	0.45	0.72	1.00	0.73	0.50	0.32	0.19	0.10	0.05	0.03	0.01												
0.78					0.01	0.05	0.13	0.27	0.47	0.73	1.00	0.74	0.51	0.33	0.20	0.11	0.06	0.03	0.01											
0.76						0.02	0.06	0.14	0.29	0.49	0.73	1.00	0.74	0.52	0.34	0.21	0.12	0.06	0.03	0.02										
0.74							0.03	0.07	0.16	0.30	0.50	0.74	1.00	0.75	0.53	0.35	0.22	0.13	0.07	0.04	0.02									
0.72								0.03	0.08	0.17	0.32	0.51	0.75	1.00	0.75	0.54	0.36	0.23	0.13	0.07	0.04	0.02								
0.70								0.01	0.04	0.09	0.18	0.33	0.52	0.75	1.00	0.76	0.54	0.37	0.23	0.14	0.08	0.04	0.02							
0.68									0.02	0.04	0.10	0.20	0.34	0.53	0.76	1.00	0.76	0.55	0.37	0.24	0.14	0.08	0.04	0.02						
0.66										0.02	0.05	0.11	0.21	0.35	0.54	0.76	1.00	0.77	0.55	0.38	0.24	0.15	0.08	0.04	0.02					
0.64											0.02	0.06	0.12	0.22	0.36	0.55	0.77	1.00	0.77	0.56	0.38	0.25	0.15	0.08	0.04	0.02				
0.62												0.03	0.06	0.12	0.22	0.37	0.55	0.77	1.00	0.77	0.56	0.39	0.25	0.15	0.08	0.04	0.02			
0.60												0.01	0.03	0.07	0.13	0.23	0.37	0.56	0.77	1.00	0.77	0.56	0.39	0.25	0.15	0.09	0.04	0.02		
0.58													0.01	0.03	0.07	0.14	0.24	0.38	0.56	0.77	1.00	0.77	0.57	0.39	0.25	0.15	0.09	0.04	0.02	
0.56														0.01	0.03	0.07	0.14	0.24	0.38	0.56	0.77	1.00	0.78	0.57	0.39	0.25	0.15	0.09	0.04	0.02
0.54															0.02	0.04	0.08	0.14	0.25	0.39	0.57	0.77	1.00	0.78	0.57	0.39	0.25	0.15	0.08	0.04
0.52																0.02	0.04	0.08	0.15	0.25	0.39	0.57	0.78	1.00	0.78	0.57	0.39	0.25	0.15	0.08
0.50																	0.02	0.04	0.08	0.15	0.25	0.39	0.57	0.78	1.00	0.78	0.57	0.39	0.25	0.15
0.48																		0.02	0.04	0.08	0.15	0.25	0.39	0.57	0.78	1.00	0.78	0.57	0.39	0.25
0.46																			0.02	0.04	0.08	0.15	0.25	0.39	0.57	0.78	1.00	0.77	0.57	0.39
0.44																				0.02	0.04	0.09	0.15	0.25	0.39	0.57	0.78	1.00	0.77	0.56
0.42																					0.02	0.04	0.09	0.15	0.25	0.39	0.57	0.77	1.00	0.77
0.40																						0.02	0.04	0.09	0.15	0.25	0.39	0.56	0.77	1.00

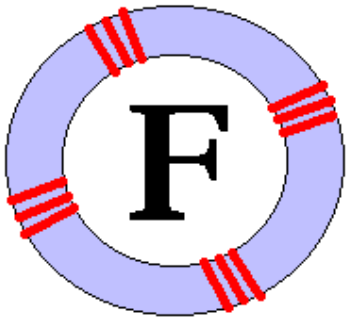
95% probability interval

99% probability interval

Table for the significant difference for two p-values p_1 from WT-1995 and p_2 from WT-1995

The numbers indicate $P(p_1 = p_2)$



Appendix 

Research results of the TIMSS Written Test, 1995 - 1999

This appendix presents research results on comparable items from the TIMSS-95 and TIMSS-99 Mathematics Written Test (WT-1995 and WT-1999). The items are paired, being considered comparable (identical or cloned). Cloned items from WT-1999 can have a different name than their counterpart from WT-1995.
Data at the level of the intended curriculum: 0 = not matching, 1 = matching.
Data at the level of the implemented curriculum: OTL rates (percentage of mathematics teachers indicating that the item was suitable for an imaginary test covering all content taught). In 1995, only 16 items were submitted. Standard errors are given in brackets.
Data at the level of the attained curriculum: p-values (percentage of students scoring correctly). Standard errors are given in brackets.

Item name		Intended curr. "1"=*matching* "0"=*not matching*		Implemented curr. OTL rate (SE)		Attained curr. p-value (SE)	
1995	*1999*	*1995*	*1999*	*1995*	*1999*	*1995*	*1999*
A01	A01	1	1		94.3 (3.8)	69.8 (3.4)	75.1 (2.6)
A02	A02	1	1	89.0 (3.2)	84.6 (5.9)	82.4 (2.8)	85.9 (2.1)
A03	A03	1	1		85.7 (5.7)	84.6 (2.6)	89.0 (1.9)
A04	A04	1	1		94.9 (3.6)	65.9 (3.5)	70.7 (2.7)
A05	A05	1	0		76.3 (6.9)	70.6 (3.3)	74.2 (2.6)
A06	A06	0	1		83.8 (6.0)	74.5 (3.2)	74.5 (2.6)
B07	B07	1	1		94.3 (3.8)	76.9 (3.1)	82.3 (2.3)
B08	B08	1	1	90.2 (3.1)	94.3 (3.8)	67.2 (3.5)	70.6 (2.7)
B09	B09	1	0		97.4 (2.6)	65.2 (3.5)	68.8 (2.8)
B10	B10	1	1	95.7 (2.1)	100 (0.0)	62.3 (3.6)	73.9 (2.6)
B11	B11	1	1	97.7 (1.5)	94.9 (3.6)	80.3 (2.9)	82.3 (2.3)
B12	B12	0	1		91.4 (4.6)	67.1 (3.5)	80.1 (2.4)
C01	C01	1	1		97.4 (2.6)	77.9 (3.1)	78.4 (2.5)
C02	C02	1	1		97.4 (2.6)	84.0 (2.7)	86.8 (2.0)
C03	C03	0	0		62.9 (7.9)	64.5 (3.5)	72.2 (2.7)
C04	C04	1	1		88.6 (5.2)	53.0 (3.7)	59.3 (2.9)
C05	C05	1	1	89.0 (3.2)	92.1 (4.4)	57.1 (3.7)	59.7 (2.9)
C06	C06	1	0		73.7 (7.2)	72.5 (3.3)	75.7 (2.6)
D07	D07	1	1		86.8 (5.5)	59.4 (3.7)	64.3 (2.9)
D08	D08	0	1		78.9 (6.7)	66.1 (3.5)	70.3 (2.7)
D09	D09	1	1		92.3 (4.3)	84.4 (2.7)	86.5 (2.0)
D10	D10	0	1		87.2 (5.4)	67.7 (3.5)	73.5 (2.6)
D11	D11	1	1		82.1 (6.3)	89.0 (2.3)	89.4 (1.8)
D12	D12	0	1		91.4 (4.6)	92.4 (1.9)	90.0 (1.8)
E01	E01	1	1		87.2 (5.4)	81.7 (2.8)	79.5 (2.4)
E02	E02	1	1		51.3 (8.2)	38.9 (3.6)	42.1 (2.9)
E03	E03	1	0		92.1 (4.4)	69.3 (3.4)	69.5 (2.7)
E04	E04	1	1		84.6 (5.9)	54.0 (3.7)	57.3 (2.9)
E05	E05	1	1	89.0 (3.2)	74.3 (7.1)	66.0 (3.5)	73.8 (2.6)
E06	E06	1	1		97.1 (2.7)	48.4 (3.7)	48.5 (3.0)

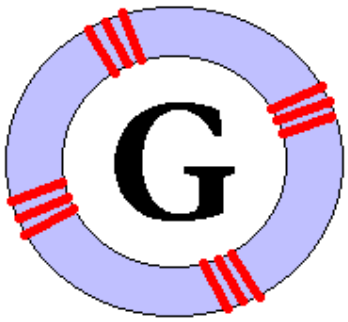
Item name		Intended curr. "1"=*matching* "0"=*not matching*		Implemented curr. OTL rate (SE)		Attained curr. p-value (SE)	
1995	*1999*	*1995*	*1999*	*1995*	*1999*	*1995*	*1999*
F07	F07	1	1		82.1 (6.3)	37.2 (3.6)	40.4 (2.9)
F08	F08	0	1		50.0 (8.2)	72.0 (3.3)	78.8 (2.4)
F09	F09	0	1		94.3 (3.8)	79.2 (3.0)	85.5 (2.1)
F10	F10	0	0		81.6 (6.3)	68.8 (3.4)	71.5 (2.7)
F11	F11	1	0		78.9 (6.7)	38.8 (3.6)	41.0 (2.9)
F12	F12	1	1	95.7 (2.1)	97.4 (2.6)	68.5 (3.4)	75.0 (2.6)
G01	G01	0	1		97.4 (2.6)	70.2 (3.3)	72.6 (2.7)
G02	G02	1	1		89.7 (4.9)	86.4 (2.5)	85.3 (2.1)
G03	G03	1	1		88.6 (5.2)	59.5 (3.6)	59.0 (3.0)
G04	G04	0	0		77.1 (6.9)	75.7 (3.1)	80.5 (2.4)
G05	G05	1	1		94.9 (3.6)	70.9 (3.3)	85.4 (2.1)
G06	G06	0	1		82.1 (6.3)	40.3 (3.6)	44.3 (3.0)
H07	H07	1	1	86.1 (3.6)	84.6 (5.9)	81.1 (2.8)	82.0 (2.3)
H08	H08	1	1		94.3 (3.8)	87.5 (2.4)	91.6 (1.6)
H09	H09	1	1		92.1 (4.4)	91.8 (2.0)	94.8 (1.3)
H10	H10	1	1		82.9 (6.1)	67.2 (3.4)	70.6 (2.7)
H11	H11	0	1		77.1 (6.9)	79.3 (2.9)	85.5 (2.1)
H12	H12	1	1		94.7 (3.6)	86.1 (2.5)	86.7 (2.0)
I01	I01	1	1		35.9 (7.8)	36.6 (3.6)	23.3 (2.5)
I02	I02	1	0		76.9 (6.9)	71.6 (3.4)	62.1 (2.9)
I03	I03	1	1		94.1 (3.8)	50.8 (3.7)	49.2 (3.0)
I04	I04	0	1		56.4 (8.1)	38.1 (3.6)	33.3 (2.8)
I05	I05	1	1		97.1 (2.7)	68.1 (3.5)	67.6 (2.8)
I06	I06	1	1		88.6 (5.2)	76.2 (3.2)	70.0 (2.7)
I08	I08	1	1		73.7 (7.2)	66.0 (3.5)	65.0 (2.8)
I09	I09	0	1		73 (7.2)	81.2 (2.9)	79.0 (2.4)

Item name		Intended curr. "1"=*matching* "0"=*not matching*		Implemented curr. OTL rate (SE)		Attained curr. p-value (SE)	
1995	*1999*	*1995*	*1999*	*1995*	*1999*	*1995*	*1999*
J10	J10	1	1		92.3 (4.3)	54.8 (3.8)	56.7 (3.0)
J11	J11	0	1		79.5 (6.6)	36.9 (3.7)	42.2 (3.0)
J12	J12	1	0		60.5 (8.0)	15.7 (2.8)	11.6 (1.9)
J13	J13	0	1		91.4 (4.6)	87.1 (2.5)	88.5 (1.9)
J14	J14	0	0		57.9 (8.1)	34.4 (3.6)	44.2 (3.0)
J15	J15	1	0		80.0 (6.5)	77.9 (3.2)	62.4 (2.9)
J16	J16	1	1		97.4 (2.6)	72.9 (3.4)	78.5 (2.5)
J17	J18	1	1		89.7 (4.9)	74.3 (3.3)	70.0 (2.7)
J18	J17	0	1		80.0 (6.5)	54.9 (3.8)	80.6 (2.4)
K01	K01	1	1		100 (0.0)	78.4 (3.0)	87.0 (2.0)
K02	K02	1	1		87.2 (5.4)	55.4 (3.6)	56.7 (3.0)
K03	K03	0	1		82.1 (6.3)	76.6 (3.1)	86.5 (2.0)
K04	K04	0	0		34.2 (7.7)	48.9 (3.7)	32.6 (2.8)
K05	K05	1	1		91.4 (4.6)	47.6 (3.7)	39.4 (2.9)
K06	K06	1	0	95.7 (2.1)	94.7 (3.6)	51.3 (3.7)	75.2 (2.6)
K07	K07	0	0		56.4 (8.1)	59.0 (3.6)	61.4 (2.9)
K08	K08	0	0		62.9 (7.9)	21.2 (3.0)	31.7 (2.8)
K09	K09	1	0		76.9 (6.9)	29.1 (3.3)	48.1 (3.0)
L08	L09	1	1	95.7 (2.1)	100 (0.0)	69.1 (3.4)	81.8 (2.3)
L09	L10	1	0		89.5 (5.0)	93.1 (1.8)	78.6 (2.4)
L10	L11	1	1		91.4 (4.6)	89.2 (2.2)	75.9 (2.5)
L11	L12	1	1		87.2 (5.4)	32.4 (3.4)	57.8 (2.9)
L12	L13	0	1		94.3 (3.8)	81.6 (2.8)	89.8 (1.8)
L14	L15	1	1		89.5 (5.0)	29.1 (3.3)	49.1 (3.0)
L15	L16	0	0		89.5 (5.0)	38.2 (3.5)	38.6 (2.9)
L16	L17	0	1		51.4 (8.2)	21.0 (3.0)	19.2 (2.3)
L17	L18	1	0		74.4 (7.1)	51.3 (3.6)	55.0 (3.0)
M01	M01	1	1		94.7 (3.6)	97.1 (1.2)	86.0 (2.1)
M03	M03	0	1		65.8 (7.7)	91.3 (2.0)	87.0 (2.0)
M04	M04	1	1		97.1 (2.7)	41.2 (3.6)	91.0 (1.7)
M05	M05	1	0		74.3 (7.1)	69.7 (3.4)	63.7 (2.9)
M06	M06	0	1		88.6 (5.2)	43.0 (3.6)	37.1 (2.9)
M07	M07	0	1		68.4 (7.6)	73.8 (3.2)	70.5 (2.7)
M08	M08	1	1		85.7 (5.7)	27.0 (3.2)	19.8 (2.4)

Item name		Intended curr. "1"=*matching* "0"=*not matching*		Implemented curr. OTL rate (SE)		Attained curr. p-value (SE)	
1995	*1999*	*1995*	*1999*	*1995*	*1999*	*1995*	*1999*
N11	N11	1	0		97.4 (2.6)	94.4 (1.6)	90.2 (1.8)
N12	N12	0	0		64.1 (7.8)	77.5 (3.0)	66.7 (2.9)
N13	N13	0	0		60.5 (8.0)	37.4 (3.5)	50.3 (3.0)
N14	N14	1	1		92.3 (4.3)	73.5 (3.2)	74.0 (2.7)
N15	N15	1	1	99.0 (1.0)	94.9 (3.6)	64.2 (3.4)	52.3 (3.0)
N16	N16	1	0		77.1 (6.9)	54.2 (3.6)	53.2 (3.0)
N17	N17	1	1		94.7 (3.6)	49.8 (3.6)	48.7 (3.0)
N18	N18	0	0		38.2 (7.9)	63.4 (3.5)	63.9 (2.9)
N19	N19	1	1	95.7 (2.1)	97.4 (2.6)	63.7 (3.5)	61.2 (3.0)
O01	O01	1	1		94.7 (3.6)	75.7 (3.2)	46.5 (3.0)
O02	O02	1	0		82.1 (6.3)	44.4 (3.7)	61.1 (2.9)
O03	O03	1	1		91.4 (4.6)	42.3 (3.7)	47.6 (3.0)
O04	O04	0	0		100 (0.0)	53.4 (3.7)	59.9 (2.9)
O05	O05	0	0		64.1 (7.8)	61.7 (3.6)	72.5 (2.7)
O06	O06	0	1		94.7 (3.6)	91.1 (2.1)	90.2 (1.8)
O07	O07	0	1		64.1 (7.8)	65.0 (3.5)	51.5 (3.0)
O08	O08	1	0		63.2 (7.9)	66.7 (3.5)	63.3 (2.9)
O09	O09	1	1	95.7 (2.1)	92.3 (4.3)	75.2 (3.2)	73.2 (2.7)
P08	P08	0	0		61.5 (7.9)	59.9 (3.5)	15.0 (2.1)
P09	P10	1	0		57.1 (8.1)	34.9 (3.4)	43.8 (3.0)
P10	P11	0	1		84.2 (5.9)	51.3 (3.6)	57.2 (3.0)
P11	P12	1	1		94.3 (3.8)	61.9 (3.4)	56.5 (3.0)
P12	P13	1	1		92.1 (4.4)	81.0 (2.8)	80.7 (2.4)
P13	P14	1	1		94.9 (3.6)	62.0 (3.4)	63.2 (2.9)
P14	P15	1	0		85.7 (5.7)	91.2 (2.0)	60.8 (2.9)
P15	P09	1	1		64.1 (7.8)	69.0 (3.3)	55.3 (3.0)
P16	P17	1	1		88.6 (5.2)	28.5 (3.2)	31.0 (2.8)
P17	P16	1	1		85.7 (5.7)	95.8 (1.4)	89.7 (1.8)

Item name		Intended curr. *"1"=matching "0"=not matching*		Implemented curr. OTL rate (SE)		Attained curr. p-value (SE)	
1995	*1999*	*1995*	*1999*	*1995*	*1999*	*1995*	*1999*
Q01	Q01	0	0		54.5 (8.1)	44.8 (3.6)	30.6 (2.8)
Q02	Q02	0	0		36.8 (7.9)	35.2 (3.5)	41.8 (3.0)
Q03	Q03	1	1		78.9 (6.7)	27.1 (3.2)	55.8 (3.0)
Q04	Q04	1	1		92.1 (4.4)	86.7 (2.5)	79.7 (2.4)
Q05	Q05	0	1		59.0 (8.0)	77.5 (3.0)	82.2 (2.3)
Q06	Q06	1	1		94.7 (3.6)	53.5 (3.6)	67.4 (2.8)
Q07	Q07	0	0		71.8 (7.3)	74.5 (3.2)	56.5 (3.0)
Q08	Q08	1	1		100 (0.0)	50.0 (3.7)	56.7 (3.0)
Q09	Q09	0	0		73.5 (7.2)	36.8 (3.5)	32.3 (2.8)
Q10	Q10	0	1		89.7 (4.9)	46.2 (3.6)	22.6 (2.5)
R06	R07	1	1		89.5 (5.0)	59.3 (3.6)	69.5 (2.8)
R07	R08	1	0	95.7 (2.1)	94.9 (3.6)	54.0 (3.6)	56.6 (3.0)
R08	R09	1	1		82.1 (6.3)	63.6 (3.5)	69.9 (2.8)
R09	R10	0	0		11.8 (5.3)	26.6 (3.2)	39.7 (3.0)
R10	R11	1	1	97.7 (1.5)	97.4 (2.6)	61.9 (3.6)	65.7 (2.9)
R12	R13	1	0		82.9 (6.1)	81.8 (2.8)	79.0 (2.4)
R13	R14	1	1		97.4 (2.6)	42.4 (3.6)	46.8 (3.0)
R14	R15	1	1		89.5 (5.0)	40.9 (3.6)	53.1 (3.0)
S02	S02	1	1		87.2 (5.4)	37.9 (3.6)	49.1 (3.0)
T01	T01	1	1		76.3 (6.9)	36.0 (3.5)	41.6 (2.9)
U01	U01	1	0		84.6 (5.9)	45.0 (3.6)	23.5 (2.5)
U02	U02	1	0		75.7 (7.0)	32.2 (3.4)	26.9 (2.7)
V01	V01	1	1		85.7 (5.7)	61.0 (3.6)	64.0 (2.9)
V02	V02	1	1	86.1 (3.6)	88.9 (5.1)	24.0 (3.1)	45.1 (3.0)
V03	V03	1	1		88.6 (5.2)	65.3 (3.5)	74.3 (2.6)
V04	U03	1	1		100 (0.0)	38.8 (3.6)	47.3 (3.0)

Appendix



Research results of the TIMSS Performance Assessment, 1995 - 2000

This appendix presents research results on the TIMSS Performance Assessment of 1995 and 2000 (PA-1995 and PA-2000). Items in PA-1995 and PA-2000 were identical.

Data at the level of the intended curriculum: item-curriculum matching indices (percentage of experts judging positively on the match with the intended curriculum).

Data at the level of the implemented curriculum are two-fold: (a) OTL-covered rates (percentage of mathematics teachers that indicated item as 'covered'), and (b) OTL-testing rates (percentage of mathematics teachers that indicated item could be included in a future test). Standard errors are given in brackets.

Data at the level of the attained curriculum: p-values (percentage of students scoring correclty). Standard errors are given in brackets.

Legend:

- ↓ item was judged together with previous item.
- \- judgement was not asked.
- x data are considered unreliable or uncomparable.

M1	Dice	G1	Shadows
M2	Calculator	G2	Plasticine
M3	Folding	S3	Batteries
M4	Around the Bend	S4	Rubber Band
M5	Packaging		

Item name	Intended curr. *Item-curr. matching index*		Implemented curr. *OTL rate (SE)*				Attained curr. *p-value (SE)*	
			OTL-covered		*OTL-testing*			
	1995	*2000*	*1995*	*2000*	*1995*	*2000*	*1995*	*2000*
M1-1	100	100	37 (11)	85 (8)	56 (11)	70 (10)	98 (1)	98 (2)
M1-2	100	100	56 (11)	80 (9)	65 (11)	80 (9)	87 (3)	69 (5)
M1-3	100	100	56 (11)	75 (10)	53 (11)	65 (11)	96 (2)	93 (3)
M1-4	↓	↓	↓	↓	↓	↓	73 (4)	70 (5)
M1-5a	100	100	59 (11)	80 (9)	56 (11)	80 (9)	88 (3)	83 (4)
M1-5b	33	60	29 (10)	47 (11)	25 (10)	55 (11)	23 (3)	32 (5)
M2-1	100	100	67 (11)	95 (5)	72 (10)	85 (8)	97 (1)	98 (2)
M2-2	67	80	56 (11)	80 (9)	56 (11)	85 (8)	42 (4)	46 (6)
M2-3	67	80	28 (10)	70 (10)	47 (11)	80 (9)	79 (3)	79 (5)
M2-4	↓	↓	↓	↓	↓	↓	61 (4)	55 (6)
M2-5	67	80	28 (10)	60 (11)	41 (11)	80 (9)	44 (4)	51 (6)
M2-6A	67	60	39 (11)	50 (11)	59 (11)	80 (9)	81 (3)	73 (5)
M2-6B	67	20	39 (11)	55 (11)	59 (11)	80 (9)	27 (4)	20 (5)
M3-1	100	80	17 (8)	32 (10)	28 (10)	75 (10)	71 (4)	80 (5)
M3-2	100	80	↓	↓	↓	↓	81 (3)	89 (4)
M3-3	100	60	↓	↓	↓	↓	76 (4)	80 (5)
M3-4	100	20	17 (8)	30 (10)	28 (10)	75 (10)	62 (4)	60 (6)
M4-1	100	100	65 (11)	90 (7)	76 (10)	90 (7)	94 (2)	96 (2)
M4-2	100	100	44 (11)	95 (5)	72 (10)	95 (5)	91 (2)	96 (2)
M4-3	100	100	39 (11)	80 (9)	6 (11)	80 (9)	88 (3)	90 (3)
M4-4	100	100	22 (9)	55 (11)	61 (11)	85 (8)	79 (3)	83 (4)
M4-5A	100	80	22 (9)	45 (11)	39 (11)	85 (8)	57 (4)	62 (5)
M4-5B	↓	↓	↓	↓	↓	↓	60 (4)	56 (6)
M4-5C	↓	↓	↓	↓	↓	↓	70 (4)	71 (5)
M4-6	100	20	17 (8)	45 (11)	33 (11)	70 (10)	5 (2)	4 (2)
M5-1	100	80	39 (11)	58 (11)	67 (11)	80 (9)	57 (4)	79 (5)
M5-2	100	80	83 (8)	80 (9)	83 (8)	90 (7)	54 (4)	53 (6)
M5-3	67	60	72 (10)	85 (8)	78 (9)	95 (5)	44 (4)	41 (6)

Item name	Intended curr. *Item-curr. matching index*		Implemented curr. *OTL-rate (SE)*				Attained curr. *p-value (SE)*	
			OTL-covered		*OTL-testing*			
	1995	*2000*	*1995*	*2000*	*1995*	*2000*	*1995*	*2000*
S3-1	–	20	–	–	–	–	x	x
S3-2	–	40	–	–	–	–	x	x
S3-3	–	20	–	–	–	–	x	x
S3-4	–	20	–	–	–	–	x	x
S4-3	–	80	–	–	–	–	x	x
S4-4	–	80	–	–	–	–	x	x
S4-5	–	80	–	–	–	–	x	x
S4-6	–	80	–	–	–	–	x	x
G1-1	100	100	39 (11)	40 (11)	56 (11)	70 (10)	x	x
G1-2	100	60	35 (11)	15 (8)	61 (11)	60 (11)	x	x
G1-3	67	60	33 (11)	25 (10)	44 (11)	60 (11)	x	x
G1-4	33	60	12 (7)	20 (9)	29 (10)	55 (11)	x	x
G1-5	33	60	33 (11)	55 (11)	44 (11)	80 (9)	x	x
G1-6	33	60	28 (10)	40 (11)	44 (11)	75 (10)	x	x
G2-1A	100	80	33 (11)	45 (11)	39 (11)	70 (10)	x	x
G2-1B	67	40	18 (9)	40 (11)	22 (9)	70 (10)	x	x
G2-2B	↓	↓	↓	↓	↓	↓	x	x
G2-3B	↓	↓	↓	↓	↓	↓	x	x
G2-4B	↓	↓	↓	↓	↓	↓	x	x
G2-2A	67	0	17 (8)	35 (11)	17 (8)	70 (10)	x	x
G2-3A	↓	↓	↓	↓	↓	↓	x	x
G2-4A	↓	↓	↓	↓	↓	↓	x	x