BENEFITS AND LIMITATIONS OF LARGE-SCALE INTERNATIONAL COMPARATIVE ACHIEVEMENT STUDIES

THE CASE OF IEA'S TIMSS STUDY

Klaas Tj. Bos

DOCTORAL COMMITTEE

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

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

Members: Prof. dr. Jan van den Akker ▪ University of Twente
Prof. dr. Kerst Boersma ▪ University of Utrecht
Prof. dr. Roel Bosker ▪ University of Twente
Prof. dr. Jan van Damme ▪ University of Leuven
Dr. Hans Wagemaker ▪ Executive director IEA

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Bos, Klaas Tj.

Benefits and limitations of large-scale international comparative achievement studies: The case of IEA's TIMSS study
Thesis University of Twente, Enschede – With refs. – With Dutch summary
ISBN 90 365 17 39 7

Layout: Sandra Schele
Press: PrintPartners Ipskamp - Enschede

BENEFITS AND LIMITATIONS OF LARGE-SCALE INTERNATIONAL COMPARATIVE ACHIEVEMENT STUDIES

THE CASE OF IEA'S TIMSS STUDY

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 13 juni 2002 om 16.45 uur

door

Klaas Tjakob Bos

geboren op 27 februari 1959
te Groningen

Promotoren: Prof. dr. Tj. Plomp
Prof. dr. J. Scheerens

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

NAWOORD

Lieve Henny, weet je nog kerst 2001 op Curaçao. Zee, strand, ruisende palmen en een overdenkstoel. "Laat ik de eindsprint inzetten voor dat proefschrift" heb ik toen gezegd. Het idee voor het onderzoekstraject dat naar een proefschrift moest leiden kwam voort uit een reis naar Sigtuna in Zweden. Daar kwamen de Nordic countries plus 'the Dutch representative' bijeen om over verbeteringen van draft TIMSS vragenlijsten te discussiëren. Er zouden nog vele reizen volgen. Naar Vancouver, Washington DC, Rome, Praag, Parijs, Portoroz en ga zo maar door. Inderdaad TIMSS is een wereldwijd project. De samenwerking met collega National Research Co-ordinators en de internationale projectleiding in Boston waren en zijn erg waardevol. Het werk voor het proefschrift heeft er soms door stil gelegen, maar ten slotte kon ik bij de reflectie op waar het proefschrift nu eigenlijk op neer moest komen, ook profiteren van de dagelijkse TIMSS werkzaamheden.

De planning tussen januari en april 2002 liep gesmeerd. Henny, thuis hielp jij me de planning te halen. Ook jij hebt het tot het eind toe volgehouden. Ik kan je daarvoor op vele manieren bedanken. Wanneer zullen we weer naar een exotisch oord gaan, wie weet wat er dan voor snode plannen te voorschijn komen (als ze maar niets met werk hebben te maken).

Maar weet je, op het werk werd ik ook erg geholpen. Het dagelijkse contractonderzoek met zijn vaste opleverings data mocht niet lijden onder mijn bemoeienis met het proefschrift. Dat is niet gebeurd, omdat ik een unieke werkrelatie heb met Martina Meelissen. Martina, zonder jouw ondersteuning was het manuscript wellicht nooit afgekomen. Ik heb je al vaak gezegd dat ik je kritiek op mijn concept-hoofdstukken niet kon missen. En dat je veel van het werk op de huidige TIMSS-2003 studies uit mijn handen nam, vraagt om een tegenprestatie. Ik hoop jou net zo goed te kunnen helpen met jouw proefschrift als je mij hebt geholpen.

Henny, je kent die andere paranimf toch ook? Dat is Marjolein Drent, zij luncht vaak met mij en dan hebben we het over dat extra werk wat proefschrift heet. Marjolein, bedankt voor je grote betrokkenheid en succes met het schrijven van jouw proeve.
Mijn promotoren heten professor Tjeerd Plomp en professor Jaap Scheerens. Zij hebben mij voorzien van stimulerende adviezen en ideeën op momenten dat ik die nodig had. Bovendien hebben ze concept-versies van hoofdstukken van commentaar voorzien waarmee ik tot definitieve teksten ben gekomen. Heren, hartelijk dank voor de rustige wijze van begeleiden. Wilmad Kuiper die optrad als assistent-promotor, bedank ik voor zijn kritische opmerkingen bij en suggesties over concept-teksten.
Charles Matthijssen heeft de multilevel analyses uitgevoerd en de resultaten met mij besproken. Roel Bosker was hierbij van grote waarde. Charles en Roel, bedankt.
De afdeling Curriculumtechnologie van de faculteit Toegepaste Onderwijskunde van de Universiteit Twente bedank ik voor het beschikbaar stellen van faciliteiten in de laatste maanden van de proefschriftactiviteiten.
De figuren, tabellen en de gehele lay-out van het proefschrift zien er uit zoals ze eruit zien, omdat Sandra Schele zich daarmee op een nauwkeurige manier heeft beziggehouden. Sandra, het maakt voor jou niet uit of ik nu met één figuur of met 60 pagina's tekst aan kom zetten. Je maakt er in korte tijd een geheel van. Erg bedankt en ik hoop nog lang met je aan onderzoeksopdrachten te kunnen werken.
De hoofdtekst is in het Engels geschreven. Karen Bogard Givvin uit Los Angeles heeft van mijn Engels op een voortreffelijke wijze beter Engels gemaakt. Karen thanks. Ik hoop nog langer met je samen te werken in het kader van de TIMSS-R Video Study.
Zou ik niemand vergeten zijn? Vast wel, maar niet mijn naaste familie die mij ook steunde als het moest, al was het alleen maar door er juist niet over te praten (en mij een boekje ter (ont)spanning te geven). En de Echoes van Pink Floyd worden ook bedankt voor de inspiratie.
Maar nu het is mooi geweest. De teksten zijn gereed en ik wens dat op basis van de volgende hoofdstukken in de toekomst nog meer gebruik kan worden gemaakt van resultaten van internationaal vergelijkend onderzoek in belangrijke schoolvakken.

Henny … en laat nu de champagne maar vloeien.

Borne, 30 mei 2002
Klaas Bos

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

FIMS	First International Mathematics Study
FISS	First International Science Study
IEA	International association for the Evaluation of educational Achievement
INES	Indicator in Education Study
ISC TIMSS	International Study Center TIMSS
LINCAS	Large-scale International Comparative Achievement Studies in education
LV	Latent Variable
MV	Manifest Variable
OECD	Organization for Economic Co-operation and Development
SIMS	Second International Mathematics Study
SISS	Second International Science Study
TIMSS	Third International Mathematics and Science Study
UNESCO	United Nations Educational Scientific and Cultural Organization

CHAPTER 1

INTRODUCTION

Large-scale International Comparative Achievement Studies in education (LINCAS) have been conducted on a regular basis since the 1960s. The main goal of these studies is to provide policymakers, educators, educational researchers, and other people interested in education with information regarding similarities and differences across systems. Stakeholders of education systems can use this information to gain insight into the strengths and weaknesses of their own education system.
The benefits and limitations of large-scale international comparative achievement studies are the central theme of this thesis. It this chapter, the usefulness of international comparative studies in education is discussed in the light of their main goals and functions (1.2). In the next section the purpose of this thesis is illuminated. The Third International Mathematics and Science Study (TIMSS) is studied in-depth from the perspective of one specified function of LINCAS: understanding cross-national differences in student achievement (1.3). In section 1.4 the problem statement and the research questions of this thesis are formulated. The chapter ends with a description of the structure of the thesis (1.5).

1.1 INTERNATIONAL COMPARISONS IN EDUCATION

Systematic monitoring of school systems and the publication of accountability reports at a national level have existed for decades. Many countries around the world evaluate their educational system on a regular basis to improve the quality of their education. Within countries, student achievement and educational processes are studied in different subjects in both primary and secondary education.

The United States' National Assessment of Educational Progress (NAEP), started in the 1980s, was a pioneer in national periodic educational evaluation studies (Beaton, 1987). From the 1980s on, several national surveys were organized periodically in England to monitor students' achievement in core subjects (see for example, Newton, Adams, et al., 2002). Since 1987, in the Netherlands the National Institute for Educational Measurement (CITO) has been conducted assessment studies in different subjects in primary and lower secondary education every four years (e.g. Bokhove, Van der Schoot & Eggen, 1996).
Besides periodic monitoring of the educational system at the 'national' level, an increasing number of countries have become interested in making comparisons between their own educational system and the systems of other countries. The increasing globalization of the world economies might be one of the reasons for this development (Howson, 1999).

The International Association for the Evaluation of Educational Achievement (IEA), the Organization for Economic Co-operation and Development (OECD), the World Bank, and the United Nations Educational Scientific and Cultural Organization (UNESCO) are four organizations that support and organize international comparative educational research. Comparative educational research has been defined by Postlethwaite (1988, p. xvii): *"Strictly speaking, to 'compare' means to examine two or more entities by putting them side by side and looking for similarities and differences between or among them. In the field of education, this can apply both to comparisons between and comparisons within systems of education"*. This definition applies to many studies supported or conducted by UNESCO, the World Bank, OECD, and IEA.
The international comparative studies in education relevant to this thesis concern large-scale achievement studies, including those on all forms of formal education from preschool through secondary education, with an emphasis on lower secondary education.
Large-scale international comparative achievement studies in education (denoted as 'LINCAS' in the remainder of this thesis) are defined as *studies in which both achievement of a certain age/grade group in one or more subjects is compared across education systems and effects of contextual factors at system, school, classroom, and student level on achievement are studied.*
International comparative studies in education that do not focus on student achievement in a school have also been conducted. An example of the latter is IEA's Second Information Technology in Education Study (SITES) which aimed at

the description of the implementation and use of information technology in primary and secondary education (Pelgrum & Anderson, 1999). Studies categorized as international studies (and not as international *comparative* studies) do not compare countries but are intended to describe, analyze, or make proposals for a particular aspect of education in a country other than the author's own (Kaiser, 1999).

Countries may have various reasons to participate in international comparative studies (Robitaille, 1994). Comparative studies may provide countries with the opportunity to examine their own implicit theories (e.g., about how children learn mathematics), values and practices. A variety of teaching practices, curriculum goals and structures, school organizational patterns, and other arrangements for education can be studied that might not be possible within in a single country or education system.
Another reason may be that setting realistic standards and monitoring success of their educational system can be facilitated for countries by using the results of comparative studies complementarily to the results of their national evaluation studies. Thus, comparative results can serve as a baseline for the evaluation of the quality of education within countries.

In general, international comparative studies have two purposes. *Descriptive* studies describe crucial aspects of educational practices and outcomes separately. An example of a descriptive study is a periodical study organized by the Organization for Economic Co-operation and Development (OECD). This study is called the Indicator in Education Study (INES) and is conducted at a predetermined interval. The INES study results in periodic publications called 'Education at a Glance' (OECD, 1992, 1993a, 1997, 2000). For these publications, OECD is using data collected in other studies like surveys conducted under the auspices of IEA. *Explanatory* studies are designed to study the relationship between educational practices and outcomes. In explanatory studies, descriptive results are provided first. For example, the relationship can be examined between characteristics of instructional practices and achievement in mathematics to find explanations for differences in mathematics achievement levels across countries. In the IEA's Second International Mathematics Study (SIMS) countries could participate in a longitudinal version which provided them the opportunity to find explanations for cross-national differences in achievement level. SIMS is reviewed in chapter 2.

Different target groups use the outcomes of comparative studies: policymakers, educational specialists, educators, and scientific researchers. Some of them might be more interested in results at the descriptive level, others would prefer explanations of the descriptive statistics.

Many questions arise from looking at the results of international comparative achievement studies in education. If the results are stated in terms of the number of countries that outperform their own, users of the data often ask for explanations. Possible questions are: Why are we outperforming these countries and not other countries? Is it, for example, because our students are different, or because the instructional practice in our country is on average different from other countries?; Do factors that matter within one education system also matter in other systems?, and Which factors can be considered as universally important (in many education systems around the world or in one region of the world)?

To address such kind of important questions, a great deal of background data must be collected in addition to achievement data. Background factors can be found at four levels of an education system: the way education is organized within a country (system/country level), and within schools (school level), the quality of the instructional practices within the classroom (classroom level), and students' characteristics (student level). At each of these four levels of an education system, variables can be studied which can contribute to the explanation of differences in achievement in school subjects across countries. Two examples of questions that can be studied are: What topics are taught within a country? (system level), Does teachers' behavior in the classroom enhance students' achievement results? (classroom level). The factors classified at different levels may interact with each other. Therefore, it is necessary that student-, classroom-, school-specific, and country-specific factors are studied simultaneously as predictors of cross-national differences in achievement. An example of a question that refers to such interactions is: What kinds of teaching activities enhance student achievement given particular students characteristics?

The international comparative achievement studies conducted by IEA are examples of LINCAS. IEA studies can be characterized as multi-purpose studies aimed at providing both descriptions and explanations (Plomp, 1998). In most IEA studies, both achievement data on a core subject and contextual/background data are collected. Contextual data are measured at different educational levels and are

supposed to be related to student achievement. A major aim of analyses on contextual educational data is to identify factors that influence student achievement. Once factors have been identified which are supposed to influence educational outcomes, possible ways to improve education can be proposed.

In the remainder of this thesis the term education system is used as a synonym of 'country,' because certain members of IEA are part of a country but operate from an educational perspective independently from the other part(s) of the country. An example is Belgium with its two education systems: Belgium Flanders and Belgium French. Both systems participate in international comparative studies in education. Yet, for the sake of variability and convenience the terms 'nation' and 'country' are used interchangeably with education system.

The next section describes the general goals and functions of large-scale international comparative studies in education conducted by IEA, in which student achievement in a core subject is measured.

1.2 LARGE-SCALE INTERNATIONAL COMPARATIVE ACHIEVEMENT STUDIES: GENERAL GOALS AND FUNCTIONS

The IEA was the first international organization that conducted international comparative studies in which student achievement was measured by the same standardized objective cognitive tests in more than one education system. The IEA was founded in 1959 as an independent, non-governmental, international co-operative of research centers of different education systems. Any system may join IEA and currently IEA has more than 60 members.
Since its establishment, IEA has been primarily interested in international comparative studies from a research perspective. The founders of IEA regarded the world as a laboratory in which the differences between education systems would provide the opportunity to examine the impact of different variables on educational outcomes. In most of the studies, 'educational outcomes' was operationalized in terms of student achievement. IEA emphasizes that their studies should not be interpreted as an international race to determine a winner. Instead, IEA studies should be considered as opportunities to explore the diversity of political, economic, cultural, and educational contexts of participating systems. Achievement

in core subjects is studied against a wide background of school, classroom, student, and societal factors *"in order to use the world as an educational laboratory so as to instruct policymakers at all levels about alternatives in educational organization and practice"* (Robitaille & Garden, 1989, p. V). Within this context, IEA considers as its *mission* to conduct comparative studies focusing on educational policies and practices in order to enhance learning within and across education systems (Plomp, 1998).

Since the 1980s, policymakers have become increasingly interested in educational indicators, regardless of their interrelationships. Consequently, IEA tried to serve these interests. For example, the OECD included achievement indicators which were measured in IEA studies, in its periodical publication 'Education at a Glance' (OECD, 1997, 2000).

General goals

LINCAS have different goals and functions. In order to reflect on the results of such studies the goals and functions must be explicit. The goals and functions of Large-scale achievement studies conducted by IEA are discussed by Postlethwaite (1988) and Plomp (1998).

The IEA recognizes two main, general goals of its achievement studies (Plomp, 1998):

(i) to provide policymakers and educational practitioners with information about the quality of their education system in relation to other relevant systems; the first step toward learning from other systems is to identify what is happening within them.

(ii) to assist in understanding the reasons for observed differences between education systems.

Each of the above goals requires its own kind of comparison. The first asks primarily for international comparisons - at a descriptive level - of effects of education in terms of total test and sub-test scores on international achievement tests. Differences in mean test scores and in the distribution of the test scores across systems can serve as indicators for the quality of education systems. Achievement data is not the only kind of information necessary to accomplish the first goal. To identify 'what is happening elsewhere' a description of indicators referring to educational inputs, conditions and processes at different levels in the school (student, classroom/teacher, and school) is needed as well.

The second goal refers to explanations of described differences across nations. This goal can be dealt with by analyzing factors of educational processes and their relationships with achievement in an international comparative context. An example of such a factor is 'curriculum': the difference between what should be taught in a particular grade (intended curriculum) and what is actually being taught in that grade (implemented curriculum) could be investigated.
The comparative analysis necessary to achieve the two general goals can also be conducted at national level, within an educational system, by comparing data from different schools by ability track (like in the Netherlands) or by geographical region (like provinces in Canada).
Given the different kinds of data available in LINCAS, specific functions can be formulated which stress the importance of international comparative studies, in particular the studies conducted under the auspices of IEA.

Functions

Within the framework of its mission and goals, IEA studies may serve a number of functions for national and international educational policymakers, practitioners and researchers. Plomp (1998) and Postlethwaite (1988; 1999) named five major specific functions of IEA's international comparative achievement studies: description, benchmarking, monitoring the quality of education, understanding of reasons for observed differences, and cross national research.

Description
This function refers to describing similarities and differences in educational phenomena between systems of education. The status of an education system in an international comparative context can be described on the basis of tables in which outcomes of an achievement test in a number of countries are ranked. Additionally, tables can be presented with descriptives of individual background factors considered determinants of national achievement. Policymakers may consider these tables as a *starting point* for an examination of their own educational system. The examination can show strengths and weaknesses of one education system compared to other systems in terms of achievement results, and background variables such as students' attitude towards school and particular subjects, and features of the instructional practices and school organization. After studying the results in greater

depth, policymakers can formulate new policies regarding teacher education programs or the curricular contents of the subject(s) that have been investigated (Postlethwaite, 1988).

Benchmarking

Education systems can compare their own achievement level on an international test with the level in other systems. Differences with other systems (positive or negative) can lead investigators to study differences with the same system. Investigators may study such variables as curricular materials, instructional processes, school variables (e.g., instructional leadership of the principal), teacher training, and in-service training. By estimating the relative effects of these variables on outcomes, the question 'what affects what' can be investigated both within and between education systems. Similarities in determinants among systems may lead to generalizations across systems about the effect of particular student, class/teacher, and school variables. Further data analyses may result in recommendations for educational change. In a future international comparative study, an education system can compare its own achievement results with the same systems to find out whether the differences have changed.

Monitoring the quality of education

If a cycle of regular assessments in certain subject areas is conducted, an education system can monitor the achievement level within an international perspective. TIMSS-Repeat is an example of an IEA study that delivered trend data as the TIMSS test was administered in 1994/1995 and the TIMSS-Repeat test, a replication study, in 1998/1999. The trend study will be continued as a four-year periodic study. The next data collection is planned for 2003.

Understanding of reasons for observed differences

Related to the second general goal of IEA studies mentioned above, policymakers might want to understand the similarities and differences between or within education systems from the perspective of national policy making. This function goes one step further than just collecting data for monitoring purposes. Understanding similarities asks for the identification of general principles concerning educational effects. To be able to reveal a possible pattern of relationships between variables within an educational system and an outcome variable (for example

achievement in mathematics) a model must be postulated. In such a model, certain variables must be held constant before examining the relationships between other variables and the outcomes. The resulting relationship is often estimated by a regression coefficient. Potentially, factors can be traced which are effective in all systems under investigation (called general principles) or in one or a few systems (education system specific effects). For instance, in IEA's SIMS the data of the United States were compared with data of other countries resulting in possible explanations of the relatively low achievement level in mathematics found in the U.S. during the SIMS data collection in 1982 (McKnight et al., 1989).

Cross-national research

IEA studies result in enormous databases. These databases can serve to answer more exploratory and in-depth questions from a comparative perspective across many participating systems. Results of cross-national research concern not necessarily the comparison between one particular country and a group of other countries. Two examples of cross-national research based on IEA databases are particularly relevant to the present discussion: Postlethwaite and Ross (1994) tried to explain differences between low and high achieving countries on the IEA Reading Literacy test (data collection in 1990-1991) by identifying indicators at school, classroom and student level. Keeves (1996) reviewed the outcomes of all IEA studies conducted before 1995, summarized ten key-findings, and presented a discussion of implications for educational planning by policymakers.

In this thesis the 'understanding' function of IEA studies is examined. The purpose of the study is described in the next section.

1.3 PURPOSE OF THE STUDY

IEA study results have often been used rather uncritically by policymakers. This has been particularly so when lists of education systems ranked on the mean achievement score were presented with little information about contextual factors (Kaiser, 1999). The determination of the relative position of one's education system in relation to other systems has some intrinsic value. Policymakers can develop an understanding of their own system and they can try to detect strong and weak features of the achievement results of their own student population compared to

other nations (function 1 described above). In order to better understand such strengths and weaknesses, information about the context or background factors within each system should be taken into account as well. In fact, understanding the relationships between achievement and all possible factors that may influence achievement is crucial for comparisons both within and across education systems (function 4). The international data sets collected in two IEA studies conducted in the 1960s (the First International Mathematics Study, FIMS) and the 1980s (SIMS) were analyzed to find explanations for cross-national differences in mathematics achievement. In chapter 2 some results of these analyses are discussed. The general conclusion about these results is that it is very hard to find explanations for differences in achievement level across nations. FIMS and SIMS showed how difficult it is to achieve this function. Both studies are predecessors of one of the most important and ambitious IEA studies, the Third International Mathematics and Science Study (TIMSS). TIMSS is an example of a large-scale international comparative achievement study with high explanatory ambitions and its two main functions are 'description' and 'understanding,' as was formulated by Robitaille (1993, p.25): *"describing the status of education systems in terms of outcomes of the international mathematics test and individual background variables"* and *"understanding of reasons for observed differences and similarities."*

The purpose of this thesis is to determine to what extent the 'understanding' function was accomplished by TIMSS. The first international results of TIMSS data analyses were published by Beaton, Mullis, et al. (1996) and Beaton, Martin, et al. (1996). In these reports, descriptive achievement results were dominant. In addition, frequencies and descriptives of background data that were collected by means of questionnaires (students and teachers) were presented.

The background information collected in TIMSS has not yet been used intensively to understand similarities and differences in student achievement across countries. Few reports are known in which researchers attempt to find explanations for cross-national differences in students' achievement level in mathematics or science (e.g., Martin, Mullis, Gregory, et al., 2000; Wossmann, 2000; Zuzovsky & Aitkin, 2000; Bos & Kuiper, 1999). What possible reasons could be given for this observation? Is there maybe a want of financial resources or are the collected data and the applied data collection methods possible reasons for the little explanatory reports on TIMSS? In this thesis the possibilities TIMSS data sets offer to find such explanations are studied in greater depth.

The utility of the background information is studied within the perspective of the 'understanding' function of TIMSS. In particular, the conceptual framework of TIMSS and its instrumentation and design are the focus of the study. As far as the instrumentation is concerned, the development and utility of the background questionnaires are investigated. In LINCAS such as TIMSS, international achievement tests are developed to measure students' knowledge and skills of the subject under investigation. The development process of such tests is an intensive endeavor, which is itself worth a study. However, in this thesis the composition of the TIMSS achievement test is not discussed. The development, international reliability and validity and the results of the TIMSS achievement test have been studied and discussed by Beaton, Mullis et al. (1996), Kuiper, Bos and Plomp (1999), Afrassa and Keeves (2001) and others and appeared to be adequate.
Reflection upon the TIMSS results will provide insight in the extent to which TIMSS fulfilled its explanatory ambitions and in lessons that could be learned. The ultimate aim and the scientific interest is to formulate recommendations for the design and components of future studies in which achievement results and influencing factors in different education systems are compared with each other to serve explanatory goals.

In the next section the problem statement and the research questions are formulated.

1.4 PROBLEM STATEMENT AND RESEARCH QUESTIONS

The dominance of descriptive reports about TIMSS in the first years after the study was completed, leads one to suspect that the data sets of this study have not often been used to find explanations for cross-national differences in achievement results. Given the results of its two predecessors, the question can be raised as to what extent TIMSS data offer researchers opportunities to examine cross-national similarities and differences in relationships between achievement in mathematics and background factors at several education levels (student, classroom and school). This is the general problem addressed in this thesis and can be formulated in the following statement:

To what extent does TIMSS meet its predefined function of understanding cross-national similarities and differences in students' mathematics achievement level related to background factors, and how can future studies be improved to optimize their results in favor of their 'understanding' function?

In TIMSS, education in mathematics and science in three populations were studied in many education systems around the world (see chapter 2). However, in this thesis the TIMSS objectives are limited to mathematics education in grade 8 (year 2 of lower secondary education; most of the students are 14-years old) in three education systems: the Netherlands and two other State members of the European Community, 'neighboring' Belgium Flanders and Germany. The subject mathematics is more univocal across the three neighboring countries than science. Grade 8 is the main grade of the second population studied in TIMSS.
The problem statement can be translated into two research questions. The first part of the problem statement requires an examination of relationships between variances in students' overall scores on the TIMSS mathematics test and variances in scores on background variables. The first research question is:

I. *To what extent can variability in the overall TIMSS mathematics test scores for grade 8 within the Netherlands, Belgium Flanders and Germany be explained by variability in the scores on variables at student and classroom/school level and to what extent are these outcomes generalizable across these three European educational systems?*

To be able to answer research question I, relational data analyses were carried out in an exploratory way. This kind of data analysis addresses the function of 'understanding of reasons for observed differences.' In chapter 4 the results of the exploratory analyses are discussed.
Notwithstanding the potential for the TIMSS data sets to compare countries, one important question is what lessons can be learned from TIMSS. Once the answers to question I are available, the conceptual foundation and instrumentation, and the design of TIMSS are reflected upon (see chapter 5). What can be concluded about the extent to which TIMSS fulfilled the function of understanding cross-national similarities and differences in background factors related to mathematics achievement and what recommendations can be made?

This results in the second research question for this thesis:

II. *What can be learned from the case of IEA's TIMSS for future international comparative achievement studies in education regarding the conceptual foundation, instrumentation and design in view of their possibilities to uncover factors related to different outcomes across educational systems on an international student achievement test?*

As the first research question concerns a concrete example of the results of TIMSS in an international comparative context, the second one concerns the feasible improvements of LINCAS more generally. Research question II can be located at a meta-level. In what way can future LINCAS be improved to provide policymakers, educators, researchers and others with better opportunities to find meaningful reasons for similarities and differences they might see in the 'international' mirror, between their own country and others? The investigation of research question II is mainly based on a reflection on the results of the three-country comparison conducted by means of analysis of TIMSS data. More generally, the appropriateness of three main components of large-scale international comparative studies, the conceptual foundation, the instrumentation, and the design is reflected upon in the light the goals and functions of LINCAS. The three components are described in chapter 2. In chapter 5 research question II is elaborated further.

1.5 STRUCTURE OF THE THESIS

In chapter 2, the structure of two predecessors of TIMSS, FIMS and SIMS, are presented in terms of their goals, conceptual foundation, the instrumentation and design and the utility of the results from the perspective of the goals. The structure of TIMSS is described more generally in chapter 2. In chapter 3, the conceptual framework is reviewed resulting in additional reviews of instructional and school effectiveness models. This chapter ends with an organizing conceptual framework that is used in the first step from the data analysis plan to address the first research question. In this step, reported in chapter 4, sets of items are explored that can be regarded as operationalizations of contextual factors at student and teacher level that are supposed to be potentially effective on student achievement.

Also in chapter 4, final results are described of exploratory path analysis on TIMSS data that was focused on factors affecting achievement in mathematics in the Netherlands, Belgium Flanders and Germany. As a follow-up of the exploratory path analysis, multilevel analysis was applied. In the latter, the hierarchical structure of the study design could be taken into account. In the multilevel analysis, variables were included which in the unidimensional path analysis turned out to have direct or indirect relationship with the dependent variable 'achievement in mathematics.'
In chapter 5, the benefits and limitations of the conceptual foundation, the instrumentation and design of TIMSS are discussed. The benefits and limitations of TIMSS are considered in greater depth in light of the predefined function of 'understanding.' Subsequently, general reflections and recommendations are given aiming at an appropriate conceptual framework and instrumentation and design for future international comparative achievement studies in education.

CHAPTER 2

IEA's TIMSS AND ITS PREDECESSORS

Each international comparative study can learn from its predecessors. As the Third International Mathematics and Science Study (TIMSS) is central in this thesis, two of its predecessors are reviewed in this chapter. These studies were conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) in the period between 1964 and 1995 and mathematics was the investigated core subject. In particular, the extent to which the results of these studies are adequate to accomplish the studies' function of 'understanding' cross-national similarities and differences in background factors related to students' mathematics achievement level, is of interest. The reviews are guided by a general study framework for Large-scale International Comparative Achievement Studies in education (LINCAS) with the components: general goals and functions of the study, the conceptual framework, the design and instrumentation, and the utility of the study results (2.1). The criticism TIMSS' predecessors received from internal and external sources is described as well.

The First International Mathematics Study (FIMS) is described in 2.2 and the Second International Mathematics Study (SIMS) in 2.3. The benefits and limitations of both studies are summarized from which the founders of TIMSS could learn (2.4). Finally, basic components of the general study framework for TIMSS are described in 2.5.

2.1 COMPONENTS OF A GENERAL STUDY FRAMEWORK FOR LINCAS

The Third International Mathematics and Science Study (TIMSS) was preceded by the First and Second Mathematics Studies (FIMS and SIMS). The latter are reviewed in 2.2 and 2.3. All of these studies were organized by the International Association for the Evaluation of Educational Achievement (IEA). The reviews are written along components of a general study framework for Large-scale International Comparative Achievement Studies (LINCAS) presented in Figure 2-1. The framework shows different components that are usually included in such studies and the way they are assumed to be interrelated (Rosier, 1997).

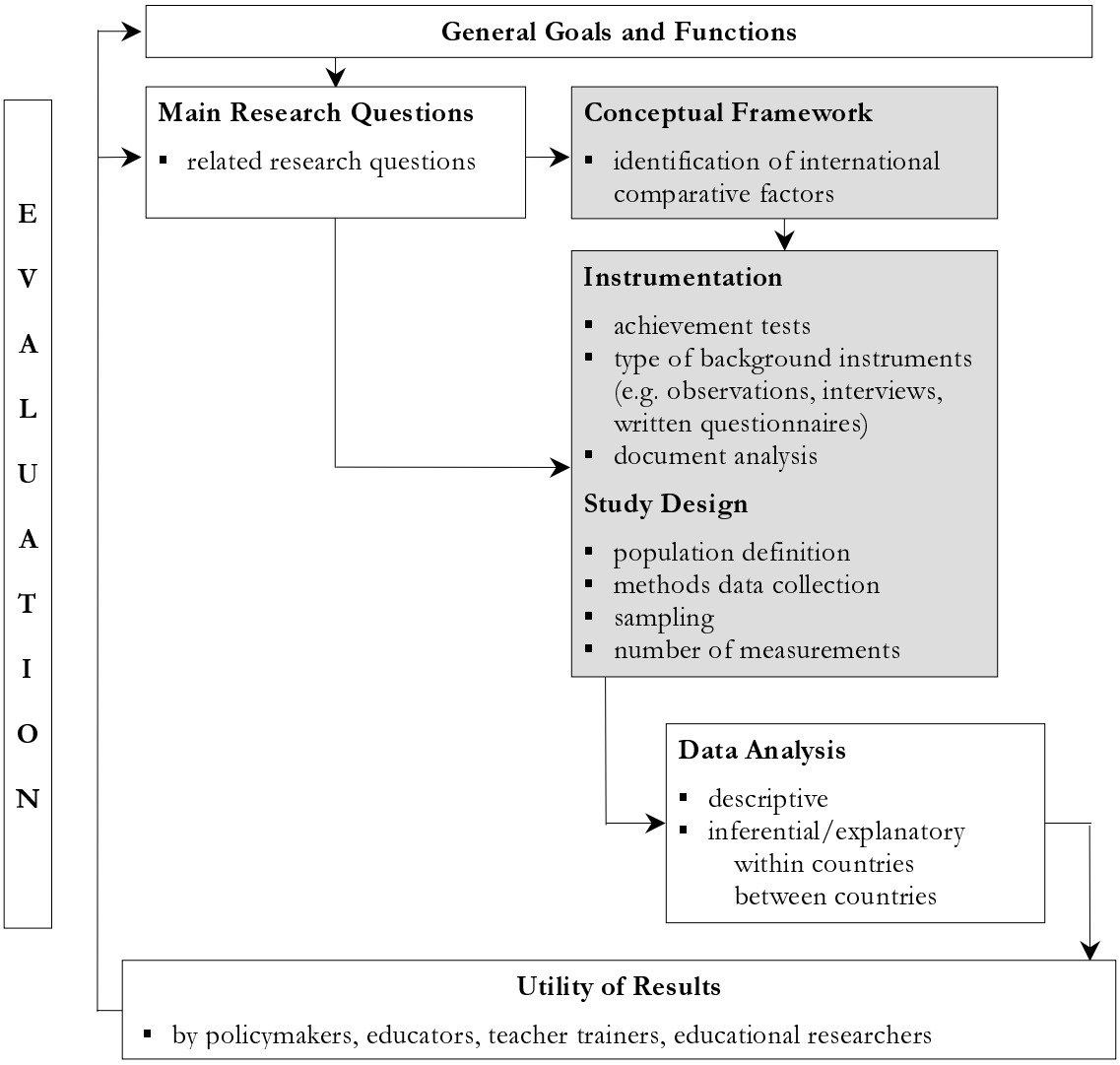


Figure 2-1
Components of a general study framework for LINCAS

The *general goals and functions* of a LINCAS are located at the top of Figure 2-1. Possible goals and functions of IEA studies were described in 1.2. The formulation of the *main research questions* of a LINCAS are based on these goals and functions and on the concrete queries of the study. A LINCAS is organized primarily around the questions that the funders (usually governments) of the study want answered. For instance, a government is interested in the extent to which its education system compare with other systems with regard to students' achievement level in core subjects and it might be interested in the educational factors associated with achievement.
To accomplish predefined goals, international comparative research questions should be formulated as concretely as possible and the collected data should be internationally comparable, reliable, and valid (Martin et al, 1999). The cross-national comparability of the data collected in LINCAS is enhanced if the *conceptual framework and instrumentation* and *the study design* are applicable to all participating education systems. More than in national studies, these two components (shaded in Figure 2-1) are essential in international comparative studies with participants from around the world. The development of a conceptual framework, a set of instruments and the study design should facilitate the identification of possible key factors (concepts) in each country that affect achievement.

The *conceptual framework* should be appropriate for studying education within and across the participating systems. In national studies, the definition and operationalization of reliable and valid key concepts is already a difficult task. In LINCAS this task is even more complicated because the definitions should be valid for all participating countries. Next, indicators are to be developed in terms of variables and further in terms of questions that will be included in survey instruments (Postlethwaite, 1999). For example, is it possible to formulate questions for students about their opinions of going to school in primary and secondary education in many countries around the world in an unambiguous way? Queries like this need an affirmative answer in order to include student opinions as valid variables in an international comparative study. Furthermore, in the conceptual framework the hypothesized interrelationships between (clusters of) key concepts must be made clear. Preferably, such interrelationships must be theory-based and findings of previous studies related to the main topic of the research questions. However, it might be questioned whether any theory is available about international comparisons in education (Postlethwaite, 1999).

The next step in designing a LINCAS is the choice of the *type of instrumentation*. International content validity is a goal of the developmental process of instruments applied in LINCAS. The reliability and validity of international data can be partly dependent on the kind of instruments used to collect the data. In large-scale studies, written questionnaires are most desirable from the perspective of time and money. However, more time-consuming and more expensive methods of data collection such as classroom observations, interviews, and videotapes can enhance international reliability and validation of the data. Students' achievement is often measured by means of international achievement tests (sometimes via a performance test). Document analysis is a method to collect data at the system level, for example about national curricula and the structure of the educational system.
The choice of the type of instruments is partly related to the *study design*. In principle, two designs can be applied to accomplish the 'description' and 'understanding' function of LINCAS: a descriptive (qualitative and/or quantitative) or a more inferential explanatory (mainly quantitative) research design. The choice of the kind of study design – quantitative, qualitative or a combination – determines the instrumentation types that will be used in the study.

The choices of the design, the instrumentation and *the techniques to analyze the data* are interrelated. Each choice provides the researcher with certain possibilities to answer complex comparative research questions. Yet, each selected design and developed instrumentation of a LINCAS set limits for the statistical analysis techniques that can be applied. In the participating education systems data are usually collected to study the influence of one factor on another factor and finally on educational achievement in a certain subject. The selection of appropriate techniques of statistical analysis is essential for revealing the relative effects of factors postulated as influencing a particular outcome. In general, the majority of techniques for analyzing quantitative data available for the social sciences result in the estimation of coefficients. Coefficients indicate the strength of the effect of one factor on other factors, relative to the strength of the effect of other factors. Examples of such statistical techniques are unidimensional path analysis and hierarchical linear modeling.

Utility of results

Potential users of LINCAS results (policymakers, educators and other educational specialists) would like to receive recommendations for improving education if

needed. Preferably, such indications are based on relationships between achievement scores and student, classroom and school factors.
Results of previous international comparative achievement studies have been used in many different ways. For instance, results from SIMS were directly used to reform curricula. Hungary and Sweden are two examples of countries which have used results in this way (Wolf, 1992). In other countries, the curriculum was modified after the results of a LINCAS were presented so that certain groups of students (for instance, lower achieving students) are better served by the educational system. Such curriculum reforms or modifications are possible in centralized systems.

In decentralized countries like the Netherlands or Germany, LINCAS do not have such impact and changes cannot be implemented in the short term. In these education systems, schools function rather autonomously and the use of results of LINCAS depends more on the level of concreteness by which the researchers are able to present the results.

Evaluation: benefits and limitations of LINCAS

The benefits and limitations of LINCAS may be evaluated in the light of their goals, functions, and research questions. Given the complicated nature of international comparative studies, an important question is to what extent researchers succeed in satisfying users of LINCAS results, by pointing out which key factors account for cross-national similarities and differences in achievement scores. On the basis of the results of such evaluations, it is possible to recommend improvements to crucial components of the study framework of future LINCAS. The recommendations can include making the benefits of the LINCAS explicit and providing guidelines for the development of an appropriate conceptual framework and instrumentation, and design.
Likewise, the LINCAS that is central in this thesis, TIMSS, could have learned from two of its two predecessors. In the next two sections FIMS and SIMS are reviewed. The First and Second International Science Studies (FISS and SISS) are also predecessors of TIMSS (Comber & Keeves, 1973; Postlethwaite & Wiley, 1992). As in this thesis influencing factors on *mathematics* achievement is studied, merely FIMS and SIMS are reviewed. In the reviews, four components of the general study framework for LINCAS presented in Figure 2-1 are discussed: 'general goals and functions,' 'conceptual framework,' 'instrumentation and the study design,' and

'utility of results.' In addition, attention is paid to criticism the studies received after their international reports were released.

2.2 FIRST INTERNATIONAL MATHEMATICS STUDY (FIMS)

One of the first international comparative studies conducted under the auspices of the IEA, took place in the mid-1960s (Husén, 1967; Wiegersma & Groen, 1968). This study is called the First International Mathematics Study (FIMS). The number of participating education systems was 12 (including Australia, Belgium Flanders, England, Israel, Japan, the Netherlands, and United States).

General goal and function

FIMS was one of the first studies organized by IEA. The general goal of FIMS was to study the feasibility of investigating schools and schooling in different education systems in a comparative way. The main objectives were to examine differences across educational systems and the relationships between differences in achievement and students' interests and attitudes. To some extent these objectives can be regarded as expressions of the 'description' and 'understanding' functions mentioned in section 1.2. Mathematics achievement served as a surrogate measure of the outcomes of schooling. Organizers chose mathematics achievement as the educational output measure as a matter of convenience (Travers & Weinzweig, 1999). They believed that it would be easier to make international comparisons in mathematics than in any other area.

Design and instrumentation

In FIMS, two populations were tested: 13-years-olds and students in the final year of secondary education. The majority of the 12 participating countries took part in the first population only. No conceptual framework is known that guided the development of FIMS quantitative study design and instruments. The set of instruments consisted of an international mathematics achievement test and four background questionnaires: a student, a teacher, a school principal and a questionnaire about system characteristics. In all questionnaires, many questions referred to students' interests and attitudes. Neither contextual classroom nor school data were collected.

Utility of the results

As FIMS was one of the first LINCAS on mathematics, the data could serve as a baseline for international comparisons. FIMS is regarded as an experiment in international comparative studies that provided useful descriptive information on mathematics achievement. The state–of–the–art of mathematics education reflected in the FIMS results within an education system, could not be compared with outcomes of previous studies. It was the first time achievement results of each system could be compared with the results of the other participating systems.

The main results of FIMS consisted of lists, ranking countries on the basis of the mean achievement test score. The results were interpreted in terms of 'on average country A performed better on the international mathematics achievement test than country B.'
It can be stated that FIMS was a huge and innovative undertaking in the mid-sixties. Howson (1999) reflected on the value of FIMS by stating that it provided guidance on what worked in education and on what required rethought. FIMS was used as a starting point in the development of international comparisons in education. Representatives of the 12 participating education systems could use the achievement results of FIMS to formulate questions about reasons for differences between the results of their own students and students' results in other systems.

Criticism

There was also criticism of FIMS. First, it became clear that a ranked list of countries means little without knowledge about the context in which schooling is taking place (Wiegersma & Groen, 1968). Information was needed about the national and the school context (including the curricular context) and about teachers and students to enhance interpretation of the test results in an international comparative perspective. The goal to relate achievement data to students' interests and attitudes in order to understand some of the differences in achievement scores across nations was not reached. This might have been caused by the fact that the design and the instrumentation of FIMS were not very well defined. Second, despite the exclusive goal of FIMS to relate achievement to students' interests and attitude, the collected student background information was hardly mentioned in the reports. Finally, one of the main questions that was raised was the composition of the

achievement test. The match with the contents of a system's curriculum was far from ideal. The item bank used to compose the FIMS achievement test was criticized for not fairly reflecting the different curricula in the participating education systems (Howson, 1999). Besides, curricular aspects were missing in the formulation of comparative results (Freudenthal, 1975). Another criticism of the test was that it consisted only of multiple-choice items. Educational output could be measured with the test, but the reliability and validity of the FIMS test were open to question (Robitaille & Travers, 1992).

2.3 SECOND INTERNATIONAL MATHEMATICS STUDY (SIMS)

The second predecessor of TIMSS reviewed here, is IEA's Second International Mathematics Study (SIMS). Experiences from FIMS were taken into account in the preparation phase of SIMS, which began in 1976. The SIMS data collection took place in the period between 1980 and 1982. The number of participating education systems increased from 12 to 20.

General goals and functions

The goals of SIMS were much more ambitious than the goals of FIMS. The overall goal of SIMS was to produce an international portrait of the teaching and learning of mathematics with a particular focus on what actually transpires in mathematics classrooms around the world (Robitaille & Garden, 1989). In SIMS, the varieties of curricula, instructional practices, and student outcomes (both cognitive and attitudinal) were studied. The founders of SIMS planned to provide participating education systems with a mirror of their own mathematics curriculum in a cross-national perspective. This refers to the 'description' function of LINCAS presented in chapter 1.
It was assumed that knowledge about instructional practices in relation to the output factor (level of mathematics achievement) in each country would prevent simplistic ranking lists. The complexity of including input and process factors in SIMS was clear from the beginning. Yet, by doing so the SIMS results would afford each system an opportunity to better understand the relative strengths and shortcomings of its own mathematics education (Travers, Garden & Rosier, 1989). This refers to a certain extent to the 'understanding' function of LINCAS.
SIMS attempted to *"search for information about what mathematics is intended to be taught, what mathematics is actually taught, how that mathematics is taught, and what mathematics is*

learned by those taught" (Travers, et al., 1989, p. 3). The conceptual framework for SIMS fits into this proposed function (see below). The plan was to have significant input of the mathematics education community at every stage of the project.

Conceptual framework

SIMS was planned as an in-depth study of the mathematics curriculum. The focus of SIMS was the attained curriculum against the background of classroom practices. Hence, the theoretical framework of the study consisted of three curriculum levels: the intended curriculum, the implemented curriculum, and the attained curriculum (Travers et al., 1989). In many countries, the curriculum is prescribed in national curriculum guides and presented in approved textbooks. This is regarded as the intended curriculum. Teachers are expected to translate these guides and textbooks into actual classroom instruction: the implemented curriculum. The third curriculum level, the attained curriculum, is defined as the student outcomes of education. Understanding the relationship between the three levels is necessary if the SIMS goal is to be accomplished.

Examples of important questions in SIMS about the relations between the three curriculum levels are: How much and what have students learned from the materials presented and instructed in the classroom? (is there a discrepancy between attained and implemented curriculum?) and How well do the teachers translate what has been mandated at national or system level? (match between intended and implemented curriculum). This 'three curriculum level conceptual framework' was adopted and developed further by TIMSS and is described in more detail in 3.3.

Design

Two populations were studied in SIMS: all students in the grade where the majority was 13-years-old by the middle of the school year (population A) and all students in their last year of secondary school who are studying mathematics as a substantial part of their academic program (population B; students of different ages). These two populations were roughly comparable to the ones that were studied in FIMS. The second population differed across participating countries. As a consequence, the interpretation of the countries' achievement results was difficult.

The international portrait for which SIMS strove had to be based on a cross-sectional data set of 20 education systems from all over the world, ranging from

highly developed industrialized countries to developing countries. Given this range, painting the international portrait was far from simple. To pursue the general goal of SIMS further, a second version of the study design was developed: a longitudinal study with 8 out of the 20 participating systems.
In the longitudinal version, pretests were administered at the beginning of the school year and post-tests at the end of it. From the two test scores a gain score was computed. In fact, this gain score can be seen as the dependent variable of the study. The pretest-posttest design made it possible to investigate *growth* in student achievement (what and how much did they learn within one school year) and its links with teaching practices. The pretest-posttest design provided SIMS researchers opportunities to address the question, "What teaching practices are utilized in the mathematics classrooms of the various systems and to what extent can these classroom processes explain differences in student achievement?" This is a research question many policymakers would like to have answered for their country in an international perspective. The pretest-posttest design offered possibilities for isolating the influence of current year instruction on student performance from prior learning and mathematical ability. In the cross-sectional design only the current *status* of education system could be studied.

Instrumentation

The attained curriculum was determined by an international mathematics achievement test. Freudenthal (1975) criticized FIMS because the bank of test items did not adequately reflect the different curricula followed in the participating countries. SIMS strove for a better fit between the item banks and the national curricula. The construction of the international test was an intensive process and a joint activity between the study centers of the participating education systems. National study centers conducted preliminary surveys to collect data of the intended national mathematics curriculum. The national data sets were analyzed to construct a content-behavior grid (Travers et al., 1989). The content dimension of the grid refers to different strands (e.g., arithmetic, descriptive statistics and measurement) which were further subdivided and refined. The behavior dimension of the test grid refers to four cognitive behavior categories: computation, comprehension, application, and analysis.
Participants of the preliminary surveys in the countries were asked to provide examples of how they interpreted the content/behavior combinations. They were

also asked to provide items fitting in each cell of the grid. Each country had also the opportunity to specify the level of importance of each cell. The final result of these surveys consisted of a topography of the international curriculum in the form of the content/cognitive behavior grid, with cells marked 'very important for most countries,' 'important for most countries,' 'important for some countries,' and 'unimportant for most countries.' The grid formed a solid basis for the selection of items for the international mathematics achievement test.

Another criticism on FIMS was not taken into account in the preparation stage of SIMS: the format of the test items was only multiple choice in FIMS as just well as it was in SIMS. Student's knowledge of mathematics can be measured by means of multiple-choice items. The free response format could provide more insight than multiple-choice format in the way students interpret text or diagrams to describe or explain procedures, processes, or mathematics concepts (Beaton, Mullis, et al., 1997). Free response items require students to construct their own answers. However, the scoring of free response items is very labor-intensive (and costs a lot of money). Therefore, a mix of free response and multiple-choice items is desirable for an international achievement test.
A thorough coverage of the curriculum across countries was more possible in SIMS than in FIMS, because SIMS used twice as many items than FIMS. However, the SIMS achievement test still contained an insufficient number of items to provide a full coverage of the curricula across countries.
At the level of the intended curriculum, the national centers were asked to complete a questionnaire containing system-level background information, including ratings of the appropriateness of each test item employed in the study for students in that system. In SIMS publications this is called 'Intended Coverage' data (Robitaille & Garden, 1989).

In order to collect data at the implemented curriculum level in SIMS, a variety of background questionnaires were developed. The questionnaires contained background questions and were directed to school principals, teachers and students (Travers et al., 1989). These instruments were administered within both the cross-sectional and the longitudinal version of the study. The school principal completed the school questionnaire concerning general characteristics of the school, teachers, the mathematics curriculum and the school and departmental policies aimed at

mathematics instruction. Teachers were asked to fill in a questionnaire with background questions about their teaching experience, training, qualifications, beliefs, and attitudes. The teachers were also asked (in both design versions) to rate whether the content needed to respond to each item on the achievement tests had been taught previous to the administration of the SIMS test. In SIMS publications this is referred to as 'Implemented Coverage' (Robitaille & Garden, 1989). It is called 'Opportunity-to-Learn' (OTL) in other IEA studies (see for example, de Haan, 1992).

In the pretest-posttest study, mathematics teachers of the tested classes filled out extra questionnaires containing questions regarding classroom processes. These questions dealt mainly with three topics (Travers, 1993): the way teachers handled subject matter during the year, features of the organization of the instructional practice by the teacher, and the beliefs of the teacher about effective teaching. One of the teacher questionnaires contained questions to assess what teaching methods would be utilized during the school year in the teaching of a selection of topics. The topics were selected from the content-behavior test grid.
The students completed a questionnaire with questions about their personal background (e.g. parents' educational level and occupation), out-of-school support (both from their families and in the form of extra tuition) and their attitudes and beliefs about mathematics in terms of importance, liking, and difficulty. The 'Implemented Coverage' questions from the teacher questionnaire were also asked of students.

Utility of results

The SIMS results consisted of three parts: the achievement test results, the descriptive results of the questionnaire, and the results of the analysis of relations between achievement and student, classroom, and school background factors (both in the cross-sectional and the pretest-posttest design). Each part of the results received some – positive and negative – criticism.
After the criticism FIMS received, the researchers in charge of SIMS were very careful to prevent a simple list of countries ranked by student achievement in mathematics as the main result of the study. In three SIMS reports (Burstein, 1993; Robitaille & Garden, 1989; Travers & Westbury, 1989) the responses to the achievement test and questionnaires collected in the cross-sectional design were

presented at a descriptive level. Achievement results were presented in the form of ranked lists of countries. However, many school, classroom/teacher, and student variables were reported as well. For example, teacher practices were presented partly in terms of the most common approach in each country and across countries (e.g., 'chalk and talk,' whole class instruction, and a heavy reliance on the textbook). Class size is another example of a classroom variable that was measured. In the report, for each country the relation between this variable and achievement level was described in terms of *"some countries with relatively large class sizes performed at the highest level, Hong Kong for example"* (Travers & Weinzweig, 1999, p. 27). Such relations were presented without any correlation coefficient or another measure for the relationship under investigation.
With regard to the appraisals of the appropriateness of the achievement test, it was concluded that the students' and teachers' responses to the Opportunity-to-learn questions matched quite well (Robitaille & Garden, 1989). Teachers agreed that most of the items in the population A achievement tests concerned topics that were part of the curriculum. Thus, the content of the mathematics items was appropriate for most of the countries, which means that the majority of the items was rated as part of the implemented curriculum of most of the countries.
One of the relations described between the country mean score on the mathematics test and students' attitudes toward mathematics, was the top performance on the international test of Japanese students and their negative opinion about the attractiveness and difficulty of the subject.

For some countries it is known that governments started to change the mathematics education in some way after the SIMS results became public. For example, in Sweden the results of the study led to a huge in-service training program for mathematics teachers. In the United States the SIMS results formed an important impulse to set standards for mathematics which were developed by the National Council on the Teaching of Mathematics. Plans were developed there to monitor the status and progress of mathematics teaching and learning through the diffusion of its results and its methodology into national and state-level educational assessment and indicator efforts. In New Zealand SIMS results influenced an ongoing curriculum revision of mathematics and the development of pretests by topic area so teachers would have a better view of what students knew coming into their classrooms (Kifer & Burstein, 1993).

Criticism

Criticism on SIMS was directed to three topics: the cross-national comparability of scores on background variables and the achievement test, the pretest-posttest design, and the explanatory power of the results.

Cross-national comparability of background variables

In SIMS – different from FIMS – one of the goals was the explication and estimation of a causal model underlying academic achievement in different countries. As stated, background factors were measured by means of written questionnaires at different educational levels: student, classroom, school and/or education system. Schmidt and Kifer (1989) tried to find relationships between achievement and characteristics of students, teachers, classrooms, and schools. They applied comparable classical least squares regression analyses for each system. Because of the hierarchical nature of the data sets multistage regression procedures were employed to examine the joint contribution of all characteristics to academic achievement. Each analysis was a replicate. The results were interpreted in two ways. First, the extent to which the relationships were generalizable across systems and second, the extent to which findings were unique within a system. The researchers stressed that these analysis were far from complete: *"it represents little more than a first pass over the data"* (Travers et al. 1989, p. 2). The number of statistically significant regression coefficients indicating relationships between student factors and achievement scores, and between classroom and school factors and achievement scores for industrialized countries ranged from 7 in England and Wales to 43 in Finland. Schmidt and Kifer (1989) suggested a few reasons for these substantial differences across countries.

First, they mention *"the quality of the data and the extent to which variables, though labeled the same, do not mean exactly the same things in different systems"* (Schmidt & Kifer, 1989, p. 215). According to the authors the most important issue that can explain the differences is the difference in context in which schools and teachers operate. Such differences can account for disparities in the way factors are statistically related to achievement levels. Factors which seemed to be appropriate for one education system with their own context do not operate in the same way in other systems with other contexts. Comparisons between developing and developed countries showed big differences across systems as well. For instance, in the Nigerian data set only two regressions coefficients turned out to be significant.

Second, Schmidt and Kifer (1989) showed far more significant factors at the student level than at the school and classroom level. The authors mentioned some methodological reasons for this discrepancy. For example, the sample size for students was larger than for schools and classrooms and the conceptualization of the student factors was better than those of the school and classroom. The student factors have shown to be stable predictors in previous studies and the school and classroom factors have not. In SIMS factors were operationalized at school and classroom level as status variables instead of process variables. Observations of classrooms or schools were not included in SIMS. Observation might be a more appropriate method to collect process data, but is relatively expensive for many participating education systems.

It can be argued that the application of multi-level techniques instead of the multistage regression procedures could have delivered more desirable results from SIMS. Other researchers applied multilevel techniques to analyze the SIMS data sets from the cross-sectional design to find generalizations across education systems of instructional and school effectiveness indicators (Scheerens, Vermeulen & Pelgrum, 1989). They found that only two potentially effective factors turned out to have clear and consistent (across countries) positive effects on achievement: opportunity to learn (test items covered in mathematics lessons) in 9 out of 20 countries and teacher expectations (estimate by the teacher of the number of students who belong to the top band in mathematics) in 13 out of 20 countries.
Scheerens and Bosker (1997) warned for a contamination effect with the achievement measure (the international IEA mathematics test) since the operationalization of the two effective factors is strongly related to the contents of the test. In a limited number of countries a few other factors were identified as affecting achievement (Scheerens et al, 1989): class size (positive effect), homework, teacher experience, time spent on teaching, and time spent on keeping order (the latter had a negative effect).
Critics also focused on background factors which were not included in the SIMS questionnaires. Howson (1999) emphasized the importance of including factors at country level such as resources (national investment figures e.g., for teachers and for classroom materials and equipment) and the status education has in a country (how important education is for people, what the status is of the teacher's job). He also pleads for student factors like motivation, percentage of hours spent at school devoted to homework, and opting for mathematics or physics at the university level.

Furthermore, Robitaille and Garden (1989) mention the translation process. Each instrument had to be translated from English into the country's language, which could – despite of the execution of quality control procedures such as back translation – cause problems in the understanding of some questions. They emphasized that the comparability of variables depends on the way they were translated and measured internationally and on the importance the variables have in each education system.

Generally, opportunities researchers in the different countries have to measure all kinds of variables are both various and limited. Time constraints are important: filling out a questionnaire should not take more than about one hour, otherwise respondents are not willing to complete it seriously. Given such constraints, the number of data collections and the number of variables that can be operationalized in a questionnaire is also limited.

Achievement test

The achievement test developed and administered in SIMS was criticized by the study directors (Robitaille & Garden, 1989) and other researchers. The extent to which the achievement was 'equally unfair' for all participating systems was not seen as optimal. In some countries the fit between the test and the intended and implemented curriculum might have been better than in others. As long as any one country is convinced that the test is as unfair for their own students as it is for students in all other participating countries, the test should be developed further. However, the results of the opportunity to learn (OTL) analysis indicated that the SIMS achievement test was rather robust across countries.

Opportunity to learn (OTL) is an important measure for interpreting achievement results across countries properly. In SIMS, the OTL data were used to select so called 'system specific items' those items of which at least 80% of the teachers judged they are appropriate for their students. The results per education system on the different sets of system specific items were more or less comparable to each other. Hence, the relative position of a system in the list of the eight participating systems in the longitudinal design was not affected by the item set that was chosen as the achievement measure of comparison (Burstein, 1993).

Cross-sectional versus pretest-posttest design

Many international studies are cross-sectional, which limits the possibilities for causal modeling. The entry-level knowledge of students has not been measured in many cross-sectional studies. Therefore, the current status in achievement can be studied, but growth cannot. In these studies it is not possible to estimate effects that are attributable to teacher or classroom characteristics from the educational experience in the grades that are being studied. Such effects are confounded with aggregated student characteristics which are themselves related to prior achievement (Schmidt & Kifer, 1989).

The dependence of the relationships between predictors on the kind of design – with or without a pretest – is stressed by Schmidt and Kifer (1989). If relationships are estimated in a cross-sectional design, the meaning can be different from the estimations in a pretest-posttest design. At the same time, the set of variables which potentially influence the output variable (either a status achievement score or a growth score) can be different.

Results of the two design versions applied in SIMS showed advantages of the inclusion of a pretest and a questionnaire with process variables (Schmidt & Burstein, 1993). Analyses were conducted on the data sets of all eight participating education systems in the longitudinal version of the study design. The researchers reported bias of regression coefficients caused by a lack of pretest scores in the cross-sectional design. On average, in each system the level of statistical significance changed for more than half of all regression coefficients calculated when the pretest scores were included in the data analysis.

Explanatory power of SIMS results

The longitudinal design in SIMS provided more opportunities than the one-shot study to describe differences across education systems in mathematics education. However, available results of analyses on SIMS data – including the pretest results – offer few possibilities for *explaining* the differences. Thus, policymakers could use SIMS results as a mirror, but not really as a basis for planning the implementation of concrete changes in their educational system.

Schmidt and Burstein (1993) stress one general constraint of international comparative achievement studies in education. Cultural traditions and specific context within systems prevent researchers from predicting what will happen if successful teacher practices (according to SIMS analysis) from a system will be

adopted by another system. The political context and the curriculum context, but also the individual pedagogical and didactical beliefs of teachers and school principals can be very different between systems. Moreover, such context factors are hard to define and to measure in a LINCAS and even if data on these factors are collected, the factors are hard to change within an education system.

One of the interesting points in discussing predecessors of TIMSS is the question "what could the founders of TIMSS learn after studying the criticism from researchers, policymakers, and other people who tried to make use of the benefits of FIMS and SIMS?" In the next section the benefits and limitations of FIMS and SIMS are summarized.

2.4 WHAT COULD TIMSS LEARN FROM ITS PREDECESSORS?

From the reviews of FIMS and SIMS a picture can be drawn of the benefits and limitations of these two international comparative studies on mathematics achievement from which TIMSS could profit.
Considering the development of the *goals and functions* of IEA studies in mathematics since 1960s, it can be stated that the goals have become more and more ambitious. It is clear that the general goal of FIMS and SIMS was to measure achievement in mathematics, cross-nationally. The most important result of FIMS was a ranked list of participating countries. The description of differences between countries in mean achievement scores was dominant. Attempts to understand the differences were hardly made. FIMS was primarily an investigation to check the feasibility of studying schools and schooling cross-nationally.
SIMS was designed in a more sophisticated way than FIMS. Nevertheless, understanding differences in mathematics achievement across education systems in terms of differences in background variables was hardly possible in SIMS. The overall goal of SIMS focused specifically on educational processes in classrooms around the world. In SIMS, not only were student features measured, characteristics of curricula, instructional practices, and schools were studied as well. All concepts investigated could be described in a cross-national perspective.
SIMS attempted to accomplish the 'understanding' function by relating instructional practices and student features to the output factor (level of mathematics

achievement) in each country. Having context data available from different curriculum levels (the intended, implemented, and attained curriculum) researchers were able to describe these data sets. The data sets from both a cross-sectional (status achievement test) and a longitudinal design (pretest-posttest) found few explanations. Student factors such as 'attitude towards school' and 'homework effort' were found to have explanatory power for differences in achievement within countries. Yet, in cross-country comparisons the explanatory power of teacher and classroom factors is of more interest to potential users of the study results. Relational analysis on SIMS data sets, including student, teacher, and school data did not reveal meaningful teacher or school factors (Robitaille & Garden, 1989).

Several reasons can be posed for this failure from which TIMSS could learn. First, IEA studies have become more complicated because the number of participating education systems from different continents have been increased since the 1960s. In FIMS, 12 countries from four continents (Europe, Northern America, Australia, and Asia) participated and in SIMS, the number of participating countries was 20 from four continents (Europe, Northern America, Asia and Africa). Studies with so many different countries have to take into account many geographical and cultural differences, while the international research questions, design, and instrumentation were uniform for all countries. This is a tough challenge, and it requires that there be a strong conceptual foundation for the research questions and the instruments. A strong conceptual foundation includes a basis to select concepts that potentially influence student achievement within and across nations.

Second, the development process of the background instruments (written questionnaires) can be seen as a limitation of SIMS. In SIMS, the process of test development was much more sophisticated than the process of questionnaire construction. In SIMS, the three curriculum level framework was developed. Nevertheless, the concepts' definition and data about their international reliability and validity are not very well documented.

Robitaille and Garden (1989) claimed in their reflection on SIMS that system features limit international comparisons. The intrinsic difference among participating systems with regard to their system characteristics (for example, the existence of a national curriculum, which some countries have but others do not) might be a factor too big to overcome in future studies.

With regard to the analysis of SIMS data, nowadays the application of multilevel techniques can be recommended, as the collected data had a hierarchical structure. Hierarchical linear models can be estimated by means of statistical techniques. Again, the conceptual foundation of the study needs to be clear (which was not completely the case in SIMS) and is crucial to achieving meaningful results.

The goals and functions predefined for the Third International Mathematics and Science Study (TIMSS) were even more ambitious than the ones of SIMS.
In the next section the design of TIMSS is described in terms of the basic components from the general study framework.

2.5 GOALS AND DESIGN OF TIMSS

In this thesis, the benefits and limitations of TIMSS are central. In this section, three components of the general study framework are described for TIMSS: the general goals and functions, the general international research questions, and the design and instrumentation. The other components of the study framework of TIMSS are elaborated in the next chapters. The planning phase of TIMSS started in 1991. The data collection took place in 1994-1995. The number of 45 participating education systems was higher than ever.

General goals and functions

The goals of SIMS were much more ambitious than the goals of FIMS. Subsequently, IEA formulated for TIMSS goals that were more ambitious than the SIMS' goals. The ultimate goal of TIMSS was *"to isolate the factors directly relating to student learning that can be manipulated through policy changes in, for example, curricular emphasis, allocation of resources, or instructional practices"* (Martin & Kelly, 1996, pp.1-2/1-3).
The first function of TIMSS reflects the 'description' function mentioned in 1.2. The data collected in TIMSS was described in terms of frequency features such as country means and standard error. For instance, the mathematics achievement results were reported in country tables (Beaton, Mullis, et al., 1996). Background data collected at student, classroom, and school level was reported in the form of country tables as well. Most tables can be studied by policymakers to compare the results of their own country with the results of other countries.

These comparisons were a starting point for attempts to understand differences and similarities between one's own country and other countries which refers to function 4 'understanding of reasons for observed differences' (see 1.2). The TIMSS study directors wanted to uncover similarities and differences between and within education systems. Once similarities and differences in educational factors across nations had been identified, researchers could attempt to reveal possible patterns of relationships between these factors. Patterns which are based on postulated theoretical models can be used to understand the identified differences and similarities.

The ambitions of TIMSS are also reflected in some innovative aspects. Compared to its predecessors, the first innovation TIMSS offered was the dependent variable of the study. Two subjects, mathematics *and science*, instead of one were the main objectives in the study. Both subjects have been assessed by means of one 'paper and pencil' achievement test to the same student sample. For the first time in IEA studies, a performance test was administered in TIMSS (Vos, 2002; Bos, Kuiper & Plomp, 2001; Harmon, Smith, et al., 1997). A sub sample of the schools and grade 8 students participating in the TIMSS achievement test was selected to complete an international mathematics and science performance test. Performance assessment refers to the use of practical tasks involving instruments and equipment. Previously, science was studied separately in the 1970s in IEA's FISS (Comber & Keeves, 1973) and in the 1980s in IEA's SISS (Postlethwaite & Wiley, 1992).
A second innovative aspect of TIMSS – from the design perspective of LINCAS – was that two parallel projects were conducted more or less at the same time as TIMSS, in some of the participating countries. First, the curriculum and textbook analysis was conducted in many of the countries (Schmidt, McKnight, et al., 1996). Second, the TIMSS Video Study was carried out in three of the TIMSS countries: the United States, Germany, and Japan (Stigler, Gonzales et al., 1999). The latter was a national option of the participation of the United States in TIMSS.
The inclusion of a third population (9-year olds) was the third innovation of TIMSS.

General international research questions

The four general international research questions of TIMSS were formulated by Robitaille and Maxwell (1993). The questions refer to the three curriculum levels of the conceptual framework developed in the SIMS study (see 2.3) which was adopted by TIMSS:

1. Intended curriculum: How do countries vary in the intended goals for mathematics and science; and what characteristics of educational systems, schools, and students influence the development of those goals?
2. Implemented curriculum: What opportunities are provided for students to learn mathematics and science; how do instructional practices in mathematics and science vary among nations; and what factors influence these variations?
3. Attained curriculum: What mathematics and science concepts, processes, and attitudes have students learned; and what factors are linked to students' opportunity to learn?
4. Relationships between curricula and social and educational contexts: How are the intended, the implemented, and the attained curriculum related with respect to the contexts of education, the arrangements for teaching and learning, and the outcomes of the educational process?

The first three questions can be addressed by *describing* scores on the different variables in each country. The description function is served by answering these questions. Addressing the fourth question can result in information needed to *understand* similarities and differences across countries with regard to the various aspects located at the three curriculum levels.

Design and instrumentation

The design of the study can be characterized as cross-sectional. From the experiences with the two design versions in SIMS, a choice for the pretest-posttest design would have made more sense and was deliberated. However, the financial costs of this design were too great for most of the education systems interested in participating in TIMSS.

Three student populations were investigated:

- population 1 consisted of the two adjacent grades with the majority of the 9-years-old students (grade 3 and 4 in most countries);
- population 2 consisting of the two adjacent grades with the majority of the 13-years-old students (grade 7 and 8 in most countries);
- population 3 consisting of students in the last year of secondary school, regardless of the type of program in which they were enrolled.

The data collection in all populations was carried out in the eighth month of the school year 1994-1995. In countries in the Southern Hemisphere the eighth month was October 1994 and in the Northern Hemisphere it was April 1995.
Data were collected from students, teachers, and school principals. Data collected at country level were regarded as necessary to accomplish the goals of the study. In order to collect data at all levels, a two-stage sample was drawn: a sample of schools out of the population of schools and a sample of one intact classroom of each grade (lower and upper) in each participating school. As part of the second stage, the teachers of each of the TIMSS subjects (mathematics and science) from the classroom involved in the study were selected.

In TIMSS, an international mathematics and science achievement test and a set of background questionnaires were developed. In the development of the instruments, the TIMSS study directors tried to meet the criticism SIMS results received (Howson, 1999).
As opposed to the SIMS test, open-ended items were included in addition to multiple-choice items. The written test was supplemented by a performance assessment test mentioned above. This substantial addition to the SIMS test was developed by the International Assessment of Educational Progress (IAEP; Foxman, 1992) and the International Study Center of TIMSS (Harmon, Smith, et al., 1997) in the period between the end of SIMS and the start of TIMSS.
The international written mathematics test consisted of 150 mathematics items, distributed across sub-scales called (1) Fractions and number sense, (2) Geometry, (3) Algebra, (4) Data representation, analysis & probability, (5) Measurement, and (6) Proportionality. A test-rotation-design was used to collect the student test data (Adams & Gonzales, 1996). This consisted in giving each student a core set of 6 math items and a rotation consisting of a certain number of items belonging to every sub-test.

In addition, background questionnaires for students, teachers and school principals were developed to collect data about factors that potentially influence student achievement in mathematics and science. Taskforces were asked to develop conceptual frameworks for the different educational levels factors are located. From experiences in SIMS it was concluded that a conceptual framework was needed to select relevant factors. In chapter 3, the conceptual framework for TIMSS is discussed.

The students filled out a student background questionnaire with questions about themselves and their perception of and attitude towards school matters. The other background questionnaires contained many questions and were filled out by mathematics and science teachers of the tested classes (teacher questionnaires) and by school principals (school questionnaire). Questions asked to the teachers refer to their background and to aspects of their instructional practice such as homework features, grouping procedures, and use of students' evaluation results. Examples of topics asked to school principals are 'school size', and 'time spent on educational leadership tasks'.
At country level, data was collected by means of participation questionnaires. For the purpose of a participation questionnaire, the National TIMSS Center in each country collected data of population characteristics, teacher training figures, and economical indicators. National information on contexts of mathematics and science education was published in an encyclopedia (Robitaille, 1997).

In Figure 2-2 the components of the general study framework presented in the first section of this chapter is completed for TIMSS. The particular issues of TIMSS that are studied in this thesis are written in bold.

The components of TIMSS are studied in the remaining chapters of this thesis, starting with the conceptual foundation in chapter 3. In chapter 4, results of secondary data analyses on mathematics data (grade 8) from three education systems – the Netherlands, Belgium Flanders and Germany – are presented. Finally, the utility of these TIMSS results and the benefits and limitations of all TIMSS components are reflected upon from the perspective of the TIMSS' goals (chapter 5).

The TIMSS study was repeated in 1999 with respect to the upper grade of population 2. The main aim of this repeat study was to measure change in mathematics and science achievement between 1995 and 1999. The same instruments were used as in 1995 (Mullis, Martin et al, 2000; Martin, Mullis et al., 2000). This repeat study served function 3 described in 1.2: 'monitoring the quality of education.'
In 1999, the IEA formulated plans for a long-term, 4-year periodic trend study on mathematics and science in the upper grade of population 1 and 2. As a result, from 2001 on, 'TIMSS' stands for *Trends* in International Mathematics and Science Study.

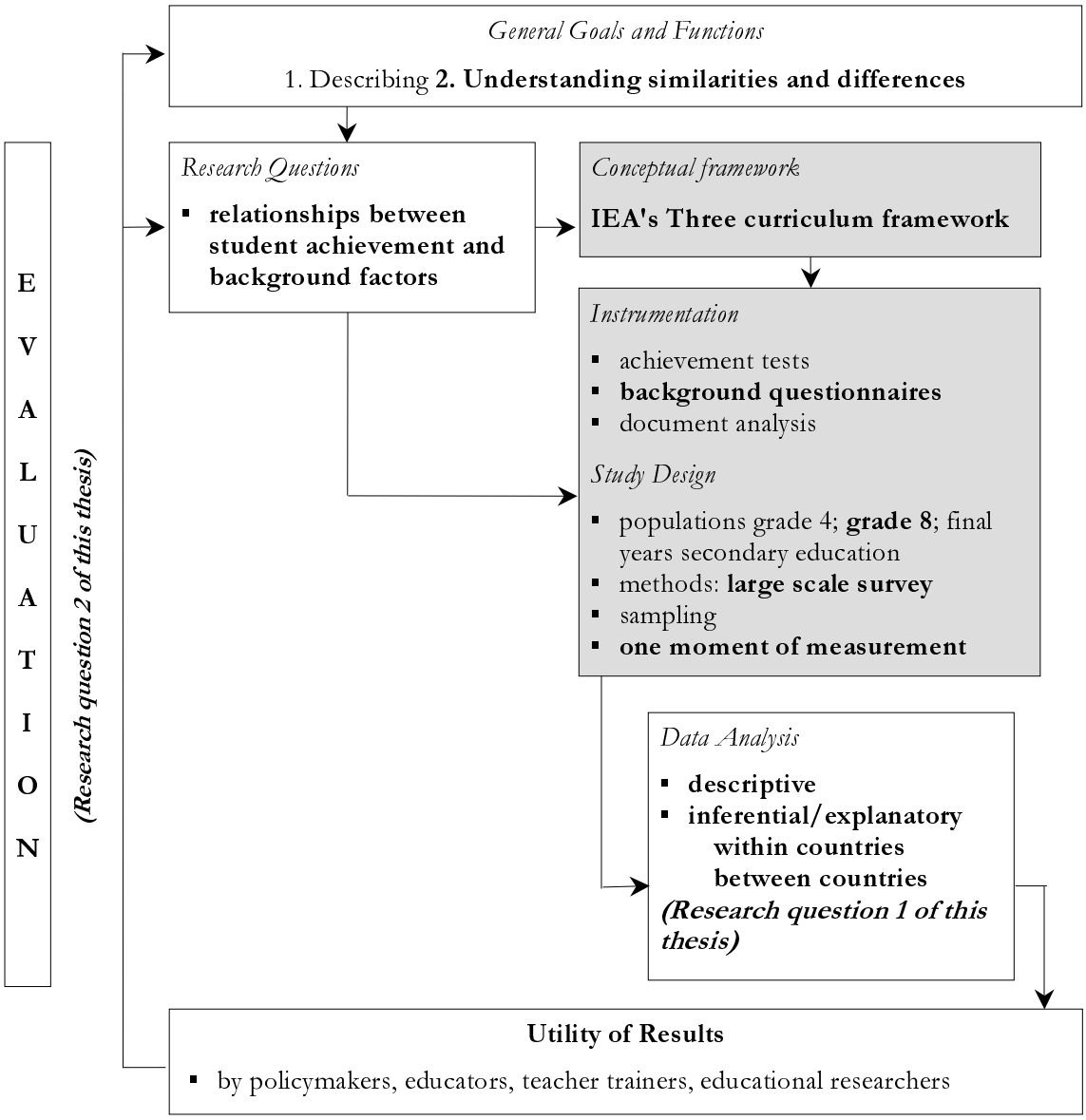


Figure 2-2
The general study framework of the case of IEA's TIMSS (the particular issues studied in this thesis are written in bold)

Every four years, TIMSS will be organized to collect trend data. The first data collection in the two trend populations, grade 4 (primary education) and grade 8 (secondary education), will take place in November 2002 (Southern Hemisphere) and May 2003 (Northern Hemisphere). Because of these changes, the first dat a collection of TIMSS in 1995 is referred to as TIMSS -1995 and the repeat in 1999 as TIMSS-1999. In this thesis, the upper grade (grade 8) of population 2 from TIMSS -1995 is central. For the sake of convenience, this study is referred to as TIMSS.

CHAPTER 3

DEVELOPMENT OF AN ORGANIZING CONCEPTUAL FRAMEWORK

One of the important predefined functions of TIMSS is 'understanding differences in students' achievement results across education systems.' The first research question of this thesis concerns the comparison of background factors measured in TIMSS that can explain differences in the overall scores on the international mathematics test across three education systems: the Netherlands, Belgium Flanders, and Germany (3.1).
To address this question, a conceptual framework is needed in which the factors that might affect mathematics achievement are classified. The conceptual framework for TIMSS was reviewed according to the four criteria formulated in 3.2. The results of this review indicated strengths and shortcomings of the TIMSS conceptual framework. The two dimensions on which the framework is based – educational levels and curricular contents (including context and antecedents) – are seen as strengths.
However, factors classified within the TIMSS framework are not very well defined and their theoretical and empirical basis is not clear. Therefore, potentially effectiveness-enhancing factors were derived from review studies on instructional and school effectiveness and included in the basic TIMSS framework (3.4). As a result, an organizing conceptual framework was formulated to guide the exploration of potentially effective factors represented in the TIMSS background questionnaires (3.5).

3.1 RESEARCH QUESTION I AND RELATED QUESTIONS

In chapter 1, five functions of large-scale international comparative achievement studies in education (LINCAS) were described. One out of these five functions is part of the scope of this thesis: understanding cross-national similarities and differences in background factors related to student achievement measured in TIMSS.
In chapter 2, the utility of results of the predecessors of TIMSS – FIMS and SIMS - was discussed in the light of this function. The two principal parts of the instrumentation used to address the goal of IEA studies are the international achievement tests and the background questionnaires. The reviews in chapter 2 concluded that for the successive IEA studies on mathematics achievement – despite the improvement of the international achievement test - it would be very difficult to fulfill the ambition of understanding differences across education systems.
The ambitions of TIMSS though, were even greater than those of its predecessors. TIMSS' most ambitious research question – aside from the research questions primarily aimed at international comparative descriptive information – was formulated from the perspective of the 'understanding' function (see also section 2.5): *"How are the intended, the implemented, and the attained curriculum related with respect to the contexts of education, the arrangements for teaching and learning, and the outcomes of the educational process?"* (Robitaille & Maxwell, 1996, p.42).
The core components of IEA's three curriculum level conceptual framework can be seen in this question (see 3.3). The authors assumed that various factors at the three curriculum levels were studied in TIMSS. The ultimate ambition of TIMSS was *"to allow researchers to apply theories about contextual factors that contribute to achievement simultaneously to systems of diverse contexts"* (Robitaille & Maxwell, 1996, p.42).
As was noted in chapter 1, the TIMSS achievement test will not be analyzed in this thesis. In this thesis, the TIMSS background questionnaires are studied in greater depth to determine the extent to which analysis of the data provides information regarding differences in achievement results across nations. The TIMSS data sets of the Netherlands and two of its neighbor education systems, Belgium Flanders and Germany, were compared. This investigation was conducted to address the first research question of this thesis formulated in chapter 1:

I. *To what extent can variability in the overall TIMSS mathematics test scores for grade 8 within the Netherlands, Belgium Flanders, and Germany be explained by variability in the scores on variables at student and classroom/school level and to what extent are these outcomes generalizable across these three European educational systems?*

The categorization, selection, and measurement of crucial background factors are important steps in understanding country differences in achievement results. Crucial background factors are factors that potentially influence achievement in mathematics and are changeable by policymakers.

Related questions to research question I are:

Ia. Which factors measured in TIMSS at student and classroom/school level are associated with mathematics achievement in lower secondary education in the Netherlands, Belgium Flanders, and Germany?

Ib. What can be learned from similarities and differences across the three education systems under review, with respect to outcomes of student and classroom/ school factors that were predictors of achievement in mathematics?

The answers to questions Ia and Ib serve the 'understanding' function formulated in chapter 1. Particularly, students' attitudes and students' background factors and factors at classroom and school level are compared. In order to address question Ib, the interrelationships between student, classroom, and school factors are explored further in relation to mathematics achievement.

Education systems can differ on a number of factors that can partially explain the differences in the mean achievement scores across systems. Consider, for instance, a possible difference between education systems A and B in the way homework is treated during a regular mathematics lesson. Here, the example is simplified but suppose, in education system A most teachers review completed homework at the beginning of most lessons and in system B homework is reviewed rarely by teachers. Further, the mean achievement in mathematics education system A is significantly higher than in system B and treatment of homework was a predictor of achievement in each system (directed positively in system A and negatively in system B). These differences can be a starting point for policymakers and educational practitioners in education system B to change education in their system at the classroom level. Of course, in reality the situation will be more complicated, as more background factors might be involved and the influence of a factor on achievement might not be direct.

In large-scale international comparative studies, comparisons between education systems rely mainly on means and variances of distributions of scale scores both

within and across systems. In the example, the mean score 'treatment of homework' of each system can be taken as the basis for the comparison across systems with respect to the relationship between achievement scores and the treatment of homework.
Similarly, in analyzing the TIMSS data sets it is necessary that scores on factors that potentially affect mathematics achievement are available. Potentially effective factors need first to be identified. Therefore, the conceptual foundation of the TIMSS background questionnaires is reviewed. The review of the conceptual framework for TIMSS provides insight in (1) the appropriateness of the framework to address research question I, and (2) the extent to which the operationalization of the framework can be used as a guide to find potentially effective factors in mathematics achievement across education systems. Preferably, the correspondence between the conceptual framework for TIMSS and the TIMSS background questionnaires contains concrete operationalizations of key factors at both student, classroom, and school level.
The final results of this review are used as a guide for the comparisons of the data sets of the three education systems. The results of the comparative explorations provide answers to research question I and are reported in chapter 4.

3.2 CRITERIA FOR AN APPROPRIATE CONCEPTUAL FRAMEWORK

In the LINCAS under review, the focus is on predictors of student achievement in mathematics. Usually, achievement is measured by means of curriculum-driven achievement tests. The test results indicate the extent to which students within the participating countries attained the curriculum that was taught. To be able to understand differences in achievement within and across countries, it is necessary to have information available on groups of background factors that are regarded as potentially affecting student achievement.
The identification of potentially effective background factors in TIMSS should be guided by the conceptual framework for this study. Four criteria were predefined to judge the appropriateness of the conceptual framework for TIMSS in this respect. Basically, the criteria are derived from the technical standards for IEA studies formulated by Martin, Rust and Adams (1999). However, the technical standards for IEA studies focus on achievement tests. Standards for developing conceptual models and background questionnaires are described briefly by Martin et al. (1999).

The first criterion for evaluating the framework deals with the selection process of important factors given the particular research questions of this study.

1. The conceptual framework can be used as a tool for classification or categorization of the key factors selected as being crucial to explain variability in students' achievement.

Factors can be classified at different dimensions of a framework. Levels of education (i.e., the system (country) level, and the school, classroom, and student levels) are considered to be one dimension for a conceptual framework. Other examples of dimensions are input-process-output and content levels of a curriculum (including the levels of intended, implemented and attained curriculum).

If the framework meets the first criterion, it can be studied further. The second criterion to judge the appropriateness of the conceptual framework refers to the extent to which it depicts the clusters of factors and their interrelationships on which TIMSS is focused (Munck, 1979). The classification of factors should include concrete definitions of the factors, so that the factors can be operationalized It must also be possible that potential relationships between clusters of these factors can be derived from the conceptual framework. The second and third criterion are formulated as follows:

2. The operationalization of the conceptual framework includes concrete definitions of the key factors.
3. The framework can be used to formulate clear assumptions about the relationships between the categorized clusters of key factors.

If the three criteria are met, the study framework can be used as a basis for the development of a measurement model for achievement studies (Munck, 1979).

In addition to the three general criteria, one extra criterion is formulated because of the *international comparative* purposes of TIMSS. As stated, the comparison between education systems aims at understanding cross-national differences in student achievement for which background data is needed. Thus, the international comparability of the background factors is important. The conceptual framework should do justice to the differences and similarities of countries participating in a study. Therefore the framework should be based on empirical studies conducted in different countries located on several continents. Also, the studies on which the framework for TIMSS is based should have a common theme associated with 'factors influencing student learning of a core subject such as mathematics.'

The fourth criterion to judge the appropriateness of a conceptual framework for TIMSS is:

4. The theoretical basis of the framework is aimed at factors influencing student learning of mathematics, and founded in a substantial body of empirical studies conducted in countries around the world.

In the next section, the two conceptual frameworks for the TIMSS study found in the literature are discussed against the background of the four criteria formulated above.

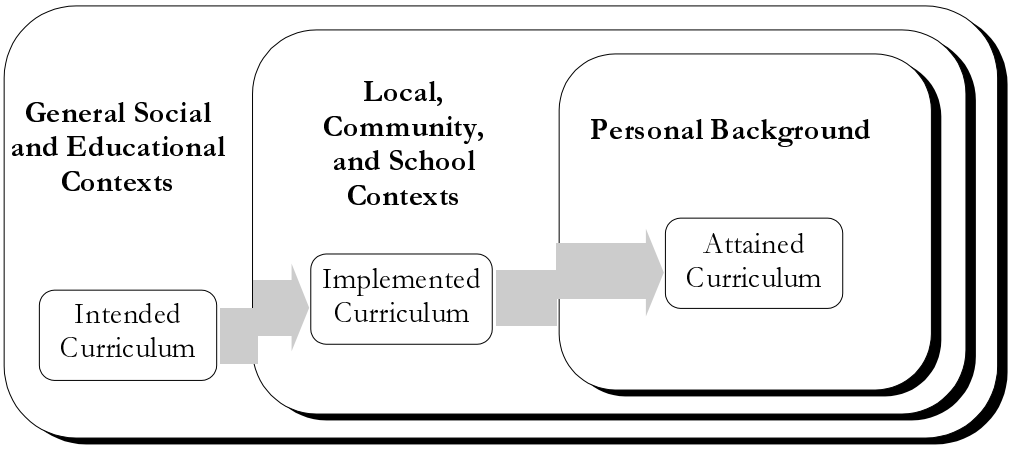
3.3 APPROPRIATENESS OF TIMSS CONCEPTUAL FRAMEWORKS

In the literature, two frameworks were found labeled as 'TIMSS conceptual framework' and both of them were published in 1996. The first publication in which a conceptual framework for TIMSS was described is the TIMSS monograph 2 (Robitaille & Maxwell, 1996) and the framework is called 'the three curriculum level framework' The second publication is a contribution from Schmidt and Cogan (1996) to the first Technical Volume of TIMSS (Martin & Kelly, 1996). The framework they developed is called 'the educational experience opportunity framework.' Both frameworks are described below.

The three curriculum level framework for TIMSS

IEA's FIMS resulted mainly in a mathematics achievement ranked list of the 12 participating countries (see 2.2). The FIMS findings could not be used to draw a picture of schooling in the different countries. One of the critical issues of FIMS was the missing link between the achievement test and the curriculum in the countries. The next IEA studies attempted to fill in this omission by developing a conceptual framework in which 'curriculum' is one of the most important factors.

The conceptual framework of the second IEA study on mathematics (Second International Mathematics Study - SIMS, see 2.3) was curriculum-based and incorporated three curriculum levels (Travers & Westbury, 1989). The intended, the implemented, and the attained curricula were adopted for TIMSS as the best means of discussing different views of curricula and addressing the contexts of education (Robitaille & Maxwell, 1996). The TIMSS framework described by Robitaille and Maxwell (1996) is shown in Figure 3-1.



Source: Robitaille & Maxwell, 1996, p.37

Figure 3-1
Conceptual framework for TIMSS: the three curriculum framework

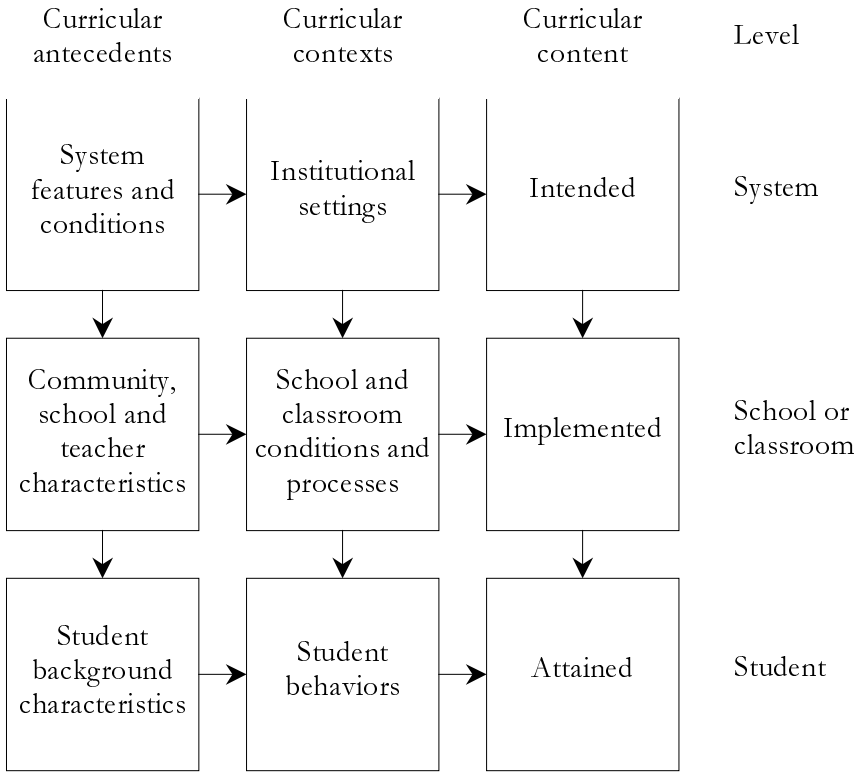
This conceptual framework for TIMSS was based not only on the three curriculum framework developed in SIMS, but also on the work of Shavelson on educational indicators (Shavelson, McDonnell, Oakes & Carey, 1987). In the model of Shavelson et al. the complex interactions of teaching and learning are conceptualized by means of the identification of input, process, and output factors. In TIMSS, the input-process-output dimensions are not literally used. TIMSS speaks of contexts and institutional arrangements instead of processes and inputs. The contents of the three curriculum levels are regarded as outputs.
The framework shown in Figure 3-1 is very simple. It shows only the three levels of curriculum content and for each level it shows one general label of context factors (general social and educational contexts; local, community and school contexts; and personal background). The description of this framework is very limited. Factors are mentioned, but not defined. Furthermore, the boundaries between the different components within the framework are not clear, because of the general labeling of the contextual factors. According to the developers of this framework, this might not be necessary (Robitaille & Maxwell, 1996). For them the important point is that (p.41) *"the variables of three different kinds of content need to be considered in the light of three different societal contexts. The content and institutional arrangements of the intended, the implemented, and the attained curricula, together with features of the society at large, the local community, and the student's own context constitute an appropriate description of the educational environment "*. However, because of its simplicity the framework does not meet any

of the criteria formulated in 3.2. The framework in Figure 3-1 was based on the IEA research framework developed in SIMS. The latter has been described in detail by Travers & Westbury, 1989. Given the detail, the IEA research framework might be more appropriate than the TIMSS framework to address research question I. The IEA research framework is presented in Figure 3-2 and a review follows.

Main characteristics

The three curriculum levels of the TIMSS conceptual framework in Figure 3-1 are central in the IEA research framework. The three levels can be described in terms of three broad sets of curricular variables: content, contexts, and antecedents. The curricular content at each of the three levels is supposed to be influenced by the context in which it occurs and the contexts themselves are supposed to be determined by a number of antecedent conditions (Travers & Westbury, 1989).

In Figure 3-2 the three curricular dimensions can be seen from left to right: the curricular antecedents, the curricular contexts, and the three curricular content levels.



Source: Travers & Westbury, 1989, p.8

Figure 3-2

An IEA research framework

At the far right side of the model the educational levels at which data are collected are included: the system or country level, the school or classroom level (combined), and the student level. The description of the framework starts with the 'contents' part, as this part is what is common to both frameworks (see Figure 3-1 and 3-2).

Curricular content

The set of curricular variables consists of the intended, implemented, and attained content. The curriculum of core subjects such as mathematics and science in a school system takes on different embodiments at each of the three content levels.

The *intended* curriculum is defined as 'a statement of a society's goals for teaching and learning' and can be regarded as the planned curriculum for a particular subject (Travers & Westbury, 1989). At system or country level it is important to know to what extent the intended curriculum is valid for the entire country. In countries with a lower degree of centralization the responsibility for curriculum design, development, and initiative is more in the hands of states, regions, or local districts. The intended curriculum for a state or region is the one that counts only for the schools within that state or region.

National authorities transmit the intended curriculum to the schools. The intended curriculum is reflected in curriculum documents such as goal statements, prescribed textbooks, syllabi and other educational resources (e.g., information communication technology tools and laboratory equipment), and examinations (Robitaille & Maxwell, 1996). The contents of the intended curriculum can be studied through an analysis of its curriculum guides, official policy statements, and regulations (e.g., identifications of mandatory and optional courses, and the definition of the diversity of school programs).

The second level of curricular content concerns the *implemented* curriculum and is regarded as the intended curriculum as interpreted and translated by teachers according to their experience and beliefs for particular classes. It is that which is actually taught. Obviously, the classroom is central to the educational process. In the classroom, students are introduced to the subjects and the teacher has the responsibility for both transmitting the knowledge to students and enabling them to acquire appropriate skills. One of the most important factors in international comparative achievement studies is 'opportunity-to-learn' which measures what is actually taught (de Haan, 1992). The opportunity-to-learn concept is defined as the

extent to which the subject contents – needed to complete each of the items included in an (international) achievement test – has been covered (Travers & Westbury, 1989). The implemented curriculum can be restricted to the school year by the end of which the international achievement test is administered, but it can also refer to the school years prior to the year of testing.

The *attained* curriculum can be described in terms of what students have learned, including their attitudes towards school and towards the particular subject(s) under investigation. The knowledge and skills of a certain subject and attitudes towards that subject, are usually the dependent variables of primary interest in international comparative studies.
The inclusion of the three curricular content levels allows researchers to analyze possible explanations for the student outcomes in the participating countries. It is assumed that to a certain extent, differences both in curricular intentions and in what is actually taught in the classroom can account for differences in student results within and across nations.

Curricular contexts and antecedents
Travers and Westbury (1989) added a *contextual* dimension to the content framework because they wanted to take into account not only curricular content variables but also contextual factors in finding explanations for differences in student outcomes. At each curricular content level of the model a set of context variables can be classified.
The contextual variables at the level of the intended curriculum are called institutional settings. An example of such a variable is the existence of central examinations (independent from individual schools in an education system). At the second content level of the model – implemented curriculum – school and classroom conditions and processes determine the curricular context. At the level of the context of the implemented curriculum, features of the instructional practices, such as the way teachers structure their lessons and the type of assessment they apply, can be classified.
The curricular context of the third level of the curriculum model consists of variables defined as 'student behaviors.' Examples of student behavior variables include perceived limitations related to disruptive student behavior and students' perceptions of the class climate.

The *antecedent* dimension of the three curriculum model contains background factors. The antecedents can be seen as non-changeable factors (system features and background characteristics of schools, teachers, and students) which can affect contextual factors, which are potentially changeable.
Background factors classified at the intended curriculum level include system features such as the wealth of the society, which may affect the retention of students within education (Travers & Westbury, 1989).
At the implemented curriculum level, community characteristics (e.g., degree of urbanization of the area where the school is located), school characteristics (e.g., school size and percentage of full-time teachers) and teacher characteristics (e.g., professional qualifications and experience, age, and gender) are regarded as antecedents.
At the third level of the model – the attained curriculum level – student background factors are classified as antecedents. Travers & Westbury (1989, p.8) described this set of variables as *"the characteristics which the students bring to the class"*. Student background characteristics, which can be classified at the level of the attained curriculum and could not be influenced by schools, are the educational and socio-economic background of the student's family. These factors are examples of 'given' antecedents.

Appropriateness
Considering IEA's three curriculum level framework presented in Figure 3-2 and the presentation of the factors, which can be classified at the different levels, some remarks can be made with regard to the appropriateness of the framework to address research question I.
Ideally, it should be possible to assign factors to one of the clusters of factors which are included in the framework. The factors should be defined concretely and it should be possible to derive the relationships between the clusters of factors from the framework (see the three criteria in 3.2).

Criterion 1. Tool for the classification of factors.
In principle, the framework can be used to classify many important factors needed to answer international comparative research questions in a descriptive way. Nevertheless, the boundaries between the different clusters within the framework are not very strict. For example, factors such as classroom climate, belonging to classroom conditions located at the implemented level, could also be assigned to student behaviors located at the attained curriculum.

Criterion 2. Concrete definitions of key factors.
Both concrete definitions of the factors and reliable and valid operationalizations of the factors in the form of scales are needed. However, no public documents could be found in which the factors classified at each level of the three curriculum level framework, were defined concretely. Nor could documents be found with complete descriptions of the correspondence between the classified factors and their operationalizations in the TIMSS background instruments.

Criterion 3. Basis for assumptions about relationships between clusters of key factors.
It seems that the distinction between the three curricular content levels is very useful in international comparative studies. Differences across countries can be found and described at each of the three levels. The three curricular content levels can be described both separately and in relation to each other as was done in IEA's SIMS (Travers & Westbury, 1989). The three curriculum level framework seems appropriate to formulate assumptions about relationships between the clusters of factors. In SIMS, the description of the outcomes per curricular content in each participating country was followed by the description of the relationship between the intended and the implemented curriculum. For instance, countries were ranked on the basis of the correspondence of the described level of the intended and the implemented curriculum (Travers & Westbury, 1989).
The relationship between the attained curriculum level and either of the other two curricular contents was not analyzed in SIMS. Results of the attained curriculum level were presented in the form of patterns of attitudes and achievement. The description of these patterns included information on backgrounds of schools, teachers, and students without explicitly paying attention to the intended and the implemented curriculum (Robitaille & Garden, 1989).

Criterion 4. Substantial theoretical basis, founded in empirical studies associated with one theme and conducted in education systems around the world.
Criterion 4 is not perfectly met by IEA's research framework. The theoretical and empirical foundation of the clusters of factors is not clear from the description of the framework in the literature (Travers & Westbury, 1989). Few empirical results are mentioned as bases for inclusion of clusters or individual factors in the framework.

The Educational Experience Opportunity conceptual framework for TIMSS

The second conceptual framework for TIMSS found in the literature was also elaborated from the three curriculum level model of SIMS. The elaboration resulted in the TIMSS conceptual framework called 'The educational experience opportunity' (Schmidt & Cogan, 1996). In TIMSS, the framework was used as a starting point and as a guide in developing the background (or context) questionnaires for students, teachers, and school principals. The intention was *"to assess, through context questionnaires, the factors at the system, school, teacher, and student level that are likely to influence students' learning of mathematics and the sciences"* (Schmidt & Cogan, 1996, p. 5-1).

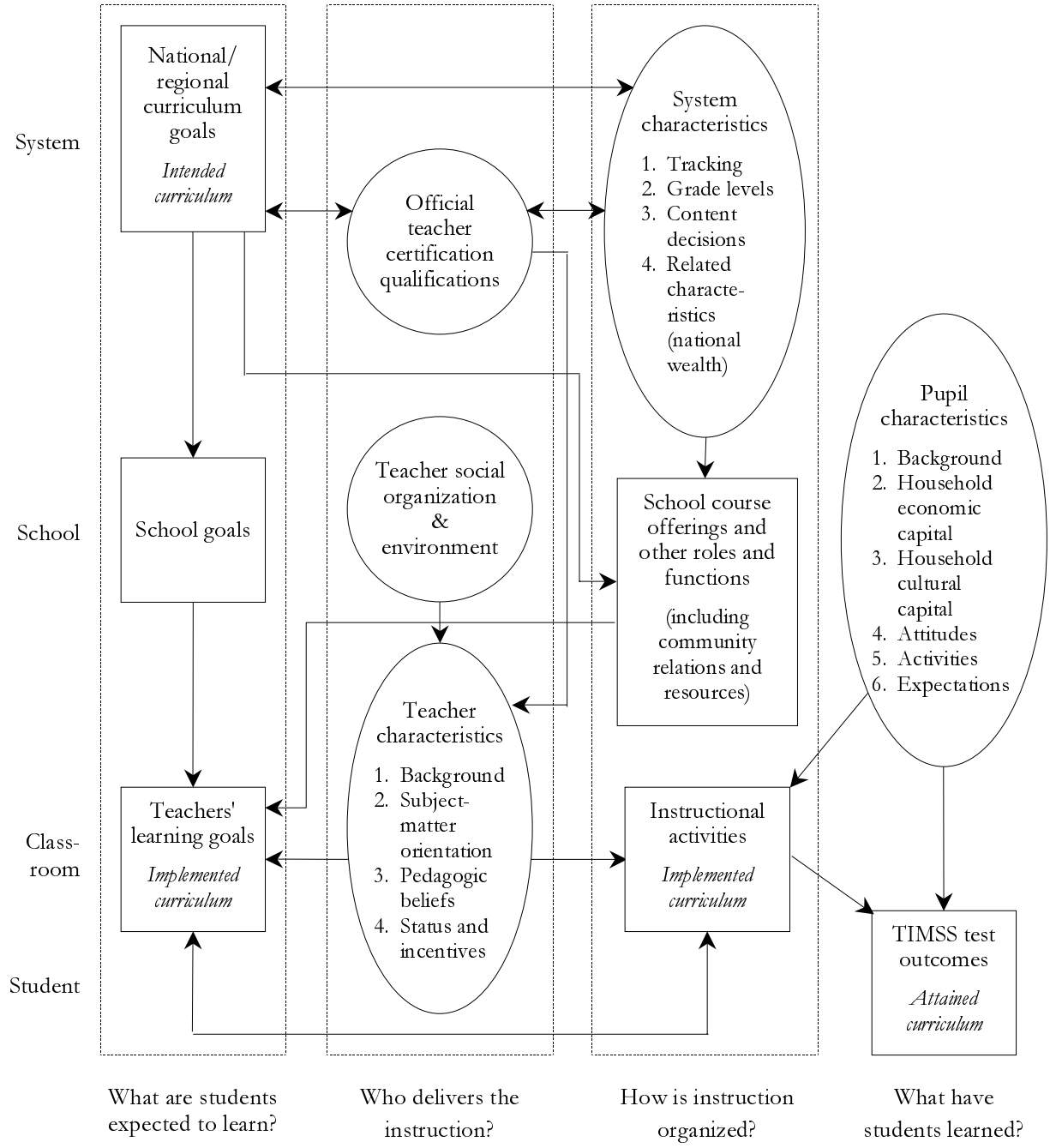
Main characteristics

In the developmental process of 'the educational experience opportunity' framework, three separate frameworks – one framework for school factors, one for classroom factors, and one for student factors- were first developed (Schmidt & Cogan, 1996). Each framework consisted of a list of factors drawn from the literature. The relationships between clusters of factors were included in the separate models. In the final conceptual framework, as shown in Figure 3-3, the three frameworks were integrated.

This TIMSS framework includes four levels of education – system, school, classroom and student – and incorporates the three curricular content levels. The latter are more or less used as the second dimension of the framework. The four levels of education are located horizontally and the curricular levels vertically. The three curricular levels are labeled by means of four questions:

- What are students expected to learn? (Intended curriculum).
- Who delivers the instruction? (Implemented curriculum).
- How is instruction organized? (Implemented curriculum).
- What have students learned? (Attained curriculum).

The first three questions refer to factors at all educational levels except student level. Only the fourth question refers to factors at student level: test outcomes and pupil characteristics. As can be seen in Figure 3-3, the level of the implemented curriculum is phrased by means of two questions referring to factors related to the teacher, and to instructional practices and their conditional components at school and system level.



Source: Schmidt & Cogan, 1996, p.5-8

Figure 3-3
TIMSS conceptual framework: the educational experience opportunity

The hypothesized interrelationships between several clusters of factors are demonstrated by arrows. Arrows are drawn between clusters of factors at the same

educational level as well as between clusters at different educational levels. For example, at classroom level, teacher characteristics are supposed to be related to teachers' learning goals and to instructional activities. The links assumed between national curriculum goals, schools goals, and teachers' learning goals are examples of links between clusters of factors located at different educational levels. Other relationships are assumed to be mutual (e.g., 'intended curriculum' and 'official teacher certification qualifications'). The implications of reciprocal relationships were not made explicit by the framework's developers.

Appropriateness

The appropriateness of the 'educational experience opportunity' framework can be judged by applying the four criteria formulated in 3.2.

Criterion 1. Tool for the classification of factors.

Various clusters of factors, which should be investigated in TIMSS, can be derived from the framework. Within each cluster a number of factors can be distinguished. The teacher and student characteristics are included explicitly in the framework.

The classification of some of the factors however, is somewhat ambiguous. For example, 'implemented curriculum' is located in two clusters at the classroom level ('teachers' learning goals' and 'instructional activities'), which are assumed to be interrelated. In the framework 'goals' are not distinguished from 'activities.' Both factors together are regarded as the implemented curriculum and provide the conceptual and practical link between various aspects of the intended curriculum and what is attained by students (Schmidt, 1993). The lack of conceptual distinction between these two factors, especially given that the two clusters are located in different columns in the framework, makes the framework complex and less appropriate as a tool for classification of key factors.

If implemented curriculum would have been separated in two levels the distinction would have been clearer. The two levels are the institutional and instructional level (Klein, 1991). 'Teacher's learning goals' seems more a factor belonging to the institutional curriculum level. The latter is defined as curriculum development at the individual school site (Klein, 1991; Goodlad, 1979). 'Instructional activities' can be located at the instructional level. This level is composed of teachers decision making about curriculum planning.

Criterion 2. Concrete definitions of key factors.
The next step in designing the TIMSS background questionnaires consisted of the formulation of concrete definitions and operationalizations of the factors. In the literature about TIMSS, few documents could be found that provide information about definitions and operationalizations of the factors. In a TIMSS project document prepared by Schmidt (1993), which was not published but only used as a working document, attempts were made to define and to operationalize factors. Factors listed in the document are described in terms of 'data collection,' 'prospective analyses,' and 'indices development.' These categories are required to get an overview of the correspondence between the factors and the data collected by means of the TIMSS background questionnaires. However, not all of the factors mentioned in Figure 3-4 could be found in this document. In addition, the development of indices for each factor was proposed by making references to the items from the TIMSS background questionnaires. Yet, the correspondence between the sets of items in the questionnaires and the factors which were classified in the TIMSS framework could not be determined precisely in the document (Schmidt, 1993). This is a consequence of the fact that the items of the questionnaires included in Schmidt's document were preliminary. The items are not exactly the same as the ones of the final background questionnaires used in TIMSS.
The majority of the factors included in the framework are described and defined in general terms and therefore the framework meets the second criterion only to some degree.

Criterion 3. Basis for assumptions about relationships between clusters of key factors.
Unlike the IEA research framework (see Figure 3-2), in the educational experience opportunity framework, the school and classroom level are no longer combined. From a theoretical perspective it is relevant to separate these two levels. Separation makes conclusions possible about factors located at school level that influence factors at classroom level. Assumptions about such relationships can be better formulated when the distinction between school and classroom level is made clear in the general conceptual framework.
However, the educational experience opportunity framework contains several complicated relationships. In the working document prepared by Schmidt (1993), only some of the arrows between clusters of key factors are described. As stated in

the example under criterion 1, the framework includes an extensive number of (reciprocal) arrows and this makes the framework complicated and difficult to understand.
On the other hand, some arrows are missing. Take for example, the cluster 'teacher social organization and environment,' which has only a link with 'teacher characteristics.' In Schmidt's document (1993), no information is provided about possible relationships between 'teacher social organization & environment' and 'instructional activities.' Such a relationship can be assumed because of their possible links: The way teachers work together with each other in their subject department might influence their instructional activities.
Finally, descriptions of the assumed relationships underlying many of the arrows cannot be found in the accompanying literature. Hence, the framework does not meet criterion 3.

Criterion 4. Substantial theoretical basis, founded in empirical studies associated with one theme and conducted in education systems around the world.
The name of the framework and the introductory text from Schmidt (1993) about the framework refers to one topic: 'learning opportunities.' However, from the little documentation available, it is not made clear whether the studies to which references were made, are aimed at one particular topic associated with the theme 'factors influencing student learning of mathematics.'
The references made in Schmidt's working document come either from different studies conducted in one country or are missing. It seems as though results of separate studies, which each investigated only a small number of factors, were combined. For example, under 'who delivers the instruction?' the cluster 'teacher social organization and environment' was included at the level of the school by referring to a study conducted in the United States in which "*the allocation of teacher time (i.e., the proportion of professional time spent during a school day in planning and teaching mathematics or science), and the amount of cross grade-level teaching was regarded as important*" (Schmidt & Cogan, 1996, p. 5-8). Some of the other clusters were described in the same way by referring to results from separate studies conducted in one country: the United States.
The complicated interrelationships between the clusters of factors (criterion 3) might also be a result of the fact that references were made from different studies

which had no substantial common theoretical basis. The lack of a clear theoretical foundation of the framework and the unilateral basis of the clusters included (empirical studies conducted in the United States only) justifies the conclusion that the framework does not meet criterion 4.

In summary, the 'educational experience opportunity' framework does not fully meet three out of the four criteria. As a consequence, the framework is regarded less appropriate as a tool for classification of key factors. It is not founded in a substantial theory on educational opportunities supported by results of empirical studies conducted in several countries around the world. Therefore, it is difficult to formulate assumptions about relationships between clusters of factors. The only criterion the framework meets somewhat is that the factors included in the framework are described. However, these descriptions are rather general as they are not classified on the basis of a substantial theory and empirical evidence.

In conclusion, IEA's research framework (Figure 3-2) is regarded as an appropriate tool for the classification of potentially effective factors on student achievement. This framework, with its curricular and educational level dimensions, is taken as the basic frame for the conceptual framework that will be used as the guide to address research question I. The lists of factors classified in each of the clusters could not be literally derived from the TIMSS frameworks because of the reasons given above. Instead, the clusters of factors distinguished in this basic framework will be filled in with potentially effective factors derived from educational effectiveness research. The reason for this is explained below.
In the next sections the basic framework and the contents of the clusters of factors are presented.

3.4 AN ORGANIZING CONCEPTUAL FRAMEWORK FOR THIS STUDY

In principle, IEA's research framework (Figure 3-2) could be used as a basic guide for the exploration of potentially effective factors on mathematics achievement in different education systems (research question I). Given the deficiencies of the TIMSS conceptual framework, 'educational experience opportunity,' it is considered inappropriate for this purpose.

One of the strengths of IEA's three curriculum level research framework is the central position of the curriculum, which has three embodiments. To interpret differences in student achievement across nations, information is needed about differences in the content of the intended, implemented, and attained curriculum (which includes achievement). In the framework, the three curricular content levels are regarded as output levels given their antecedents and contexts.
The distinction between curricular content and its influencing context and antecedents is useful. Curricular antecedents are regarded as input factors that cannot be changed by policymakers or any other group of professionals in education. Curricular context factors are those that can be seen as process or intermediate factors between curricular antecedents and curricular content factors. The distinction between the levels of education in IEA's research framework is useful as well, yet the combined classroom and school level should be separated in two levels. The implemented curriculum content at school level is different from the curriculum content at classroom level. At school level, the curriculum content is called 'institutional' curriculum, and at classroom level it is called 'instructional' curriculum. The institutional level is defined as the curriculum planning at the individual school site. The instructional level is composed of what the classroom teacher decides in his or her planning about curriculum (Klein ,1991).

The conceptual framework for addressing research question I of this study is adapted from IEA's research framework and is called an *organizing* conceptual framework (see Figure 3-4). The addition 'organizing' indicates the emphasis of the framework as a classification tool. The framework is primarily meant and used as an organizer of theoretical and empirical relevant factors that potentially influence student achievement.

Clusters of factors

The organizing framework includes the curricular dimension (antecedents, contexts, and content) as well as the four levels of education. In contrast to IEA's research framework, school and classroom level were separated.

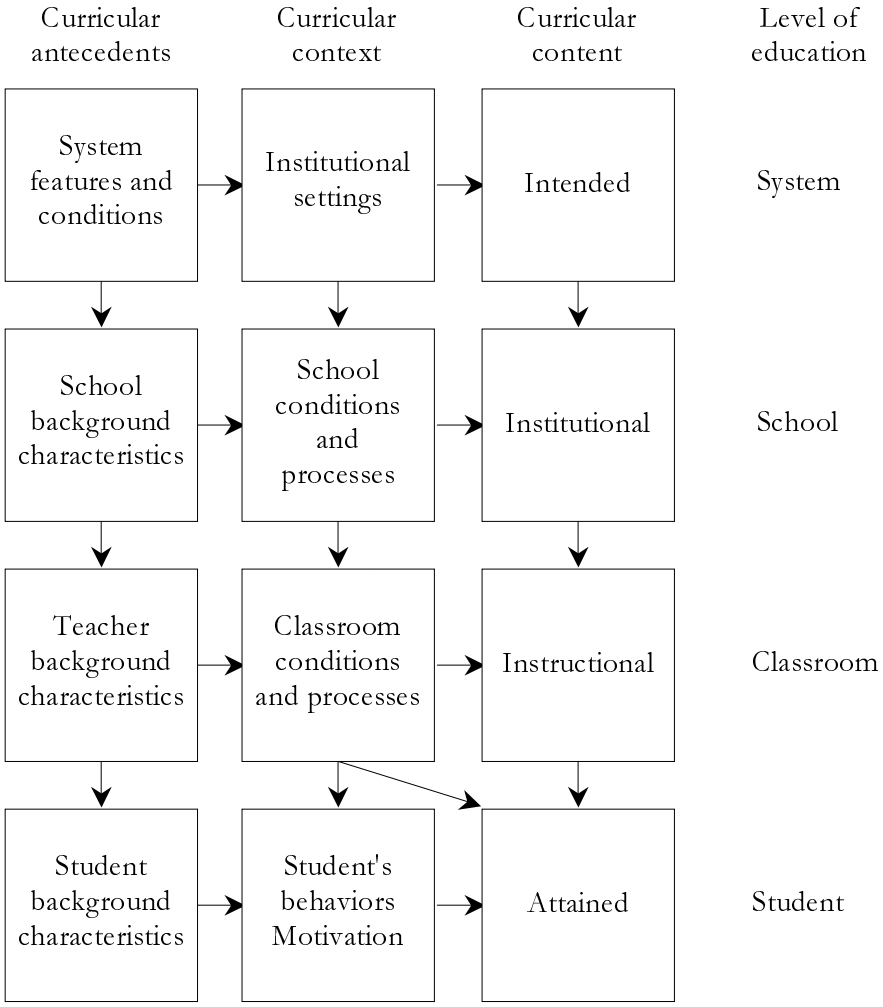


Figure 3-4
An organizing conceptual framework for research question I

The list of clusters of factors at the four levels of education within each curricular dimension looks as follows:

System/country level

- System features and conditions (curricular antecedents)
- Institutional settings (curricular context)
- Content of the intended curriculum

School level

- School background characteristics (curricular antecedents)
- School conditions and processes (curricular context)
- Content of the institutional curriculum

Classroom level

- Teacher background characteristics (curricular antecedents)
- Teaching conditions and practices (curricular context)
- Content of the instructional curriculum

Student level

- Student background factors (curricular antecedents)
- Student behavior and motivation (curricular context)
- Content of the attained curriculum

In a paper of the International Academy of Education, five broad groupings of explanatory factors are mentioned which are important in other international studies (Beaton, Postlethwaite et al., 1999): student home background factors; student motivation; teacher (background) characteristics; teaching conditions and practices; school characteristics. These five groupings are reflected in the organizing conceptual framework presented in Figure 3-4.

The TIMSS data explorations are guided by clusters of factors at the four levels of education (see chapter 4). In the exploration of the TIMSS data, the list of factors will be divided into two groups: factors that can be manipulated by policymakers, teachers, or others and factors that cannot be manipulated. The non-changeable factors are located at the curricular antecedent level, the changeable ones at the curricular context and content level. The changeable factors are of more interest to professionals just mentioned who would like to benefit from the results of international comparative research in education. That is, the professionals who would like to learn from other education systems in order to improve their own. If the differences in educational outcomes across nations can be explained mainly by the non-changeable factors, the professionals will not be able to improve their education based on such study results.

Assumed relationships

In this study, the framework is not primarily meant to be a tool to generate assumptions about relationships between factors within and across clusters of factors. However, the organizing framework includes assumed relationships between clusters of factors and takes into account the review of IEA's research framework. Assumed relationships concern all clusters of factors at all curricular dimensions and all educational levels.

The three levels of curriculum content found within the model are assumed to be related. Moreover, the content levels are supposed to be influenced by curricular contexts and antecedents.

The intended curriculum is located at the system level and can be influenced by institutional settings and features and conditions of the education system. The implemented curriculum located at school level (school goals for core subjects written in documents) is supposed to be directly related to school processes and conditions and indirectly to school background characteristics. The curriculum content which is actually taught (implemented curriculum) is located at classroom level. Teacher background characteristics are classified as antecedents that are supposed to influence classroom conditions and processes. The attained curriculum content is located at the student level and this content is assumed to be influenced by the curricular context factors (student's behaviors and students' motivation). The latter are assumed to be influenced by student background characteristics.

The clusters of curricular context factors are assumed to be interrelated as well. Institutional settings are related to school conditions and processes, which are in turn related to classroom conditions and processes. Classroom conditions and processes are assumed to be related to students' behaviors and to the attained curriculum. Curricular antecedents at the system level are assumed to be related to school background characteristics. Relationships between school background characteristics and teacher or student background characteristics could also be assumed.

Theoretical and empirical foundation and definition of factors

As stated, the theoretical and empirical basis for the factors included within the clusters remained unclear from the literature (criterion 4). Factors included in the framework are labeled only, a clear definition is missing (criterion 2). These two shortcomings of IEA's framework are overcome by additionally using results of educational effectiveness studies.

Research on educational effectiveness attempts to identify factors that 'work' in education either at classroom (instructional) or at school level. Studies in many countries around the world have been carried out to identify influencing factors on student achievement. Many of these studies were reviewed. These review studies resulted in models of instructional effectiveness (Creemers, 1994) and models of school effectiveness (Scheerens, 1990; Scheerens & Bosker, 1997).
The theoretical and empirical basis of the clusters of factors in educational effectiveness models is more profound and more internationally oriented than the ones from IEA's research framework and the TIMSS framework. Also, the provided definitions of factors are more concrete than the factor definitions in IEA's research framework and can be used to operationalize the factors in terms of questionnaire items. For these two reasons, lists of factors classified and defined within educational effectiveness models are used to fill in the organizing conceptual framework for this study.

The next section describes student, classroom, school, and system factors that were derived from a comprehensive model of educational effectiveness (Creemers, 1994) and an integrated model of school effectiveness (Scheerens, 1992; Scheerens & Bosker, 1997).

3.5 POTENTIALLY INFLUENCING FACTORS DERIVED FROM INSTRUCTIONAL AND SCHOOL EFFECTIVENESS MODELS

IEA's three curriculum model is founded on the assumed relations between three broad sets of curricular factors (curricular content and their contexts and antecedents). In the literature on educational research, two other conceptual models take levels of education into account: the educational effectiveness model and school effectiveness model.
The basic assumption of effectiveness models is that the higher levels in the model provide the conditions for what happens at the lower levels. Creemers (1994, p.117) stated that *"(…) factors at the higher levels contribute to the outcomes or are conditional for what happens at the lower levels. This means that not just one level induces results, but a combination of levels."* From this multilevel perspective, differences in achievement results across countries, including factors at four levels of education, can be studied.

School effectiveness research and instructional effectiveness research are integrated in educational effectiveness research. In all kinds of educational effectiveness research, outputs (outcomes of schooling) are associated with antecedent conditions (inputs, processes, or contexts). Scheerens (1999) calls this a basic system model of schooling functioning.

Reviews of educational effectiveness studies resulted in a comprehensive model of educational effectiveness (Creemers, 1994) and in an integrated model of school effectiveness (Scheerens, 1992; Scheerens & Bosker, 1997). As stated in the previous section, these models have two advantages over IEA's research framework. First, the factors included in the models are considered to be potentially effective on student achievement and their theoretical and empirical basis are more profound and more internationally oriented (studies were conducted in many countries around the world) than are the ones from the IEA framework. Many of these studies were reviewed. Second, the definitions of factors are more concrete than the factor definitions in IEA's research framework.

The main clusters of factors included in the organizing framework (Figure 3-4) have been filled in by lists of factors derived from models of instructional and school effectiveness.

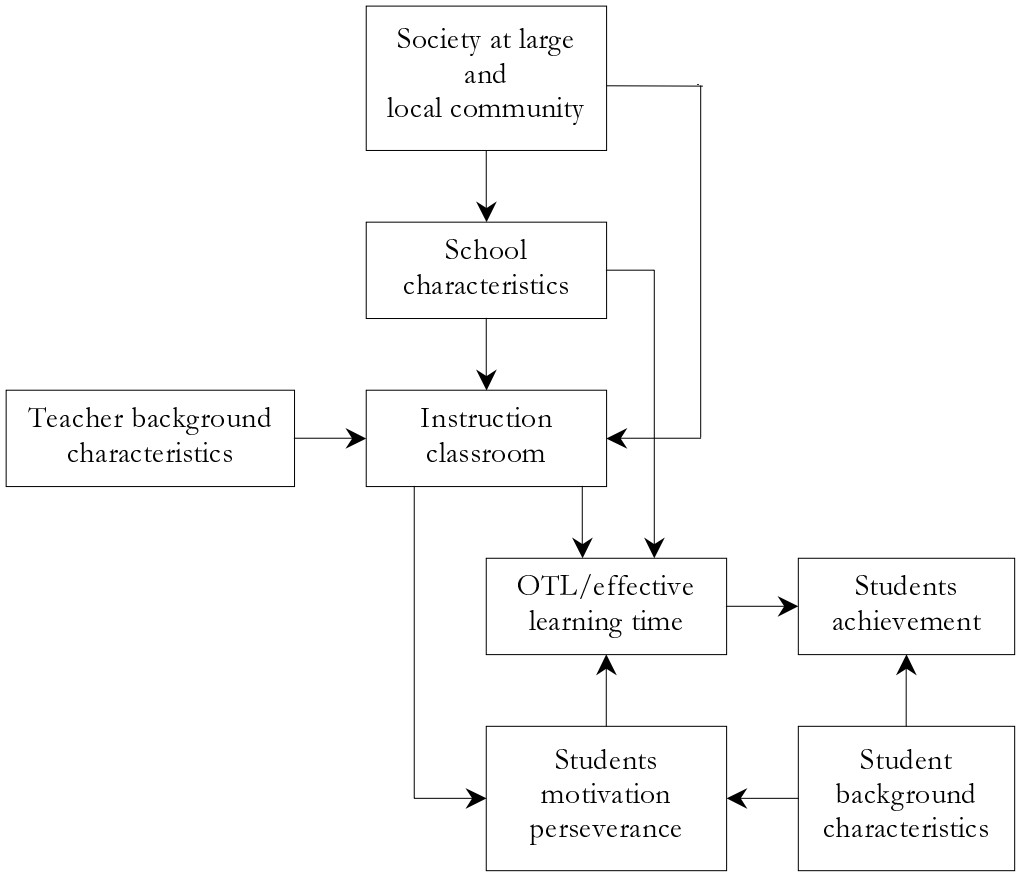
Factors from a model of instructional effectiveness

The research questions addressed by the instructional effectiveness research focus on the classroom level. The main factors involved in these studies are located at classroom, teacher, and student level (Creemers, 1994) and the main topic is the improvement of educational output in terms of student achievement and student attitude towards learning. It is obvious that the key concept in studies focusing on instructional effectiveness is instruction. Creemers (1994) tried to answer questions concerning this concept. What is instruction? How can instruction influence student performance? He reviewed results of many empirical studies that were conducted in United States and in other industrialized countries (e.g. United Kingdom, Germany, the Netherlands). He also used results of review studies. The main goal was to develop a theoretical framework that allows studies aimed at improving student performances to be given an appropriate classification (Creemers, 1994).

The identification of key factors at all educational levels, particularly at classroom and student level, and the classification of those factors is seen as a very useful

filling in IEA's research framework. As stated, cross-national comparisons of achievement results are made by many nations because they want to learn what the benefits and weaknesses are of their own educational system. Many nations want to improve student performances and are looking for ways to accomplish this. In this respect, 'ways' should be read as 'changing factors that are changeable.'

Creemers' model (Creemers, 1994) can be seen as an extension and refinement of Edmonds' five-factor model of (1) strong educational leadership, (2) emphasis on basic skills achievement, (3) safe and orderly climate, (4) high expectations of students' achievement, and (5) frequent evaluation of students' progress (Edmonds, 1979). Figure 3-5 presents the basic conceptual framework from Creemers (1991). The framework of the model was based on Carroll's model of school learning (Carroll, 1963; 1989).



Source: Creemers, 1991

Figure 3-5
Basic conceptual framework of school learning

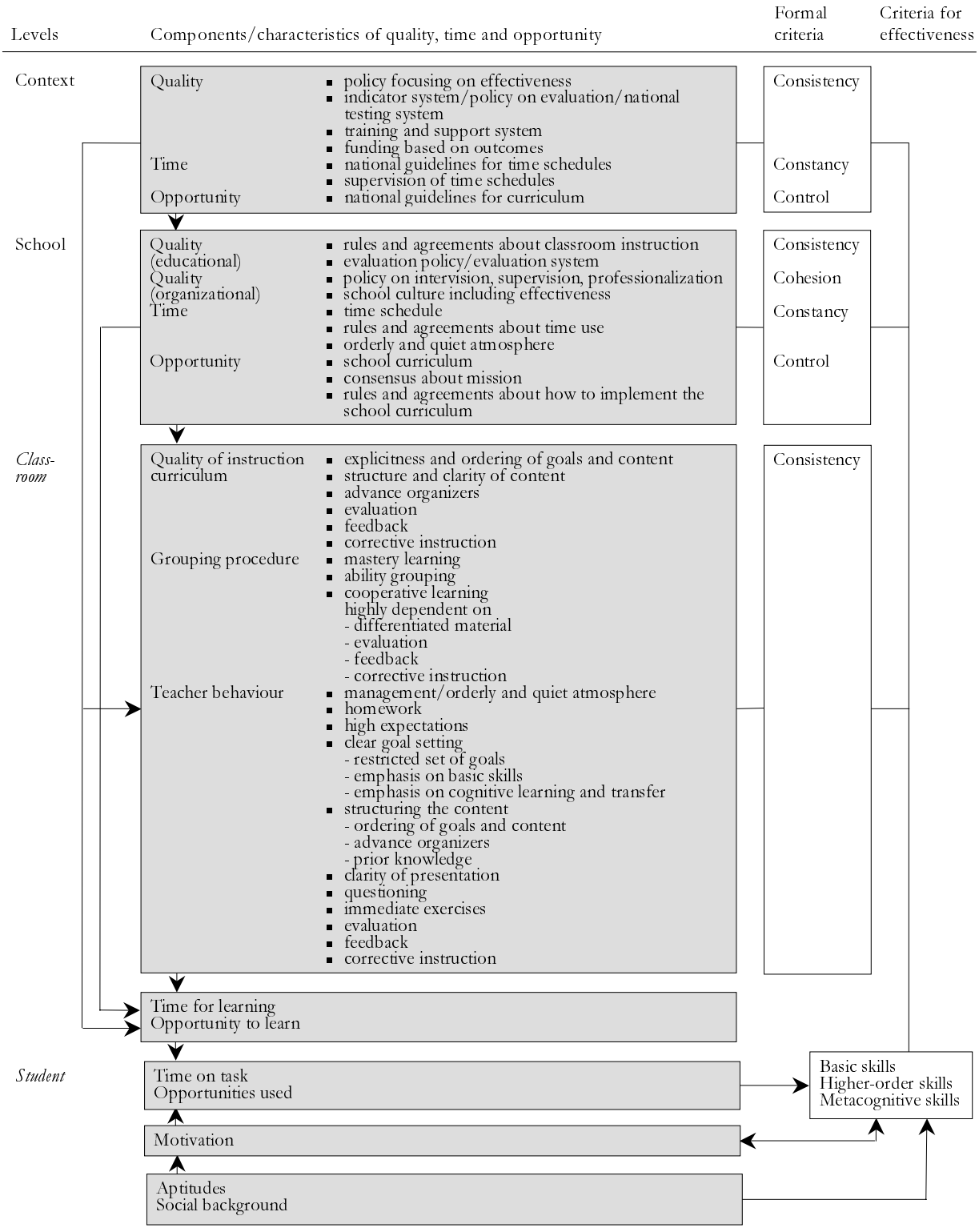
Carroll's model is theoretical, and includes determinants of learning in school. The essential factors in Carroll's model are time for learning and quality of instruction, and the aptitude of students. As a result of his review, Creemers (1994) presented a comprehensive model of educational effectiveness with three essential components. The model consists of four educational levels: student, classroom, school, and context (country) level (see Figure 3-6). The three essential components are quality, time, and opportunity. Creemers included for each level a list of potentially effective factors related to these key characteristics, based on theoretical notions and empirical evidence.

Classroom, teacher and student factors

At the classroom level, the availability of time and opportunity is emphasized. At the student level, the time used and opportunity to learn are meant. Creemers (1994) distinguished three main components within 'quality of instruction' at the classroom level: curricular materials, grouping procedures, and teacher behavior. Creemers presented an integration of separate characteristics of the three main components at classroom level, and included their empirical evidence. Instructional effectiveness is enhanced by integrations of characteristics of curricula, grouping procedures, and teacher behavior. For example, curricula are interpreted by teachers and the choice of grouping procedures requires suitable curriculum material (Creemers, 1994).

For each component, several characteristics and combinations of characteristics which could be effective are distinguished. Isolated instructional characteristics are not effective without taking into account other characteristics. From a theoretical point of view, one can say that the characteristics which constitute quality of instruction are related to the time and the opportunities for learning offered to the students.

The extent of empirical evidence differs between the three components. For instance, the explicit effectiveness of curricular materials is difficult to investigate in educational practice because the contribution of curricular materials can only be studied together with teacher behavior. Therefore, the empirical evidence for the importance of curricular materials is only moderate. In contrast, the empirical evidence of the effects of grouping procedures and teacher behavior is strong (Creemers, 1994).



Source: Creemers, 1994, p. 119

Figure 3-6
A comprehensive model of educational effectiveness

The list of characteristics of the three components Creemers (1994) composed on the basis of his review study is adopted and included in the organizing conceptual framework at the level of the classroom and curricular contexts: classroom conditions and processes. In Figure 3-6, the adopted components and characteristics are shaded. The characteristics of the three classroom components are defined by Creemers (1994). The list of classroom factors adopted at the different levels of the curriculum dimension of the organizing framework is as follows:

Curricular antecedents

Teacher background characteristics (see Figure 3-5; gender, age, teaching experience).

Curricular context

Class size

Curricular materials:

- explicitness and the ordering of goals and content
- structure and clarity of content
- advance organizers.

Grouping procedures:

- mastery learning
- ability grouping
- cooperative learning.

Teacher behavior:

- Teaching style
- Management and orderly and quiet atmosphere
- Homework
- High expectations
- Clear goal setting
- Structuring the content
- Clarity of presentation
- Questioning
- Immediate exercise after presentation of new content
- Evaluation, feedback and corrective instruction.

Curriculum content

Opportunity to learn, instructional curriculum content.

At the *student* level, Creemers recognized four main factors in his model. These factors are included in the organizing conceptual framework: students' aptitude and social background are student background factors (curricular antecedents) and students' motivation and 'time on task and opportunities used' are located at the curricular context level.
In this study, the attained curriculum content is student's achievement in mathematics.

School and system factors
Creemers' model provides factors at the school and system level as well. He defined all school and system level factors in his model as conditions for classroom level factors. Thus, in his model only those school and system level factors were selected that are conditional for and directly related to quality of instruction time or opportunity to learn (Creemers, 1994).

At the *school* level, the following factors were identified which were included in the organizing conceptual framework:

Curricular antecedents
School culture, including effectiveness

Curricular context
Rules and agreements about classroom instruction
Evaluation policy/evaluation system
Policy on supervision
Professionalization
Time schedule
Rules and agreements about time use
Orderly and quiet atmosphere
Rules and agreements about how to implement the school curriculum.

Curriculum content
Content of school curriculum (institutional curriculum).

In the organizing framework the following *system* level factors are derived from Creemers' model:

Curricular antecedents
Resources, funding
Training and support systems
National guidelines for time schedules.

Curricular context
Policy focusing on effectiveness
Policy on evaluation/national testing system.

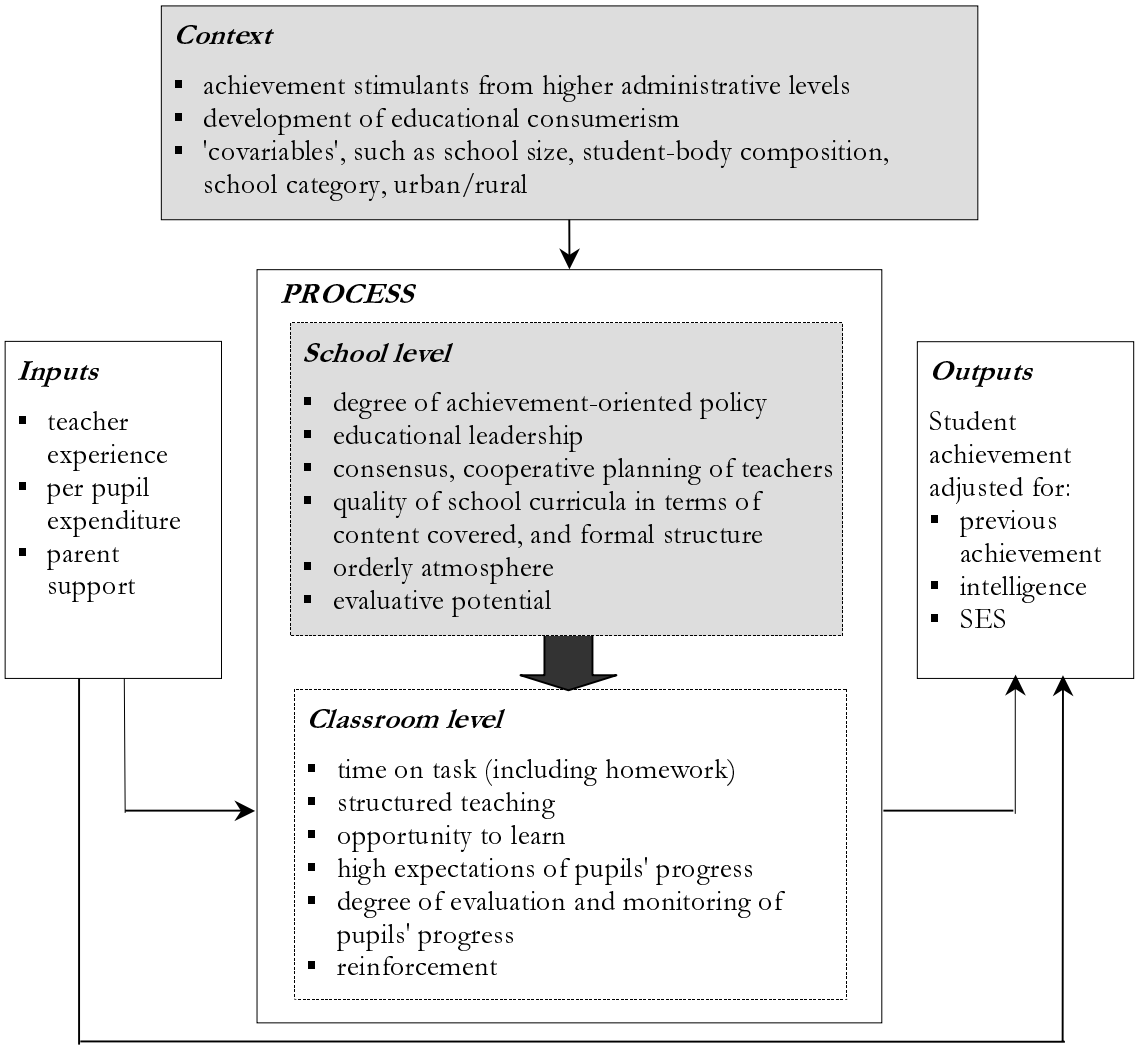
Curricular content
National guidelines for curriculum.

Next to instructional effectiveness model, an integrated model of school effectiveness was studied to add the list of school and system factors in the conceptual framework under construction in this chapter.

School and system factors from an integrated model of school effectiveness

Scheerens (1990) developed an integrated model of school effectiveness, which is presented in Figure 3-7. The model is based on the four levels of education and the input-process/context-output dimension.

Several approaches to educational effectiveness were integrated in this model with a focus on school effectiveness. A 'review of reviews' on school effectiveness research resulted in the selection of variables. The empirical evidence of the variables is documented by Scheerens (1990) and Scheerens and Bosker (1997).
Scheerens (1999, p.6) concluded on the basis of a review of recent studies in different countries that *"gradually school effectiveness research has lead to the development of causal models in which specific characteristics of school organization or instruction are related to each other and the effect criterion."* He emphasized the uncertainty of the unidimensionality of school effects found in studies involving just a few subject-matter areas, sub-systems of the school, and one point in time when effects are measured.



Source: Scheerens, 1990

Figure 3-7
An integrated model of school effectiveness

International studies such as TIMSS usually are not planned as school effectiveness studies. In relation to Scheerens' warning, the comparison between countries on the basis of a one-shot assessment should be carried out with great care (see also chapter 5).

Scheerens and Bosker (1997) considered an indirect influence of school level characteristics via class teaching techniques on student achievement (e.g., quality of school curricula). A direct influence on s tudent achievement is recognized from other school characteristics (e.g., school climate, as indicated by an orderly atmosphere).

The factors at the school level form the central cluster of factors of Scheerens' model: degree of achievement-oriented policy, educational leadership, consensus, cooperative planning of teachers, quality of school curricula in terms of content covered and formal structure, orderly atmosphere (school climate) and evaluative potential. Creemers (1994) included four out of these five factors from Scheerens' model in his model. *Educational leadership* is added to the list of curricular context school factors from the organizing conceptual framework.

At the system level, Scheerens' model provides a list of 'covariables' which is added to the list curricular antecedents at school level in the organizing framework. The covariables are *school size, student body composition, and school category (ability tracks or location of the school in urban/rural area).*

With the additions from Scheerens' model, the lists of potentially effectiveness enhancing factors assigned to the clusters from the organizing conceptual framework (Figure 3-4) are considered complete. In Figure 3-8, the organizing conceptual framework is filled in with factors listed above.

In the next chapter the completed organizing conceptual framework is taken as the starting point for the analysis of the TIMSS data sets. An important question is: which sets of items form scales which can be regarded as an operationalization of potentially effective factors on student achievement?
After the operationalization of factors, interrelationships of factors are explored. Previously, it was concluded that IEA's research framework is limited in the sense that it provides an overview of clusters of factors which potentially affect students' achievement in a core subject without providing clues for the interactions between factors within and between these clusters. After the modification of IEA's model, resulting in the organizing conceptual framework, it is unknown whether this limitation still holds.
The final goal of the exploratory analysis is to find meaningful relationships between background factors, and overall achievement in mathematics.

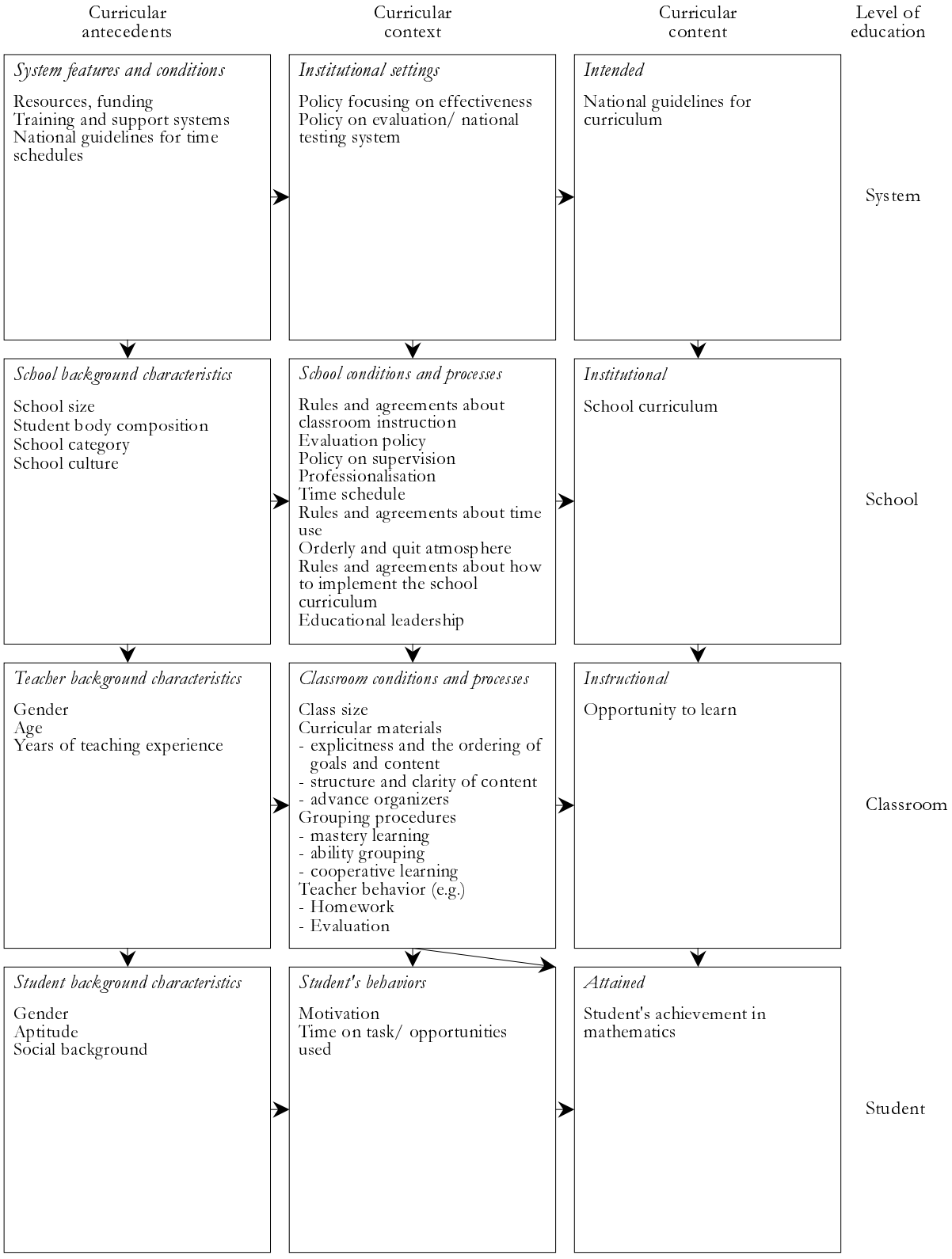


Figure 3-8
Organizing conceptual framework for research question I with factors

CHAPTER 4

UNDERSTANDING CROSS-NATIONAL DIFFERENCES IN MATHEMATICS ACHIEVEMENT IN TIMSS

In this chapter, results are presented of exploratory analyses on TIMSS data sets from the Netherlands and two other European education systems: Belgium Flanders and Germany. The explorations were conducted to investigate the TIMSS goal of understanding differences in mathematics achievement across education systems by means of collected background data. The TIMSS background questionnaires and the background data sets were examined closely in a three-stage analysis plan.

First, the organizing conceptual framework developed in the previous chapter was used as a guide for the identification of possible indicators of potentially effective factors in the TIMSS background questionnaires. Second and third, the interrelationships across the identified indicators were explored empirically within and between the three education systems by means of two consecutive steps: (1) unidimensional partial least squares path analysis (PLS) to estimate a path model per system including student and classroom variables, and (2) the estimation of hierarchical linear models (HLM) by means of multilevel analysis on the separate data sets and on the pooled data set.

In all education systems, aggregated student variables at the classroom level such as parents' educational background and students' attitude towards mathematics explain more variance in achievement scores than individual student variables. The multilevel analysis results on the pooled data set show as well that (aggregated) student variables contribute more to the explained variance in student achievement scores than do classroom (school) variables. The vertical organization of the three education systems is reflected in these results. Moreover, the TIMSS data contain scores on student and classroom variables that are related differently to student achievement across countries.

4.1 TIMSS MATHEMATICS ACHIEVEMENT SCORES IN THREE EDUCATION SYSTEMS

In chapter 1, the main goals and functions of achievement studies conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) were described. Two important functions of IEA studies are describing and understanding similarities and differences in student achievement. In chapter 2, the use of the results of two predecessors of the Third International Mathematics and Science Study (TIMSS) were discussed in light of these two functions. It was concluded that for successive IEA studies on mathematics achievement it was very difficult to fulfill the ambition of understanding cross-national differences. It turned out that in the First and Second International Mathematics Study the 'description' function was accomplished to a greater extent than the 'understanding' function. The ambition of TIMSS was *"to allow researchers to apply theories about contextual factors that contribute to achievement simultaneously to systems of diverse contexts"* (Robitaille & Maxwell, 1996, p.42). With this ambition, TIMSS emphasized the 'understanding' function as well. In this thesis, the understanding function of TIMSS is examined as reflected in research question I and its two related research questions (explained in 3.1). To address these research questions, the categorization, selection, and measurement of potentially effectiveness enhancing factors on student achievement is important. Such factors are located at different education levels: system, school, classroom, and student. In the next sections, background factors indicated in TIMSS population 2 (grade 8) and their usefulness in serving the 'understanding' function are investigated. Results of exploratory analysis on *background* data collected in three neighboring systems are compared: Belgium Flanders, Germany and the Netherlands.

The design of the TIMSS study was explained in 2.5. The main components of the design of the study in grade 8 (second year of lower secondary education in most systems) are repeated here.

Design of the TIMSS Study in grade 8

In each education system that participated in TIMSS, a random sample of secondary schools was selected and within each school an intact grade 8 classroom (the TIMSS

class) took part. At the end of the school year (1994/1995), the international TIMSS achievement test (including the mathematics test) and the student background questionnaire were administered to all students of the selected intact classroom. The mathematics teacher of the TIMSS completed a teacher background questionnaire and the school's principal completed the school background questionnaire. All instruments were developed under the auspices of IEA and were provided in the English language to all participating systems. Each system translated the instruments into their own language according to prescribed procedures. The data collection in the schools took place according to prescribed procedures as well. The uniformity of the execution of the study was controlled by the International Study Center of TIMSS located at Boston College (Boston, US). More information about the design of the TIMSS Study can be found in Martin and Kelly (1996).

TIMSS mathematics achievement scores and research group

As stated in chapter 1, the performance on the TIMSS *mathematics achievement test* is taken as the operationalization of the variable 'mathematics achievement' without discussing its conceptual and curricular foundation.

The three systems under investigation scored differently on the TIMSS mathematics test. The weighted means of the student scores on the international TIMSS mathematics test are presented in the final columns of Table 4-1. In TIMSS, the total student weight is applied in the data sets of the separate education systems. The sum of these weights within a sample per education system provides an estimate of the size of the population in the system (Gonzales and Smith, 1997). Each participating student from grade 8 in each country received an international mathematics score based on his score on the TIMSS test. The scores were standardized with a mean of 500 and a standard deviation of 100 (for more information on the scaling of the scores on the items of the TIMSS test, see Adams and Gonzales (1996)). Belgium Flanders (mean achievement score=565) performed significantly better ($p<.05$) on the mathematics test than did the Netherlands (mean achievement score=541), and Germany (mean achievement score=509) performed significantly less well than did Belgium Flanders and the Netherlands. For more information on the TIMSS mathematics achievement results of the three systems, the reader is referred to the international report on this topic (Beaton, Mullis et al., 1996).

Table 4-1

Composition of research group, and mean (s.e.) of TIMSS mathematics weighted test scores in grade 8 from Belgium Flanders, Germany, and the Netherlands; Spring 1995

Education system	Number of students	Number of teachers (schools)	Mathematics achievement	
			M	*s.e.*
Belgium Flanders (Bfl)	2748	147 (147)	586	5.7
Germany (Ger)	2020	94 (94)	515	4.5
Netherlands (Nld)	1814	88 (88)	551	6.7
Total (pooled dataset)	6582	329 (329)	555	5.6

Note: M = mean; s.e.= standard error ; the column 'students' contains the total number of students that could be linked to the teachers who are presented in the next column

For the three education systems, Table 4-1 also includes the number of students, mathematics teachers, and schools included in the exploratory analyses. Per school, only one intact classroom with only one mathematics teacher took part in the study. The group of students involved in the analyses consists of all students who completed both the TIMSS achievement test and the student background questionnaire and who could be linked to the mathematics teacher who completed the mathematics teacher questionnaire.

The abbreviations of the three education systems used in the first column of Table 4-1 are used throughout this thesis (Bfl=Belgium Flanders; Ger=Germany; Nld=Netherlands).

4.2 POTENTIALLY INFLUENCING FACTORS INDICATED IN TIMSS INSTRUMENTS

To address research question I, it is important to identify which factors measured in TIMSS potentially influence factors on mathematics achievement in the three education systems. Particularly, the search is directed to the factors that can be manipulated by policymakers and teachers in order to enhance student achievement. In chapter 3, the detective process started with a close look at the conceptual foundation of the TIMSS study. The appropriateness of the TIMSS conceptual frameworks known from the literature was judged as not adequate

enough to select and operationalize relevant background factors. Therefore, an organizing conceptual framework was developed (see Figure 3-4). IEA's research framework, with its curricular dimensions, was taken as the basic frame and this frame was filled in by factors derived from instructional and school effectiveness models (Creemers, 1994; Scheerens, 1990; Scheerens & Bosker, 1997). In educational effectiveness studies, factors at the four education levels are studied which potentially influence student learning. Definitions of these factors were available. Moreover, the theoretical and empirical foundation of these studies are considered more substantial and more internationally-oriented than are the studies on which the conceptual frameworks for TIMSS were built.

The framework is identified as 'organizing,' because it was primarily used to categorize theoretically and empirically important factors in clusters which can guide the search for operationalizations of factors. The clusters were organized (categorized) in such a way that assumptions about relationships between the clusters can be derived from them. The framework is not meant to formulate hypotheses about the relationships of *individual* factors within and across the clusters.

Possible indicators in TIMSS instruments

The student, teacher, and school background questionnaires from TIMSS were scrutinized to identify items or sets of items which, as regards the content, could possibly be operationalizations of factors categorized in the organizing conceptual framework. The results of this activity are presented in Table 4-2 in the form of the list of all potentially effective educational factors from the framework and their possible indicators available in TIMSS instruments. Indicators are single items or sets of items included in one of the TIMSS instruments, which can be explored further as operationalizations of factors.

The format of the list is in accordance with the two dimensions of the organizing framework (curricular dimension and level of education). Factors are identified by letters referring to the two dimensions. For example, curricular antecedents at student level are identified by "SA" plus a sequence number and curricular contextual factors at student level are identified by "SC" plus a sequence number.

Table 4-2

Potentially effectiveness enhancing educational factors and their possible indicators available in TIMSS instruments

Level of education and curricular dimension	**Factors in organizing framework** [1]		**Indicators in TIMSS instruments** [2]
Student			
Curricular			
Antecedent	SA_1	Gender	Student's sex
	SA_2	Aptitude	▪ n.i.
	SA_3	Social background	a. Out-of-school activities b. Number of books in the home c. (Educational level mother and father)
Curricular			
Context	SC_1	Motivation	a. Attitude towards mathematics b. Success attribution c. Friends' academic expectation
	SC_2	Time on task/opportunities used	a. Number of minutes math/week b. Amount of homework per day
Curricular			
Content	SO_1	Attained curriculum (Achievement in mathematics)	Test score on entire international TIMSS mathematics achievement test
Classroom			
Curricular			
Antecedent	CA_1	background characteristics	a. Teacher's gender b. Teaching experience in years c. Teacher's workload
Curricular			
Context	CC_1	Class size	Number of students in tested class
	CC_2	Curricular materials: ▪ explicitness and the ordering of goals and content ▪ structure and clarity of content ▪ advance organizers	Textbooks used for mathematics ▪ n.i. ▪ n.i. ▪ n.i.
	CC_3	Material for evaluation of student outcomes, feedback and corrective instruction	Assessment features (standardized test vs. more subjective types of assessment)
	CC_4	Grouping procedures: ▪ mastery learning ▪ ability grouping ▪ cooperative learning	 ▪ n.i. ▪ n.i. Frequency of 'working in pairs or small groups'
	Teacher behavior:		
	CC_5	Teaching style	Teacher's teaching style as perceived by students
	CC_6	Management and orderly and quiet atmosphere	a. Perceived class climate (is it an orderly and quiet atmosphere) b. Perceived school climate (safety) c. Limitations to teach the tested class related to student/resources/ parental features
	CC_7	Homework	a. Frequency of homework b. Treatment in next lesson

Table 4-2

Potentially effectiveness enhancing educational factors and their possible indicators available in TIMSS instruments (continued)

Level of education and curricular dimension	**Factors in organizing framework** [1]		**Indicators in TIMSS instruments** [2]
Classroom			
Curricular Content *(continued)*	CC_8	High expectations	▪ n.i.
	CC_9	Clear goal setting	▪ n.i.
	CC_10	Structuring the content	▪ n.i.
	CC_11	Clarity of presentation	▪ n.i.
	CC_12	Questioning	▪ n.i.
	CC_13	Immediate exercise after presentation of new content	▪ n.i.
	CC_14	Evaluation, feedback and corrective instruction	Use of assessment results for different goals
Curricular Content	CO_1	Implemented curriculum content	Content coverage mathematics
School			
Curricular Antecedent	ScA_1	School size	Total number of students in the school
	ScA_2	Student body composition	Proportion of girls in school
	ScA_3	School category	Urban/rural area of school site
	ScA_4	School culture, including effectiveness	▪ n.i.
Curricular Context	ScC_1	Rules and agreements about classroom instruction	▪ n.i
	ScC_2	Evaluation policy/evaluation system	▪ n.i
	ScC_3	Policy on supervision	Cooperation and collaboration
	ScC_4	Professionalization	▪ n.i
	ScC_5	Time schedule	Time schedule math, grade 8
	ScC_6	Rules and agreements about time use	▪ n.i.
	ScC_7	Orderly and quiet atmosphere	Safety as perceived by the student (see CC_4b)
	ScC_8	Rules and agreements about how to implement the school curriculum	▪ n.i.
	ScC_9	Educational leadership	Number of hours per month principal spends on educational tasks
Curricular Content	ScO_1	School curriculum contents	Written school curriculum mathematics
Country/System			
Curricular Antecedent	SysA_1	Resources, funding	▪ n.i.
	SysA_2	Training and support systems	▪ n.i.
	SysA_3	National guidelines for time schedules	▪ n.i.
Curricular Context	SysC_1	Policy focusing on effectiveness	▪ n.i.
	SysC_2	Policy on evaluation/national testing system	▪ n.i.
Curricular Content	SysO_1	national guidelines for curriculum (i.e., intended curriculum content)	Curriculum questionnaire

Note: [1] SA = student curricular antecedent; SC = student curricular context; SO = student curricular content; CA = classroom curricular antecedent; CC = classroom curricular context; CO = classroom curricular content; ScA = school curricular antecedent; ScC = school curricular context; ScO= school curricular content; SysA = system curricular antecedent; SysC = system curricular context; SysO = system curricular content;
[2] n.i. = no indicators available in TIMSS instruments

Table 4-2 shows each of the potentially effectiveness enhancing factors from the organizing conceptual framework which were indicated by (sets of) items in the TIMSS instrumentation. Factors for which no items could be found are labeled 'n.i.' (no indicators available in the TIMSS instrumentation). For example, at the classroom level, no indicators were available in TIMSS instruments for factors about characteristics of the curricular materials such as 'explicitness and the ordering of goals and content,' 'structure and clarity of content,' 'use of advance organizers,' and teacher behavior (e.g., high expectations, clear goal setting, structuring the content, clarity of presentation, questioning and immediate exercise after presentation of new content). For a few other factors, only proxy items could be found in the TIMSS instruments.

Description of explored indicators

The explored indicators can be described in view of the concrete items of the TIMSS background questionnaires. In Appendix A, an overview is presented of each factor and the TIMSS questionnaire items.

Student's gender

The values of boys and girls were respectively '2' and '1.' Thus, a positive link between 'gender' and another variable means boys 'do better' or 'more' than girls.

Social background

a. Out-of-school activities

Students can have many out-of-school activities, such as working at a paid job, watching TV or videos, and reading a book for enjoyment. Two composites refer to this indicator of the 'student's social background':

- job-related activities: working at a paid job and doing jobs at home;
- leisure time related activities: being with friends, watching TV or videos and playing computer games.

High scores on these factors mean that the student spends a lot of time on them.

b. Number of books in the home

The indicator 'number of books in the home' reflects the educational level of the home of the student. This is a proxy indicator. In the TIMSS student questionnaire, other items of student's home educational background were available: 'educational level of mother and father.' However, in all countries the percentage of missing

values for these items was too high (more than 20%) to allow some kind of imputation to replace the missing values (Rubin, 1987).

Motivation

a. Attitude towards mathematics

A student's attitude towards mathematics can be regarded as a predictor for achievement in mathematics but also as a dependent variable. The TIMSS student questionnaire contains 10 items that potentially refer to attitude. All variables were re-coded such that a high score means a positive attitude towards mathematics and a low score means a negative attitude. Five manifest variables refer to 'liking mathematics,' and the other five refer to 'the perceived importance of mathematics for his/her school career and future.' An example of a 'liking' item is "I need to do well in mathematics to please myself" and an example of an 'importance' item is "I need to do well in mathematics to get the job I want." A high score on attitude means the student likes mathematics and thinks mathematics is important.

b. Success attribution mathematics

Students were presented four qualities and asked to what extent they think each is required to do well in mathematics: lots of natural talent, to have good luck, to undertake lots of hard work studying at home and to memorize the textbooks or notes. Data analysis resulted in the selection of only one variable indicating the extent to which success in mathematics is – according to the student – a consequence of lots of hard work studying at home. This variable showed the highest correlation with mathematics achievement, and the internal consistency across the four items was low. Also, as regards the contents, the other three variables were less appropriate.

c. Maternal academic expectation

This indicator reflects the student's perception of the extent to which his/her mother thought it important for him/her to do well at school in mathematics, science, and native language. A high score means the student perceives a great pressure from his mother to do well at school.

d. Friends' academic expectation

'Friends' academic expectation' has essentially the same contents as 'maternal academic expectation,' but it is the student's perception of the academic expectations his/her friends have of their own school career.

Time on task/opportunities used

a. Total number of minutes spent on mathematics per week

The total number of minutes mathematics is scheduled per week is considered to be an indicator of the concept 'time on task/opportunities used.' It is the only indicator that refers to learning time. This is not the best indicator for 'time on task,' but a better indicator is not available in the TIMSS data set.

b. Amount of homework

The mathematics teacher was asked how many minutes homework s/he usually assigns to students. The teacher was asked to consider the time it would take an average student in the tested class to complete the homework.

Teacher background

a. Teacher's gender

The values of men and women were respectively '2' and '1.' Thus, a positive link between 'teacher's gender' and another variable means male teachers 'do better' or 'more' than female teachers.

b. Teaching experience

The experience a teacher has can be seen as an indicator of the professional background. In TIMSS, teaching experience was expressed by the number of years the teacher had been teaching by the end of the school year in which TIMSS was conducted. The number of teaching years were not separated for different systems of education (primary, secondary or tertiary).

c. Teacher's workload

A third indicator of the background of the mathematics teachers of the tested classes is their workload for mathematics. Workload was measured as the percentage of the total number of appointed lessons per week that were math. The lower this percentage is, the more lessons the teacher was appointed for a subject other than mathematics or for another task in school.

Material for evaluation

The teacher can apply different materials in assessing students' work. In TIMSS, the teacher was asked to indicate how much weight s/he gives different types of assessment (e.g., standardized tests produced outside the school, teacher-made tests, and observations of students in the classroom). The score on this indicator of 'material for evaluation' was calculated by subtracting the weight the teacher gives standardized tests from the mean score of the weight the teacher gives non-

standardized tests. It was expected that students from the teachers who prefer the standardized tests achieved better on the TIMSS test than students from teachers who stressed the non-standardized types of assessment. National standards provide teachers a more objective insight in their students' progress than do non-standardized test means.

Class size

The class size is defined as the total number of students (girls and boys) in the tested class of each school that participated in TIMSS. The mathematics teacher of the tested classes provided these figures. In the international literature the direction of the relationship between class size and student achievement is disputed. Some studies show a negative relationship (small classes outperform bigger classes; (Scudder, 2000) and other studies a positive one (Robitaille and Garden, 1989). A common assumption behind these contradicting results is that other variables play an intermediate role in the relationship between class size and student achievement. For example, in secondary education in the Netherlands, the ability level of the classes differ because of the tracked system. Lower ability students are assigned to smaller classes than higher ability students. In this case, the intermediate variable is 'student's ability level.'

Grouping procedures: cooperative learning

The extent to which students work in pairs or small groups according to the mathematics teacher of the tested classes is taken as the manifest variable reflecting cooperative learning. Cooperative learning is regarded as a potentially effective instructional format.

Teaching style

Student oriented teaching style is described as the extent to which the mathematics teacher directs his or her teaching towards the needs of individual students (Smeets, 2000). The indicator for this factor available in the TIMSS data sets is the perception of students of the frequency teacher activities. The variables which reflect 'student oriented teaching style' to some extent are: students work from worksheet on their own, students work on a mathematics project, students work in pairs or in small groups, students use daily problems when problem-solving, the class discusses practical problems, and students are asked what they know related to a new topic, and solve an example related to a new topic.

Management and orderly and quiet atmosphere

a. Mathematics lesson climate

In TIMSS, the indicator 'mathematics lesson climate' was operationalized by a student perceptual measure. Students were asked about their perception of the climate during mathematics lessons. Three items reflect this concept: 'students often neglect their schoolwork' (scores were inverted to mean that 'students did not neglect schoolwork but took it seriously); 'students are orderly and quiet'; 'students do exactly as the teacher said.' A high score means the student perceived an orderly and quiet atmosphere during mathematics lessons, which is seen as a positive contributor to student's achievement.

b. Perceived school climate

This indicator was measured by means of perceptions of the student about wrong student behavior affecting not only himself but also other students (both those within their own classroom but also from other classrooms) during the last month: something was stolen and someone was hurt. The raw frequency scores of these variables were coded from high to low. Hence, a high score means the school climate was safe and a low score means a poor and unsafe school climate in the perception of the student.

c. Perceived limitations in teaching mathematics

The TIMSS, mathematics teacher questionnaire contains 16 items asking to what extent the teacher perceives the item as a limitation to teach the tested class. The items were categorized into three groups. Each of these groups refers to different kinds of limits: limits having to do with resources, limits related to student characteristics and limits related to parental behavior. For each group, an example is given.

(1) Student limits

An example of a limit related to students is 'students come from a wide range of backgrounds (e.g., economic, language).'

(2) Resource limits

An example of limits related to (lack of) resources is: the teacher was asked to give his opinion regarding the extent to which a shortage of computer hardware serves as a limitation to his/her mathematics teaching.

(3) Parental limits

'Parents uninterested in their children's learning and progress' is an example of a possible limit a teacher can encounter in teaching mathematics.

Homework

Three aspects of homework were represented in the TIMSS mathematics teacher questionnaire: the number of times the teacher assigns homework, the amount of homework, and the teacher's treatment of the completed written homework in the next lesson.

a. Frequency per week

The number of times per week the teacher assigns mathematics homework to the tested class is regarded as an indicator of the factor 'homework.'

b. Amount of homework

The amount of homework was assessed by how many minutes a student spends on average on the assigned homework.

c. Treatment of completed homework during next lesson

The teacher was asked how often he conducted four possible actions with regard to the completed written homework of his students during the next lesson. An example of such an action is 'give feedback on homework to whole class.'

Evaluation, feedback, and corrective instruction

One aspect of the concept 'Evaluation, feedback, and corrective instruction' was indicated in the TIMSS teacher data set by a composite score. This composite refers to six different goals of 'use of evaluation results,' and its sum score consists of the frequency evaluation information the teacher used, on average, for each goal. The different goals are:

- provide students' grades or marks;
- provide feedback to students;
- diagnose students' learning problems;
- report to parents;
- assign students to different programs or tracks;
- plan for future lessons.

A high score on the indicator means the teacher uses assessment information often, which is seen as a positive contributor to achievement results.

Content of the implemented curriculum

In the three curriculum level model IEA developed since the SIMS Study (see above) the content of the implemented curriculum is one of the three levels. In TIMSS, the content of the implemented curriculum at classroom level (instructional

curriculum level) was measured by a question regarding the coverage of several mathematics topics (in total 21 main topics were distinguished). If teachers answered that a topic was taught in the current school year and before the administration of the TIMSS test or that it was taught in a previous year, the topic was marked as 'covered before the TIMSS test was administered.' The percentage of the 21 topics marked as covered was taken as the measure of 'mathematics content coverage.' It was expected that the higher this content coverage percentage, the better the achievement results of the students.

School size and student body composition

The school questionnaire contains items to collect data on the total number of students in school and in grade 8 (the target grade of this study). From these items the proportion of boys in school can be derived.

School category

The school principal was asked to indicate the area the school is located in terms of level of urbanization (number of inhabitants). Schools could be assigned to a group located in a rural area or to a group located in an urban area.

School's policy on supervision

Questions were asked to the principal about the opportunity teachers have to discuss their teaching with other teachers within and across subject departments.

School's time schedule

From the school questionnaire it is known how many minutes per week mathematics was scheduled for grade 8.

Principal's educational leadership

The principal of secondary schools usually has administrative as well as educational tasks. In the TIMSS school questionnaire a list of both kind of tasks was included. The principal was asked to indicate for each task the average number of hours per month (s)he spends on it on.

School curriculum contents

The content of implemented mathematics curriculum at the school level (the institutional curriculum level) is indicated by the availability of documents about the goals and the content of the mathematics curriculum and the ways the school (i.e., the mathematics department) is planning to accomplish the curriculum goals. The availability of such a document was asked about in TIMSS.

The potentially effectiveness enhancing factors for which no indicator could be found in the TIMSS instruments, were necessarily not included in the data explorations and are not reported here. In Table 4-2, factors at *school* and *country or system* level are presented, including their possible indicators available in one of the TIMSS instruments. However, none of the school factors were inserted in the data analyses. The most important reason for this is that the school questionnaire data of the Netherlands contains too much missing data (less than 75% of the principals returned the school questionnaire).

At system level, only one factor is indicated by TIMSS: national guidelines for the curriculum contents. The guidelines provide insight into the intended curriculum. The other system factors listed in Table 4-2 were not indicated in TIMSS, but information on these factors can be found in documents about national education systems. Nevertheless, system factors are not variables within an education system and therefore they are not taken into account in the subsequent exploratory data analysis.

In this thesis, the emphasis is on the student and classroom/teacher factors. The factors at student and classroom/teacher level for which indicators were found in TIMSS are explored further in three stages. In the next section, the three-stage data analysis plan is explained.

4.3 THREE-STAGE DATA ANALYSIS PLAN

Data analysis conducted to address research question I consisted of three stages. In the first stage, the data sets were explored to find scales which can be regarded as operationalizations of factors listed in Table 4-2. The results are expressed in distributions of the scores on the explored composites and singletons added by their statistical reliability. Also in this first stage, correlation matrices were calculated

for the bivariate correlations of the explored scales and mathematics achievement and for the intercorrelations of explored scales. In second stage, direct and indirect relationships between background variables and mathematics achievement were explored by means of exploratory path analysis (unidimensional). The final stage consisted of estimating a hierarchical linear model by means of multilevel analysis in which the variance in mathematics achievement explained by variables at the student level could be separated from the variance explained by variables at the classroom level.

The three stages of data analysis were executed on both the separate data sets of the three education systems under review and the combined (pooled) data sets of these systems.

In this section, the three stages of data analysis, including the requirements of the data sets, are described in general terms. The reasons for conducting these three stages are explained by presenting the principles of the selected techniques and showing their interrelationships. In the following sections, the outcomes of each stage are presented, including more specific decisions made on the basis of the consecutive results.

Stage A: Exploring data sets to find scales

The first exploration of the data sets was carried out to find sets of items and single items which indicate the factors listed in Table 4-2. The definition of a factor was taken as a starting point and was compared with the contents of the items from a TIMSS background questionnaire. A set consisting of more than five items was first analyzed by means of a principal component analysis. The outcomes of these analyses were studied to identify subscales and scales. The resulting (sub)scales from principal components analysis which seemed to make sense from a contents perspective were analyzed further by calculating the reliability coefficient, Cronbach α. For each set of less than five items identified as an operationalization of a factor, the reliability coefficient Cronbach α was calculated.

The lower bound for the Cronbach α to keep a scale in the data analysis was .50 (in data sets of all education systems). This limit is rather low. However, if in the international data set of TIMSS, a limit of .70 or higher was used (usual limits in *national* surveys), a great deal of data would have been deleted from the exploratory data analysis (see Table 4-3).

As the exploratory character of the data analysis is stressed in this thesis and given the international character of the TIMSS instruments, bounds lower than .70 were allowed (Pelgrum & Anderson, 1999). In this study the statistical reliability itself of explored scales was reviewed critically after step B and C of the data analysis, aiming at improvements of the questionnaires (see chapter 5).
Additionally, the weighted descriptives (mean and standard deviation) of the identified singletons and composite variables were analyzed. The difference between the mean scores on the composite variables across the three education systems was examined by means of Bonferroni's pair wise multiple comparisons test between each pair of countries at an alpha level of 0.01.

Finally, in step A the bivariate Pearson product-moment correlation coefficient r was calculated between scores on the explored scales of the factors and the scores on the international TIMSS mathematics achievement test. Also, the Pearson product moment correlation coefficients were calculated for the interrelationships between all selected variables at student and classroom level.

Stage B and C: exploring interrelationships

Stage A can be seen as a preparatory step for stages B and C. The TIMSS data were collected at three different levels of education (student, classroom/school, and country level). The most appropriate techniques to analyze the data are the ones in which this nested design of the data sets are taken into account: hierarchical linear modeling (HLM) techniques.
The major advantage of HLM techniques (e.g., multilevel analysis) over unidimensional ones such as partial least squares techniques (PLS), is the estimation of the effects of variables on the dependent variable at one level (for example student level) taking into account at the same time the effect of variables on the dependent variable at another level of the hierarchical data structure (for example classroom level). Multilevel analysis results in better estimation of the amount of variance in the output variable that each variable in the model can tie. The direct effects of each variable at each level can be estimated, and direct effects of classroom/school variables on student variables can be estimated. Furthermore, indirect interaction effects of classroom/school variables on the effect of a student variable on mathematics achievement can be specified. An example of the latter is the effect of the relative amount of mathematics topics a teacher has covered in his

lessons (classroom level variable) on the relationship between student's attitude towards mathematics and mathematics achievement.
However, HLM techniques have limitations that can be regarded as benefits of single-level path analysis techniques. There are two important limitations: (1) multilevel techniques do not permit the estimation of scores on latent variables, and (2) they do not permit the modeling of hierarchical covariance structures (indirect effects between variables).
The measurement model in HLM requires variable scores (latent variables) which cannot be estimated by means of HLM techniques. The first limitation implies the calculation of scores on the latent variables outside the HLM technique, for example by means of principal components analysis. In this study, the scores on the majority of the latent variables selected as input for modeling latent variables by means of HLM techniques were sum scores on composites explored in stages A and B of the data analysis plan.
The second limitation of HLM can technically only be overcome through subsequent analysis by applying a stepwise procedure (Lietz, 1996). Independent HLM runs should be conducted with different mediating variables, with the whole model specified as the outcome variable. The results of these independent runs must be compared with each other to determine how variables are interrelated, including indirect relationships with the outcome measure (the achievement variable). Kotte (1992) is one of the few researchers known who has tried this procedure. He showed that this procedure is far from ideal and very laborious. Recently, a study was published in which multilevel structural equations models were explored. Causal relationships were tested among variables on more than one level (Shalabi, 2002).

Beyond the limitations of HLM techniques, there is a more dominant reason to explore the TIMSS data first by means of a single level technique. To model the TIMSS data at more than one level, some theoretical basis must be available (Glaser and Strauss, 1967). Important direct and indirect (mediated) relationships between student and classroom/school variables should be known before a more advanced technique such as multilevel analysis, is applied.
As stated in previous sections, there is little relevant research available to serve as a sound theoretical and empirical basis for the specification of a hierarchical model of student- and classroom/school variables influencing mathematics achievement of

grade 8 students *in different countries*. Hence, the interrelationships between student and classroom/school variables had to be explored in stage B. This was done at two separate levels: at level-1 interrelationships between student variables have been explored. The level-2 analyses concerned interrelationships between classroom and teacher variables combined with the student data which were aggregated to the classroom level. The path analysis results at both level-1 and level-2 were studied for inclusion in a hierarchical linear model (Lietz, 1996; Bryk & Raudenbush, 1992).

Stage B: Partial Least Squares (PLS) path analysis

The factors listed in Table 4-2 for which indicators were found in either the student or the teacher questionnaire of TIMSS, can have both a direct and an indirect influence, mediated by other concepts, on the dependent variable 'performance on international mathematics test.' An appropriate tool to uncover direct and indirect effects of different sets of variables is path analysis (Campbell, 1996). In general, path analysis is an outgrowth of regression analysis. It involves determining the order of the variables (indicators of concepts) and their direction in a hypothesized model.
Because decisions concerning instruments and associated variables were made in TIMSS before a model was developed, the path model has been developed post hoc. Consequently, the nature of the analysis must be seen as more exploratory than confirmatory. In the TIMSS data analysis, the Partial Least Squares (PLS) approach has been applied. PLS has been developed especially for research situations that require a great deal of exploratory analyses (see for example Lietz, 1996; Sellin & Keeves, 1994; Sellin, 1992 and 1990; and Wold, 1982). Other approaches like LISREL (Jöreskog & Sörbom, 1989) and AMOS (Hox, 1995), on the other hand, were designed primarily for situations that require confirmatory tests of theoretically well-established path models.

PLSpath program
The latest version of the program 'PLSpath' was used (PLSpath version 3.01; Sellin, 1989). Applying PLS consists of two main steps:

1. Estimation of latent variables (LVs) at student/classroom level as linear composites of their associated manifest variables (MVs, items from the TIMSS questionnaires) by means of either principal component analysis or by means of regression analysis. This is called the *outer* PLS model.

2. Estimation of the direction and strength (path coefficients) of links between latent variables. This estimation is conducted by means of ordinary least squares regression applied to each equation (endogenous latent variables predicted by two or more other latent variables) separately and results into the estimated recursive *inner* PLS model.

In step 1 of PLSpath, clustering of manifest variables (i.e., items from the questionnaires) should result in the estimation of meaningful latent variables (outer model). Postlethwaite and Wiley (1992) stressed that clustering of items should reflect meaningful homogeneity within clusters, both conceptually and empirically. *Conceptually* means the latent variable must make sense and has a meaning in the literature; *empirically* means the manifest variables must have meaningful loadings on one factor in a principal component analysis (or meaningful weights in a regression analysis) and a correlation higher than 0.10 (absolute value) with the dependent variable. For example, clustering items with respect to student's attitude towards mathematics can form a latent variable if the meaning of each of these items can be linked to student's attitude (conceptual homogeneity), if the loadings of each of these items are high enough, and if the correlation of the sum score of the individual item scores with 'mathematics achievement' is higher than 0.10 (empirical homogeneity).
An important feature of exploratory path analysis by means of PLSpath is trimming the results in order to achieve a parsimonious path model for each of the three education systems. First, the outcomes of the outer PLS model were trimmed to finalize the outer model. Thereafter, the inner model was trimmed. For both 'model trimming' steps PLSpath provides indices.
The criteria for keeping a manifest variable in the outer model have been set in advance. It is assumed that MVs with low (regression) weights or (factor) loadings harm the predictive power of the LV (Campbell, 1996; Keeves, 1992). If, per nation, the absolute value of the weight of an MV was lower than .10 (this is the inward mode in PLSpath in which MVs form an LV) or the absolute value of the loading was lower than .30 (this is the outward mode in PLSpath in which the MVs reflect the LV), the MV was removed from the model. Sellin and Keeves (1994) stated that if the contributing MVs are theoretically and empirically homogeneous then a loading of .30 would seem acceptable. A loading of .30 indicates that an LV contributes 10% to explaining an observed MV. Regarding the weights, Sellin and

Keeves (1994) indicated the lower bound of .10 because this value means that the MV contributes approximately 1% to the explanation of an LV. Beyond the limits of loadings and weights as selection criteria for keeping MVs in the outer model, PLSpath provides three other statistics to judge the appropriateness of the explored outer path model: redundancy, tolerance, and communality.
Checks on multicollinearity should be carried out. All MVs that are supposed to reflect (outward) or form/produce (inward) the same LV should intercorrelate highly. Intercorrelations must be studied between an MV and LVs other than the one to which the MV is assigned. If it turns out that an MV correlates highly with more than its 'own' LV, the allocation of the MV to another LV can be reconsidered. A measure of multicollinearity is *redundancy*. Redundancy is defined as the squared correlations between a particular MV and the set of LVs linked indirectly through inner model relationships (Afrassa, 1999; Kotte, 1992; Sellin & Keeves, 1994).
The PLS output of the estimation of the outer model contains a second measure of multicollinearity: *tolerance*. Low tolerance indicates low intercorrelations between MVs that are supposed to reflect or form one LV. For each MV, the PLS output of the outer model presents a tolerance value. This value is the squared multiple correlation of a particular MV with all remaining MVs in the set. They provide information about possible multicollinearity within a block of MVs (a pattern in PLSpath). If the inward mode is applied, multicollinearity within a block of MVs is indicated by a tolerance value higher than .50 (Sellin, 1990).
A measure of the explained variance of a particular MV with respect to the LV it reflects is called *communality*. Communality is defined as the squared correlations between MVs and their corresponding LV (Sellin, 1989). The average of the communalities of all MVs included in the outer model is taken as the criterion for the strength of the outer model (Afrassi, 1999; Falk, 1987). The higher this average, the better the outer model fits the data. Usually, an average communality value of .30 is taken as the lower bound.

After finalizing the outer PLS model, the inner PLS model, in which relationships are hypothesized between the latent variables, is trimmed. The inner model is a recursive one: each LV depends only on previous or on no latent variables.
For endogenous latent variables, direct effects and indirect effects can be distinguished. The strength of both kind of effects is indicated by a path coefficient beta (β). The total effect is the sum of the direct and the indirect path coefficients

(effects) of latent variables on another latent variable. Sellin and Keeves (1994) suggest that an absolute value of beta coefficient of .05 or higher is to be considered significant in large samples (n > 200, like in the student samples of TIMSS). Beforehand, they demand a strong theoretical case for inclusion of a particular path in a model to accept the value of .05 as a minimum for the beta. For smaller samples (n < 200, like the classroom sample of TIMSS) the absolute value of beta coefficient should be .10 or higher, indicating approximately 1% explained variance of an endogenous latent variable in a model (Sellin & Keeves, 1994).
The product-moment correlation between an endogenous latent variable and the latent variables that have a direct relationship with it, is shown in the PLSpath output of the inner model estimation. These intercorrelations between two LVs must be compared with the beta coefficient. The beta and correlation coefficient must be in the same direction and of the same magnitude; if not, a suppressor effect might exist as a result of measurement error or multicollinearity (Keeves, 1997). Consequently, the relationship should generally be removed from the path model.
The PLS output of the estimation of the inner model also contains a measure of multicollinearity of LVs within blocks of the inner model: the *'tolerance'* index. This value is the squared multiple correlation between a predictor LV with all remaining LVs in the set. Multicollinearity is indicated by tolerance values higher than .50 (Sellin, 1989).

Statistical test of significance of the PLSpath results

Exploring the hypothesized path model by means of PLSpath does not include a statistical test of significance in terms of goodness-of-fit measures. Such measures are absent in PLSpath. Instead, PLSpath employs a jackknifing method which is inappropriate if the dataset consists of more than a few hundred cases (Sellin, 1989). Jackknife procedures are primarily aimed at obtaining appropriate estimates of standard errors and requires a simple random sample. However, TIMSS databases are not founded on simple random samples but on cluster samples. It can be stated that the student sample is not a simple random sample, but a two stage stratified cluster sample (schools and intact classes within each school; the data have a nested structure). Consequently, the jackknife procedures included in the PLSpath program are not appropriate. Then, the statistical test of significance of the PLSpath results should be conducted by means of another program or it can not be done at all. The WesVar module for complex samples available in SPSS could have been

used. However, PLSpath was only applied to explore indirect and direct links of LVs estimated from the TIMSS data sets. The final PLSpath results were not used to fully address research question I, but to find indications for influencing variables and their interrelationships which make sense and could be included in multi-level analysis. The final answers to the research question are formulated on the basis of the multi-level analysis and not on the basis of the exploratory path analysis. Hence, the purely exploratory character of the application of PLSpath within the total data analysis plan was a reason to judge the PLSpath results without a straightforward significance measure. Besides, no "objective" significance measures are available in PLSpath. Thus, a certain degree of subjectivity in judging the appropriateness of the model cannot be avoided.

Following Lietz (1996), who applied PLSpath intensively on IEA Reading Literacy data, the determination of possible misspecification of the model is left to the researcher rather than relying on tests of statistical significance. Lietz (1996) used the average of the R^2 (explained variance) of all LVs (including mathematics achievement) in the model as an indication of the fit of the model to the data sets given. The average multiple R^2 for the endogenous variables in the model is regarded as an indication of the predictive power or strength of the inner and outer relationships and therefore it can be used for the evaluation of the path model (see also Afrassi, 1999, Falk, 1987).

The size of R^2 values for each LV is given in PLSpath output and indicates the part of the variance of a construct explained (reproduced) when all preceding LVs are taken into account. The R^2 of the outcome variable (mathematics achievement) indicates what percentage of the variance in the outcome variable is explained or reproduced by the latent variables (predictors) in the model.

Step C: Multilevel analysis

The outcomes of the PLSpath are used to specify a hierarchical linear model for each education system. Building hierarchical linear models by means of multilevel analysis must be done carefully. The models are representations of complex, multivariate relationships operating at different levels simultaneously with different units of analysis. Several multi-level data analysis techniques are available. In this study the MLn program (Woodhouse, 1995) was used to specify two-level country models: students at level-1 and schools (classrooms) at level-2. In TIMSS, schools are coincided with classrooms because within each selected school one intact

classroom was selected. The nesting structure of the TIMSS data can be described as students within classes (schools) and the variability in scores on the international TIMSS mathematics achievement test between students and between classes (schools) is to be explained.
A hierarchical random intercept (effects) model was specified for each country. Thereafter, a hierarchical linear model was estimated for the pooled data set. The results of the latter model indicate whether student and classroom variables have different effects in the three systems.

The specification of the model starts with the fully unconditional model. The fully unconditional model, with two levels and one dependent variable Y which is the sum of a general mean (γ_{00}), a random effect at classroom level (U_{0j}), and a random effect at student level (R_{ij}) can be expressed as follows (Snijders & Bosker, 1999):

$$Y_{ij} = \gamma_{00} + U_{0j} + R_{ij}$$

The test score of student i from school j is expressed by 'Y_{ij}'. This model is also called the 'empty' model. An important assumption of using the random coefficient model is that the random coefficients U_{0j} and R_{ij} are normally distributed with a mean of '0' given the values of the explanatory variables. Both coefficients are assumed to be mutually independent, and to have variances var(U_{0j}) = τ_0^2 *(intercept variance)* and var (R_{ij})= σ^2 *(between student variance)*. These variances are important features in multilevel modeling and are random effects which are equal to the unexplained variability or variance.

The empty model provides partition of the variability in the data between the student- and classroom-level. The total variance of Y (students scores on the TIMSS mathematics test) can be decomposed as the sum of the level-2 and the level-1 variances (Snijders & Bosker, 1999):

$$\text{var}\ (Y_{ij}) = \text{var}(U_{0j}) + \text{var}\ (R_{ij}) = \tau_0^2 + \sigma^2$$

Further specification of the hierarchical linear model is aimed at reduction of the random effects. The empty model is further specified by adding one variable at a

time, starting with level-1 variables (student variables). First, variables which are less or not changeable by schools or policymakers are inserted. This method is called *forward steps upward from level-1 method* (Snijders & Bosker, 1999). Only if the added variable showed a significant effect it was kept in the model. Non-significant effects were immediately excluded from the model. First, an attempt is made to explain within-group variability. Next, an attempt is made to explain between-group variability. Finally, possible interaction effects between level-1 and level-2 variables were included in the models.

The formula of the extended empty model with one explanatory variable from level-1 (variable *x*) is written as:

$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + U_{0j} + R_{ij}$$

Variables from level-2 are indicated by z and can be inserted in the model:

$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + \gamma_{01} z_{ij} + U_{0j} + R_{ij}$$

With regard to interaction effects, whereby a level-2 variable influences the effect of a level-1 predictor on the outcome variable, the level-1 effect should be random. However, this rule could not be interpreted too strictly (Snijders & Bosker, 1999). Even if effects of level-1 variables are fixed without a random effect, there might be reasons to explore interaction effects from level-2 on level-1 variables. The interaction term in the formula is $z_i x_{ij}$

$$Yij = \gamma_{00} + \gamma_{10} x_{ij} + \gamma_{01} z_{ij} + \gamma_{11} z_i x_{ij} + U_{0j} + U_{1j} x_{ij} + R_{ij}$$

Test of significance

The *fixed effects* were tested on their difference from zero (significant effect) by means of a *t-ratio* for the γ-coefficients, which can be interpreted as standardized path coefficients. The t-ratio is defined as the proportion of the estimated γ-coefficient and its standard error. In this exploratory study, two-tailed tests were applied, which means that the absolute t-value should be greater than 1.96, with a p-value less than .05, to keep the corresponding effect in the analysis.

The *deviance test* is used for tests concerning the *random part* of hierarchical linear models. In this study, the parameters of the models were estimated by the maximum likelihood method. The likelihood can be transformed into deviance defined as minus twice the natural logarithm of the likelihood (Snijders & Bosker, 1999). This deviance is seen as a measure of lack of fit between model and data. However, the values of the deviance cannot be interpreted directly. Only *differences in deviance values* for several models fitted to the same data set provide information regarding the model fit. The difference of the deviance of two models determines whether the second model is an improvement compared to the first model. The difference in deviance can be used as a test statistic having a chi-squared distribution with 'number of parameters in model 1 minus number of parameters in model 2' degrees of freedom (Snijders & Bosker, 1999).
In this study, the deviance difference is calculated (1) between the fully unconditional model (the empty model) and the unconditional level-2 model (the model with all selected student level variables included), (2) between the unconditional level-2 model and the model with all selected student and classroom variables, and (3) between the model with all selected student and classroom variables and the model with all selected variables and all selected interaction effects between level-2 and level-1 variables.

Explained proportion of variance by a two-level model
The results of the MLn analysis provide estimates of the proportion of variance associated with each level: the final estimation of variance components. The comparison of these estimates belonging to a particular model with figures of the fully-unconditional model can provide an indication of the amount of variance explained by the predicting variables at each level (Snijders & Bosker, 1999).

For each level, the proportion of variance explained by adding level-1 predictors and level-2 predictors can be calculated given the estimates regarding the partitioning of variance at the two levels from *the empty model or the fully unconditional model*. In other words, the calculation of the percentage of the variance that is explained by the model can be obtained by calculating the proportional reduction of the unexplained variance from the fully unconditional model.

In this study, for level-1 and level-2, the proportion of variance that is explained by the model was calculated by means of different formulae. The level-1 explained proportion of variance in the individual scores on the outcome variable is defined as the mean squared prediction error, which equals the proportional reduction in the value of $\hat{\tau}_0^2 + \hat{\sigma}^2$ due to including predictors in the model (Snijders & Bosker, 1999). This can be expressed in a formula:

$$1 - (\hat{\sigma}^2_{final} + \hat{\tau}_0^2{}_{final} / (\hat{\sigma}^2_{empty} + \hat{\tau}_0^2{}_{empty}))$$

in which the denominator contains the sum of the unexplained between student variance and the unexplained variance between schools of the empty model and the numerator contains this sum for the model in which one or more predictors were included.

The level-2 explained proportion of variance is defined by Snijders & Bosker (1999) as the proportional reduction in mean squared prediction error, for the prediction of the group mean $\overline{Y}_{.j}$ for a randomly drawn, level-two unit *j*. The level-2 explained proportion of variance is estimated as the proportional reduction in the value of $\hat{\sigma}^2 / n + \hat{\tau}_0^2$. n is a representative value of the group size. The formula looks as follows:

$$1 - [((\hat{\sigma}^2_{final} / n) + \hat{\tau}_0^2{}_{final}) / ((\hat{\sigma}^2_{empty} / n) + \hat{\tau}_0^2{}_{empty})].$$

Requirements of the data sets

The basic requirements for the data sets that are explored in stage A are that the data do not contain any missing value and that the data is weighted.

Missing Values

As the PLSpath program (see stage B) does not allow for missing values in the database, the first step in the procedure is to deal with the cases that have missing values on the variables included in the analysis. The cases with a missing value on the dependent variable, the mathematics score on the international TIMSS test are removed from the data set. Furthermore, there were some cases that had a valid answer on the dependent variable, but missing values on a majority of the other variables. These cases were removed as well.

For some cases, imputation was applied: the mean or median score of all the cases on the variable (or the class mean or class median of the variable) was used to replace the missing value(s). The statistical accuracy of applying an imputation procedure can be discussed (and the different types of imputation possible as well). However, this simple procedure is justifiable when the number of cases to which it was applied is not very large. In the case of more than 20% missing values, a solution is not available. Examples of variables with more than 20% missing values are the educational level of mother and father, which are regarded as indicators of the social economic background of the student (see previous section). Such variables were not included in further analyses.

Weights

The student data sets were analyzed after weighting the data. In TIMSS, two weighting variables were used in these analyses: the total student weight and the senate weight (Gonzales & Smith, 1997). The total student weight is applied in the data sets of the separate education systems. The sum of these weights within a sample per education system provides an estimate of the size of the population in the system. If a weighted estimate of the mean score on a certain variable (for example student's home background) is required for the total population of three systems, it is desirable that each system contributes equally to the international estimate. For this purpose, the senate weight was developed by TIMSS (Gonzales and Smith, 1997). The senate weight is proportional to the total student weight by the ratio of 1000 divided by the size of the population. The sum of the senate weights within each system is 1000.

Standardized data

In international comparative studies like TIMSS, country model comparison is an important aim. Therefore, the choice between the use of standardized or metric data is a difficult one. Metric coefficients can only be compared across models (countries) and standardized coefficients only within a model. Nevertheless, in this study, standardized data were employed. Lietz (1996, p. 137) wrote about this issue that *"an examination of the results obtained from the different analyses* (i.e., separate analysis per country) *might reveal similar patterns of the interrelationships between the different variables in the models. Although such results could not be used to prove that the observed processes were*

generalizable across systems, they would provide an indication of interesting relationships for further investigation." Reports on secondary analysis of IEA data (Keeves, 1992; Kotte, 1992) showed consistency of results across samples on mean effect sizes and corresponding standard deviations. These researchers concluded that although the use of standardized coefficients across samples might theoretically be inappropriate, it might provide useful additional information when investigating patterns of relationships across education systems.

As stated, in this thesis, the general purpose of exploring path models by means of path analysis was to find indications of relationships between potentially influencing factors on mathematics achievement. The resulting path coefficients are not used to draw *final* conclusions about the differences across education systems with regard to influencing factors. Instead, the indicated relationships were used as input for multi-level analysis.

In stage B, the TIMSS data sets were explored without weighting because in PLSpath no weighting option is available. The estimation of a hierarchical linear model was conducted on the standardized and weighted data sets. The weights described above were applied. The total student weight was applied in the data sets of the separate education systems and the senate weight in the pooled data set (Gonzales & Smith, 1997).

4.4 STAGE A: RESULTS OF FIRST DATA EXPLORATIONS

The TIMSS student and teacher data sets of the three education systems were explored by applying different techniques to find indicators of potentially effective educational factors from the organizing conceptual framework listed in Table 4-2. In Table 4-3, the outcomes of the first data explorations are shown for each indicator listed in the first column.
The content of each column of Table 4-3 will be explained, including a discussion about the results of the reliability tests of the explored indicators, the frequencies of the explored indicators and the bivariate correlation coefficients of the indicators with mathematics achievement.

Table 4-3

Outcomes of first explorations on TIMSS student and teacher questionnaire data

Potentially effective educational factors and explored indicators in TIMSS data sets **(number of items and range of scale)**	**Statistics per educational system and pooled data set (weighted data)** [1]											
	Standardized Cronbach alpha [2]				M (sd)				Pearson r with math achievement [2]			
	Bfl	*Ger*	*Nld*	*Pooled data set*	*Bfl*	*Ger*	*Nld*	*Pooled Data set*	*Bfl*	*Ger*	*Nld*	*Pooled data set*
SA_1 Student's gender	n.a.	n.a.	n.a.	n.a.	girl: 50%	girl: 52%	girl: 50%	girl: 50%	ns	ns	.05	ns
SA_3 Social background												
a. Out-of-school activities												
1. paid jobs (2 items; 0 - 4) [3]	r=.20	ns	r=.19	r=.15	0.9 (0.7)	0.9 (0.7)	1.0 (0.8)	0.9 (0.7)	- .17	- .09	- .15	- .15
2. leisure time (3 items; 0 - 4) [3]	.50	.36	.49	.48	1.4 (0.6)	1.8 (0.7)	1.8 (0.7)	1.6 (0.7)	- .19	- .07	- .27	- .24
b. number of books in the home (1 item; 5 categories) [3]	n.a.	n.a.	n.a	n.a	3.3 (1.2)	3.6 (1.3)	3.4 (1.2)	3.4 (1.2)	.13	.34	.30	.22

Table 4-3 (*continued*)

Outcomes of first explorations of TIMSS student and teacher questionnaire data

Potentially effective educational factors and explored indicators in TIMSS data sets **(number of items and range of scale)**	**Statistics per educational system and pooled data set (weighted data)** [1]											
	Standardized Cronbach alpha [2]				M (sd)				Pearson r with math achievement [2]			
	Bfl	*Ger*	*Nld*	*Pooled data set*	*Bfl*	*Ger*	*Nld*	*Pooled data set*	*Bfl*	*Ger*	*Nld*	*Pooled data set*
SC_1 Motivation												
a. Attitude towards math (10 items; 1 - 4) [3]	.84	.80	.78	.80	2.8 (0.5)	2.8 (0.6)	2.7 (0.4)	2.8 (0.5)	.22	.04	.12	.14
importance (5 items) [3]	.72	.64	.68	.68	2.9 (0.6)	3.0 (0.6)	2.8 (0.5)	2.9 (0.5)	.17	ns	ns	.04
like (5 items) [3]	.77	.78	.76	.76	2.7 (0.6)	2.6 (0.7)	2.6 (0.6)	2.6 (0.6)	.22	.06	.20	.18
b. Success attribution (1 item: lots of home-work; 1 - 4) [5]	---	----	----	---	3.2 (0.7)	3.1 (0.9)	3.2 (0.7)	3.2 (0.8)	.15	.24	.17	.13
c. Maternal academic expectation (3 items; 1 – 4) [3]	.74	.67	.87	.76	3.4 (0.5)	3.4 (0.6)	3.4 (0.5)	3.4 (0.5)	ns	- .04	- .06	ns
d. Friends' academic expectation (3 items; 1 – 4) [3]	.85	.82	.93	.86	3.0 (0.6)	2.6 (0.8)	3.1 (0.6)	2.9 (0.7)	ns	- .18	- .07	ns
SC_2 Time on task/ opportunities used												
a. instructional time mathematics in minutes per week (1 item) [3]	n.a.	n.a.	n.a.	n.a.	224 (22)	181 (34)	149 (17)	192 (40)	.24	ns	.17	.29
b. amount of homework per day (1 item; 5 categories) [4]	n.a.	n.a.	n.a.	n.a.	2.1 (0.5)	1.9 (0.4)	1.8 (0.5)	2.0 (0.5)	ns	.30	ns	.21

Table 4-3 (*continued*)
Outcomes of first explorations of TIMSS student and teacher questionnaire data

Potentially effective educational factors and explored indicators in TIMSS data sets **(number of items and range of scale)**	**Statistics per educational system and pooled data set (weighted data)** [1]											
	Standardized Cronbach alpha [2]				M (sd)				Pearson r with math achievement [2]			
	Bfl	*Ger*	*Nld*	*Pooled data set*	*Bfl*	*Ger*	*Nld*	*Pooled data set*	*Bfl*	*Ger*	*Nld*	*Pooled data set*
CA_1 Teacher background												
a. Teacher's gender	n.a	n.a	n.a	n.a	female: 63%	female: 33%	female: 22%	female: 43%	ns	ns	- .21	- .15
b. Teaching experience in years [6]	n.a.	n.a.	n.a	n.a.	21.0 (10.0)	19.0 (8.2)	15.4 (8.6)	18.9 (9.4)	ns	ns	ns	ns
c. Teacher's work load Perc. of math lessons of total number of lessons per week [3]	n.a.	n.a.	n.a.	n.a.	72.5 (23.7)	51.5 (26.5)	86.2 (16.9)	70.2 (26.4)	ns	.43	.21	.30
CC_1 Class size [7]	n.a.	n.a.	n.a.	n.a.	19.3 (4.7)	24.5 (4.6)	24.8 (4.8)	22.3 (5.4)	.43	.33	.56	.16
CC_2 Material for evaluation (2 items; 1 - 4) [3]	n.a.	n.a.	n.a.	n.a.	- 0.7 (0.8)	- 1.1 (0.5)	- 0.2 (0.9)	- 0.7 (8.0)	ns	ns	ns	.10
CC_4 Grouping procedures.												
c. Cooperative learning (2 items; 1 - 4) [3]	r= .58	r=.51	r=.54	r=.60	1.5 (051)	1.9 (0.5)	2.3 (0.9)	1.9 (0.7)	ns	ns	ns	ns

Table 4-3 (*continued*)
Outcomes of first explorations of TIMSS student and teacher questionnaire data

Potentially effective educational factors and explored indicators in TIMSS data sets **(number of items and range of scale)**	**Statistics per educational system and pooled data set (weighted data)** [1]											
	Standardized Cronbach alpha [2]				M (sd)				Pearson r with math achievement [2]			
	Bfl	*Ger*	*Nld*	*Pooled data set*	*Bfl*	*Ger*	*Nld*	*Pooled data set*	*Bfl*	*Ger*	*Nld*	*Pooled data set*
CC_5 Teaching style student oriented (7 items; 1 - 4) [3]	.56	.55	.48	.53	1.3 (0.5)	1.2 (0.5)	1.0 (0.4)	1.2 (0.5)	- .16	- .14	ns	- .08
CC_6 Management and orderly and quiet atmosphere:												
a. Class climate (3 items; 1 - 4) [3]	.67	.74	.69	.71	2.6 (0.6)	2.3 (0.7)	2.3 (0.6)	2.5 (0.7)	.14	ns	.ns	.13
b. Safety at school (4 items; 1 - 4) [5]	.65	.75	.59	.68	3.7 (0.4)	3.6 (0.5)	3.7 (0.4)	3.7 (0.5)	.16	.12	.14	.17
c. Perceived limits in teaching mathematics												
1. Resource limits (6 items; 1 - 4) [5]	.78	.79	.73	.78	3.60 (.47)	3.31 (.57)	3.53 (.46)	3.50 (.51)	ns	ns	ns	.09
2. Student limits (6 items; 1 - 4) [8]	.81	.73	.68	.79	2.97 (.61)	2.81 (.50)	3.36 (.38)	3.03 (.56)	.27	.38	.29	.28
3. Parental limits (2 items; 1 - 4) [8]	ns	r=.21	r=.21	r=.19	3.38 (.57)	3.48 (.53)	3.81 (.37)	3.52 (.54)	.25	ns	ns	ns

Table 4-3 (*continued*)

Outcomes of first explorations of TIMSS student and teacher questionnaire data

Potentially effective educational factors and explored indicators in TIMSS data sets **(number of items and range of scale)**	**Statistics per educational system and pooled data set (weighted data)** [1]											
	Standardized Cronbach alpha [2]				M (sd)				Pearson r with math achievement [2]			
	Bfl	*Ger*	*Nld*	*Pooled data set*	*Bfl*	*Ger*	*Nld*	*Pooled data set*	*Bfl*	*Ger*	*Nld*	*Pooled data set*
CC_7 Homework												
a. frequency (1 item; 6 categories) [7]	n.a.	n.a.	n.a.	n.a.	3.1 (0.7)	4.1 (0.8)	3.9 (0.7)	3.6 (0.9)	ns	.33	.19	ns
b. amount of homework (1 item; 1-4) [9]	.40	.41	.55	.43	2.5 (0.5)	2.7 (0.5)	2.7 (0.6)	2.6 (0.5)	ns	ns	.18	ns
CC_14 Evaluation, feedback and corrective instruction												
Use of evaluation results (6 items; 1 - 4) [9]	.73	.60	.73	.64	2.9 (0.4)	2.8 (0.4)	2.8 (0.5)	2.8 (0.5)	ns	- .28	.23	ns
CO_1 Implemented curriculum content												
Content coverage mathematics (21 items) [7]	.73	.73	.79	.75	49.8 (14.7)	59.7 (15.0)	59.4 (16.6)	55.2 (16.0)	.31	ns	.37	.09

Notes: [1] number of students: Belgium Flanders (*Bfl*) = 2748; Germany (*Ger*) = 2020; Netherlands (*Nld*) = 1814; pooled data set = 6582 students

[2] n.a = reliability coefficient not applicable; ns = correlation coefficient non-significant ($p<.10$)

[3] differences in mean scores across the three systems are significant ($p<.01$)

[4] Bfl > Nld and Ger & Nld = Ger ($p<.01$)

[5] Bfl and Nld > Ger & Nld = Bfl ($p<.01$)

[6] Bfl > Nld & Bfl = Ger & Nld = Ger ($p<.01$)

[7] Bfl < Nld and Ger & Nld = Ger ($p<.01$)

[8] Bfl and Ger < Nld & Bfl = Ger ($p<.01$)

[9] None of the cross-national differences in mean scores are significant ($p<.01$)

Statistical reliability of the scales

The standardized Cronbach α was used as the measure for the psychometric reliability of each indicator operationalized in TIMSS by more than 2 items (in case of 2 items a Pearson product-moment correlation coefficient r was calculated). In Table 4-3, the standardized reliability coefficients are presented per education system and for the pooled data set. As can be seen, the results of the reliability analysis are satisfactory for some sets of items in all countries, for some sets of items the standardized Cronbach α differs across nations, and for other sets the coefficients are not satisfying in two or all three data sets.

The internal consistency of the scales of the three attitude indicators is sufficient in each of the three education systems (standardized Cronbach $\alpha > .64$). The same goes for the scales of the student (perceptual) indicators 'maternal academic expectation' (standardized Cronbach $\alpha > .67$), 'friends' academic expectation' (standardized Cronbach $\alpha > .82$), and 'class climate' (standardized Cronbach $\alpha > .67$).

The scales at the teacher level that showed a sufficient reliability coefficient in all countries are 'mathematics content coverage' (standardized Cronbach $\alpha > .73$), 'limits in teaching the tested class related to resources' (standardized Cronbach $\alpha > .73$), 'limits related to student characteristics' (standardized Cronbach $\alpha > .68$), and 'use of evaluation results' (standardized Cronbach $\alpha > .60$).

For some indicators, the reliability coefficient Cronbach α differs substantially across the three education systems. The range of the coefficients across the three systems is at least .10 for 'maternal academic expectation,' 'friends' academic expectation' (in each country the coefficient of these two indicators was sufficient), 'safety at school as perceived by the student,' 'limits the teacher experienced in teaching mathematics related to student characteristics,' 'treatment of completed homework in next lesson,' and 'the frequency evaluation results are used on average for six different goals.'

Rather low reliability coefficients (standardized Cronbach α lower than .60) were found in 2 or all 3 of the systems under review for two student variables, 'out-of-school activities related to leisure time' and 'student-oriented teaching style as perceived by the student,' and for the classroom variable 'treatment of homework in next lesson'.

Because of the low internal consistencies of these scales, it is doubtful whether it is worthwhile to insert these indicators in the exploratory path analysis. However, the selection of the indicators for the PLS path models was not based only on the reliability coefficient. The most important criterion for selection was the bivariate correlation coefficient of the indicator with the dependent variable 'mathematics achievement' (see below).

Descriptives

In Table 4-3, the weighted mean and standard deviation of each explored indicator (variable) are presented per system and for the pooled data set. The results of the tests of the differences in mean scores across the three systems are also presented (pair wise comparisons and significance determined by Bonferroni adjustment). The mean score on the majority of variables measured at student level differ across all pairs of countries. For example, the mean score of out-of-school activities related to leisure time differs significantly ($p < .01$) across Belgium Flanders and Germany and the Netherlands and across Germany and the Netherlands. For two variables, the mean scores do not differ across all pairs of countries: 'success attribution' and 'perceived safety at school'. In Germany, the mean score on these two variables is significantly lower than in Belgium Flanders and the Netherlands. The latter two do not differ on the average score for the two variables.

Some significant differences are more relevant than others. A significant difference of .10 in mean score on a four-point scale on variable SA_3a (paid jobs) between Germany and the Netherlands is less relevant (yet statistically significant) than a difference of .40 (Belgium Flanders and Germany) on a four-point scale of variable SA_3b (leisure time activities).

The mean scores of some of the indicators of factors at classroom level differ significantly across all education systems (e.g., time on task and teacher's workload). For other variables, the mean score differs across countries. In Table 4-3 the various patterns of pair wise country differences in mean scores are indicated by notes. For instance, the mean score on 'amount of homework' is in Belgium Flanders significantly higher than in the Netherlands and Germany, while the Netherlands and Germany do not differ on this variable. Another example is 'perceived limitations in teaching mathematics due to student features'. In the

Netherlands, teachers perceive on average less limitations than in the other two countries. The mean score on this variable does not differ significantly across Belgium Flanders and Germany

Correlational analysis

The data explorations of stage A were finalized by considering the Pearson product moment correlation coefficients calculated between the variables listed in Table 4-3. The bivariate correlations of all variables with mathematics achievement were analyzed first to gain insight in the potential strengths of direct relationships. Next, the intercorrelations between all variables which showed a significant correlation with mathematics achievement in at least two out of the three education systems ($p < .10$) were calculated. The intercorrelation coefficients were studied to select potential indirect relationships between variables and mathematics achievement which could be inserted in the PLSpath model (see stage B).

The bivariate correlation coefficients between the identified indicators (variables) and the dependent variable 'mathematics achievement' are presented in the final four columns of Table 4-3. The majority of these coefficients are lower than .20 in the separate education systems and in the pooled data set. Variables with a relatively high correlation with mathematics achievement ($r > .20$) in two or three countries are 'number of books at home,' 'teacher's workload,' 'class size,' 'mathematics content coverage,' 'limits in teaching related to student characteristics,' and 'use of evaluation results.'

As stated, the problem of too many missing values for the most appropriate indicator of 'student's home educational background' available in the TIMSS data set – the educational level of the parents – was given in by replacing this indicator with a proxy indicator: 'number of books in the student's home.' The meaning of this indicator is not equal to 'educational level of the parents,' but a better one was not available in the data set. In Table 4-3, it can be seen that the indicator 'number of books' correlates relatively highly with 'mathematics achievement' in Germany ($r = .34$) the Netherlands ($r = .30$), but the correlation coefficient is modest in Belgium Flanders ($r = .13$).

Intercorrelations between all selected variables at student and classroom levels with $|r > .15|$ were posed as indirect links in the initial path model in stage B (Bos, 2001).

4.5 STAGE B: RESULTS OF EXPLORATORY PATH ANALYSIS (PLSPATH)

Two unidimensional path models were explored by means of PLSpath, one at the student level and one at the aggregated classroom level. The first model depicts the interrelationship of students' mathematics achievement outcomes and student variables: the between-student model. In the second model, some classroom variables were added to the student model at the aggregated level: between-classroom model. Both models were explored separately for the three education systems and the pooled data set. The final between-classroom models are presented. The interim results of PLSpath analyses, including the results of the exploration of the student path model, can be found in Bos (2001).
In Figure 4-1, the final recursive, between-classroom model is presented, which was tested for all education systems.

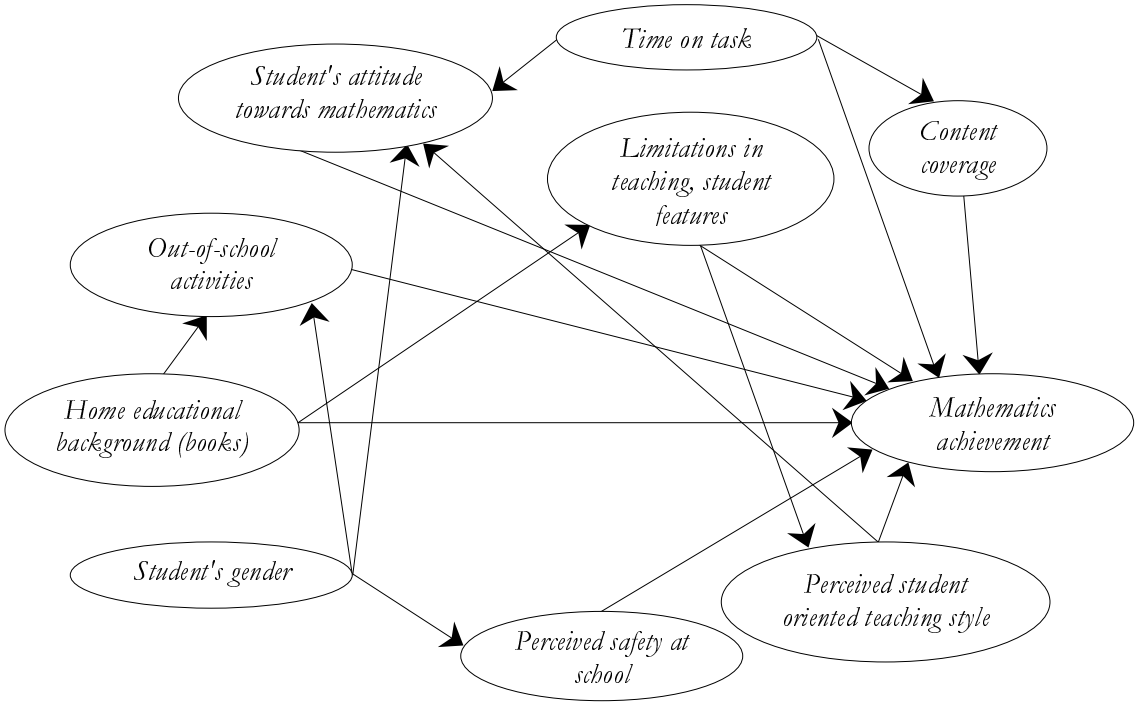


Figure 4-1
Final recursive between-classroom path model (including aggregated student variables)

In PLSpath the proportion of latent variables and number of cases should be equal to 1 : 15 (Campbell, 1996; Sellin, 1989). In the Dutch data set of mathematics teachers in grade 8, the number of cases is limited to 88. This means that for 88 classrooms, both teacher and student data were available. For the other two

education systems, the number of classrooms in the data sets are 94 (Germany) and 147 (Belgium Flanders). Consequently, the number of latent variables to be inserted in the between-classroom model could not exceed nine. The final between-classroom model contains six student variables (aggregated at classroom level) and three classroom variables. In fact, the aggregated student variables can be regarded as classroom characteristics. The aggregated scores were per classroom assigned to each student of that classroom. In the between-classroom model, mathematics achievement is the aggregated score per classroom on the TIMSS mathematics test.
The arrows in the model show the relationships across factors (latent variables, LVs) that were explored. It is assumed that some of the student input factors are interrelated. For example, 'home educational background' (indicated by number of books in the home) is assumed to be related to 'out-of-school activities related to leisure time' and 'student's gender' (percentage of girls in the classroom) is assumed to be related to both 'out-of-school activities' and 'student's attitude towards mathematics'. 'Student's gender' is the only factor which is not directly related to mathematics achievement, indicated by the very low correlation coefficient between these two variables found in all data sets.
Bivariate correlations between classroom variables and mathematics achievement of at least r=.15 and common sense rationales, were the foundations for the selection of three teacher variables:

- Time on task (opportunities used) operationalized as the total number of minutes mathematics scheduled per week.
- Coverage of the mathematics topics tested in the TIMSS test to the students before the date of test administration.
- Perceived limits in teaching mathematics related to student characteristics as perceived by the teacher.

These three LVs are thought to be directly linked to mathematics achievement. In addition, the effect of 'time on task' on achievement is also thought to be indirect. The intermediate variables are 'student's attitude towards mathematics' and 'content coverage.' The total effect of 'perceived limits in teaching related to student features' is assumed to be composed of an indirect effect as well. Here the intermediate variable is the 'perceived student oriented teaching style'. If the teacher experiences limitations in his or her teaching related to student behavior, (s)he might be less likely to structure lessons from a student-oriented point of view. In these cases, the teacher-centered style would be more appropriate, according to many teachers.

The outer between-classroom model

The results of the outer between-classroom model for the separate data sets are relatively straightforward. The loadings of the two manifest variables (MVs) which reflect the LV 'student's attitude towards mathematics' (aggregated at classroom level) are in each data set above .40 (see Appendix B) and thus sufficient. The other LV which was estimated through more than one MV is 'time on task.' In two data sets, the loading of either one of these two is too low. In Belgium Flanders the loading of 'amount of homework per time' is .15 and in Germany the loading of 'number of mathematics minutes scheduled per week' is -.04. In the next step of PLSpath analyses, these outcomes were taken into account. In all countries except Germany, 'number of mathematics minutes scheduled per week' was selected as the MV for 'time on task.' In Germany, 'amount of homework' was selected as the MV for 'time on task'. The other LVs were estimated by one MV (unity).

The final outer between-classroom model met all criteria set in advance. The percentage of explained variance of one LV by all MVs included in the estimation of that LV (communality) is high enough in each data set. The redundancy is low enough for each MV in each country, meaning that multicollinearity between an MV and the LVs to which the MV is indirectly linked is low. The tolerance index is not higher than .50 for any of the MVs included. This means that in none of the blocks of MVs multicollinearity exists.

The inner between-classroom model

The final inner model results for the three education systems and for the pooled data set are presented in Table 4-4 and in Table 4-5. In Appendix C, Figure 4-1 is completed with path coefficients shown in Table 4-4 for the three systems and the pooled data set.

In Table 4-4, the direct and total effects of the LVs on mathematics achievement are given. The strength of the effects differs across nations per LV. For instance, the effect of average student's attitude towards mathematics is non-existent in the Netherlands (path coefficient β is lower than .10), moderate in Germany (with a negative direction: β =-.19) and somewhat larger in Belgium Flanders (β =+.23). The negative direction of the effect of 'attitude' in Germany indicates that the more positive students' attitude within one classroom is, the lower the average TIMSS mathematics test score of the classroom. The opposite direction of this effect would have been expected in all countries.

Another example of different effect sizes across countries concerns the LV 'home educational background.' In Germany and the Netherlands, the effect of this LV on mathematics achievement is much larger (total effect β >.60) than in Belgium Flanders (total effect β =.25). This difference might be caused by the different ways the question in the student questionnaire was interpreted by students in Belgium Flanders as opposed to the students in the other two countries.

Table 4-4

Direct effects (path coefficient β) and total effects on mathematics achievement in final between-classroom path model

	Direct effects/total (direct + indirect) effects							
	Pooled data set		*Belgium Flanders*		*Germany*		*Netherlands*	
Latent Variable	Direct	Total	Direct	Total	Direct	Total	Direct	Total
- * Percentage of girls in classroom	----	- .14	----	.14	----	n.s.	----	.10
- * Out-of-school leisure time activities	- .41	- .41	- .33	-.33	n.s.	n.s.	- .19	- .20
- * Number of books at home	.27	.42	.15	.25	.65	.67	.48	.63
- * Student's attitude towards mathematics	n.s.	n.s.	.23	.23	- .19	-.19	n.s.	n.s.
- Time on task/ opportunities used	.23	.20	n.s.	.14	n.s.	n.s.	.13	.12
- Content coverage mathematics	.10	.10	.19	.19	n.s.	n.s.	.13	.13
- * Level of student oriented teaching style as perceived by the students	- .17	- .17	- .24	-.22	n.s.	n.s.	n.s.	n.s.
- * Safety at school as perceived by students	.15	.15	.19	.19	.15	.15	.27	.27
- Limitations in teaching class related to student behavior	.14	.18	n.s.	n.s.	.13	.14	n.s.	n.s.
Percent variance explained in mathematics achievement	63.7		63.2		72.5		73.0	
Average percent explained variance by model	18.0		18.2		22.8		21.6	

Notes: * = aggregated student variable; n.s.= non-significant path coefficient (β <.10)

The final two rows of Table 4-4 contain information regarding the fit of the models to the data sets (after Lietz, 1996; see 4.3). The final models explain between 63.2% (Germany) and 73.0% (the Netherlands) of the variance in mathematics achievement scores. In the final row of Table 4-4, an indication is given of the predictive power (strength) of the inner and outer relationships in the between-student model: the average multiple R^2. The average multiple R^2 does not differ much across the models, as it varies between 18.0 and 22.8. These outcomes can be interpreted as a rather moderate model fit to the data.
The direct effects of aggregated student and classroom LVs on endogenous factors in the final between-classroom path model are presented in Table 4-5.

Table 4-5
Direct effects of student LVs on endogenous factors in final between-classroom path model per data set

	Effects on endogenous latent variable:					
Effects of Influencing Latent Variable	*Students' attitude*	*Perceived safety*	*Out-of-school activities*	*Experienced limits in teaching*	*Content coverage mathematics*	*Student oriented teaching style*
* Percentage of girls in classroom	-.17 Pool -.16 Bfl -.16 Ger -.21 Nld	.47 Pool .57 Bfl .36 Ger .30 Nld	-.17 Pool -.23 Bfl -.36 Ger -.20 Nld	n.a.	n.a.	n.a.
* Home educational background	n.a.	n.a.	-.29 Pool -.26 Bfl -.29 Ger -.72 Nld	.15 Pool .21 Bfl .30 Ger .18 Nld	n.a.	n.a.
Time on task/ opportunities used	.32 Bfl	n.a.	n.a.	n.a.	-.24 Pool -.10 Nld	n.a.
* Level of student oriented teaching style as perceived by the students	.29 Pool .55 Ger .11 Nld	n.a.	n.a.	n.a.	n.a.	n.a.
Limitations in teaching class related to student features	n.a.	n.a.	n.a.	n.a.	n.a.	-.24 Pool -.11 Bfl -.12 Ger .10 Nld

Notes: * = aggregated student variable; β >.10 are presented only; n.a. = not applicable;
pool = pooled data set; Bfl = Belgium Flanders, Ger = Germany, Nld = the Netherlands

Many effects of LVs on endogenous LVs have a β coefficient of higher than .10 (the lower bound above which the path coefficient is regarded as different from zero) in all data sets. The strength of the effects differs across the three education systems. For example, the effect of 'percentage of girls in the classroom' on 'out-of-school activities related to leisure time' in Germany is greater (β = -.36) than in Belgium Flanders and the Netherlands (β is around -.20). The difference in the coefficient of the effect of 'percentage of girls in the classroom' on 'safety at school as perceived by the students' between Belgium Flanders on the one side and Germany and the Netherlands on the other side, is great: the difference is about .20. In Table 4-5, it can also be seen that some assumed effects exist in only one or two data sets. This is particularly true for the effect of the classroom variable 'time on task.' The effect of this LV on 'students' attitude' is different from zero in Belgium Flanders (β = .32) but does not exist in the models of Germany and the Netherlands. The effect size of 'home educational background' on 'out-of-school activities' in the Netherlands is stronger than in the other two systems. This could also be seen in the final inner between-*student* models, because in the inner between-classroom model the LVs on student level have been aggregated. The aggregation of the student scores to classroom scores does not change the direction and differences in magnitude of the effect sizes across the systems.

Selected variables from PLSpath results

The PLSpath results of each education system were studied to select variables for inclusion in the multilevel analysis. The latent variables (LVs) which showed a direct effect (i.e., path coefficient β >|.10|) on mathematics achievement in the final between-student and between-classroom model, respectively, were selected. Thus, the differences in PLSpath outcomes across the three systems regarding the direct effects on mathematics achievement are reflected in the different sets of selected LVs. In Table 4-6 (in section 4.6 which shows results of multilevel analysis), per education system the non-selected LVs are indicated by 'a.'
For instance, PLSpath results showed that 'out-of-school activities related to leisure time' did not have an effect on mathematics achievement for Germany, but it did have an effect for Belgium Flanders and the Netherlands. Consequently, in the MLn analysis for Germany this variable was not inserted, although it was inserted in the analysis for the other two systems.

Possible interaction effects of level-2 variables on level-1 variables were selected from the PLSpath results as well. Particularly, the direct effects of LVs on *endogenous* LVs within the between-students model and within the between-classroom model were selected. An example of a possible interaction effect is the effect of 'percentage of girls in the classroom' (level-2 variable) on the effect of 'attitude towards mathematics' on mathematics achievement. From the final between-classroom model, it was concluded that in all countries the percentage of girls in the classroom has a negative effect on the attitude variable. In addition, in two countries (Belgium Flanders and the Netherlands) 'attitude' turned out to have a direct effect on mathematics achievement in the between-student path model (Bos , 2001). These two results from PLSpath analysis are the reason to include the possible interaction effect in the MLn analysis of Belgium Flanders and the Netherlands.

4.6 STAGE C: RESULTS OF MULTILEVEL ANALYSIS (MLN)

In stage C, the multilevel analysis was carried out per education system and on the pooled data set. The analysis on the three separate country data sets addresses research question Ia, and the analysis on the pooled data set addresses research question Ib. These questions asks for the identification of student and classroom factors measured in TIMSS, which are associated with mathematics achievement in each of the three countries (see 3.1).

In the analysis per education system, factors were included which were selected from the PLSpath results. For each system, different factors resulted from the unidimensional path analysis. As a consequence, the multilevel analyses show different results per education system as well. The results of the analyses on the *weighted, standardized* data are presented. From the perspective of research question Ia, these results are most relevant. The coefficients of the predictors from the 2-level model per system based on the weighted, *standardized* data can be compared *within* a system but not across systems (Hox, 1994).

In a subsequent step, a hierarchical linear model was estimated for the pooled data set. The variables included in this analysis are the ones which show bivariate correlations with mathematics achievement higher than .15 in at least two out of the three countries. Hence, the list of variables both at student and classroom level is longer than the list of selected variables after the PLSpath results.

The results of the pooled model are presented in the final part of this section and indicate whether student and classroom variables have effects within and across the three systems.

Hierarchical linear model per education system

Table 4-6 presents, for each system, the intercepts and the γ-coefficients of the predictors inserted in level-1 and level-2.

Table 4-6

Final estimation of fixed effects in 2-level model on mathematics achievement per education system; weighted, standardized data; γ-coefficient

		Coefficient per system		
Level	**Fixed effect**	*Bfl*	*Ger*	*Nld*
1. Student	*Intercept*	.01 n.s.	.00 n.s.	.03 n.s.
	Student's gender	a	a	a
	Out-of-school activities related to leisure time	n.s.	a	-.02
	Home educational background	.03	.10	.04
	Attitude towards mathematics	.14	a	.16
	Student oriented teaching style as perceived by the student	-.06	n.s.	a
	Level of safety in the school as perceived by the student	.02	n.s.	n.s.
2. Classroom	* Percentage of girls in classroom	a	a	a
	* Out-of-school activities related to leisure time	-.23	a	-.11
	* Home educational background	.08	.39	.34
	* Attitude towards mathematics	.09	-.11	a
	Time on task	a	a	.08
	Content coverage mathematics	.13	a	.09
	* Student oriented teaching style as perceived by the student	-.13	a	a
	* Level of safety in the school as perceived by the student	.12	.10	.20
	Limits in teaching related to student features	a	.10	a

Notes: * aggregated student variables at classroom level;
a not included in the model due to PLSPath results of the education system;
γ-coefficients significant ($p<.05$; two tailed tested); n.s. effect not significant in final model

In Belgium Flanders, four out of the five selected student-level variables have a fixed effect on mathematics achievement: home educational background, attitude towards mathematics, student oriented teaching style as perceived by the students, and safety at school. At the classroom level, the model of Belgium Flanders shows six variables (out of six) with a significant effect (e.g., out-of-school activities, attitude, and content coverage).
In Germany, only one out of three student-level variables has a significant effect: home educational background. At classroom level, four variables were selected from the PLSpath results and all of them turned out to have a significant effect on mathematics achievement.
In the Netherlands, three out of the four selected student-level variables have a significant effect: out-of-school activities, home educational background and attitude towards mathematics. The five selected classroom variables have a significant effect. Examples of these effective variables are average home educational background of the classroom and the level of safety in the school as perceived by all students within the classroom.
As for the possible interaction effects which were inserted in the multilevel analysis, none of these were meaningful. Therefore interaction effects are regarded as non-existent in the hierarchical models of the three countries.

Explained proportion of variance by the two-level model per education system
The first columns of Table 4-7 show the proportion of variance that can be explained from the fully unconditional models at both student and classroom level.
The explained variance at student level is labeled as 'total' explained variance because students are nested in classrooms. The variance at the classroom level is not only related to variances in classroom variables. Students within the same classroom are more similar to one another than they are to students from different classrooms. Thus, student background variables such as 'home educational background' and 'perceived safety in school' are also to be seen as classroom variables. The effects of variables located at the student level on mathematics achievement are combined student and classroom effects.
Furthermore, Table 4-7 shows, per education system, the proportion of variance that was explained after all of the selected *student* variables were added (the resulting model is called the unconditional level-2 model). Columns six and seven show the

added proportion of variance that was explained after adding the level-2 variables to the unconditional level-2 models. Finally, the proportion of explained variance that was explained by the final level-2 model is presented.
The difference in deviance of the successive models is significant (p<.001) for all education systems. In all three systems, in the final model the proportion of explained variance between classrooms is extensive. The percentage of explained variance by variables at classroom level varies from 65% (Belgium Flanders) to 72% (the Netherlands).

Table 4-7

Proportion of variance in mathematics achievement explained at student and classroom level in fully unconditional 2-level model and final level-2 model per education system

	PROPORTION OF VARIANCE							
	To be explained from fully unconditional model		**Explained by adding level-1 variables**		**Explained by adding level-2 variables**		**Explained by final level-2 model**	
Education system	*between students*	*between class-rooms*	*'total'*	*between class-rooms*	*'total'*	*between class-rooms*	*'total'*	*between class-rooms*
Belgium Flanders	58%	42%	10%	15%	21%	50%	31%	65%
Difference in deviance	127.2 (df = 4)		122.2 (df = 6)				249.4 (df = 10) [1]	
Germany	56%	44%	7%	13%	27%	58%	34%	71%
Difference in deviance	37.0 (df = 1)		105.9 (df = 4)				142.9 (df = 5) [1]	
The Netherlands	48%	52%	6%	8%	34%	64%	42%	72%
Difference in deviance	96.9 (df = 3)		106.0 (df = 5)				202.9 (df = 8) [1]	

Notes: all differences in deviance significant (p < .001); df = degrees of freedom

[1] difference in deviance between fully unconditional and final level-2 model

In Figure 4-2, the total variance of students' scores on the TIMSS mathematics test in each education system is decomposed into an explained and unexplained part at both student and classroom level in the final model.

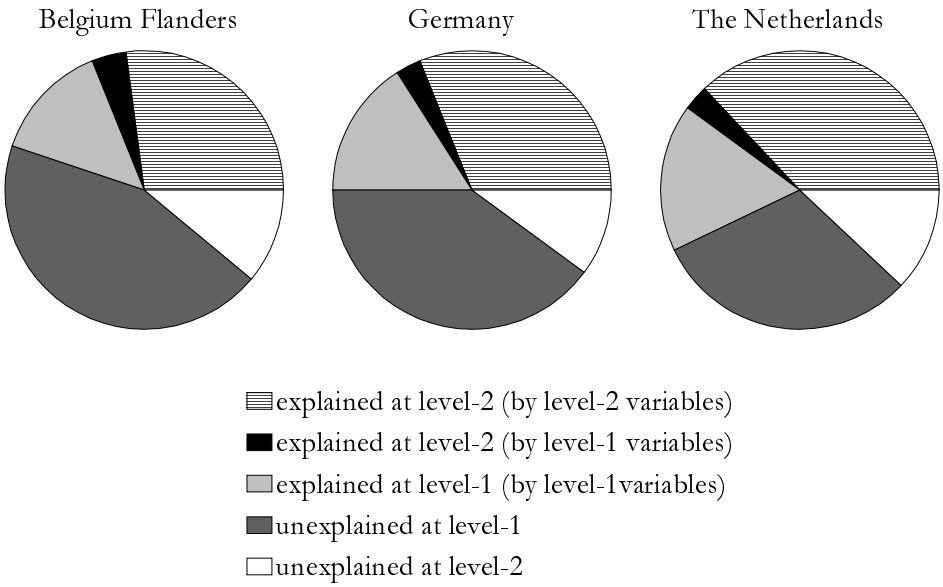


Figure 4-2
Proportion of variance in students' mathematics achievement scores explained respectively unexplained at student (level-1) and classroom level (level-2) in final level-2 models; Belgium Flanders, Germany and the Netherlands

The fully unconditional model (empty model) provides partition of the variability in the data between student- and classroom-level. For example, the empty model for Belgium Flanders showed that 58% of the variance in student ac hievement scores is located at student level and 42% at classroom level. The proportion of variance at classroom level that was explained in the final model for Belgium Flanders is 65%. Thus, 27% (65% x 42) of the variance at classroom level was explained by level-2 variables and the remaining part of 15% was partly explained by student level variables and the rest was unexplained by the model. The other percentages that are shown in Figure 4-2 for Belgium Flanders and the other countries were calculated in the same way.
It can be seen that the proportion of explained respectively unexplained variances at student and classroom level differ slightly across the three countries. In the model for the Netherlands, level-2 variables tie a relatively greater amoun t of the variance in achievement scores (37%) than in the model for the other two countries (31 % in Germany and 27% in Belgium Flanders). At student level, the total proportion of explained variance is almost the same in the three countries (18%, 19% resp ectively 20%).

Hierarchical linear model estimated on pooled data set

The pooled data set includes all three countries and was weighted by the senate weight (see 4.3) in order to equalize the contribution of students from each country in the data set. Table 4-8 shows the results of the estimation of hierarchical linear models on the pooled data set. First, a model was estimated without identification of students by their country (model 1). Thereafter, models 2 and 3 were estimated in which students were identified by their country by means of a dummy variable (e.g., the students from Germany were assigned code '1' on their dummy variable and all of the others students were assigned code '0' on the same dummy variable). The results of the final models 1 and 3 were compared to find variables that possibly could explain differences in mathematics achievement across the three education systems.

Model 0 contains none of the student and classroom variables and is called the fully unconditional model. In model 1, the final set of student and classroom variables are included that have a significant effect just after they were included within the step up method on which the model was built. Some of the variables in the final model turned out to be non-significant ($p<.05$) after adding one or more of the other variables. In Table 4-8, the coefficients of these non-effective variables are indicated by 'n.s.' between brackets. Variables that have no significant effect *just after* their inclusion to the model are indicated by 'n.s.' without brackets.

The student variable with the greatest effect in pooled model 1 (without dummy variables for countries) is 'attitude towards mathematics' (γ-coefficient = .16). Other student variables with an effect on mathematics achievement are 'student's gender,' 'home educational background,' 'friends' academic expectation,' 'perceived safety in school,' 'working hard doing homework,' and 'student oriented teaching style as perceived by the student.' Three student factors had no significant effect on mathematics achievement of students in the pooled data set: 'out-of-school activities related to leisure time,' 'maternal academic expectation,' and 'class climate.' At the classroom level, the variables with relatively strong effects (γ-coefficient $> |.10|$) are three aggregated student variables: 'home educational background,' 'out-of-school activities related to leisure time,' and 'perceived safety in school.' The influence of the composition of the classroom with respect to student social background variables ('home educational background' and 'out-of-school activities') on mathematics achievement seems more important than the scores of individual students on these variables. Genuine classroom variables with a relatively strong effect are 'time on task,' 'teacher's workload,' and 'content coverage mathematics.'

Table 4-8

Final estimation of fixed effects in 2-level models on mathematics achievement in pooled data set; weighted, standardized data; γ-coefficient

	Model				
Level	*Model 0*	*Model 1*	*Model 2*	*Model 3*	*Model 1a*
Intercept	-.05	-.01	-.12	-.10	.02
Belgium Flanders			.40	.45	
Germany			-.36	-.27	
1. Student					
Student's gender		.07		.07	.07
Home educational background		.05		.05	.05
Out-of-school activities related to leisure time		-.01 (n.s.)		-.01 (n.s.)	-.01 (n.s.)
Maternal academic expectation		n.s.		n.s.	
Friends' academic expectation		-.07		-.07	-.07
Perceived safety in school		.02		.02	.02
Class climate		n.s.		n.s.	
Attitude towards mathematics		.16		.16	.16
Working hard doing homework		.06		.06	.06
Student oriented teaching style as perceived by the student		-.04		-.04	-.04
2. Classroom					
Class size		n.s.		.13	.03 (n.s.)
* Percentage of girls in classroom		-.05 (n.s.)		-.04 (n.s.)	
* Home educational background		.16		.19	.14
* Out-of-school activities related to leisure time		-.28		-.19	-.37
* Perceived safety in school		.12		.15	.12
* Class climate		.09		n.s.	
* Attitude towards mathematics		-.08		n.s.	
* Student oriented teaching style as perceived by the student		-.08		n.s.	
Teacher's gender		n.s.		n.s.	
Limits in teaching related to student features		.06		.07	.07
Time on task		.11		n.s.	
Teacher's workload		.15		.10	.16
Content coverage mathematics		.08		.09	.06
Homework frequency		n.s.		n.s.	
Amount of homework		.07		n.s.	
Kind of tests		n.s.		n.s.	
Use of assessment results		n.s.		n.s.	

Notes: * aggregated student variables at classroom level;

γ-coefficients significant ($p < .05$; two tailed tested); n.s. = fixed effect not significant

Model 2 is the empty model in which two dummy variables were included for the education systems 'Belgium Flanders' and 'Germany.' The final model with these dummy variables is model 3. Model 3 was estimated with the same list of variables used for the estimation of model 1 in which dummy variables for the countries were not included.

In final model 3, the list of student and classroom variables with a significant effect ($p<.05$) is essentially the same as the one of final model 1. In both models the same *student* variables turned out to have an effect on mathematics achievement with identical coefficients. Hence, in every country, each of the effective student variables contributes to the explanation of variance in student achievement scores. The effect of some of the classroom variables are different across model 1 and model 3. In model 3, 'class size' is a relatively strong factor (.13), and in model 1 this factor has no significant effect. This difference indicates that the effect of class size on achievement exists within each of the countries. To interpret this result the frequency table must be consulted (see Table 4- 3). The mean class size in Belgium Flanders is significantly lower than in Germany and the Netherlands. The two latter have equal mean class sizes in grade 8. In Table 4-3, it can also been seen that the bivariate Pearson correlation coefficient between class size and mathematics achievement is rather high in all countries. It is also known that Belgium Flanders outperformed the other two countries on the TIMSS mathematics achievement test (see Table 4-1). Considering the effect of class size and the frequency and correlation results of the three countries, it can be assumed that class size influences mathematics achievement and that smaller classes (Belgium Flanders) might be enhancing achievement more than larger classes. As stated in a previous section, a third variable might influence the relationship between class size and student achievement.

The estimated models show more variables that are important to consider within and across countries. Six other classroom level variables than class size show an effect in model 1, but they show no effect in model 3: 'class climate' (aggregated student variable), 'attitude towards mathematics' (aggregated student variable), 'student oriented teaching style as perceived by the student' (aggregated student variable), 'time on task,' and 'amount of homework.' With respect to the influence of these variables on mathematics achievement, the country in which the students live seems unimportant. For example, the results of model 3 show that within countries, the average score per classroom on 'class climate as perceived by all students' has no effect on mathematics achievement.

The classroom variables 'time on task' and 'amount of homework' have an effect in the pooled data set without the identification of the countries. The effect disappears in model 3. However, the mean 'minutes of instructional mathematics time' differs greatly across the three systems (see Table 4-3). The mean score is the lowest in the Netherlands (149 minutes), and the highest in Belgium Flanders (224 minutes). Also, in Belgium Flanders and the Netherlands, the bivariate correlation between 'time on task' and 'mathematics achievement' is significantly different from '0' and positive, and Belgium Flanders outperformed the Netherlands on the TIMSS mathematics test. What could these results mean for the enhancement of mathematics achievement in the countries? It might be assumed that an increase in the number of minutes of mathematics per week effects student's achievement in mathematics positively.

For some other classroom variables an effect is shown in both model 1 and model 3. This indicates that the variables seem effective within separate countries. These variables are 'home educational background,' 'out-of-school activities related to leisure time,' 'perceived safety in school,' 'limitations in teaching related to student features,' 'teacher's workload,' and 'content coverage mathematics.'

Five out of these six variables show an 'increasing' effect from model 1 to model 3, varying from .01 to .09, indicating that within the three countries these variables might be even more effective than across the countries. For example, 'limitations in teaching related to student features' show an effect of .06 in model 1 and of .07 in model 3. Both within and across countries, this variable shows an effect on mathematics achievement.

One other variable shows a decrease in coefficient from model 1 (.15) to model 3 (.10): 'teacher's workload.' The decrease indicates that in each country, the percentage of mathematics lessons a teacher is assigned relates positively to students' achievement level in mathematics. The interpretation of this result can differ across the three systems if the frequency table in Table 4-3 is taken into account. Table 4-3 shows different means for the Netherlands (86.2), Belgium Flanders (72.5), and Germany (51.5). In all countries, increasing the teacher's mathematics workload relates positively to student achievement. The separate countries can decide to stimulate teachers to spend their time in school mainly to

mathematics, particularly in Germany. In the Netherlands and Belgium Flanders the mean 'teacher's load' is rather high. For these two systems, this statistic might be a reason not to stimulate the teacher's load with regard to mathematics.
The comparison between final model 1 and final model 3, in relation to model 0, can be made in a more accurate way if model 1 is estimated with only the variables that turned out to have a significant effect in model 3. The resulting model is labeled 'model 1a' (see final column of Table 4-8).
With respect to the student variables, the results of model 1a and model 1 are the same. The differences in effect size of the seven variables inserted at classroom level between model 1a and model 3 differ slightly from the differences described between model 1 and model 3, but only in strength not in direction.

Proportion of variance explained by the two-level model 1, model 1a, and model 3
Table 4-9 shows for model 1, model 1a, and for model 3, respectively, the proportion of variance that was explained at the student and classroom levels. The percentage of variance that could be explained at the two levels is presented first. Thereafter, the percentage of variance explained after the dummy variables were included is presented (only applicable for model 3). Next, the percentage of variance explained after all of the selected *student* variables were added is shown (the resulting model is called the unconditional level-2 model). Columns eight and nine show the proportion of variance at the student and classroom levels that was explained after adding the level-2 variables to the unconditional level-2 model. Finally, the proportion of variance at the two levels that was explained by the final level-2 model is presented.
The variance at the classroom level is not only related to variances in classroom variables. Students within the same classroom are more similar to one another than they are to students from different classrooms. Thus, student background variables such as 'home educational background' and 'perceived safety in school' are also to be seen as classroom variables. The effects of variables located at the student level on mathematics achievement are combined student and classroom effects. Therefore, the label 'total' is inserted for the variance explained by level-1 (student) variables.

Table 4-9

Proportion of variance in students' mathematics achievement scores explained at student and classroom level in fully unconditional 2-level model and final level-2 model of model 1 and model 1a (without dummy variables for countries) and model 3 (with dummy variables for countries); pooled data set

	PROPORTION OF VARIANCE									
	To be explained from fully unconditional model 0		**Explained by adding dummy variables for countries [1)]**		**Explained by adding level-1 variables**		**Explained by adding level-2 variables**		**Explained by final level-2 model**	
Pooled data set	*between students*	*between classrooms*	*'total'*	*between classrooms*	*'total'*	*between classrooms*	*'total'*	*Between classrooms*	*'total'*	*Between classrooms*
Model 1 (without dummy variables)	49%	51%	n.a.	n.a.	8%	9%	33%	60%	41%	69%
Difference in deviance	n.a.		493.8 (df=8)		364.6 (df=12)				858.4 (df=20) [2)]	
Model 1a (without dummy variables)	49%	51%	n.a.	n.a.	8%	9%	30%	56%	38%	65%
Difference in deviance	n.a.		493.8 (df=8)		234.8 (df=7)				728.6 (df=15) [2)]	
Model 3 (with dummy variables)	49%	51%	10%	19%	0%	11%	33%	44%	43%	74%
Difference in deviance	69.9 (df=2)		508.6 (df=8)		327.9 (df=8)				905.4 (df=18) [2)]	

Notes: all differences in deviance significant ($p < .001$); df = degrees of freedom

[1)] n.a. = not applicable; [2)] difference in deviance between fully unconditional model 0 and final level-2 model

Model 1, model 1a, and model 3 were improved by adding level-1 variables and were improved again by adding level-2 variables. These improvements are indicated per model by the significant differences in deviances ($p<.001$) shown for model 1, model 1a, and model 3 in Table 4-9.

The total percentage of variance explained by the final model 1 at the classroom level is 69% and at the student level it is 41%. The inclusion of student variables in the model explains only 8% of the variance at the student level and 9% at the classroom level. By adding the aggregated student-classroom variables and the genuine classroom variables, the percentage of explained variance is increased by 33% at student level and 60% at the classroom level. In Figure 4-3, the proportions of variances at the two levels that could be explained by the final model are shown (see below). Obviously, the classroom variables and the aggregated student variables contribute a great deal to the proportion of explained variance in mathematics achievement in the pooled data set. The proportions of variance explained at the different levels by model 1a (with only the variables which showed a significant effect in model 3) are a little smaller than in model 1.
Table 4-9 also presents the proportion of variances for model 3, that includes the dummy variables for the countries. The total percentage of variance explained by the final model 3 is 74% at the classroom level and at the student level it is 43%. The increase in proportion of explained variance from final model 1a to final model 3 indicates that country specific factors might play a role in relation to cross-national differences in student achievement. The identification of the country the students live in tie variance in achievement scores between classrooms.
The fully unconditional model (model 0) provides partition of the variability in the data between student-level (49%) and classroom-level (51%). In Figure 4-3, the total variance of students' scores on the TIMSS mathematics test in model 1a and model 3 is decomposed into an explained and unexplained part at both student (level-1) and classroom level (level-2). For example, the proportion of variance at classroom level that was explained in the final model 1a without dummy variables for the countries is 65%. Thus, 33% (65% x 51%) of the variance at classroom level was explained by level-2 variables The remaining part of 17% (51 – 33) was partly explained by level-1 variables and the rest was unexplained by the data. The other parts that are shown in Figure 4-3 were calculated in the same way. It can be seen that the proportion of explained respectively unexplained variances at student and classroom level differ across the two models.

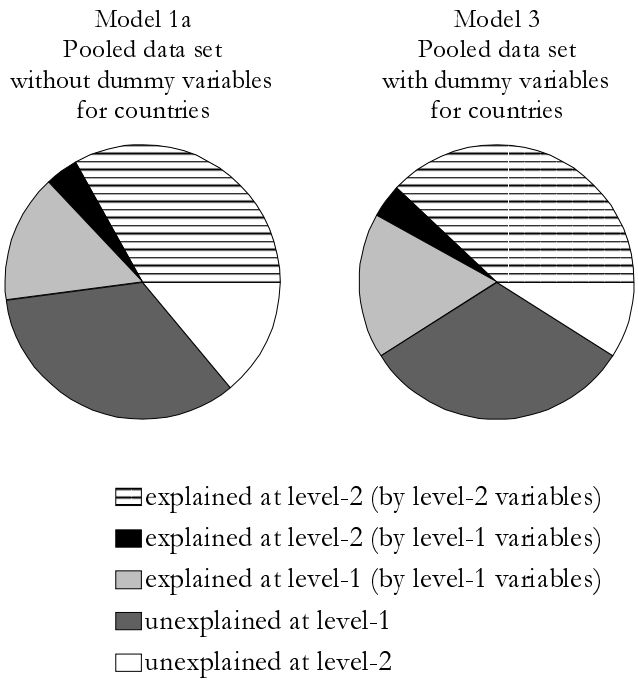


Figure 4-3
Proportion of variance in students' mathematics achievement scores explained respectively unexplained at student (level-1) and classroom level (level-2) in final level-2 pooled models; Pooled data set *without* dummy variables for countries (model 1a) and *with* dummy variables for countries (model 3)

The identification of students by their country (model 3) results in a greater percentage explained variance at classroom level (38% as opposed to 33% in the model without the country identification) and in a greater proportion of explained variance at student level as well.

These results confirmed that students from different countries performed differently on the TIMSS mathematics achievement test (see Table 4 -1) and that classrooms differ more across countries than students. The description of the results of Table 4-8 show some examples of student and classroom variables that contribute to the explanation of country differences in mathematics achievement.

4.7 UNDERSTANDING SIMILARITIES AND DIFFERENCES ACROSS EDUCATION SYSTEMS

In this chapter, results of exploratory TIMSS data analysis of three education systems were compared and both similarities and differences with regard to predictors at

student and classroom level of mathematics achievement in grade 8 were found. The results of the execution of the three-stage data analysis plan can be reflected upon to find answers to the question regarding whether TIMSS can serve the "understanding" function of IEA studies. Given the similarities and differences in predictors of mathematics achievement across the three systems, the question is why the differences exist and whether they can be explained from the TIMSS data sets.

Influencing variables on mathematics achievement within countries

In the multi-level analysis on the separate data sets, a hierarchical linear model was estimated for each country (see Table 4-6). The selection of variables was based on the results of the separate path models resulting from the PLSpath analysis. A maximum of six different student and three different classroom variables plus six aggregated student variables were included in a country model.
The results of these multi-level analysis show, per education system, which variables are associated with mathematics achievement (research question Ia, see 3.1). The results are inappropriate for comparison across countries because the analysis were conducted per country. They mainly indicate which student and classroom variables are important within each system.

Belgium Flanders
In Belgium Flanders, relevant student variables are 'home educational background' (positively), 'attitude towards mathematics' (positively) and 'student oriented teaching style' (negatively). At classroom level, one classroom variable turned out to be relevant: 'content coverage of mathematics' (positively) The other relevant classroom variables are aggregated student variables: 'out-of-school activities related to leisure time,' 'home educational background,' 'attitude towards mathematics,' 'student oriented teaching style,' and 'level of safety in school.'

Germany
In Germany, the average 'home educational background' of the classroom determines the achievement scores to a great extent. 'Attitude towards mathematics' (negative relationship), 'level of safety in school,' and 'limits in teaching related to student features' are the other relevant classroom variables in Germany (both positively related to achievement). Only one relevant student variable remained after PLS and multilevel analysis. This variable is 'home educational background'

The Netherlands

The aggregated student variables 'home educational background' and 'level of safety in school' determine achievement scores to a great extent in the Netherlands. An additional variable at classroom level which turned out to be relevant in the Dutch system is 'content coverage of mathematics.'

At the student level, two relevant variables resulted form the analyses: 'home educational background' and 'attitude towards mathematics.'

Strength of the country models

The multilevel analysis on separate data sets revealed for all three systems that the classroom (school) variance component is greater than the one at student level, which is usually the case (Scheerens and Bosker, 1997). For instance, in Germany, at the student level the explained percentage of variance in achievement scores is 34% and at the classroom level it is much larger (71%; see Table 4-7). Consequently, for all countries the effects of variables located at classroom level which have been explored first by means of PLSpath (including aggregated student variables), are greater than the effects of the selected student variables located at student level. These results might reflect the vertical organization of the education systems with ability tracks within schools. Indeed, in the three countries 'ability grouping' is applied in lower secondary education. Hence, classrooms (coincidental with schools in the TIMSS design) differ more from each other than individual students do. Students within a classroom are more similar regarding background variables than students from different classrooms. In addition, the relatively great contribution of the aggregated student variables to the explained variance at classroom level implies that classroom variances are mainly explained by student background variables.

Therefore, from these results it cannot simply be concluded that in order to achieve well on the TIMSS mathematics test in the separate countries, it matters more to which classroom a student belongs than which individual background and other characteristics a student has.

Influencing variables on mathematics achievement across countries

In order to enable country comparisons, the data sets of the three education systems were pooled and two hierarchical linear models were estimated. In the previous section, results were presented of these models.

It turned out that student background variables influence mathematics achievement in the three countries both at student and classroom level. From the TIMSS data analysis, a relatively low number of classroom variables could be identified as variables with different influencing strength in the three countries. Examples of the latter are 'limits in teaching related to student features as experienced by the mathematics teacher,' 'time on task,' 'mathematics teacher's workload,' and 'content coverage mathematics.'

These results are considered to be an indication of the appropriateness of the TIMSS data to facilitate understanding of cross-national differences in student achievement. From the start of the relational data analysis, it was known after stage A that a number of potentially effective factors were covered by the TIMSS background questionnaires and another series was not. The data analysis showed that the number of resulting factors which can be influenced by policymakers to improve mathematics education is rather small.

The number of minutes of mathematics scheduled per week for grade 8 (indicator for 'time on task') and the percentage of lessons a teacher is assigned to teach mathematics (indicator for 'mathematics teacher's workload') are two concrete examples of the few changeable factors which account for some part of the variance in student achievement within and across countries. At the student level, a few changeable factors resulted from the pooled data analysis. In all countries, the 'perceived safety in school' and 'attitude towards mathematics' contribute to the explained variance at the student level and at the classroom level (aggregated scores of student variable). For example, the level of safety as perceived by students can be improved by measures taken by the school board. The improvement could enhance mathematics achievement in the countries. However, the mean safety level across schools in all of the three countries is relatively high. The schools with low safety level are likeliest to benefit from such measures.

Another set of changeable factors (included in the organizing framework, see chapter 3) did not show effects on mathematics achievement in any country. Examples of such factors measured in TIMSS are factors regarding 'evaluation, feedback and corrective instruction' such as 'kind of tests used' and the 'use of assessment results.' Possible causes of these results are that countries and classrooms within countries do not differ on these factors or that the operationalization of these factors in TIMSS is not appropriate. If the operationalization of a factor is not internationally valid and/or not reliable, inclusion of such factors in country comparative data analysis can result in invalid results. Teachers in one country can interpret the questions about

these factors in a different way than teachers from another country. As a consequence, the results could be less comparable across countries. In chapter 5, the appropriateness of TIMSS items forming a scale to measure potentially effective factors are discussed further.
The set of potentially effective (changeable) factors categorized in the organizing conceptual framework for which no indicators could be found in the TIMSS questionnaires, was substantial. Particularly, factors located at the level of the curricular context were not covered in TIMSS. Examples are characteristics of curricular materials (explicitness and ordering of goals and content, structure and clarity of content, and advance organizers) and some features regarding teacher behavior (e.g., high expectations about student achievement, clear goal setting, and immediate exercise after presentation of new content). If such classroom factors would have been covered in TIMSS, country comparisons could have been more meaningful for policymakers and educational practitioners. This issue is reflected upon in section 5.3.2 (conceptual foundation and instrumentation).

Variance between students and between classrooms
In both the individual country multilevel models and the pooled models, the proportion of variance explained at student level is smaller than the explained proportion at the classroom level. In fact in all countries, it seems to matter in which classroom a student is taught. However, the TIMSS analyses provided only a limited number of genuine classroom variables with an effect on student achievement. The analyses resulted in relatively more effective student variables. Students within a country differ less in mathematics achievement than students across countries (see Table 4-1). In comparing countries by means of pooled data analysis, this point must be taken into account. The most important outcomes for the comparison is the difference of the list of resulting variables with an effect in the pooled model without dummy variables for the countries, on one hand, and in the pooled model with the dummy variables, on the other hand.

In the final chapter of this thesis, the possibilities TIMSS data offer to understanding cross-national differences in potentially effective factors on mathematics achievement are reflected upon based on the results presented in this chapter. Important topics of this reflection are the appropriateness of the TIMSS design and instruments to address international comparative research questions such as 'Why do countries differ in student achievement?'

CHAPTER 5

SUMMARY, REFLECTIONS, AND RECOMMENDATIONS

In this thesis, IEA's Third International Mathematics and Science Study (TIMSS) was taken as the case of large-scale international comparative achievement studies in education (LINCAS). The ambition of TIMSS and other IEA studies is to understand similarities and differences in student achievement across education systems (5.1). The extent to which IEA reached this goal in TIMSS was investigated by means of exploratory analysis on data collected in three education systems (research question I). The results of this analysis are summarized in 5.2. The summary is followed by a reflection on the benefits and limitations of TIMSS (research question II; 5.3.1). Two main components of LINCAS are discussed in connection to the 'understanding' function: the conceptual foundation and instrumentation (5.3.2), and design issues (5.3.3). For each component, reflections on the TIMSS case are followed by some general reflections and some recommendations. In the final section, a six-stage plan is presented to improve the 'understanding' function of large-scale international comparative achievement studies in education (5.4).

5.1 THE UNDERSTANDING FUNCTION OF IEA STUDIES

In chapter 1, it was stated that since 1964, the IEA has organized several large-scale international comparative achievement studies different core subjects such as mathematics, science, and reading. IEA studies can be characterized as multi-purpose studies not aimed at one specific goal.

IEA recognizes two main, general goals of its achievement studies (Plomp, 1998):

(i) to provide policymakers and educational practitioners with information about the quality of their education system in relation to relevant reference systems. By identifying what is happening elsewhere, an education system can learn from other systems.

(ii) to assist in understanding the reasons for observed differences between education systems.

Each goal requires its own kind of comparison. In chapter 1 it was recognized that one of the main functions of a LINCAS is the description of the status of an education system in an international comparative context. The first purpose asks primarily for international comparisons at a descriptive level, of effects of education in terms of total test and sub-test scores on international achievement tests. Differences in mean test scores and in the distribution of the test scores across systems can serve as indicators for the quality of education systems. Such descriptions can form the basis for policymakers to have a 'look in the mirror.' Countries can choose their own way of looking at and comparing countries' results. Some countries might be interested in the most important factors that affect achievement in top-performing countries anywhere in the world, while other countries might prefer to look particularly at the results of all countries of their own (geographical) region.

Achievement data are not the only kind of information needed to accomplish the first goal. The identification of 'what is happening elsewhere' requires a description of indicators referring to educational processes at different levels in the school (student, classroom/teacher, and school).

The second goal refers to explanations for described differences in achievement and its potential influential factors at several educational levels across nations. This 'understanding' goal can be dealt with by analyzing the many variables (indicators) of educational processes and their relationship with achievement in an international comparative context. An example of such relationships is the one between a student-oriented teaching style and student achievement.

In chapter 2, the utilization of results of two successive IEA studies on mathematics were discussed in the light of the two general goals. An important prerequisite to describe and understand country differences in achievement results is that all collected data are internationally valid and reliable. For instance, the

operationalization of 'student-oriented teaching style' must be univocal for respondents across all participating education systems. The selection of background variables to be studied should be based on their potential to enhance achievement in a core subject and to be useful they must be changeable by policymakers.
In chapter 2, for successive IEA studies on mathematics achievement, it was concluded that it was very difficult to fulfill the ambition of understanding cross-national differences in student achievement. The conceptual foundation of the selection and operationalization of background factors – i.e., factors other than 'student achievement in mathematics' - in these studies, was found to be incomplete.

The problem statement of this thesis focuses on the benefits and limitations of large-scale international comparative achievement studies in education. The first research question was addressed by investigating the case of IEA's TIMSS from the perspective of the 'understanding' function. The ambition of TIMSS was to provide participating education systems with data to find explanations of cross-national differences in student achievement.
The results of investigating the TIMSS case are summarized in the next section. In the subsequent section, the results of each of the three stages from the data analysis plan are reflected upon to address the second research question of this thesis (see 5.3).

5.2 SUMMARY OF THE CASE OF IEA'S TIMSS

The TIMSS study was taken as the case to investigate the first research question of this thesis, which refers to the 'understanding' function of IEA studies:

I. *To what extent can variability in the overall TIMSS mathematics test scores in grade 8 in the Netherlands, Belgium Flanders, and Germany be explained by variability in the scores on background variables at student and classroom/school level, and to what extent are these outcomes generalizable across these three European education systems?*

In grade 8, most of the students are 14 years old at the time of testing by the end of the school year. The performance on the TIMSS *mathematics achievement test* was taken as the operationalization of the dependent variable 'mathematics achievement.' Belgium Flanders outperformed Germany and the Netherlands on the TIMSS mathematics test and the Netherlands outperformed Germany (Beaton, Mullis, et al., 1996).

The analysis of the TIMSS case consisted of four stages. First, the basic conceptual framework for TIMSS was reviewed, which resulted in an organizing conceptual framework of potentially influencing factors (chapter 3). Second, the contents of the items included in the TIMSS background questionnaires were scrutinized to find indicators for the factors categorized in the organizing framework (chapter 4). Third, scores on the sets of items were analyzed to create scales (variables) for the identified indicators (chapter 4). Fourth, differences and similarities regarding relationships between the revealed variables were analyzed across the three education systems under review by means of one-dimensional path analyses and multilevel analyses (chapter 4). Each of these four steps is summarized below.
Particularly, the TIMSS case investigation was aimed at potentially influencing background variables at student and classroom level that are associated with mathematics achievement in grade 8 in the Netherlands, Belgium Flanders, and Germany. Cross-national differences and similarities in influencing background factors could be taken as a starting point to address the question what can be learned from the international comparative results regarding the effects of background factors on mathematics education in neighboring countries which differ significantly in student achievement in mathematics.

Conceptual foundation

The process of identifying potentially influencing factors on mathematics achievement started with a close look at the conceptual foundation of TIMSS. The conceptual framework that was developed by IEA during SIMS and adopted by the TIMSS study, was reviewed first. The strengths and weaknesses of this so-called three curriculum level conceptual framework (Travers & Westbury, 1989) were related to the 'understanding' function of TIMSS. It was concluded that the three curriculum level framework was basically appropriate to guide the search for potentially influencing factors. The basic framework was adapted into an organizing conceptual framework by adding one education level (the school level). The contents of each cluster of factors of the basic framework was judged less appropriate to function as a guide to identify potentially influencing factors. For many factors mentioned in the accompanying literature of the conceptual framework for TIMSS (Schmidt, 1993), no concrete definitions were available. Furthermore, the framework was not developed on the basis of empirical evidence from previous studies conducted in countries around the world. In international

comparative studies in education, the empirical basis of the conceptual framework and the selection of factors should, preferably, be internationally oriented.
To overcome the two weaknesses of the basic conceptual framework mentioned, models of educational effectiveness were studied. The basic conceptual framework of IEA was filled in by factors derived from models of school effectiveness (Scheerens, 1990) and from models of instructional effectiveness (Creemers, 1994). The models of educational effectiveness were developed as a means for classification of key factors that potentially influence student achievement. Particularly, the lists of factors at the school, classroom and student level were selected to be used to cover blocks in the organizing conceptual framework. The factors included were clearly defined by Scheerens (1990) and Creemers (1994) and can be mapped onto the components of the IEA conceptual framework.
The theoretical and empirical foundations of the educational effectiveness models were appropriate. The factors were included in the effectiveness models, mainly because of empirical evidence found in studies conducted around the (industrialized) world. Scheerens' model focused primarily on school level factors. The precise construction of Creemers' model focused on a set of factors related to 'quality, time, and opportunity.'

Indicators within TIMSS background questionnaires

The TIMSS data explorations started with a search for empirically and theoretically relevant factors in the TIMSS student, teacher, and school background questionnaires, guided by the organizing conceptual framework developed in this thesis. Factors from this organizing framework were listed that, as regards the *contents*, were indicated by an item or a set of items in one of the TIMSS background questionnaires.
Examples of important student and classroom factors that were indicated in one of the TIMSS questionnaires include: student's motivation (such as 'student's attitude towards mathematics' and 'perceived maternal expectation'), student's social background (such as 'out-of-school activities' and 'home educational background'), and management and orderly and quiet atmosphere (such as 'level of safety in school as perceived by students,' 'classroom climate as perceived by the students,' and 'experienced limitations in teaching related to student behavior').
Examples of student, classroom, and school factors for which no indicators could be found in TIMSS questionnaires include: student's aptitude, characteristics of

curricular materials ('explicitness and the ordering of goals and content,' 'structure and clarity of content,' and 'availability of advance organizers'), grouping procedures ('mastery learning,' 'ability grouping'), and school's evaluation policy.

From indicators towards variables

The sets of items that were identified as indicators for potentially influencing factors from a content perspective, were explored further statistically, by means of the calculation of a reliability coefficient (Cronbach α). The TIMSS data under review in this thesis were collected at two educational levels: student level and classroom/teacher level (school background data could not be used in the analysis because in the Netherlands less than 75% of the school questionnaires were returned).

The internal consistency of the scales differed across countries within a certain range (see Table 4-3). Sets of items with a relatively high reliability coefficient in one country have also a relatively high reliability coefficient in the other two countries. The same is true for items with a relatively low reliability coefficient. However, the range of differences in the Cronbach α-coefficient across countries varies from .05 ('perceived limits in teaching mathematics due to resources') to .20 ('maternal academic expectation').

Thereafter, the bivariate correlation coefficient was calculated between the explored student and classroom background variables, on one hand, and the student's scores on the TIMSS mathematics achievement test, on the other hand. It appeared that the bivariate correlations differed across the countries in strength, not in direction. The correlation coefficients between mathematics achievement and the majority of the student background variables ranged from less than .10 to around .20. Some background variables correlated significantly with mathematics achievement in one country, but not in one or both of the other countries.

The results of the bivariate correlations determined the selection of variables that were inserted in the next steps of the data analysis: the exploratory path analysis and multilevel analysis. The main selection criterion for inclusion of background variables in the path analysis was that the bivariate correlation coefficient between an indicator and the dependent variable 'TIMSS mathematics achievement test' was higher than |.10| in at least two out of the three countries under investigation.

One-dimensional explorative path analysis and multilevel analysis

The most appropriate techniques to analyze relationships between different background variables and mathematics achievement are the ones in which the nested design of the data sets is taken into account: hierarchical linear modeling (HLM) techniques. The major advantage of HLM techniques (e.g., multilevel analysis) over unidimensional ones such as Partial Least Squares techniques (PLS), is the estimation of the effects of variables on the dependent variable at one level (for example student level), taking into account at the same time the effect of variables on the dependent variable at another level of the hierarchical data structure (for example classroom level).

However, relationships were first explored by means of a single level technique. To model the TIMSS data at more than one level, some theoretical basis must be available. Important direct and indirect (mediated) relationships between student and classroom factors should be known before a more advanced technique such as multilevel analysis, could be applied. There is little relevant research available to serve as a sound theoretical and empirical basis for the specification of a hierarchical model of student and classroom factors influencing mathematics achievement of grade 8 students *in different countries*.

Partial Least Squares path analysis (using the program PLSpath) was conducted to explore relationships. This was done at two separate levels: at level-1, relationships between student factors were explored. The level-2 analyses concerned relationships between classroom and teacher factors combined with the student data (which were aggregated to the classroom level). The path analysis results at both level-1 and level-2 were studied to select variables for inclusion in hierarchical linear models, which were estimated by means of multilevel analysis.

Results Partial Least Squares path analysis

The PLS path analysis resulted in a classroom model per education system (see Figure 4-1). Due to selection criteria and the limited number of cases in the data sets, the list of background variables included in the exploratory path analysis per system was rather small. As a result of path analyses, the list of indicators was reduced further. The resulting path model differs across the three systems. For each system, the PLS path analysis resulted in a list of influencing student and classroom variables on mathematics achievement (see Table 4-6). The lists are different. However, as the analyses were conducted on separate data sets collected per

education system, the lists could not be compared. The results were regarded as indications for the influencing variables for each system separately.

The lists of influencing factors per country provided some insight into the background variables which influence mathematics achievement *directly*. The lists of variables for Germany is smaller than the ones for Belgium Flanders and the Netherlands. In the path model for Germany, three student variables remained ('home educational background,' 'students' attitude towards mathematics,' and 'safety at school as perceived by students') and just one classroom variable ('limitations in teaching mathematics related to student behavior'). In Belgium Flanders, five student variables and one classroom variable remained. In the Netherlands, three student variables and two classroom variables remained. With respect to the strength of *indirect* relationships between student and classroom variables and achievement in mathematics resulting from the country models, countries differed as well. Some indirect links are shown in only one country model. For example, the relationship between 'time on task' and 'content coverage' is shown only in the Dutch model.

One possible explanation for cross-national differences in influencing factors could be that other classroom factors (not included in the model) or country-specific factors influence instructional practices. Another reason might be related to the international reliability and validity of the operationalization of the factors. This possible reason is reflected upon in section 5.3.

Results multilevel analysis

For each country, a hierarchical linear model was estimated using the variables that showed a direct relationship with mathematics achievement in the path model. The indirect relationships found in the path model between student and classroom variables were taken into account as well.

The multilevel analysis results, per country, revealed a much larger proportion of explained variance at the classroom level than at the student level (see Table 4-7). However, the variables inserted in the hierarchical linear country models at the classroom level were mainly aggregated student background variables (for example, 'home educational background' and 'attitude towards mathematics'). From the PLSpath analysis, per country, only a few genuine classroom variables ('content

coverage mathematics' and 'limits in teaching related to student features') were identified as influencing variables on student achievement in mathematics.

The country models revealed influencing variables that could not be compared across countries, because analyses were run separately for each country. To enable country comparisons in order to understand differences in mathematics achievement across nations, a multilevel analysis was run on the pooled data set. In the pooled data set, all students and teachers from all three education systems were included. The variables inserted in this analysis were not merely the ones that resulted from the exploratory path analysis. Other factors that showed bivariate correlation coefficients with mathematics achievement higher than |.10| in the pooled data set, were included as well.

The final results of the analysis on the pooled TIMSS data sets of the three countries showed some similarities and differences across the three education systems (see Table 4-8). In all three systems, individual student variables turned out to influence mathematics achievement. These variables are: 'student's gender,' 'home educational background,' 'friends' academic expectations,' 'perceived safety in school,' 'attitude towards mathematics,' 'working hard doing homework,' and 'student oriented teaching style as perceived by the student.' All of these student variables related positively to mathematics achievement, except 'friends' academic expectation' and 'student oriented teaching style as perceived by the student.' The more students are convinced that friends are motivated to achieve well at school, the lower their own achievement level is. Students who perceive the instructional behavior of their teacher as more a student oriented perform less well in mathematics than students who perceive the opposite.

In the list of variables presented above, some are changeable by the school and teachers: 'perceived safety in school,' 'attitude towards mathematics,' 'working hard doing homework,' and 'student oriented teaching style as perceived by the student.' However, countries cannot learn much from each other when changeable student variables are considered. In all countries, the student factors are related to mathematics achievement in the same way.

Countries could learn from each other however, by looking at the final results at the classroom level. Three student aggregated factors are important predictors in the pooled data set: 'home educational background' (positively), 'out-of-school activities related to leisure time' (negatively), and 'perceived safety in school' (positively). In

all countries, mathematics achievement could be enhanced by taking into account the average 'home educational background of the classroom' and by taking measures to increase the safety in schools. The first factor cannot easily be influenced by schools, but 'safety' is an issue schools can change (e.g., by means of taking measures to prevent needling behavior.

Three classroom factors influence mathematics positively in at least two countries: 'limitations in teaching related to student features,' 'teacher's workload,' and 'content coverage of mathematics.' In Germany, the mathematics achievement level was lower than in Belgium Flanders and the Netherlands. The distribution of the scores on the three classroom factors that enhance student achievement in all countries provide some clues for Germany. In Germany, teacher's assignment to teach mathematics is, on average, around 50% of their total assignment at school. In Belgium Flanders and the Netherlands this percentage is, on average, 72% respectively 86%. Hence, in Germany, the mathematics achievement level of grade 8 students might profit from measures to increase teacher's workload of mathematics. With regard to the other effectiveness enhancing factors found at the classroom level, it is more difficult to conclude what a lower performing country could learn from a higher performing country. In Belgium Flanders, 'limitations in teaching related to student features' is, on average, experienced on a lower level by teachers than by teachers in Germany. Germany could take measures to prepare teachers better, aiming at coping with students who come from different backgrounds (economic, language, and of different academic abilities). From the results it can also be seen that Dutch teachers experience fewer limitations in teaching related to student features than their colleagues from Belgium Flanders. As Belgium Flemish' students outperformed Dutch students on the TIMSS test, it seems that Belgium teachers can cope better with classrooms with students with different backgrounds than their Dutch colleagues.

The third effectiveness enhancing classroom factor is 'content coverage in mathematics.' All of the mathematics topics listed in the TIMSS teacher questionnaire (see Appendix A, under CO_1 'contents implemented curriculum') refer to (sets) of items included in the international mathematics test. On average, in the highest performing country, Belgium Flanders, the teachers covered the smallest number of topics previous to the administration of the TIMSS test in grade 8. The question why Flemish grade 8 students outperform Dutch and German students, while their mathematics teachers covered fewer topics than teachers in the other two countries is hard to answer.

In Germany, many topics have been covered. However, the time spent per topic could be less than in Belgium Flemish classrooms. On the other hand, in Belgium Flanders some topics were not covered at all. The comparison between Germany and the Netherlands, considering the relationship across 'content coverage mathematics' and student's achievement in mathematics, is unclear as well. The distribution of scores on 'content coverage mathematics' for the Netherlands is about the same as for Germany. Yet, the Dutch students outperform their German peers on the TIMSS test.

The differences in proportion of explained variance across the pooled model without country identification of the students, on the one hand, and the pooled model with country identification, on the other hand, indicated that country specific factors might play a role in relation to cross-national differences in student achievement. In the model with 'country identification of students' the proportion of explained variance at classroom level is smaller than in the model without 'country identification' (see Table 4-9).

In the next section, results from the analyses conducted in the TIMSS case are reflected upon.

5.3 RESEARCH QUESTION II: REFLECTIONS AND RECOMMENDATIONS

5.3.1 Research question II

The answers to research question I reported in chapter 4 and summarized and discussed in the previous section, indicated that TIMSS fulfilled the function of understanding cross-national similarities and differences in background factors related to mathematics achievement only to a limited extent. The TIMSS case showed that the aim of *describing* similarities and differences across countries has better been met than the aim of *explaining and understanding* differences in achievement level. The exploratory data analysis procedures conducted in this thesis did not result in clear recommendations for countries (i.e., policymakers and educational practitioners) to improve their education in mathematics. Countries are interested in cross-national differences in *changeable* factors at student or classroom level that turned out to be strong predictors of variances in achievement scores. To uncover such predictors on the basis of TIMSS turned out to be difficult in the countries selected.

Limitations in descriptions and explanations of cross-national differences in achievement and in predictors of variance in achievement scores, can lead to dissatisfaction among policymakers (i.e., the funders of studies in many countries) and educational practitioners about the relevance of participation in an international comparative study. For instance, policymakers can hesitate to take part in a next study if they are not satisfied with merely country-ranked lists based on mean achievement scores without explanations for differences across countries. Teachers and school principles might in turn opt not to participate in a large-scale and time-consuming study, if they cannot be convinced by researchers that the study has the potentials of providing them with information relevant to improving their educational practice.

Hence, it is expedient to increase the benefits of TIMSS and to downsize its limitations as much as possible. Directors of future studies as well as the National Research Coordinators who are responsible for the execution of such studies within their country, will be interested from their own position.

The big question though is in what way the benefits of large-scale comparative achievement studies could be enlarged and how the limitations could be overcome. In chapter 1, this question was presented as research question II of this thesis and was formulated as follows:

II. *What can be learned from the case of IEA's TIMSS for future international comparative achievement studies in education regarding the conceptual foundation, instrumentation, and design in view of their possibilities to uncover factors related to different outcomes across educational systems on an international student achievement test?*

While research question I was answered by an examination of the conceptual framework for TIMSS followed by exploratory analyses on the contents of the TIMSS background questionnaires and the background data, research question II can be characterized as a 'meta' question. Research question II is primarily addressed by reflecting upon the results of the TIMSS case (research question I). Based on the reflections, benefits and limitations are identified and recommendations are formulated to improve the usefulness of future large-scale international comparative achievement studies.

One of the future studies is the continuation of TIMSS. As stated in section 2.5, after the 1994/1995 TIMSS study (the study from which data were analyzed in this

thesis), TIMSS was continued under the auspices of IEA as the Trends in International Mathematics and Science Study (the acronym TIMSS was kept). Trends will be analyzed every four years. In 1999, TIMSS was repeated in grade 8 only. In 2003, 2007 and so on, trends in student achievement in mathematics and science and in background factors will be analyzed in both grade 4 and grade 8.

In this thesis, the TIMSS case gave rise to recommendations for the improvement of the conceptual foundations and instrumentation of future studies in order to enhance appropriate comparisons across countries which can serve the function of understanding differences in student achievement results. In Figure 2-1 (see chapter 2) a general study framework for LINCAS was presented. This framework shows a close connection between the conceptual framework and instrumentation issues on one hand, and design issues, on the other hand. From the TIMSS case, lessons can be learned regarding these two main components from the general study framework.

5.3.2 Conceptual foundation and instrumentation

Descriptive TIMSS results

One of the benefits of the TIMSS study is the provision of descriptions of differences and similarities in achievement results across many education systems from around the world. The descriptions are provided by TIMSS, giving expression to country-ranked lists. Systems can compare their achievement level with all other participants. These comparisons are fair, as the TIMSS achievement test was robust across countries at the intended curriculum level. This was reflected in the way the ranking lists hardly changed after, for each participating country, the most appropriate set of items was selected as the basis for international comparisons. The rankings of countries barely differed and seemed not dependent on a country's set of most appropriate items that was taken as the basis for comparisons. Nor did the average percentage correct scores on sub-scales of the achievement test (Beaton, Mullis et al., 1996).

The international reports of TIMSS contain descriptives of background questions as well (e.g. Beaton, Mullis, et al., 1996). The majority of the descriptions are features of the distribution of scores on *separate* background items of the participating countries. These descriptions provide a global insight in cross-national differences

on an item-by-item basis. For instance, the level of 'safety at school as perceived by the students' (indicated by separate items from the student questionnaire, see Appendix A) or 'the amount of homework per day' can differ within and across nations. The global, comparative descriptions themselves are informative and interesting, and can be seen as a benefit of TIMSS. Countries can make use of the descriptive results in the way they choose. Nevertheless, the risk of 'ecological fallacy' must be mentioned here.

The ecological mistake can be made easily if descriptive results of achievement tests are presented together with descriptive results of potentially effective factors at an aggregated level (Postlethwaite, 1999). If individual student data on two variables are aggregated to a country level – e.g., a country mean – the relationship between these two variables can be misinterpreted. For instance, in the descriptive, international TIMSS report on mathematics in the middle school years (Beaton, Mullis et al., 1996), the mean mathematics achievement score per country is presented next to the percentage of students having a computer at home (Beaton, Mullis et al., 1996; p. 163, Table 5.13). For countries in which more students have a computer at home than in other countries and the mean mathematics score is above the international average, one might conclude that having a computer at home encourages the student to achieve well in mathematics. However, the data of both variables are reported at country level. As a consequence, the results can not be assigned to individual students. The country level scores cannot tell the reader whether 'having a computer at home' has an influence - positively or negatively – on mathematics achievement. Readers who do so are making ecological mistakes.

In order to interpret hierarchical data properly, the influence of many more factors measured at more than one level should be inserted in the analysis of relationships with mathematics achievement. The application of multilevel analysis techniques will save researchers, and the users of its results in particular, from ecological fallacy. To make such relational analysis possible, the item-by-item analyses of background data is only a start. Item-by-item analysis cannot provide information (neither conceptually nor empirically) about background *factors* because factors usually cannot be measured in a valid way by one item. Therefore, sets of items indicating one factor should be analyzed. Sets of items could form a scale.

In the international TIMSS reports (e.g. Beaton, Mullis, et al., 1996; Beaton, Martin, et al., 1996) compared to item-by-item descriptions, scores on scales were barely

reported. As a consequence, no information is available about the internal consistency (reliability) and content validity of scales for the participating countries. The reliability and validity of scales in international comparative studies are called *international* reliability and validity. The addition, 'international' refers to the fact that a scale should be both reliable and valid for each of the participating countries.
Together with the criticism on the conceptual framework for TIMSS (theoretical and empirical evidence is unclear and not derived from international literature and lists of concepts are not very well defined) the lack of scales can be seen as a limitation of the first TIMSS reports. In this thesis, the TIMSS case analyses included a review of the conceptual framework for TIMSS and a search for operationalizations of background factors that consist of sets of items instead of one single item. The internal consistency of sets of items, forming scales, was analyzed for the three countries under review.

In the first reports of the repeat of TIMSS in 1999, more scales (indices) for student and classroom factors were reported than in the reports from TIMSS-1995 (eg., Mullis, Martin, et al., 2000). However, information about international reliability and validity of the indices was not available.

Conceptual foundation

The selection of relevant background factors in an international comparative achievement study is a difficult task. An appropriate conceptual framework is necessary. The TIMSS case showed that the conceptual framework for TIMSS was appropriate as a tool for the classification of potentially effectiveness enhancing factors and less appropriate as regards the contents of the respective factors. In chapter 3, the components (or clusters of factors) of the framework were filled in with factors derived from the educational effectiveness literature. As noted, the resulting organizing conceptual framework was used as a guide to identify groups of indicators for the factors in the TIMSS background questionnaires.
As a consequence, the analyses conducted to address research question I are *secondary* in nature. The international research questions of the TIMSS study were formulated in a general way and an elaborated conceptual framework for TIMSS could not be found in the literature. The method applied to develop the international background questionnaires in TIMSS remained unclear. The TIMSS background questionnaires were not distinctly based upon a conceptual framework

(see chapter 3). The organizing conceptual framework facilitated the identification of potentially effective factors on (mathematics) achievement in grade 8 that are both theoretically and empirically found in international literature. Inevitably, secondary analysis on existing TIMSS databases resulted in lists of factors identified at different curricular and educational levels of the organizing conceptual framework, but that are missing in TIMSS (see 4.2). Some predictor variables found to be important in educational effectiveness research, were only partially covered or not covered at all in the TIMSS data sets. Depending on the research questions, in future large-scale international comparative achievement studies, these important missing factors can be inserted.

One of the factors that was missing in TIMSS is 'Opportunity-to-learn as part of the implemented curriculum at classroom level.' The Dutch national center for TIMSS inserted, as a national option, a so-called 'turbo measure' to indirectly assess the content of implemented curriculum at the classroom level (Bos & Vos, 2000; Kuiper, Bos & Plomp, 1997). Teachers of the classes tested were asked to judge every TIMSS achievement test item on its appropriateness for their own students, both from a content perspective ('content of item covered?') and a format perspective ('item suitable to be included in a test on the topic?'). These two indicators for 'opportunity-to-learn' were not part of the international TIMSS instruments.

For future studies, it is recommended to insert this (or a similar) measure in the international component of the study, because the factor provides information regarding the students' opportunity to learn the topics which are tested by means of the achievement test used in the study. Country comparisons can benefit from these data. Differences in achievement results can directly be linked to 'opportunity-to-learn' data. If in one country a topic of the core subject under investigation is taught to a greater extent and more in-depth than in another country, and the countries differ in achievement results on this topic, the 'opportunity-to-learn' results can provide insight into reasons for the difference. As stated previously, the TIMSS test turned out to be robust across participating countries from the perspective of the intended curriculum. The 'opportunity-to-learn' results provide additional information regarding the appropriateness of the test at the level of the implemented curriculum.

Other factors that are not part of the TIMSS questionnaires are listed in Table 4-2. Some examples were mentioned in 5.2 and are repeated here. Examples of student, classroom, and school factors for which no indicators could be found in TIMSS questionnaires are: student's aptitude, characteristics of curricular materials ('explicitness and the ordering of goals and content,' 'structure and clarity of content,' and 'availability of advance organizers'), grouping procedures ('mastery learning,' 'ability grouping'), high expectations of teachers regarding the learning capacity of their students, and school's evaluation policy. In educational effectiveness research, empirical evidence was found that indicated the potentials of these factors to enhance student achievement (Creemers, 1994). Therefore, such factors could be labeled as 'white spots' in the organizing conceptual framework. It is recommended that, taking the organizing framework as a starting point, these 'white spots' be filled in during the preparation phase of future studies.

Recommendations regarding conceptual framework

The TIMSS case clearly showed the importance of a well-developed conceptual framework in selecting the key factors to be examined in an international comparative study. The organizing conceptual framework developed and used in this thesis (Figure 3-8) can be used as a basic framework for future studies in which influencing factors on student achievement will be studied. The two dimensions of the framework, consisting of curricular and educational levels, form a good basis for the study of effectiveness enhancing factors on student achievement that are measured by means of curriculum-driven tests. At each of the curricular levels, potentially influencing factors can be identified in the framework. In chapter 3, it was argued that reviews of educational effectiveness studies could provide useful information on factors that potentially enhance student achievement in different countries.

Extensive literature searches on reviews concerning instructional effectiveness and school effectiveness are recommended to find appropriate operationalizations of the factors for which no indicators could be found in the TIMSS background instruments. The review studies could also explicitly concern investigations in the field of the curriculum. In fact, Creemers (1994) included results of many of such studies in a comprehensive model of educational effectiveness (see Figure 3-6). In

the organizing conceptual framework of this thesis, different curricular aspects were derived from Creemers' model. Most of these aspects were not found in TIMSS instruments. Reviewing of separate studies on curriculum aspects (e.g., van den Akker, 1988; Kulik and Kulik, 1987; Fraser, Walberg, et al., 1987) could result in appropriate operationalizations (instruments) that can be used in future studies.
If the 'white spots' of the organizing conceptual framework are filled in, the relationships between 'current' factors and potentially influencing factors that were missing in the TIMSS case, can be taken into consideration as part of the analyses. Then, the effectiveness of more factors on student achievement can be estimated in a more comprehensive way.

Minimalizing cultural bias in the conceptual framework
The selection and definition of key factors in international comparative studies could be related to cultural bias, particularly if countries from different parts of the world are participating in one study, like in TIMSS. Cultural bias with respect to the meaning of key factors must be low or negligible.
Most international comparative studies are led by industrialized countries (Western Europe, Northern America). Therefore, there is evidence that the definition of key factors that are studied are culturally biased. What is important with regard to achievement in schools in an industrialized country might be of less importance in, for example, an African country. A factor that turned out to have statistical predictive power with regard to achievement in one education system (either a developed or a developing system) can be of no importance in another system. Previous studies showed no evidence for the existence of one best education method for achieving high test scores (Scheerens, 1999).
In a review of research evidence, Scheerens (1999) found many fundamental factors (mostly at the level of resources such as availability of textbooks and pencils and furniture) which are important to include in studies in developing countries. Similar conclusions were drawn by Howie (in press) on the basis of a secondary analysis on TIMSS data from South Africa. Such factors cannot be considered as variables in developed countries and therefore they are not useful to statistical predictions of achievement in certain subjects.
In future studies, the conceptual framework and its operationalization might be adapted to the variety of developing and developed countries participating in the

studies. A subset of the selected set of key factors (or all of them) might need different operationalizations for certain groups of countries within the developed and the developing world. Hence, the design of a worldwide study might be adapted into a study having not only core components, but also differential ones at a regional level (see below under 'design issues').

Instrumentation

Factors related to achievement found in this thesis by means of the exploratory analyses on TIMSS data, particularly, the classroom factors identified as influencing factors in different countries (e.g., content coverage mathematics and teacher's workload) can be implemented in future studies.
The operationalization of the factors in background variables measured by means of TIMSS background questionnaires were analyzed in the TIMSS case. The quality of the operationalization of some factors could be judged as good and that of other factors as doubtful, both from a statistical and a content perspective (see chapter 4). For the majority of the factors for which indicators could be found, the contents of the items corresponds to a certain extent with the definition of the factor. For some factors, the statistical internal consistency of the identified sets of items varies across countries.

For instance, in TIMSS, the student background factor 'social background' (a curricular antecedent factor) was indicated by 'out-of-school activities related to leisure time.' The internal consistency of the 3-item scale (expressed by a Cronbach α coefficient) is low and varied in the three education systems under review from .36 in Germany to .50 in Belgium Flanders. The three education systems studied are neighboring countries. Nevertheless, the three items indicating out-of-school activities (see Appendix A for the contents of the items) might not be understood unambiguously by grade 8 students within and across systems. For instance, students from separate countries might understand the question 'on a normal school day, how much time do you spend before or after school watching television and videos?' in a different way. In particular, the first part of the question might be ambiguous. Some students might think of 'yesterday' as a normal school day (and yesterday they were off the whole afternoon) and other students might think of a school day they usually have. Different interpretations can lead to different answers.

Another example is the way students in different countries interpret questions referring to 'safety in school' (see Appendix A for the 'safety' items from the TIMSS student questionnaire). Students in one country might think that hurting people or stealing from other students are severe acts and they will complete the items accordingly. In another country, students might perceive hurting and stealing as more normal and therefore their answers might be more moderate. In the latter country, 'unsafe' situations may be reported less frequently than in the country where students think these actions are severe ones.

'Student's attitude towards mathematics' is another example of a factor for which an operationalization was found in the TIMSS student background questionnaire. Yet, the operationalization can be criticized. In the literature, several components of 'student's attitude' are distinguished (Marinot, Kuhlemeier & Feenstra, 1988), such as liking, importance, self-confidence, and anxiety. To be able to draw appropriate conclusions – both nationally and cross-nationally – about the relationship between achievement in mathematics and student's attitude, information should be collected about how different attitudinal aspects are related to achievement in different countries.
In TIMSS, two subsets of five attitude items were administered to students. One subset of attitude items refers to 'liking the subject' and a second set refers to 'importance of the subject for student's school career and future life' (see Appendix A). The total set of 10 questions form the scale 'student's attitude towards mathematics' (a curricular contextual factor). In national studies, it is recommended to measure attitude by means of more than 5 items (Marinot, et al., 1988). This recommendation is even more applicable in international comparative studies, to enhance the internal consistency and validity of the scales in all participating countries. Students from different countries have different cultural backgrounds. To minimize cross-national differences on student attitude scores, it is necessary to include more attitude items in the international questionnaire.
In the TIMSS case, the internal consistency of both the sub-scales and the total scale is adequate in the three education systems (Cronbach α coefficient for the total scale varies from .78 in the Netherlands to .84 in Belgium Flanders). However, it is not exactly known whether students within and across countries understand each attitude question in the way the question was meant. What does the answer 'yes' mean to the question of whether the student likes mathematics? The question and its

answer seem rather straightforward and unambiguous. Nevertheless, the meaning of the answer can be interpreted in different ways. The interpretation depends on the circumstances the student had in mind while answering the question. The student's opinion about the mathematics teacher and about the composition of the classroom can influence his answer. In the PLS path analyses, for Germany the direction of the relationship between 'attitude towards mathematics' and 'achievement in mathematics' was different from the one for Belgium Flanders (see 4.5). This might be related to the ambiguity of the attitude items.
Therefore, the operationalization of the student's attitude by only a few items is insufficient. The respondents in all countries must be directed to the circumstances which must be kept in mind while answering a set of questions (see also Meelissen & Bos, 2001).

In TIMSS, no data are available about the level of ambiguity of the sets of items within and across countries. However, it can be assumed that differences in internal consistency of scores on sets of items might be related to the different ways the questions in the student questionnaire were interpreted by students in the three countries. The latter is related to *international content validity*. International content validity if items from survey instruments (e.g., questionnaires) means that in each participating country, every respondent should understand each item in the way it was meant. The international *content validity* of the items related to one factor and included in the international background questionnaires, should be known in advance. Only then can the level of ambiguity of the items across nations be taken into consideration in interpreting the data.
As stated, the results of the exploratory path analysis reported in chapter 4, point the attention to content validity issues as well. However, content validity was not examined explicitly. The resulting list of factors from the PLS analyses which related directly or indirectly to mathematics achievement differs across countries. One possible reason for different lists of influencing factors on mathematics achievement (see also 5.2 under PLS results) could be the international content validity of the operationalization of the factors. Perhaps there is an international validity problem with the contents of the set of items used as the operationalization of the factor(s). If there is a cross-national difference in the reliability coefficient for a variable, the level of international comparability might be problematic. As stated above, as far as is known, in TIMSS content validity of variables was not examined.

Recommendations with respect to instrumentation issues

The clarity of all items in the background questionnaires must be optimized so that the international content validity (and international reliability) is maximized. The topic of international content validity of factors is one of the most important and complicated ones in large-scale international comparative achievement studies. Cultural differences can cause multiple interpretations of the same questions across nations. In addition, translating the questions from English into the country language is also a possible cause for multiple interpretation problems (however, in TIMSS translation verification is an integral part of quality monitoring). To accomplish the 'understanding' function of IEA studies, it is necessary to optimize international validity of factors. How could this possibly be achieved?

First, the definition of a selected factor should be made uniform across the participating countries, and this process should be followed by a pilot administration of the items in each country. The definition of a factor refers to a measure of characteristics of units in education (education system, school, teacher and classroom, and student).

Second, pilot administration of background items in each participating country is a crucial step in preventing validity problems. It took place in TIMSS, but the collected data were not analyzed to develop internationally valid and reliable scales. Pilot data were reported only item-by-item. As a consequence, in the international TIMSS reports only a few scales (indices) were reported (Beaton, Mullis, et al., 1996; Beaton, Martin, et al., 1996). As noted above, almost all of the questionnaire items were reported separately.

In future studies, the pilot should consist of both quantitative and qualitative methods. Analysis of the quantitative pilot data of sets of items representing a factor results in indices per country of the statistical reliability (internal consistency) of the set of items. During interviews with a limited number of individual respondents or groups of respondents (school principals, teachers, and students), valuable in-depth information can be collected about the way they interpret each question. In each country, the interviews could be based on semi-structured questionnaires provided by the international coordinators of the study. The results of all interviews reported in a country report can be compared across countries. As a result of this comparison, the proposed items can be revised to increase the level of clarity not only within each country, but also across countries.

Regardless, of the way questionnaire items are developed, it should be kept in mind that studies like TIMSS are worldwide studies with more than 40 participating countries. Achieving perfect, valid and reliable scales across all participants might be a utopia. International validity problems within the collected data across (subsets) of countries could hardly be avoided. Yet, the validity problem may be reduced by keeping the conceptual framework as the guide and having a design for the study that addresses this point. Some reflections upon the design of the study that takes into account the validity problem are presented in the next section.

5.3.3 Design issues

In the previous section, some recommendations were formulated with respect to the development of the conceptual foundation and instrumentation of background factors in large-scale international comparative achievement studies. In addition, two design issues could be discussed which can contribute to the accomplishment of the 'understanding' function of such studies.
The first issue concerns the determination of a core part (worldwide, involving all participating countries) and a differential part (involving a subset of reference countries that are to be compared with each other) for the study, and the organization of the study. The second issue concerns the number of measurements in a single study.

Determination of a core and differential part of the study

In TIMSS, the number of participating education systems is large and systems are located around the world. In this thesis, three neighboring countries from Western-Europe were compared with each other. Policymakers of the systems are interested in results of comparative analysis, because the systems are part of the European Community. The TIMSS case indicated that the instruments used may not be perfectly valid and reliable for these countries. Improvement of the international comparability of the data could be reached by means of pilot studies within these three countries.
Similarly, other groups of countries might be interested in regional comparisons. For instance, developing countries in Africa would probably like to compare themselves with other developing countries and not with industrialized countries such as the countries belonging to the OECD. The current situation in TIMSS is

that one uniform design including uniform instruments (achievement test and background questionnaires) is applied. However, countries have the possibility to add national options to their participation in the international comparative study and are free to get other countries involved as well. An example of a national option was given in 5.3.2. (the 'opportunity-to-learn' measure the Netherlands included in TIMSS). Another example is the national option Belgium Flanders attached to their participation in TIMSS-1999. The so-called 'Vlaams Luik' (van Damme, 1998) included extra instruments for student, parental, and teacher factors.

The possibility to insert national options is restricted in terms of time and money. Usually, the completion of the compulsory part of the instruments used in IEA studies takes a lot of time from respondents. Adding as many national options as a country would like, could result in instruments that are too long. The ultimate consequence would be more non-response of the invited schools than is allowed in IEA studies to get accurate estimations of population parameters. A better solution than adding national options to the compulsory instruments is to limit the compulsory part (the core). If IEA would adopt this recommendation, a future study can have a limited international core with regional options: options that are shared by groups of countries. In order to make these options appropriate for cross-national comparisons, future studies could split-up into regions. Regions could be based on selections of countries which choose to compare study results with each other and are not necessarily geographical ones.

What is the preferred approach of participating countries in comparative educational research to choose their reference countries? In principle, the assumption in this thesis is that the countries participating in a large-scale international comparative achievement study want to learn from other countries with regard to the way education in a core subject is being organized in schools and taught in classrooms. When considering the world as an educational laboratory (Husén, 1967) this assumes a great amount of between country variance in educational factors.

Given cultural differences between countries, no matter their geographical location on the globe, each country can have its own preferences for countries to which to compare. Many of such cultural differences are hard to measure in a comparative study. Policymakers often base their choice on the geographical region of countries (neighbor countries, for example), on the economic competitiveness of the countries, and on historical connections with the countries. Educational researchers

and specialists usually base their choice of reference countries more on aspects of the contents of the education in the subject under investigation.
For instance, the selection can be based upon the intended curriculum of the subject. Countries with national curriculum guides may be more interested in comparing each other than in comparing themselves with countries without national curriculum guides. Educational specialists might also be interested in comparing themselves with countries where the curriculum is being changed in the same way as in their own country.

Splitting-up the participating countries in regions has consequences for the research questions to be addressed and the instruments to be used in the study. Some of the international research questions of the study can still be applicable to all participants and for these questions a uniform core set of instruments is necessary. Furthermore, regions can formulate regional research questions that could be addressed by regional instruments. These instruments form the differential part of the study and need to be developed within each region. The optimalization of the international reliability and validity of these instruments is restricted to the countries belonging to the same region, which makes the developmental task easier. The number of countries for which the instruments should be as valid and reliable as possible is smaller than in the current IEA studies.

It is obvious that regional studies within a worldwide comparative study require more work from the study's international coordinating center than studies without regional options. To restrict the workload of the international center, it is recommended to set up regional centers responsible for addressing and answering the regional research questions (including monitoring the quality of the study's differential part in each country). The international center would be responsible for the quality of the entire study and for addressing the core research questions.

Cross-sectional studies versus pretest-posttest design studies

The number of measurements included in a study determines the extent to which educational processes can be measured. TIMSS and his predecessors FIMS and SIMS (see chapter 2) are examples of so called cross-sectional or 'one-shot' studies. The achievement level and the influencing factors are only measured at one point in time. One-shot studies cannot be seen as sources for comparing the influence of

educational *processes* in different countries. A pre-test post-test design (not necessarily an experiment) is more appropriate to make causal inferences, that is, determining the curricular antecedents, curricular context, and curricular content factors which influence student achievement within each country. The relationships between the determinants is of particular interest because these can reveal some of the reasons for different achievement levels within and across education systems.
In SIMS, countries had the opportunity to take part in an international option that had a pretest-posttest design (see 2.2.1). In this optional part of SIMS, the 'teaching process variables' questionnaire was added to the set of instruments. The process variables were meant as reasonable alternatives to the direct observation of classroom practices. From a content perspective they can be regarded as more valid than variables that are measured at only one moment, but they are far from perfect. The process factors showed more relationships with achievement than the more status-oriented measures (measuring the situation at a certain moment) typically available in survey studies. An example of a process variable measured in SIMS is 'teachers' strategies' (Cooney, 1993). Teachers were asked how often certain activities were used.
Results of an analysis of the teachers' answers resulted in dimensions of mathematics teaching. Countries were grouped into three main clusters based on cluster analysis on the teachers' response patterns. Robin (1993) divided the eight participating countries into cultural zones (e.g., France and Belgium versus Canada and the United States). Teachers within a zone could be distinguished from teachers in another zone, by considering their teaching strategies and practices employed in teaching mathematics. The relationship between the resulting teaching strategies and the gain scores in achievement is interesting. After an extensive analysis of the growth of results in various classes in the different countries, it was concluded that "*cultural and didactic choices of teachers have a decisive influence upon the fields considered in those items and on the activities related to these fields*" (Robin, 1993, p. 257). Following Robin's recommendation it seems necessary to check cultural zones in other comparative studies (see also above).
The application of a pre-test post-design is regarded as preferable above the one-shot survey studies. However, Beaton, Postlethwaite et al. (1999, p. 16) expressed some doubts about its potential: *"In most large-scale studies it has not been possible to identify teacher methods and behaviors that have a strong effect on achievement (...) when these have been based on cross-sectional studies or longitudinal studies of one year duration. The potential influence of these variables is usually better examined in replicated small experiments."*

A replicated small experiment could also be regarded as an alternative study design for cross-sectional large-scale international comparative achievement studies. It might be true that small experiments have greater potential to reveal determinants of educational achievement in different countries. From a research point of view, it is necessary to select variables for such experiments, in particular process factors. Therefore, first a large-scale survey is needed within the education systems belonging to the same region (as far as regional options are concerned in the proposed regional studies). The conclusions of such a survey with regard to differences and similarities across the participating systems should primarily be based on the analysis of the databases. Beyond the databases, sources at the system level could be used to interpret the results of the data analysis. Examples of these sources are cultural aspects and other features of education systems. Beaton, Postlethwaite et al. (1999; p. 15) emphasized that *"recommendations for policy changes in a country need to take account of not only the results of the international analyses, but of the educational and cultural context in which that country operates."*

The feasibility of an international study with a core and a differential part (regional options) with a pretest-post design is related to the number of years the study can take. The TIMSS study with the main data collection in 1994/95 took about four years. A regional study with the proposed design probably will take more than four years. The development of the conceptual framework and its operationalization of the core and the regional portion will, in particular, take longer than in cross-sectional, worldwide studies. However, the continuation of TIMSS as a trend study could make the operationalization process easier. In subsequent TIMSS studies, scales can be improved instead of being developed as new ones which can save time.

5.4 SIX-STAGE PLAN TO IMPROVE THE 'UNDERSTANDING' FUNCTION OF LARGE-SCALE INTERNATIONAL COMPARATIVE ACHIEVEMENT STUDIES

The TIMSS case showed the difficulty of providing users (policymakers and educational practitioners) of the results of (secondary) data analyses with concrete recommendations to improve their education system (see also Meelissen and Bos, 2001; Martin, Mullis, et al., 2000; Zuzovsky and Aitkin, 2000). The ambition of

TIMSS to explain differences in student achievement across a large number of countries could not be fulfilled to a great extent. The large number of participating countries from many different regions around the world, the general international research questions and the conceptual foundation and instrumentation are seen as hindrances in accomplishing this ambition.
The reflection on the TIMSS case resulted in some recommendations regarding the conceptual foundation, instrumentation, and the design of future large-scale international comparative achievement studies, including TIMSS trend studies. The conceptual foundation, the developmental process of the instrumentation, and the choice of the study design should be seen as interrelated, key components of the studies, given the international research questions.
The reflections and recommendations described in the previous section can be summarized in a plan for setting up a LINCAS, consisting of six stages. The plan assumes a study with an international core and regional options as proposed above.

Six-stage plan for setting-up a LINCAS

Stage 1: Specification of the understanding function
Given the international research questions, the 'understanding' function needs to be specified per region. In TIMSS, the number of participating countries has been increased from around 40 in 1995 to more than 55 countries that will be participating in 2003. This growth makes the recommendation to split-up the TIMSS study into regions (see under 'design') even stronger. It can be expected that none of the countries would like to understand differences in student achievement between their own country and *all* of the other participating countries. Probably, each country will choose its own reference countries. On the basis of countries' preferences, IEA should decide how many regions will be established and to which region each country will be assigned.
Once the regions have been established, all countries from one region can select their set of background factors, above and beyond the set that belongs to the international core of the study. The 'regional' set can be selected by the regional group of countries based on recent developments related to education. For instance, countries from the same region might be interested in comparing each other with regard to the relationship between the availability and use of Information and Communication Technologies in the classroom and student achievement.

Beyond the set of background factors, which can be different across regions, some factors could be part of the *core* of the study in all participating countries. Examples of international core factors could be 'coverage of the content of the international achievement test', 'time on task' and 'teacher's workload in the subject under investigation.' IEA might be interested in publishing trends regarding non-regionally specific factors, merely at a descriptive level. Understanding differences in student achievement by means of relational analysis will particularly be useful within regions. The international validity and reliability issues discussed in the previous section could also be reduced to 'regional' validity and reliability for the regional set of variables.

The remaining stages will mainly be conducted within each region of reference countries. At the same time, in each stage attention will be paid to the core part of the study.

Stage 2: Elaboration of the basic conceptual framework: review studies
On the basis of the concrete research questions within a region, the conceptual framework can be developed. This task will be simpler than in previous worldwide studies with their general research questions. For each region, the organizing conceptual framework from chapter 3 (see Figure 3-8) can be taken as the basic conceptual framework. Each regional group of reference countries could gear this framework to their own needs and questions. Review studies within and across countries belonging to the same region could be used or conducted to develop the framework further. Relevant factors located at all curricular and educational levels of the framework, including the country level, need to be selected.
The elaborated conceptual framework forms the foundation of the instruments to be used in the regional part of the study as well as in the core part.

Stage 3: Operationalization of the conceptual framework: in-depth case studies and large-scale pilot
Depending on the results of the review studies proposed in stage 2, it might be useful to arrange a few in-depth case studies within the reference countries to check the relevance of some factors. These case studies can result in final indications for relevant factors to be included in the conceptual framework both for the regional and the core part of the study.

Once the list of factors is selected, each factor should be operationalized. First, the appropriateness of existing operationalizations (scales) of factors developed in previous studies can be judged against the meaning factors have in the new study. If existing scales are judged as inappropriate or if no scales are available, scales could be developed in a second round of case studies within all countries belonging to the same region. Interviews or lesson observations are possible methods to find out how respondents interpret sets of items referring to a factor. The final goal of these case studies is the maximalization of clarity of questions across the reference countries (i.e., international content validity). The number of cases is dependent on the research question(s) to be addressed and the number of factors that is to be examined. The scales resulting from the case studies should be piloted in a large-scale survey. Analysis of the pilot data should provide indications of the internal consistency of scales in the individual countries (international reliability). Preferably, the pilot should be continued in countries where the reliability is insufficient.

Stage 4: Large-scale survey (e.g., a trend study)

The large-scale main study addresses the main regional and 'worldwide' international research questions. In this study, the instruments are used which were developed in stage 3. The collected data will be analyzed to determine the extent to which the study reached its function of understanding cross-national (within a region) differences in student achievement.

Stage 5: Selected case-studies to deepen, validate, and to complete the survey results

The survey results might give rise to in-depth analyses. Depending on the research questions and the outcomes of stage 4, respondents in the participating countries (schools, teachers, and/or students) from the survey could be selected to investigate the research questions further, in a more qualitative way. Cross-national comparisons could reveal weaknesses in students' achievement which were not expected by some countries. These countries will be interested in causes for low student performance. For instance, relationships between some school and classroom factors, and student achievement could be investigated in case studies to provide a better understanding of such relationships as possible reasons for low student performance.

Case studies could also be used as further validation of scales (measures of selected factors from the conceptual framework) applied in the survey. Moreover, case

studies could result in the appearance of new factors, not yet present in the conceptual framework, that resulted from stage 2.
Generally, case studies can be of exploratory, descriptive, or explanatory nature (Yin, 1994). Possible criteria to select cases are best practices and outliers with respect to the objective of the study. For example, the objective could be to examine lesson structures in classrooms with students with relatively low social economic status that, as shown from the survey data, influenced student achievement in mathematics positively.

Qualitative methods (e.g., interviews with experts) could also be applied to check the feasibility of recommendations resulting from the survey study.

Beyond case studies it is recommended to invest in secondary analysis on the data sets collected in large-scale international comparative achievement studies. Results of secondary analyses can provide further indications for understanding cross-national differences in student achievement and they can provide pointers for improvement of the international background questionnaires. IEA is stimulating funding of secondary analysis on TIMSS data by organizing workshops for interested researchers.

Stage 6: Evaluation: recommendations to improve the (trend) study
The answers to the research question should be evaluated to determine whether the ambition of the international comparative study was fulfilled and how future studies could be improved. The review of TIMSS and its two predecessors (see chapter 2) demonstrate that studies could learn more from each other than has happened to date. For instance, in the set-up of the TIMSS study, the experiences from the SIMS study with regard to the conceptual foundation of the background factors was taken into account inadequately.
The evaluation of the usefulness of the results of a large-scale international comparative achievement study should not only be done by the researchers, but also by the potential users of the results, i.e., policymakers and educational practitioners.

The proposed six-stage plan will cost much more money than many of the international comparative achievement studies so far. If budgets for future studies

are too small to carry out all six steps, it is recommended to focus on stages 1, 2, 3, 4, and 6. The case studies proposed in stage 5 could be omitted if finances are limited. Also, in stage 3, the number of case studies could be limited. However, in stage 3 a pilot of developed instruments is necessary in all participating countries to determine their potential for international comparability.

Two examples

One example of an IEA study in which quantitative and qualitative were combined is the TIMSS-1999 Video Study. The TIMSS-1999 Video Study was conducted parallel to TIMSS-1999 with independent samples. In the Video Study, large-scale international comparative observations were combined with individual case studies (Stigler, Gallimore and Hiebert, 2000). Random selected mathematics and science lessons were taped and the teachers and students of the lessons taped completed background questionnaires. The background questionnaires included questions about the contents of the lessons and about teacher's and student's background. The main research questions of the Video Study could be answered by means observations from the video screen. An example of a research question is 'How is the lesson structured?' and one of the related questions is 'How much time is spent studying mathematics?' This specific question can be inserted in a written questionnaire to be administered in a survey. However, in the Video Study the lesson structure can be observed more accurately than was possible by means of the questionnaire. Activities during the videotaped lessons were defined as follows: *Time devoted to mathematical work (e.g., solving problems), mathematical organization (e.g., collecting homework), and non-mathematical activity (e.g., discussing an upcoming field trip) as recorded for each lesson.* Using these categories, the researcher is more able than on the basis of survey data, to estimate per lesson a score on 'time-on-task.'

Within the estimation, non-learning activities can better be excluded than would ever be possible by means of data collected by means of a questionnaire. Similarly, observations from videotaped lessons can provide scores on other important instructional practice variables. Even without having any comparative analysis available concerning data collected in survey-studies like TIMSS and data collected in a video study it can be recommended to measure classroom background variables by means of Video Studies.

It seems obvious that applying observation methods will lead to more internationally valid and accurate data than methods with written questionnaires

only, can make possible. The 'written' database tells the researchers little about the international validity, while videotapes of lessons can be analyzed by a group of researchers from all countries involved in the large-scale international comparative study. Within this international group, the codes for the various variables can be developed in close cooperation, as was done in the TIMSS-1999 Video Study. The final goal of the code development process was to reach consensus about the definition of the factors and variables and to reach an inter-rater reliability for each code of 85% across all participating countries (Stigler, Gallimore and Hiebert, 2000; Hiebert, Gallimore, Garnier, and Stigler, forthcoming). As a result of this labor-intensive process, the meaning of many scores was clear for each participating country and the scores could be used for comparative purposes.
The comparisons across countries with regard to similarities and differences in instructional practices based on video data will be richer than comparisons based merely on questionnaire data. Measuring instructional practices in a useful way by means of questionnaires is difficult. At the same time, the extra costs of a Video Study must be stressed. Designing a large-scale international comparative Video Study, and developing and analyzing raw observational data costs much more than a large-scale study that is designed in the 'TIMSS' way. Nevertheless, the results of a combined video study – 'written' study (which was not the case in 1999) are much more internationally valid. Therefore, it is recommended to consider the possibility of a video study as part of a large-scale international comparative achievement study to enhance the accomplishment of the 'understanding' function of differences across countries on the achievement in a core subject and influencing background variables.

A second example of an IEA study which could be considered as a study that takes more than two out of the six proposed stages into account is the Second Information Technology in Education Study (SITES). The SITES study started in 1997 and will last until 2005. In this study, countries were not separated by regions. However, this would be preferable for the same reasons as were discussed for TIMSS. The main objective of SITES is the availability and the use of Information and Communication Technology (ICT) in different sections of education (Pelgrum & Anderson, 1999).
The SITES study was set up into three modules. In module 1, a general survey was conducted to make an inventory of ICT in primary education and in lower and higher secondary education (see stage 4 of the six-stage plan). In module 2,

emerging practices were examined by means of selective case studies (see stage 5). Results of the investigation of these best practices are also used to improve the survey instruments from SITES module 1 (see stage 3). In module 3, a survey will be organized, making use of results of module 1 and module 2.

5.5 EPILOGUE

The six-stage plan proposed in the previous section, reflects the recommendations formulated on the basis of the study conducted in this thesis. The results of the investigation of the TIMSS case showed that the 'describing' function could better be fulfilled than the 'understanding' function.
The description of achievement data from many education systems participating in TIMSS could be seen as a benefit of the study. The international achievement test is reliable and valid across countries. The background data, collected to understand cross-national differences in student achievement, can be described as well. However, the way it was done in TIMSS – by means of aggregated data at country level – can not prevent users from making the ecological mistake.
Moreover, information about the statistical reliability and content validity from each country of sets of items indicating potentially effectiveness enhancing factors needs to be available to conduct relational analysis properly. Relational analysis on background and achievement data of participating countries in large-scale studies serves the 'understanding' function. From the TIMSS case it was concluded that relational analysis on background and achievement data could benefit from a more profound development of three important components of large-scale international comparative achievement studies.
First, the conceptual framework of such studies needs to be well-developed in terms of definition, and theoretical and empirical foundation of potentially effectiveness enhancing factors at all educational levels (student, classroom, school, and country) and curriculum levels (intended, implemented, and attained). The framework should be appropriate to facilitate selection of key factors for all countries the international research questions will be addressed to.
Second, the operationalization of all selected factors should be developed precisely within and across all participating countries. In each country case studies and a pilot study are methods that should be applied aiming at valid formulations of questionnaire items and reliable sets of items forming scales. After the main survey

of the study, case studies are appropriate methods to deepen, validate and to complete the survey results.

Third, the design of international comparative studies could be adapted from a worldwide study (uniform instrumentation for all countries) to a core worldwide study with regional options. All countries participate in the relatively small core part. The regional options consist of 'separate' studies in groups of countries that would like to be compared with each other (reference countries). Within each region (not necessarily geographical ones) understanding cross-national differences in student performances on the international achievement test would be better possible than within a worldwide study. Background factors can be selected and operationalized in the way proposed above, but more sophisticatedly than in worldwide studies with countries with different cultural and economic background.

Addendum

Finally, some personal remarks are made. In this thesis, TIMSS was criticized on important components. I have been personally involved in TIMSS since 1997 and in TIMSS-Video Study since 1998, as National Research Coordinator for the Netherlands. Between 1993 and 1996, I was already one of the TIMSS researchers. The analyses conducted in this thesis and the reflections are intended to contribute to the improvement of the utility of TIMSS results. TIMSS is a unique project in the world of educational research which makes the exchange possible of professional experience with researchers from various countries and different backgrounds around the world.
The way TIMSS is organized under the auspices of IEA (secretariat is based in Amsterdam, the Netherlands) is an excellent example of how international comparability can be guaranteed as far as the preparation of the instruments, the data collection procedures, and data management are concerned. The study is coordinated by the International Study Center (ISC) at Boston College (Boston, USA). This Center is the hub of the study, steering the participating education systems from all over the world. The ISC controls the quality and uniformity of the sampling plans and the instruments (translations from English) used in each country, the uniformity of the data collection procedures in each country, and the quality of the international data processing and data analysis.

Two other Centers are part of the international coordination and contribute to the international comparability of the data. Statistics Canada is responsible for the quality of the sampling frames and the sampling procedures of each country. The International Data Processing Center of IEA located in Hamburg (Germany) is responsible for the international comparability of the data sets.

It is my sincere hope that this thesis fulfills its intention of increasing the utility of TIMSS results in a way the participating education systems would prefer.

References

Adams, R.J., & Gonzales, E.J. (1996). The TIMSS test design. In: M.O. Martin, & D.L. Kelly, *Third International Mathematics and Science Study. Technical Report. Volume I: Design and Development*. Boston: Center for the Study of Testing, Evaluation and Educational Policy.

Afrassa, T.M. (1999). *Mathematics achievement at the lower secondary school stages in Australia and Ethiopia: a comparative study of standards of achievement and student-level factors influencing achievement* (Doctoral dissertation). Melbourne: Flinders University.

Akker, J.J.H., van den, (1988). *Ontwerp en implementatie van natuuronderwijs* [The design and implementation of science]. Lisse: Swets & Zeitlinger.

Anderson, L.W., Ryan, D.W., & Shapiro, B.J. (Eds.). (1989). *The IEA Classroom Environment Study (CES)*. New York: Pergamon Press.

Beaton, A.E. (1987). *Implementing the new design: The NAEP 1983-84. Technical report. National Assessment of Educational Progress*. Princeton New Jersey: Educational Testing Service.

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: Center for the Study of Testing, Evaluation and Educational Policy.

Beaton, A.E., Postlethwaite, T.N., Ross, K.N., Spearritt, D., & Wolf, R.M. (1999). *The benefits and limitations of international educational achievement studies*. Perth: International Academy of Education.

Bokhove, J., Schoot, S. ,van der, & Eggen, T. (1996). *Balans van het rekenonderwijs aan het einde van de basisschool 2: Uitkomsten van de tweede peiling rekenen/wiskunde einde basisonderwijs* [Evaluation of mathematics education at the end of the primary school 2: Results of the second measurement mathematics at the end of primary school]. Arnhem: Cito.

Bos, K.Tj., & Kuiper, W.A.J.M. (1999). Modelling TIMSS data in a European comparative perspective: exploring influencing factors on achievement in mathematics in grade 8. *Educational Research and Evaluation. An international journal on theory and practice, 5*(2), 157–179.

Bos, K.Tj., Kuiper, W.A.J.M., & Plomp, Tj. (1999). National profiles. Student performance and curricular appropriateness in the Netherlands. *Studies in Educational Evaluation, 25*(3), 269–276.

Bos, K.Tj. & Vos, F.P. (2000). *Nederland in TIMSS-1999. Exacte vakken in leerjaar 2 van het voortgezet onderwijs.* [The Netherlands in TIMSS-1999. Mathematics and science in year 2 of secondary education]. Enschede: OCTO, University of Twente.

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*(1), 79–94.

Bos, K.Tj. (2001). *Results of Partial Least Squares analysis on TIMSS-1995 data from Belgium Flanders, Germany, and the Netherlands.* Enschede: OCTO, University of Twente.

Bosker, R.J., & Scheerens, J. (1999). Openbare prestatiegegevens van scholen: nuttigheid en validiteit [Publishing school performance indicators: usefulness and validity. *Pedagogische Studiën, 76*(1), 61–73.

Bryk, S., & Raudenbush, S.W. (1992). *Hierarchical linear models: applications and data analysis methodes.* Newbury Park, CA: Sage publications.

Burstein, L. (Ed.). (1993). *The IEA study of Mathematics III: student growth and classroom porcesses.* Oxford: Pergamon Press.

Campbell, J.R. (1996). *PLSPath Primer, 2nd edition*. New York: St. John's University.

Carroll, J.B., (1963). A model of school learning. *Teachers College Record*, *64*, 723–33.

Carroll, J.B. (1989). The Carroll Model: a 25-year retrospective and prospective view. *Educational Researcher*, *18*, 26–31.

Cheng, Y.C., & Cheung, W.M. (1999). Lessons from TIMSS in Europe: An observation from Asia. *Educational Research and Evaluation, 5*(2), 227–236.

Comber, L.C., & Keeves, J.P. (1973). *Science education in nineteen countries.* Stockholm: Almqvist and Wiksell.

Cooney, T.J. (1992). Classroom processes: conceptual considerations and design of the study. In: L. Burstein (Ed.), *The IEA study of Mathematics III: student growth and classroom processes (pp. 15-27).* Oxford: Pergamon Press.

Creemers, B.P.M. (1991). *Effectieve instructie. Een empirische bijdrage aan de verbetering van het onderwijs in de klas [Instructional effectiveness: An empirical contribution to the improvement of education in the classroom]*. 's Gravenhage: SVO.

Creemers, B.P.M. (1994). *The effective classroom.* London: Cassell.

Damme, J., van, (1999). *TIMSS-R in Vlaanderen. Het Vlaams Luik.* [TIMSS-R in Belgium Flanders. The Flemish option]. Leuven: Katholieke Universiteit.

Edmonds, R.R. (1979). *A discussion of the literature and issues related to effective schooling.* Cambridge, MA: Center for Urban Studies, Harvard Graduate School of Education.

Elley, W. B. (1992). *How in the world do students read?* Amsterdam: IEA.

Fraser, B.J., Walberg, H.J., Welch, W.W., & Hattie, J.A. (1987). Synthesis of educational productivity research. *International Journal of Educational Research*, *11*(2), 145–252.

Falk, R.R., & Miller, N.B. (1991). *A primer for soft modeling.* Akron, Ohio: University of Akron Press.

Foxman, D. (1992). *Learning mathematics and science. The second International Assessment of Educational Progress in England.* Slough: National Foundation for Educational Research.

Freeland, B. (2000). *International comparisons of vocational education and training.* Leabrook, Australia: National Centre for Vocational Education Research.

Freudenthal, H. (1975). Pupils' achievement internationally compared – the IEA. *Educational studies in mathematics*, *6*, 127–86.

Fuller, B., & Clarke, P. (1994). Raising school effects while ignoring culture? Local conditions and the influence of classroom tools, rules and pedagogy. *Review of educational research, 64,*119–157.

Glaser, B.G. & Strauss, A.L. (1967). *The discovery of grounded theory. Strategies for qualitative research.* Chigago: Aldine Publishing Company.

Goldstein, H., & Lewis, T. (Eds.). (1996). *Assessment: problems, developments and statistical issues.* Chichester: John Wiley & Sons Ltd.

Gonzales, E.J., & Smith, T.A. (Eds.). (1997). *User guide for the TIMSS international database.* Boston: Boston College.

Goodlad, J.I. (Ed.). (1979). *Curriculum inquiry: The study of curriculum practice.* New York: McGraw-Hill.

Goodlad, J.I., Klein, M. F., & Tye, K.A. (1979). The domains of curriculum and their study. In: J.I. Goodlad (Ed.), *Curriculum inquiry: The study of curriculum practice* (pp. 43–76). New York: McGraw-Hill.

Hauser, R.M. (1973). Disaggregating a social psychological model of educational attainment. In: A.A. Goldberger & O.D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 255–284). New York: Seminar Press.

Hanushek, E.A. (1995). Interpreting recent research on schooling in developing countries. *The world bank research observer, 10,* 227–246.

Haan, D., de, (1992). *Measuring test-curriculum overlap (Doctoral dissertation).* Enschede: University of Twente.

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 College: Center for the Study of Testing, Evaluation, and Educational Policy.

Hiebert, J., Gallimore, R., Garnier, H. & Stigler, J.W. (forthcoming). *Mathematics teaching in seven countries: Results of the TIMSS-R Video Study.* Washington, D.C.: National Center for Education Statistics (NCES).

Howie, S.J. (in press). *English language proficiency and contextual factors influencing mathematics achievement of secondary school pupils in South Africa.* (Doctoral dissertation*)* Enschede: University of Twente.

Howson, G. (1999). The value of comparative studies. In: G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (pp. 165–188). London: Falmer Press.

Hox, J.J. (1994). *Applied Multilevel Analysis.* Amsterdam: TT-Publikaties.

Hox, J.J. (1995). AMOS, EQS, and LISREL for windows: a comparative review. *Structural equation modeling. 2,* 79–91.

Husén, T. (Ed.). (1967). *International Study of Achievement in Mathematics: a comparison of twelve countries (Vols. I and II).* New York: Wiley.

International Association for the Evaluation of Educational Achievement (1995). *Test and background questionnaires TIMSS-1995.* Amsterdam: IEA Secretariat.

Jöreskog, K.G., & Sörbom, D. (1989). *LISREL 7, user's reference guide.* Mooresville: Scientific software.

Kaiser, G. (1999). International comparisons in mathematics education under the perspective of comparative education. In: G. Kaiser, E. Luna, & I. Huntley (Eds.), *International comparisons in mathematics education* (pp. 3–15). London: Falmer Press.

Keeves, J.P. (1996). *The world of school learning: selected findings from 35 years of IEA research.* Amsterdam: IEA.

Keeves, J.P. (Ed.). (1997). *Educational Research, Methodology, and Measurement: an International Handbook* (2nd edn). Oxford: Pergamon Press.

Kelly, G.P. & Altbach, P.G. (1988). Alternative approaches in comparative education. In: T.N. Postlethwaite (Ed.), *The encyclopedia of comparative education and national systems of education* (pp. 13-–19). Oxford: Pergamon Press.

Kifer, E., & Burstein, L. (1993). Concluding thought: what we know, what it means. In: Burstein, L. (Ed.), *The IEA study of Mathematics III: student growth and classroom porcesses.* (pp. 329–341). Oxford: Pergamon Press.

Klein, M.F. (1991). A conceptual framework for currciculum decision making. In: M.F. Klein (Ed.), *The politics of curriculum decision-making. Issues in centralizing the curriculum.* (pp. 24–41). New York: State university of New York.

Knuver, J.W.M., & Doolaard, S. (1997). *Rekenen-wiskunde en natuuronderwijs op de basisschool. Nederlands aandeel in TIMSS in populatie 1* [Mathematics and science in primary education: The Dutch contribution to TIMSS population 1]. Enschede: OCTO, University of Twente.

Kotte, D. (1992). *Gender differences in science achievement in 10 countries – 1970/71 to 1983/84.* Frankfurt, Germany: Verlag Peter Lang.

Kuiper, WA.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 in populatie 2* [Mathematics and science in year 1 and 2 of secondary education: The Dutch contribution to TIMSS population 2]. Enschede: OCTO, University of Twente.

Kuiper, WA.J.M., Bos, K.Tj., & Plomp, Tj. (1999). Mathematics achievement in the Netherlands and appropriateness of the TIMSS mathematics test. *Educational Research and Evaluation. An international journal on theory and practice. 5*(2), 85–104.

Kulik, J.A., & Kulik, C.L.C. (1987). Effects of ability grouping on student achievement. *Equity and Excellence*, *23,* 22–30.

Levine, D.U., & Lezotte, L.W. (1990). *Unusually effective schools. A review and analysis of research and practice.* Madison (USA): National Center for Effective Schools Research and Development.

Lietz, P. (1996). *Changes in reading comprehension across cultures and over time.* (Doctoral dissertation). New York: Waxmann.

Lundberg, I., & Linnakyla, P. (1992). *Teaching reading around the world.* Amsterdam: IEA.

Marinot, M.J., Kuhlemeier, H.B., & Feenstra, H.J.M. (1988). Het meten van affectieve doelen: de validering en normering van de belevingsschaal voor wiskunde [The measurement of affective goals: the validation and specification of the self-perception scale for mathematics]. *Tijdschrift voor Onderwijsresearch, 13*(2), 65–76.

Martin, M.O., & Kelly, D.L. (1996). *Third International Mathematics and Science Study. Technical Report. Volume I: Design and Development.* Boston: Center for the Study of Testing, Evaluation and Educational Policy.

Martin, M.O., Rust, K., & Adams, R.J. (Eds.). (1999). *Technical standards for IEA studies.* Amsterdam: IEA Secretariat.

Martin, M.O., Mullis, I.V.S., Gregory, K.D., Hoyle, C., & Shen, C. (2000). *Effective schools in science and mathematics. IEA's Third International Mathematics and Sicence Study.* Boston: International study Center, Lynch School of Education, Boston College.

McKnight, C.C., Crosswhite, F.J., & Dossey, J.A. (1997). *The underachieving curriculum: assessing U.S. school mathematics from an international perspective.* Champaign, Il: Stipes Publishing Company.

McLean, L.D. (1996). Large-scale assessment programmes in different countries and international comparisons. In: H. Goldstein, & T. Lewis (Eds.), *Assessment: problems, developments and statistical issues* (pp. 189–207). Chichester: John Wiley & Sons Ltd.

Meelissen, M.R.M., & Bos, K.Tj. (2001). *Gender differences in mathematics achievement in relation to test, student, and teacher characteristics. Secondary analyses on TIMSS-1995 data on mathematical performances of Dutch 10- and 14 year old students.* Enschede: OCTO, University of Twente.

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

Munck, I.M.E. (1979). *Model building in comparative education. Applications of the LISREL method to cross-national survey data.* Stockholm: Almqvist & Wiksell International.

Newton, P., Adams, E., Sewell, J., & Whetton, C. (2002). *An evaluation of the year 7 progress tests and years 7 and 8 optional tests in english and maths. A report for the qualifications and curriculum authority.* London: National Foundation for Educational Research (NFER).

Nitsaisook, M., & Postlethwaite, T.N. (1986). Teacher effectiveness research: An example from Thailand. *International Review of Education*, *32*(4), 423–438.

Noah, H.J. (1988). Methods of comparative education. In T.N. Postlethwaite (Ed.), *The encyclopedia of comparative education and national systems of education* (pp. 10–13). Oxford: Pergamon Press.

OECD. (1992). *Education at a glance I*. Paris: OECD Center for Educational Research and Innovation.

OECD. (1993a). *Education at a glance II*. Paris: OECD Center for Educational Research and Innovation.

OECD. (1993b). *Handbook on International Educational Indicators*. Paris: OECD Center for Educational Research and Innovation.

OECD. (1997). *Education at a glance*. Paris: OECD, Center for Educational Research and Innovation.

OECD. (2000). *Education at a glance*. Paris: OECD, Center for Educational Research and Innovation.

Peak, L. (Ed.). 1996). *Pursuing excellence. A study of U.S. eighth-grade mathematics and science teaching, learning, curriculum, and achievement in international context*. Washington: National Center for Education statistics.

Pelgrum, W.J., & Plomp, Tj. (Eds.). (1993). *The IEA study of computers in education: implementation of an innovation in 21 education systems*. Oxford: Pergamon Press.

Pelgrum, W.J., & Anderson, R.E. (Eds.). (1999). *ICT and the emerging paradigm for life long learning: a worldwide educational assessment of infrastructure, goals and practices*. Amsterdam: IEA.

Plomp,Tj. (1998). The potential of international comparative studies to monitor the quality of education. *Prospects. 28*(1), *45–59*.

Postlethwaite, T.N. (Ed.). (1988). *The encyclopedia of comparative education and national systems of education*. Oxford: Pergamon Press.

Postlethwaite, T.N., & Wiley, D.E. (1992). *The IEA study of science II: science achievement in twenty-three countries*. Oxford: Pergamon Press.

Postlethwaite, T.N., & Ross, K. (1994). *Effective schools in reading: implications for educational planners*. Amsterdam: IEA.

Postlethwaite, T.N. (1999). *International studies of educational achievement: methodological issues*. China Hong Kong: Comparative Educaton Research Centre, University of Hong Kong.

Reezigt, G.J., Guldemond, H., & Creemers, B.P.M. (1999). Empirical validity for a comprehensive model on educational effectiveness. *School Effectiveness and School Improvement, 10*(2), 193–216.

Reynolds, D., & Farrell, S. (1996). *Worlds apart?* London: OFSTED.

Robin, D. (1993). Teachers' strategies and students' achievement. In: L. Burstein (Ed.), *The IEA study of Mathematics III: student growth and classroom processes (pp. 225–231)*. Oxford: Pergamon Press.

Robitaille, D.F., & Garden, R.A. (1989). *The IEA study of Mathematics II: Contexts and outcomes of school mathematics*. Oxford: Pergamon Press.

Robitaille, D.F. (1993). *Curriculum frameworks for mathematics and science. TIMSS Monograph no.1*. Vancouver: Pacific Educational Press.

Robitaille, D.F., & Travers, K.J. (1992). International studies of achievement in mathematics. In: J.S. Grouws (Ed.), *Handbook of research on mathematics learning and teaching (pp. 687–709). New York: Macmillan,.*

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

Robitaille, D.F., & Garden, R.A. (1996). *Research questions & study design. TIMSS Monograph no. 2*. Vancouver: Pacific Educational Press.

Robitaille, D.F. (Ed.). (1997). *National contexts for mathematics and science education. An encyclopedia of the education systems participating in TIMSS*. Vancouver: Pacific Educational Press.

Rosier, M.J. (1997). Survey research methods. In: J.P. Keeves (Ed.), *Educational research, methodology, and measurement: an international handbook* (pp. 154–61). Cambridge, UK: Pergamon Press.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.

Scheerens, J., Vermeulen, C.J.A.J., & Pelgrum, W.J. (1989). Generalizibility of instructional and school effectiveness indicators across nations. *International journal of educational research, 13,* 789–800.

Scheerens, J. (1990). School effectiveness and the development of process indicators of school functioning. *School effectiveness and school improvement*, *1*(1), 61–80.

Scheerens, J. (1992). *Effective schooling. Research, theory and practice*. London: Cassell.

Scheerens, J. (1999). *School effectiveness in developed and developing countries: a review of the research evidence*. Washington, D.C.: The World Bank.

Scheerens, J., & Bosker, R.J. (Eds.). (1997). *Foundations of educational effectiveness*. London: Routledge.

Schieber H. (1983). *PLSPATH Version a: program manual*. Hamburg: University of Hamburg, Department of Education.

Schmidt, W.H. (1993). *TIMSS educational opportunity model. Detailed instrumentation and indices development*. Project document (Doc.Ref.: ICC713/NPC276). Amsterdam: IEA.

Schmidt, W.H., & Kifer, E. (1989). Exploring relationships across population A systems. In: D.F. Robitaille & R.A. Garden, *The IEA study of Mathematics II: Contexts and outcomes of school mathematics* (pp. 209–231). Oxford: Pergamon Press.

Schmidt, W.H., & Burstein, L. (1993). Concomitants of growth. In: L. Burstein (Ed.), *The IEA study of Mathematics III: student growth and classroom porcesses (*pp. 309–327). Oxford: Pergamon Press.

Schmidt, W.H., & Cogan, L. (1996). In: M.O. Martin & D.L. Kelly. *TIMSS Technical Report Volume I: Design and Development* (pp. 5-1–5-13). Boston: Center for the Study of Testing, Evaluation and Educational Policy, Boston College.

Schmidt, W.H., McKnight, C.C., Valverde, G.A., Houang, R.T., & Wiley, D.E. (1997). *Many visions, many aims, Volume I: A cross-national investigation of curricular intentions in school mathematics*. Dordrecht: Kluwer Academic Publishers.

Scudder, D.F. (2000). *Class-size reduction evaluation, 1999-2000. A report to the North Carolina Department of public instruction*. Releigh: NC. Department of evaluation and reserach.

Sellin, N. (1989). *PLSPath Version 3.01. Application manual*. Hamburg: University of Hamburg, Department of Education.

Sellin, N. (1990). *PLSPath Version 3.01. Program manual*. Hamburg: University of Hamburg, Department of Education.

Sellin, N. (1992). Partial Least Squares Path Analysis. In: J. Keeves (Ed.), *The IEA Technical Handbook* (pp. 397–412). The Hague: IEA.

Sellin, N., & Keeves, J.P. (1994). Path analysis with latent variables. In: T. Husén & T.N. Postlethwaite (Eds.), *The International Encyclopedia of Education* (pp. 4352–4359). Oxford: Pergamon Press.

Shalabi, F. (2002). *Effective schooling in the West Bank* (Doctoral Dissertation). Enschede: Twente University Press.

Shavelson, R., McDonnell, L., Oakes, J., & Carey, N. (1987). *Indicator systems for monitoring mathematics and science education: a sourcebook*. Santa Monica: Rand Corp.

Smeets, E.F.L. (2000). *Krachtige leeromgevingen en ICT in het onderwijs*. [Powerful learning environment and ICT in education]. Nijmegen: ITS.

Snijders, T.A.B.. & Bosker, R.J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London: SAGE Publications.

Stevens J. (1986). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Erlbaum.

Stigler, J.W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study. Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States*. Washington D.C: National Center for Education Statistics (NCES).

Stigler, J.W., Gallimore, R. & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist, 35*(2), 87–100.

Travers, K.J., & Westbury, I. (1989). *The IEA study of mathematics I: international analysis of mathematics curricula*. Oxford: Pergamon Press.

Travers, K.J. (1993). Overview of the longitudinal version of the second international mathematics study. In: L. Burstein (Ed.), *The IEA study of Mathematics III: student growth and classroom processes.* (pp. 1–14). Oxford: Pergamon Press.

Travers, K.J., & Weinzweig, A.I. (1999). The second international mathematics study. In: G. Kaiser, E. Luna & I. Huntley (Eds.), *International comparisons in mathematics education* (pp. 19 – 29).. London: Falmer Press.

Travers, K.J., Garden, R.A., & Rosier, M.J. (1989). Introduction to the study. In: D.F. Robitaille & R.A. Garden (Eds.), *The IEA Study of Mathematics II: Contexts and Outcomes of School Mathematics* (pp. 1–16). Oxford: Pergamon Press.

Tuijnman, A.C., & Postlethwaite, T.N. (Eds.). (1994). *Monitoring the standards of education. Papers in honor of John P. Keeves.* Oxford: Pergamon Press.

UNESCO (1991). *World Education Report 1991*. Paris: Author.

Vari, P. (Ed.). (1997). *Are we similar in math and science? A study of grade 8 in nine Central and Eastern European countries.* Hungary: IEA.

Vos, F.P. (in press). *Like an ocean liner changing course. The grade 8 mathematics curriculum in the Netherlands* (Doctoral dissertation). Enschede: University of Twente.

Walberg, H.J. (1984). Improving the productivity of America's schools. *Educational leadership*, *41*(8), 19–27.

Wiegersma, S., & Groen, M. (1968). *Resultaten van wiskundeonderwijs. Een verslag van een onderzoek door het Nederlandse Instituut voor Preventieve Geneeskunde TNO uitgevoerd in het kader van het International Educational Achievement Project* [Results of mathematics education: A report of the study by the Dutch institute for preventive medical science, TNO, conducted within the framework of the International Educational Achievement project]. Groningen: Wolters-Noordhoff.

Wold, H. (1982). Soft modelling: The basic design and some extensions. In: K.G. Joreskog & H. Wold (Eds.), *Systems under indirect observation.* (part II, Chapter 1). Amsterdam: North Holland Press.

Wolf, R.M. (1992). International test performance. In: M.C. Alkin (Ed.), *Encyclopedia of Educational Research* (pp. 642–652). New York: Macmillan Publishing Company.

Woodhouse, G. (Ed.). (1995). *A guide to MLn for new users.* London: Institute of Education, University of London.

Wossmann, L. (2000). *Schooling resources, educational institutions, and student performance: the international evidence.* Kiel (Germany): Kiel Institute of World Economics.

Yin, R.K. (1994). *Applications of case study research*. Newbury Park, Calif.: Sage.

Zuzovsky, R., & Aitkin, M. (2000). Multilevel longitudinal analysis of IEA studies on science achievement using SISS and TIMSS data.. *International journal of educational policy, research and practice.. 1*(2) 218–259.

SAMENVATTING (DUTCH SUMMARY)

MOGELIJKHEDEN EN BEPERKINGEN VAN GROOTSCHALIGE INTERNATIONALE VERGELIJKENDE STUDIES NAAR LEERLINGPRESTATIES IN HET ONDERWIJS:

IEA'S TIMSS STUDY

PROBLEEMSTELLING EN ONDERZOEKSVRAGEN

De International Association for the Evaluation of Educational Achievement (IEA) heeft sinds 1964 vele grootschalige internationale vergelijkende studies naar leerlingprestaties georganiseerd in de kernvakken wiskunde, natuurwetenschappelijke vakken en lezen. Deze IEA studies hebben twee algemene doelen:

(i) informatie verschaffen aan onderwijsbeleidsmakers, mensen uit de onderwijspraktijk en andere onderwijsdeskundigen over de kwaliteit van het onderwijssysteem in vergelijking tot relevante referentie landen. Een land kan van andere landen leren door te bekijken hoe het onderwijs er in die andere landen uit ziet.

(ii) bijdragen aan het begrijpen van geconstateerde verschillen tussen landen.

Het eerste doel vereist een beschrijving van de status van een onderwijssysteem in een internationale context in termen van totale test en sub-test scores op internationale prestatietoetsen. Verschillen in gemiddelde scores en in de verdeling van de scores tussen landen kunnen dienen als indicaties voor de kwaliteit van een onderwijssysteem. Dergelijke beschrijvingen vormen voor beleidsmakers de basis voor een 'internationale' spiegel. Landen kunnen op hun eigen manier hun resultaten vergelijken. Ze kunnen zich bijvoorbeeld afvragen wat in de landen (waar ook ter wereld) met de hoogste leerlingprestaties de belangrijkste factoren zijn die samenhangen met die prestaties. Andere landen willen zichzelf vergelijken met landen uit hun eigen (geografische) regio.

Prestaties van leerlingen op een toets zijn niet de enige bron waaruit geput wordt om het eerstgenoemde doel te bereiken. Om te kunnen nagaan wat er in andere landen gebeurt in het onderwijs is een beschrijving nodig van onderwijsprocessen op verschillende niveaus in de school (leerling, klas/leraar en school).
Het tweede doel betreft het vinden van verklaringen voor beschreven verschillen in leerlingprestaties en de factoren op verschillende niveaus die daarop mogelijk van invloed zijn. Aan deze 'verklarings' functie kan worden voldaan door vele variabelen die te maken hebben met het onderwijsproces in een internationale vergelijkende context te analyseren inclusief hun relaties met leerlingprestaties. Een voorbeeld van een relatie is die tussen de wijze waarop een leraar lesgeeft en de toetsprestaties van leerlingen.
Een belangrijke voorwaarde voor het beschrijven en begrijpen van landverschillen in leerlingprestaties is dat alle verzamelde gegevens internationaal betrouwbaar en valide zijn. De definitie en operationalisatie van bijvoorbeeld 'student-oriented teaching style' moet eenduidig zijn voor alle respondenten in alle landen die aan een studie deelnemen. De selectie van achtergrondfactoren moet bij voorkeur gebaseerd zijn op hun potentiële samenhang met leerlingprestaties in een kernvak. Bovendien moeten factoren, om bruikbaar te zijn voor beleidsmakers, veranderbaar zijn.
Voor de eerste en tweede internationale vergelijkende IEA studie naar wiskundeprestaties (FIMS en SIMS) is op basis van een review geconcludeerd dat het erg moeilijk is om aan de 'verklarings' functie te kunnen voldoen (hoofdstuk 2). De conceptuele basis van de selectie en operationalisatie van achtergrond-kenmerken bleek in beide studies onvolledig te zijn.

De probleemstelling van dit proefschrift is gericht op de mogelijkheden en beperkingen van grootschalige internationale vergelijkende studies naar leerlingprestaties in een kernvak. De probleemstelling is vertaald in twee onderzoeksvragen. De eerste onderzoeksvraag luidt:

I. *In hoeverre kan variatie in de scores op de gehele TIMSS wiskundetoets in het tweede leerjaar voortgezet onderwijs in Nederland, Vlaanderen en Duitsland worden verklaard door variatie in scores op achtergrondkenmerken op leerling- en klas/school niveau en in hoeverre kunnen deze uitkomsten worden gegeneraliseerd over de drie landen?*

Onderzoeksvraag I is beantwoord door de derde internationale studie die IEA heeft georganiseerd naar wiskundeprestaties – de Third International Mathematics and Science Study (TIMSS) – te bestuderen vanuit het perspectief van de 'verklarings' functie. TIMSS heeft de ambitie om aan landen (onderwijssystemen) gegevens te verschaffen waarmee ze in staat worden gesteld verklaringen te vinden voor verschillen tussen landen in leerlingprestaties op een internationale toets. De TIMSS gegevens ten aanzien van wiskunde van drie buurlanden zijn nader geanalyseerd: Vlaanderen, Duitsland en Nederland.

Op basis van een reflectie op de uitkomsten van de eerste onderzoeksvraag is nagegaan in hoeverre TIMSS aan de 'verklarings' functie kan voldoen. De tweede onderzoeksvraag heeft hierop betrekking en luidt:

II. *Wat kan op grond van de resultaten van de TIMSS case worden geleerd over de conceptuele basis, de instrumentatie en het ontwerp van grootschalige internationale vergelijkende studies naar leerlingprestaties in een kernvak die als doel hebben factoren op te sporen die samenhangen met verschillen tussen landen in leerlingprestaties op een internationale prestatietoets?*

De vraag is hoe de mogelijkheden van grootschalige internationale vergelijkingsstudies vergroot kunnen worden en hoe de beperkingen ervan kunnen worden verkleind.

RESULTATEN VAN DE TIMSS CASE ANALYSES

In TIMSS zijn in elk deelnemend land bij een representatieve steekproef van leerlingen uit het tweede leerjaar voortgezet onderwijs een internationale wiskunde- en sciencetoets en een achtergrondvragenlijst afgenomen. De wiskundeleraren van de getoetste klassen vulden een leraarachtergrondvragenlijst in. Eén van de leden van de leiding van de scholen waarvan de getoetste klassen deel uitmaakten werd gevraagd een achtergrondvragenlijst over kenmerken van de schoolorganisatie in te vullen.
De leerlingen in Vlaanderen presteerden significant beter op de TIMSS wiskundetoets dan de leerlingen in Duitsland en Nederland en datzelfde deden de leerlingen in Nederland ten opzichte van hun leeftijdsgenoten in Duitsland.

De analyse van de TIMSS case bestond uit vier fasen. In de eerste fase is het basis conceptueel raamwerk van TIMSS gereviewed. Dit resulteerde in een 'organizing' conceptueel raamwerk gevuld met factoren die in potentie van invloed kunnen zijn op leerlingprestaties (hoofstuk 3). De inhoud van de achtergrondvragenlijsten die in TIMSS zijn ontwikkeld en afgenomen, is in de tweede fase bestudeerd om indicatoren te vinden voor de in het raamwerk gecategoriseerde factoren (hoofdstuk 4). In de derde fase zijn scores op item sets geanalyseerd om schalen te ontwikkelen (variabelen) voor de geïdentificeerde indicatoren (hoofdstuk 4). In de laatste fase van het analyseplan zijn overeenkomsten en verschillen tussen de drie landen onderzocht met betrekking tot samenhangen tussen de variabelen uit de derde fase en leerlingprestaties op de TIMSS wiskundetoets. Dit is gedaan door middel van eendimensionale padanalyses en meerniveau analyses (hoofdstuk 4). Van elke fase wordt een korte samenvatting gegeven.
In de TIMSS data analyses is gebruik gemaakt van leerling- en klasniveau variabelen. Schoolvariabelen zijn niet in de analyses opgenomen vanwege het ontbreken van een relatief groot aantal schoolvragenlijsten in de Nederlandse data set.

Conceptuele basis

Teneinde factoren te kunnen identificeren die in potentie van invloed kunnen zijn op wiskundeprestaties is eerst het conceptueel raamwerk van TIMSS nader bestudeerd. Dit raamwerk is gebaseerd op het drie curriculum niveau raamwerk uit het tweede internationale wiskunde onderzoek van IEA. Naast drie curriculumniveaus omvat dit IEA model drie onderwijsniveaus (leerling, klas/school en land). Geconcludeerd werd dat het TIMSS kader als raamwerk voldoet om als leidraad te kunnen dienen voor de zoektocht naar potentieel effectieve factoren. De inhoud ervan bleek echter te weinig te zijn uitgewerkt in termen van definities van factoren en de theoretische en empirische basis daarvan. In internationale vergelijkende onderwijsstudies is het van belang dat de empirische basis van het conceptueel raamwerk internationaal georiënteerd is en dat de factoren voor alle landen eenduidig gedefinieerd zijn.
Het raamwerk van het IEA model is voor deze studie overgenomen, met een uitsplitsing van het klas/school niveau in een klas- en een schoolniveau. De inhoud van de clusters van factoren van het basisraamwerk is ingevuld met factoren uit modellen voor instructie- en schooleffectiviteit. Deze modellen zijn tot stand gekomen op basis van review studies naar gedefinieerde sleutelfactoren die in eerder

onderzoek in verschillende (geïndustrialiseerde) landen samen bleken te hangen met leerlingprestaties. Het resulterende 'organizing' conceptueel raamwerk heeft als leidraad gediend in de volgende stappen van de bestudering van de TIMSS case.

Indicatoren in de TIMSS achtergrondvragenlijsten

Aan de hand van het 'organizing' conceptueel raamwerk zijn de TIMSS achtergrondvragenlijsten inhoudelijk onderzocht. Voorbeelden van belangrijke factoren waarvoor in een TIMSS vragenlijst item sets of individuele items zijn gevonden zijn: motivatie van de leerling (bijvoorbeeld leerling attitude ten opzichte van wiskunde en perceptie van de verwachting van de moeder ten aanzien van schoolprestaties), sociale achtergrond van de leerling en klasmanagement en klasklimaat.
Voor een aantal factoren uit de literatuur over onderwijseffectiviteit die in het 'organizing' raamwerk zijn opgenomen, zijn geen indicatoren aangetroffen in de TIMSS vragenlijsten. Voorbeelden hiervan zijn de leercapaciteit van de leerling, kenmerken van het leermateriaal en het beleid van een school ten aanzien van evaluatie van leerlingprestaties.

Van indicatoren naar variabelen

De sets van items in de TIMSS vragenlijsten die vanuit een inhoudelijk perspectief als indicatoren zijn geïdentificeerd, zijn verder geëxploreerd door middel van statistische analyses. Allereerst is de interne consistentie van de scores per item set geanalyseerd door de berekening van de betrouwbaarheidscoëfficiënt Cronbach α. Deze bleek in meer of mindere mate tussen landen te verschillen. Het interval waarin de verschillen in Cronbach α-coëfficiënten vielen, varieerde van .05 tot .20.
De bivariate produkt-moment correlatiecoëfficiënten tussen enerzijds de leerlingscores op de TIMSS wiskundetoets en anderzijds de scores op de achtergrondfactoren bleek tussen landen wel in sterkte, maar niet in richting te verschillen. De meeste bivariate correlaties varieerden tussen .10 en .20. Het belangrijkste criterium voor de opname van een achtergrondvariabele in de volgende stap van de analyses (padanalyses) was dat de bivariate correlatie coëfficiënt met de TIMSS wiskundetoetsscores groter was dan |.10| in tenminste twee van de drie onderzochte landen.

Eendimensionele exploratieve pad analyses en meerniveau analyses

De meest geschikte technieken om relaties tussen verschillende achtergrondvariabelen en wiskundeprestaties te onderzoeken zijn die technieken die rekening houden met het geneste design van de TIMSS data sets (leerlingen binnen klassen en scholen; in TIMSS vallen de laatste twee samen omdat per school één intacte klas is getoetst): hiërarchische lineaire modeling (HLM) technieken. Het belangrijkste voordeel van HLM technieken (bijvoorbeeld meerniveau analyses) ten opzichte van eendimensionele technieken zoals 'partial least squares techniques (PLS),' is dat in de schatting van de effecten van variabelen op de afhankelijke variabele op een niveau (bijvoorbeeld leerlingniveau), tegelijkertijd rekening wordt gehouden met het effect van variabelen op een ander niveau (bijvoorbeeld klasniveau) van de hiërarchische datastructuur.
In deze studie zijn echter eerst relaties tussen variabelen geëxploreerd door middel van PLS (programma PLSpath). Daarmee is getracht directe en indirecte relaties tussen achtergrondvariabelen en wiskundeprestaties op het spoor te komen in de drie verschillende landen, omdat hiervoor geen duidelijke theorie voorhanden bleek te zijn. Indicaties voor dergelijke relaties zijn nodig om meerniveau analyses uit te kunnen voeren. Door middel van PLS analyses zijn per land een padmodel op leerlingniveau en een padmodel op klasniveau geëxploreerd. In het klasmodel zijn naast klasvariabelen, de geaggregeerde scores op de leerlingvariabelen uit de leerlingmodellen opgenomen. Op basis van de PLS resultaten zijn de variabelen geselecteerd voor opname in hiërarchische lineaire modellen, die zijn geschat door middel van meerniveau analyses.

Resultaten Partial Least Squares pad analyses

De PLS pad analyses hebben uiteindelijk geresulteerd in een klas model met daarin opgenomen zes geaggregeerde leerlingvariabelen en drie klasvariabelen. Het aantal cases per land in de geanalyseerde data sets stond opname van meer variabelen niet toe. De lijst met variabelen met een *direct* verband met wiskundeprestaties is per land verschillend. Voor Duitsland is de lijst korter dan voor de andere twee landen. In het padmodel voor Duitsland komen drie leerlingvariabelen voor ('indicatie opleidingsniveau ouders,' 'houding van de leerling ten opzichte van wiskunde' en 'door leerling waargenomen niveau van veiligheid op school') en een klasvariabele ('door de leraar ervaren beperkingen in het onderwijzen van wiskunde vanwege

kenmerken van de leerlingen'). In het model voor Vlaanderen komen vijf leerlingvariabelen en een klasvariabele voor met een direct effect op wiskundeprestaties en in het model voor Nederland drie leerling- en twee klasvariabelen.
De drie landmodellen verschillen eveneens ten aanzien van de factoren met een *indirect* verband met de afhankelijke variabele. Sommige indirecte relaties komen in slechts één landmodel voor. Een voorbeeld hiervan is de relatie tussen 'leertijd' en 'dekking van de getoetste wiskundestof in het onderwijs' in het model voor Nederland.
Een mogelijke verklaring voor de verschillen tussen de padmodellen is dat andere – niet gemeten – klasfactoren of landspecifieke factoren van invloed zijn op het leerproces in de klas. Anderzijds kunnen de verschillen te maken hebben met de verschillen tussen landen in betrouwbaarheid en inhoudsvaliditeit van de schalen waarmee factoren zijn gemeten.

Resultaten meerniveau analyses

Voor elk land is een hiërarchisch lineair model geschat met op basis van de PLS uitkomsten geselecteerde variabelen. In de landmodellen wordt op klasniveau meer variantie in wiskundetoetsscores gebonden dan op leerlingniveau. De meeste variabelen op klasniveau zijn echter geaggregeerde leerlingvariabelen. Slechts een klein aantal in TIMSS gemeten klasvariabelen bleek in de pad- en meerniveau modellen een rol te spelen.
De datasets van de drie landen zijn gecombineerd in een 'pooled' dataset. Voor deze data is een HLM model geschat om factoren op te sporen waarvan de invloed op wiskundeprestaties tussen landen verschillen. In deze 'pooled' modellen is gebruik gemaakt van alle factoren die in TIMSS zijn gemeten en die in minimaal één van de landen bleken samen te hangen met wiskundeprestaties. De identificatie van de leerlingen naar land bleek de verklaarde variantie tussen klassen ten opzichte van het model zonder landidentificatie, te vergroten. Mogelijk spelen landspecifieke factoren hierbij een rol.
In de HLM modellen bleek een aantal leerlingvariabelen positief gerelateerd te zijn aan de afhankelijke variabele. Voorbeelden hiervan zijn 'sexe van de leerling,' 'houding van de leerling' en 'huiswerk maken is belangrijk om goede cijfers te halen'. Landen kunnen wat leerlingvariabelen betreft niet veel van elkaar leren, omdat de relatie ervan met wiskundeprestaties in alle landen in dezelfde richting wijst.

Klasvariabelen die veranderbaar zijn bieden hiertoe meer gelegenheid. Drie van deze variabelen hangen in het 'pooled' model positief samen met leerprestaties: 'door de leraar ervaren beperkingen in het onderwijzen van wiskunde vanwege kenmerken van de leerlingen', 'de werklast van de leraren (in termen van het percentage van de aanstellingstijd dat de leraar lesgeeft in het onderzochte vak)' en 'dekking van de getoetste wiskundestof in het onderwijs'. Duitsland kan, bij wijze van voorbeeld, iets van de andere twee landen leren door de werklast van de wiskundeleraren tegen het licht te houden. In Duitsland presteerden de leerlingen significant minder goed op de TIMSS wiskundetoets dan in de twee andere landen. De wiskundeleraren in Duitsland zijn gemiddeld voor de helft van hun werktijd bij een school aangesteld als wiskundeleraar. In Vlaanderen en Nederland is dit gemiddelde percentage hoger dan 70%. Omdat 'werklast' in het totale HLM model een positief effect laat zien op leerlingprestaties kan de beleidsmakers in Duitsland worden aangeraden het aanstellingsbeleid van leraren nader te beschouwen.

REFLECTIES EN AANBEVELINGEN

De analyses van de TIMSS case laten zien dat het beschrijven van overeenkomsten en verschillen tussen landen in prestaties van leerlingen op een internationale toets eenvoudiger is dan het vinden van verklaringen hiervoor. Een belangrijke voorwaarde voor de beschrijving van de verschillen tussen landen op de internationale prestatietoets is dat die toets voor alle deelnemende landen voldoende betrouwbaar en valide is. In TIMSS bleek dit tot op zekere hoogte het geval te zijn. Voor het begrijpen en verklaren van prestatieverschillen tussen landen is het vervolgens erg belangrijk dat de context waarin het leren plaatsvindt in alle landen zo betrouwbaar en valide mogelijk wordt meten. De context moet zo breed mogelijk worden opgevat en omvat factoren op leerling-, klas-, school- en landniveau.
Uit de analyse van de TIMSS case komt naar voren dat een aantal belangrijke componenten van TIMSS verder moet worden ontwikkeld om beter aan de verklaringsfunctie te kunnen voldoen (vraagstelling II). Deze componenten betreffen het conceptueel kader, de betrouwbaarheid en validiteit van de operationalisaties van factoren uit het conceptueel kader en het ontwerp van de studie.

Conceptuele basis

De conceptuele basis van de selectie van achtergrondfactoren kan in twee stappen worden versterkt. Allereerst kunnen de internationale vraagstellingen worden toegespitst op een set kernfactoren die mogelijk van invloed zijn op leerlingprestaties en die door de deelnemende landen zijn geselecteerd. Meer specifieke vraagstellingen kunnen beter worden vertaald in betrouwbare en valide instrumenten dan brede vraagstellingen zoals de TIMSS vragen. Vervolgens kan het conceptueel kader worden uitgewerkt voor de set kernfactoren die moet worden onderzocht om de gespecificeerde vraagstellingen te kunnen beantwoorden. Het 'organizing' raamwerk dat in hoofdstuk 3 van dit proefschrift is ontwikkeld op basis van het IEA onderzoeksmodel en onderwijseffectiviteitsstudies, kan als uitgangspunt voor het te specificeren raamwerk worden gehanteerd. Nader literatuuronderzoek kan leiden tot verdere invulling van het raamwerk. Hierin moeten ook school- en systeem(land)factoren uitdrukkelijk worden betrokken.

Operationalisatie van achtergrondfactoren

Uit de TIMSS case bleek dat bij verschillen tussen landen in de samenhang tussen achtergrondfactoren en leerlingresultaten de betrouwbaarheid van de gebruikte schalen mogelijk een rol heeft gespeeld. Verschillen in betrouwbaarheid kunnen gerelateerd zijn aan gebrek aan eenduidigheid van de betekenis van de factoren. De operationalisatie van de geselecteerde factoren dient daarom op een zorgvuldige wijze te gebeuren in alle landen. De definitie van elke factor moet voor alle landen eenduidig zijn.
Uitgaande van een methode van gegevensverzameling waarin vooral gebruik wordt gemaakt van schriftelijke achtergrondvragenlijsten voor leerlingen, leraren, schoolleiders en land vertegenwoordigers, is het sterk aan te bevelen de operationalisatie van elke factor in gevalstudies voor te bereiden. Door middel van interviews met betrokkenen kunnen vragenlijst vragen tussen landen zo eenduidig (valide) mogelijk worden geformuleerd. Na een proefonderzoek bij een beperkte random steekproef kunnen betrouwbaarheidsanalyses op de verzamelde gegevens uitwijzen in hoeverre item sets (elk bedoeld als operationalisatie van een geselecteerde factor) in elk land voldoende betrouwbaar kunnen worden gemeten. Voor landen waar dit niet het geval is zal de formulering moeten worden aangepast en zal bij voorkeur een nieuw proefonderzoek moeten plaatsvinden.

Na vaststelling van de internationale achtergrondvragenlijsten kan het grootschalige hoofdonderzoek worden uitgevoerd. Op basis van de resultaten van het survey kan per land worden besloten scholen (klassen) te selecteren waar verdiepende studies worden uitgevoerd. De selectiecriteria zijn afhankelijk van de onderzoeks-vraagstellingen. Er kunnen bijvoorbeeld 'best practices' of 'outliers' worden geselecteerd op grond van de survey resultaten die betrekking hebben op een aantal onderzochte kernfactoren. In het geval van 'outliers' kunnen bijvoorbeeld kenmerken van de lespraktijk in klassen met leerlingen die (gezien hun persoonlijke achtergrond) boven verwachting presteerden op de internationale toets worden vergeleken met de lespraktijk in klassen met leerlingen die beneden verwachting presteerden.
Het doel van deze gevalstudies is meervoudig. Verdiepende studies kunnen per land meer inzicht geven in de validiteit van de schalen (item sets) waarmee kernfactoren zijn gemeten. Verder kunnen de survey resultaten worden aangevuld met de uitkomsten van de gevalstudies. Toekomstig grootschalig onderzoek zou ook gebruik kunnen maken van de resultaten van de gevalstudies in de vorm van verbetering van de conceptualisering van de achtergrondvragenlijsten en van de formulering van vragenlijst items.
Naast de uitvoering van verdiepende gevalstudies is het sterk aan te bevelen te investeren in secundaire analyses op de in grootschalige studies verzamelde data sets. Dergelijke analyses kunnen niet alleen leiden tot meer verklaringen voor prestatieverschillen tussen landen, ze kunnen ook leiden tot indicaties voor verbetering van de onderzoeksinstrumenten.

Ontwerp van de studie

De derde onderzoekscomponent die verder kan worden ontwikkeld om beter aan de verklaringsfunctie te kunnen voldoen, is het ontwerp van de internationale vergelijkende studie. Hierbij kan vooral worden gedacht aan opsplitsing van de studie in een kern- en een differentieel deel. Alle aan het onderzoek deelnemende landen doen mee aan het kerndeel, waarvan bijvoorbeeld alleen leerlingprestaties en enkele kernfactoren deel uitmaken. In het kerndeel is *verklaren* van leerlingprestatieverschillen tussen landen niet het primaire doel. De IEA zou voor dit deel met name kunnen kiezen voor het *beschrijven* van landenverschillen. Naast toetsprestaties en de hoeveelheid getoetste leerstof die in elk land is onderwezen voorafgaande aan de afname van de internationale toets, kunnen klasniveau

factoren zoals 'effectieve leertijd' en de 'werklast van de leraren' worden opgenomen in het kerndeel.
In het differentiële deel van de studie kunnen landen aangeven met welke landen ze zich willen vergelijken. De IEA kan op basis van de voorkeuren van landen regio's samenstellen die niet noodzakelijkerwijs geografisch van aard zijn. De hierboven geschetste aanbevelingen voor de verdere ontwikkeling van het conceptueel kader en de instrumentatie van de studie kunnen per regio worden toegepast. Hiervoor is het tevens aan te bevelen per regio een apart coördinatiecentrum in te richten, omdat het huidige internationale studiecentrum van TIMSS haar handen vol zal hebben aan het kerndeel en de kwaliteitsbewaking van de gehele studie.
Ten opzichte van het kerndeel waarin veel landen (in TIMSS meer dan 40) uit alle werelddelen participeren, zal binnen elke regio het aantal landen geringer zijn. Bovendien zullen de culturele en economische verschillen tussen de landen per regio kleiner zijn dan die tussen alle deelnemende landen. Beide gegevens kunnen het bereiken van het verklaringsdoel beter mogelijk maken. Op 'regionaal' niveau kan de op zichzelf gecompliceerde selectie van relevante kernfactoren en de operationalisatie ervan in betrouwbare en valide onderzoeksinstrumenten tot betere resultaten leiden dan op wereldwijd niveau.
Een tweede aspect van het ontwerp van de studie betreft het opnemen van een voormeting van zowel de leerlingprestaties als achtergrondfactoren die een proceskarakter hebben (bijvoorbeeld kenmerken van de instructiepraktijk). TIMSS is een voorbeeld van een 'one-shot' studie waarin alle metingen op één en hetzelfde moment plaatsvinden. Het leggen van relaties tussen enerzijds leerlingprestaties en anderzijds achtergrondfactoren op alle onderwijsniveaus veronderstelt strikt genomen dat prestaties worden beïnvloed door die factoren. Voor zover de achtergrondfactoren betrekking hebben op het onderwijzen van de leerstof is het gewenst een leerwinst te kunnen bepalen. Op grond van een vergelijking tussen de uitkomsten op de voor- en natoets kan beter worden nagegaan in hoeverre de gemeten onderwijsfactoren van invloed zijn op de leerprestaties. Deze uitbreiding van het onderzoeksontwerp is met name in internationale vergelijkende studies van belang. Immers, daarin willen landen van elkaar leren teneinde hun onderwijs en de prestaties van hun eigen leerlingen te kunnen optimaliseren. De trend study Trends in International Mathematics and Science Study die vanaf 2003 als opvolger van TIMSS in een vierjaarlijkse cyclus wordt uitgevoerd, kan hiervoor reeds meer mogelijkheden bieden.

APPENDIX A

OVERVIEW OF EXPLORED FACTORS AND TIMSS QUESTIONNAIRE ITEMS (VERSION 1995)

Educational and curricular level	Factor label [1]	Items in TIMSS instruments (version 1995) [2]
Student Curricular antecedent	*SA_1 Gender*	Sq2: Are you a girl or a boy? *Circle either A or B.* girl A boy B
	SA_3 Social background a. Out-of-school activities	sq5: During the week, how much time before or after school do you usually spend … *Circle one letter, A, B, C, D, or E, for each line.*

	no time	*less than 1 hour*	*1-2 hours*	*3-5 hours*	*more than 5 hours*
d. working at a paid job?	A	B	C	D	E

sq6: On a normal school day, how much time do you spend before or after school doing each of these things?
Circle one letter, A, B, C, D, or E, for each line.

	no time	*less than 1 hour*	*1-2 hours*	*3-5 hours*	*more than 5 hours*
a. watching television and videos?	A	B	C	D	E
b. playing computer games	A	B	C	D	E
c. playing or talking with friends outside of school	A	B	C	D	E
d. doing jobs at home	A	B	C	D	E

b. Number of books in the home

sq11: About how many books are there in your home?
(Do not count magazines, newspapers, or your school books.)
Circle one letter, A, B, C, D, or E.

none or very few (0 - 10 books)	A
enough to fill one shelf (11 - 25 books)	B
enough to fill one bookcase (26 - 100 books)	C
enough to fill two bookcases (101 - 200 books)	D
enough to fill three or more bookcases (more than 200)	E

Educational and curricular level	Factor label [1]	Items in TIMSS instruments (version 1995) [2]
	c. (Educational level mother and father)	sq9: How far in school did your mother and father go? How far do you expect to go? *Circle ONE letter in each column.*

	Mother	***Father***	***Yourself***
a. <finished primary school>	A	A	A
b. <finished some secondary school>	B	B	B
c. <finished secondary school>	C	C	C
d. <some vocational/technical education after secondary school>	D	D	D
e. <some university>	E	E	E
f. <finished university>	F	F	F
g. I don't know	G	G	G

Curricular context	*SC_1 Motivation* a. attitude towards mathematics	sq21: How much do you like … *Circle one letter, A, B, C, or D, for each line*

	dislike a lot	***dislike***	***like***	***like a lot***
a. mathematics?	A	B	C	D

sq23: What do you think about mathematics?
Circle one letter, A, B, C, or D, for each line

	strongly agree	***agree***	***disagree***	***strongly disagree***
a. I enjoy learning mathematics	A	B	C	D
b. Mathematics is boring	A	B	C	D
c. Mathematics is an easy subject	A	B	C	D

sq24: I need to do well in mathematics …
Circle one letter, A, B, C, or D, for each line

	strongly agree	***agree***	***disagree***	***strongly disagree***
d. to please myself	A	B	C	D

Educational and curricular level	Factor label [1]	Items in TIMSS instruments (version 1995) [2]				
		sq23: What do you think about mathematics? *Circle one letter, A, B, C, or D, for each line*				
			strongly agree	***agree***	***disagree***	***strongly disagree***
		e. I would like a job that involved using mathematics	A	B	C	D
		sq24: I need to do well in mathematics … *Circle one letter, A, B, C, or D, for each line*				
			strongly agree	***agree***	***disagree***	***strongly disagree***
		a. to get the job I want	A	B	C	D
		c. to get into the <secondary school> or university I prefer	A	B	C	D
		sq23: What do you think about mathematics? *Circle one letter, A, B, C, or D, for each line*				
			strongly agree	***agree***	***disagree***	***strongly disagree***
		d. Mathematics is important to everyone's life	A	B	C	D
		sq16: I think it is important to … *Circle one letter, A, B, C, or D, for each line*				
			strongly agree	***agree***	***disagree***	***strongly disagree***
		b. do well in mathematics at school	A	B	C	D
	b. success attribution	sq20: To do well in mathematics at school you need … *Circle one letter, A, B, C, or D, for each line*				
			strongly agree	***agree***	***disagree***	***strongly disagree***
		c. lots of hard work studying at home	A	B	C	D

Educational and curricular level	Factor label [1]	Items in TIMSS instruments (version 1995) [2]
	c. Perceived maternal academic expectation	sq13: My mother thinks it is important for me to … *Circle one letter, A, B, C, or D, for each line* ***strongly agree*** / ***agree*** / ***disagree*** / ***strongly disagree*** a. do well in science at school — A / B / C / D b. do well in mathematics at school — A / B / C / D c. do well in <language of test> at school — A / B / C / D
	d. Perceived friends' academic expectation	sq15: Most of my friends think it is important to … *Circle one letter, A, B, C, or D, for each line* ***strongly agree*** / ***agree*** / ***disagree*** / ***strongly disagree*** a. do well in science at school — A / B / C / D b. do well in mathematics at school — A / B / C / D c. do well in <language of test> at school — A / B / C / D
	SC_2 Time on task/opportunities used a. Number of minutes math/week	tqB3: How many minutes per week do you teach mathematics to your mathematics class? Minutes: _____
	b. Amount of homework per day	tqB13d: Did you assign homework after the class <hour/period>? *Check one box.* Yes ❑ No ❑ tqB13e: If yes, how long would it take a typical student to complete this homework? *Please write in a number.* _____ minutes
Curricular content	*SO_1 Attained curriculum (Achievement in mathematics)*	TIMSS international mathematics achievement test

Educational and curricular level	**Factor label** [1]	**Items in TIMSS instruments (version 1995)** [2]
Classroom	*CA_1 Teacher background characteristics*	
Curricular antecedent	a. Teacher's gender	tqA2: Are you female or male? *Check one box only.* female A male B
	b. Teaching experience in number of years	tqA7: By the end of this school year how many years will you have been teaching altogether? *Please round to the nearest whole number.* _____
	c. Teacher's workload	tqA9: For how many single <hours/periods> are you formally <scheduled/time-tabled> to teach each of the following subjects during the school week? NCR Note: <List only the generic science courses appropriate for your country> *Count a double <hour/period> as two single <hours/periods>.* *Write zero if none.* ***Number of single <hours/periods>*** a. mathematics _____ b. <general/integrated science> _____ c. <physical science> _____ d. <earth science> _____ e. <life science> _____ f. <biology> _____ g. <chemistry> _____ h. <physics> _____ i. other subjects _____
Curricular context	*CC_1 Class size*	tqB1: How many students are in your mathematics class? *Write in a number for each. Write 0 (zero) if there are none.* boys _____ girls _____

Educational and curricular level	Factor label [1]	Items in TIMSS instruments (version 1995) [2]
	CC_3 Material for evaluation of student outcomes, feedback and corrective instruction	tqB22: In assessing the work of the students in your mathematics class, how much weight do you give each of the following types of assessment? *Check one box in each row.* *none* / *little* / *quite a lot* / *a great deal* a. standardized tests produced outside the school ❑ ❑ ❑ ❑ b. teacher-made short answer or essay tests that require students to describe or explain their reasoning ❑ ❑ ❑ ❑ c. teacher made multiple choice, true-false and matching tests ❑ ❑ ❑ ❑ d. how well students do on homework assignment ❑ ❑ ❑ ❑ e. how well students do on homework assignments ❑ ❑ ❑ ❑ f. observations of students ❑ ❑ ❑ ❑ g. responses of students in class ❑ ❑ ❑ ❑
	CC_4 Grouping procedures: c. cooperative learning	sq25: How often does this happen in your mathematics lessons? *Circle one letter, A, B, C, or D, for each line.* *almost always* / *pretty often* / *once in a while* / *never* h. We work together in pairs or small groups A B C D

Educational and curricular level	Factor label [1]	Items in TIMSS instruments (version 1995) [2]
	CC_5 Teaching style (student oriented)	sq25: How often does this happen in your mathematics lessons? *Circle one letter, A, B, C, or D, for each line.*

	almost always	*pretty often*	*once in a while*	*never*
d. We work from worksheets or textbooks on our own	A	B	C	D
e. We work on mathematics projects	A	B	C	D
i. We use things from everyday life in solving mathematics problems	A	B	C	D

sq26: When we begin a new topic in mathematics, we begin by … *Circle one letter, A, B, C, or D, for each line.*

	almost always	*pretty often*	*once in a while*	*never*
b. having the teacher explain the rules and definitions	A	B	C	D
c. working together in pairs of small groups on a problem or project	A	B	C	D
d. having the teacher ask us what we know related to the new topic	A	B	C	D
f. trying to solve an example related to the new topic	A	B	C	D

CC_6 Management and orderly and quiet atmosphere

a. Perceived class climate (is it an orderly and quiet atmosphere)

sq14: In my mathematics class … *Circle one letter, A, B, C, or D, for each line*

	strongly agree	*agree*	*disagree*	*strongly disagree*
a. students often neglect their school work	A	B	C	D
b. students are orderly and quiet during <lesson>	A	B	C	D
c. students do exactly as the teacher says	A	B	C	D

Educational and curricular level	Factor label [1]	Items in TIMSS instruments (version 1995) [2]
	b. Perceived school climate (safety)	sq18: How often did any of these things happen last month in school? *Circle one letter, A, B, C, or D, for each line* ***strongly agree*** / ***agree*** / ***disagree*** / ***strongly disagree*** b. Something of mine was stolen — A / B / C / D c. I thought another student might hurt me — A / B / C / D e. Some of my friends skipped classes — A / B / C / D f. Some of my friends were hurt by other students — A / B / C / D
	c. Limitations to teach the tested class related to student features	tqB7: In your view to what extent do the following limit how you teach your mathematics class? *Circle one box in each row.* ***not at all*** / ***a little*** / ***quite a lot*** / ***a great deal*** a. students with different academic abilities — ❑ / ❑ / ❑ / ❑ b. students who come from a wide range of backgrounds (e.g. economic, language) — ❑ / ❑ / ❑ / ❑ c. students with special needs (e.g. hearing, vision, speech impairment, physical disabilities, mental or emotional/ psychological impairment) — ❑ / ❑ / ❑ / ❑ d. uninterested students — ❑ / ❑ / ❑ / ❑ e. disruptive students — ❑ / ❑ / ❑ / ❑ o. low morale among students — ❑ / ❑ / ❑ / ❑
	CC_7 Homework a. Frequency of homework	tqB18: How often do you usually assign mathematics homework? *Check one box.* never ❑ less than once a week ❑ once or twice a week ❑ 3 or 4 times a week ❑ every day ❑

Educational and curricular level	Factor label [1]	Items in TIMSS instruments (version 1995) [2]
	b. Amount of homework	tqB19: If you assign mathematics homework, how many minutes of mathematics homework do you usually assign your students? (Consider the time it would take an average student in your class.) *Check one box.* I do not assign homework ❑ less than 15 minutes ❑ 15-30 minutes ❑ 31-60 minutes ❑ 61-90 minutes ❑ more than 90 minutes ❑
	c. Treatment in next lesson	tqB21: If students are assigned written mathematics homework, how often do you do the following? *Check one box in each row.* d. give feedback on homework to whole class — *never* ❑ *rarely* ❑ *some-times* ❑ *always* ❑ *I do not assign home-work* ❑ e. have students correct their own assignments in class — ❑ ❑ ❑ ❑ ❑ f. have students exchange assignments and correct them in class — ❑ ❑ ❑ ❑ ❑ g. use it as a basis for class discussion — ❑ ❑ ❑ ❑ ❑
	CC_14 Evaluation, feedback and corrective instruction	tqB23: How often do you use the assessment information you gather from students to … *Check one box in each row.* a. provide students' grades or marks? — *none* ❑ *little* ❑ *quite a lot* ❑ *a great deal* ❑ b. provide feedback to students? — ❑ ❑ ❑ ❑ c. diagnose students' learning problems? — ❑ ❑ ❑ ❑ d. report to parents? — ❑ ❑ ❑ ❑ e. assign students to different programs or tracks? — ❑ ❑ ❑ ❑ f. plan for future lessons? — ❑ ❑ ❑ ❑

Educational and curricular level	Factor label [1]	Items in TIMSS instruments (version 1995) [2]
Classroom Curricular content	*CO_1 Implemented curriculum content*	tq12B: How long did you spend teaching each of these topics to your mathematics class this year? Will you cover any of these topics in future <periods>? *Check as many boxes as apply for each topic listed.*

	have taught this year <periods> completed				*will teach later this year*	*not taught this year*	*taught a previous year*
	1-5	*6-10*	*11-15*	*> 15*			
a. Whole numbers	❑	❑	❑	❑	❑	❑	❑
b. Common & Decimal Fractions	❑	❑	❑	❑	❑	❑	❑
c. Percentages	❑	❑	❑	❑	❑	❑	❑
d. Number sets & concepts	❑	❑	❑	❑	❑	❑	❑
e. Number theory	❑	❑	❑	❑	❑	❑	❑
f. Estimation & number sense	❑	❑	❑	❑	❑	❑	❑
g. Measurement units & processes	❑	❑	❑	❑	❑	❑	❑
h. Estimation & error of measurements	❑	❑	❑	❑	❑	❑	❑
i. Perimeter, Area & volume	❑	❑	❑	❑	❑	❑	❑
j. Basics of one & two dimensional geometry	❑	❑	❑	❑	❑	❑	❑
k. Geometric congruence & similarity	❑	❑	❑	❑	❑	❑	❑
l. Geometric transformations & symmetry	❑	❑	❑	❑	❑	❑	❑
m. Constructions & three dimen-sional geometry	❑	❑	❑	❑	❑	❑	❑
n. Ratio & proportion	❑	❑	❑	❑	❑	❑	❑
o. Proportionality: slope, trigonometry & interpolation	❑	❑	❑	❑	❑	❑	❑
p. Functions, relations & patterns	❑	❑	❑	❑	❑	❑	❑
q. Equations, inequalities & formulas	❑	❑	❑	❑	❑	❑	❑
r. Statistics & data	❑	❑	❑	❑	❑	❑	❑
s. Probability & uncertainty	❑	❑	❑	❑	❑	❑	❑
t. Sets & logic	❑	❑	❑	❑	❑	❑	❑
u. Problem solving strategies	❑	❑	❑	❑	❑	❑	❑
v. Other mathematics content	❑	❑	❑	❑	❑	❑	❑

Educational and curricular level	Factor label [1]	Items in TIMSS instruments (version 1995) [2]
School	*ScA_1 School size*	
Curricular antecedent	Total Number students in the school	sq17: The students in your school: *Write in the answer for each of the following. Write 0 (zero) is there are none.* **boys** **girls** a. What is the total enrollment (number of students)? _____ _____ *Concerning upper grade students …* k. How many students are in upper grade? _____ _____ o. How many students in upper grade study mathematics? _____ _____
	ScA_2 Student body composition	
	Proportion of boys in school	See school size (ScA_1)
	ScA_3 School category (urban/rural)	
	Urban/rural area of school site	sq1: In what type of community is your school located? *Check one box only.* A geographically isolated area ❑ Village or rural (farm) area ❑ One on the outskirts of a town/city ❑ One close to the center of a town/city ❑
Curricular context	*ScC_3 Policy on supervision*	
	cooperation and collaboration	sq10: Cooperation and collaboration: *Check only one box* **Yes** **No** a. Does your school have an official policy related to promoting cooperation and collaboration among teachers? ❑ ❑ b. Are teachers in your school encouraged to share and discuss instructional ideas and materials? ❑ ❑ c. Do teachers in your school meet regularly to discuss instructional goals and issues? ❑ ❑

Educational and curricular level	Factor label [1]	Items in TIMSS instruments (version 1995) [2]
	ScC_5 Time schedule	
	time schedule math, grade 8	schq8: During the school week, about how many hours of scheduled school time does a mathematics teacher usually have for … *Write in a numeric value.* *Please write in 0 (zero) if no time is scheduled.* a. tasks related to teaching mathematics (e.g., lesson preparation, grading homework, etc.) _____ hours/week b. teaching mathematics classes _____ hours/week
	ScC_7 Orderly and quiet atmosphere	see safety as perceived by the student (see CC_6b)
	ScC_9 Educational leadership	
	Number of hours per month principal spends on educational tasks	schq11: As principal of this school, about how many hours per month do you usually spend on each of the following activities? Please indicate the approximate number for each item. Please write 0 (zero) if no time is spent on an activity. ***hours per month*** e. Teaching (including preparation) _____ f. Giving a demonstration lesson _____ g. Discussing educational objectives with teachers _____ h. Initiating curriculum revision and/or planning _____ j. Counseling and disciplining of students _____ l. Training teachers _____ m. Professional development activities _____ n. Other activities _____

Educational and curricular level	Factor label [1]	Items in TIMSS instruments (version 1995) [2]
School	*ScO_1 School curriculum content*	
Curricular content	written school curriculum mathematics	schq14. Does your school have its own written statement of the curriculum content to be taught (i.e., other than the national or regional curriculum guides)? *Check one box in each line.*
		Yes ***No***
		a. For mathematics ❑ ❑
		b. For science ❑ ❑
Country/ System		
Curricular antecedent	*SysA_1 Resources, funding*	national information
	SysA_2 Training and support systems	national information
	SysA_3 National guidelines for time schedules	national information
Curricular context	*SysC_1 Policy focusing on effectiveness*	national information
	SysC_2 Policy on evaluation/ national testing system	national information
Curricular content	*SysO_1 national guidelines for curriculum (i.e., intended curriculum content)*	national information

Source: TIMSS-1995 instruments, IEA secretariat, Amsterdam

Notes: [1] SA = student curricular antecedent; SC = student curricular context; SO = student curricular content;
CA = classroom curricular antecedent; CC = classroom curricular context; CO = classroom curricular content; ScA = school curricular antecedent; ScC = school curricular context; ScO = school curricular content;
SysA = system curricular antecedent; SysC = system curricular context; SysO = system curricular content;
[2] sq = student questionnaire;
tq = teacher questionnaire mathematics;
schq = school questionnaire.

APPENDIX B

RESULTS PLS OUTER BETWEEN-CLASSROOM MODELS

Loadings for manifest variables in outer between-classroom model (PLSpath analyses)

Latent Variable *Manifest variable*	**Pooled data set**	**Belgium Flanders**	**Germany**	**Netherlands**
Percentage of girls in classroom	1.00	1.00	1.00	1.00
Out-of-school leisure time activities	1.00	1.00	1.00	1.00
Number of books at home	1.00	1.00	1.00	1.00
Attitude towards mathematics				
Liking	.74	.83	.88	.59
Importance	.96	.94	.92	.97
Time on task/opportunities use				
Mathematics time scheduled per week	.94	.97	**- .04**	.79
Amount of homework	.45	**.15**	.99	.61
Content coverage mathematics	1.00	1.00	1.00	1.00
Level of student oriented teaching style as perceived by the students	1.00	1.00	1.00	1.00
Safety at school as perceived by students	1.00	1.00	1.00	1.00
Limitations in teaching class related to student features	1.00	1.00	1.00	1.00
Mathematics achievement	1.00	1.00	1.00	1.00

Note: figures in bold are non-significant

APPENDIX C

FINAL RECURSIVE BETWEEN-CLASSROOM PATH MODEL FOR THE POOLED DATA SET, BELGIUM FLANDERS, GERMANY, AND THE NETHERLANDS

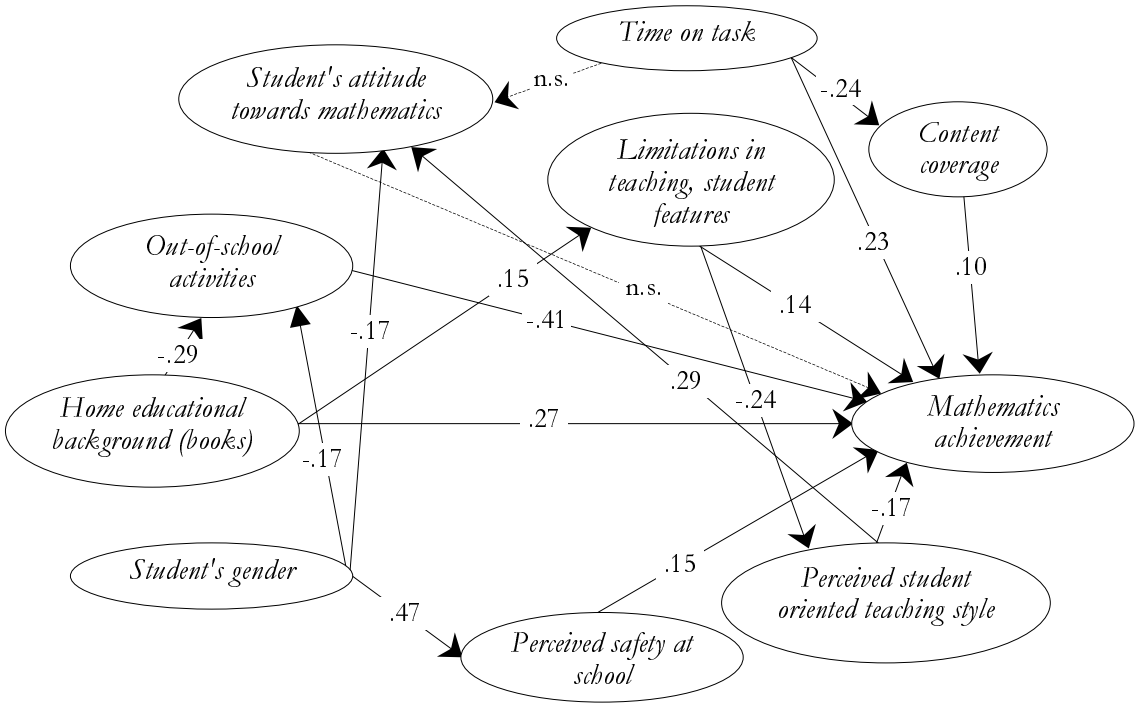


Figure 4-1a (Pool)

Final recursive between-classroom path model (including aggregated student factors)

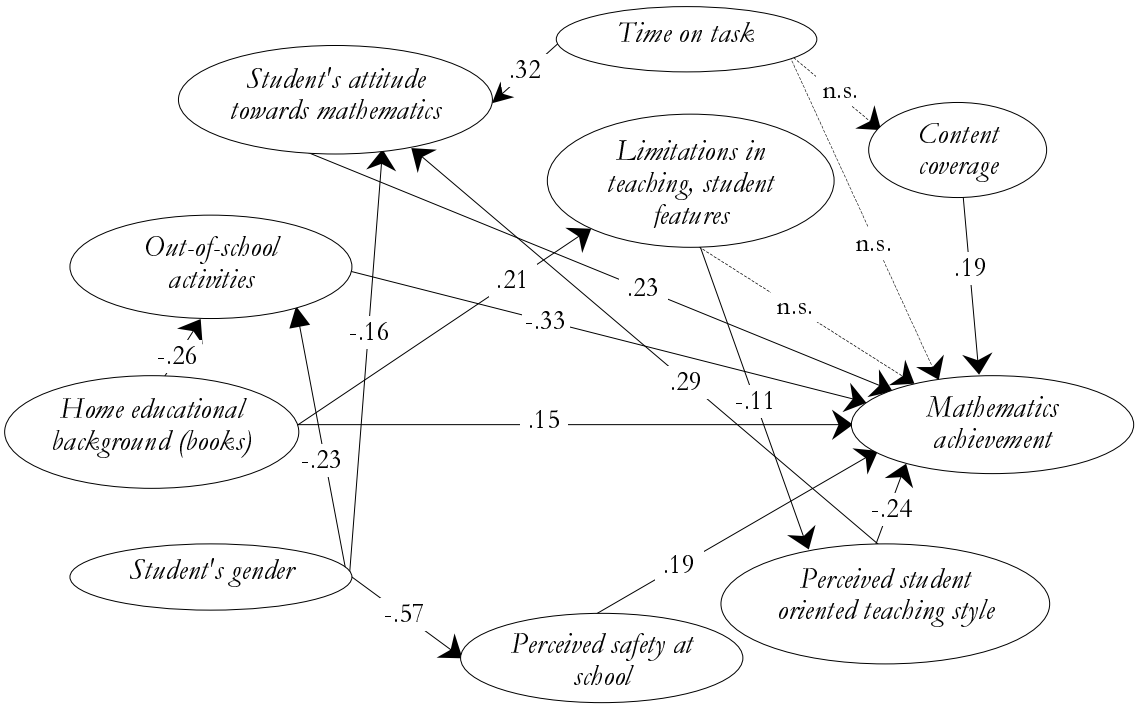


Figure 4-1b (Belgium Flanders)

Final recursive between-classroom path model (including aggregated student factors)

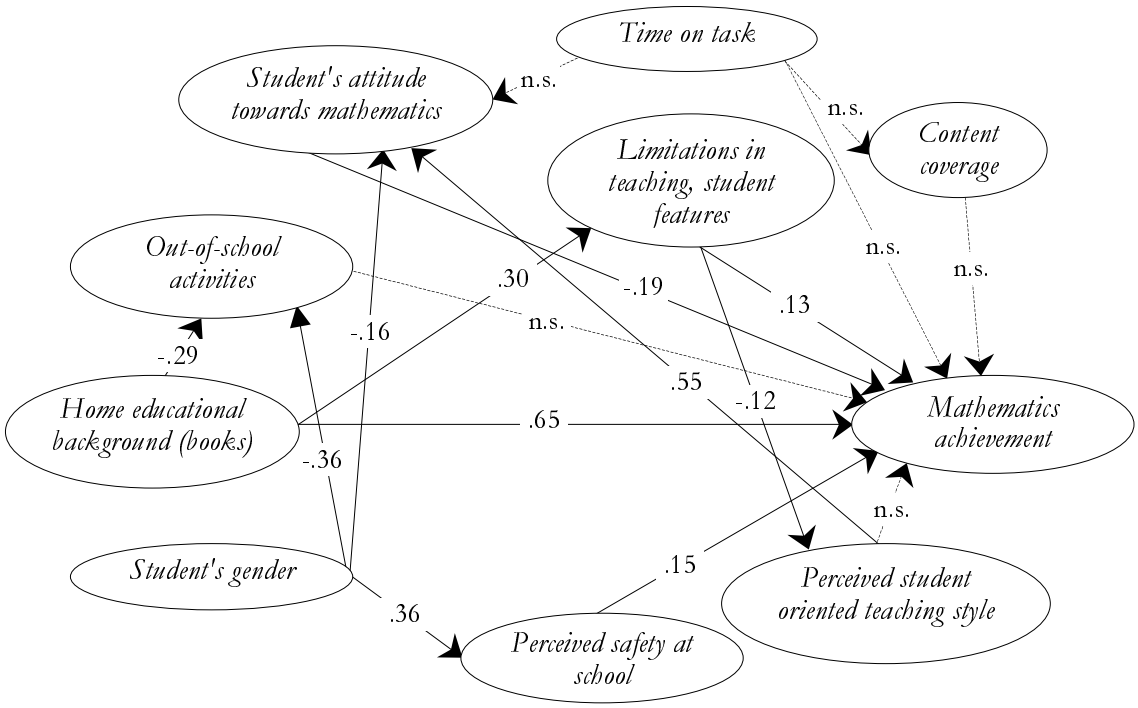


Figure 4-1c (Germany)

Final recursive between-classroom path model (including aggregated student factors)

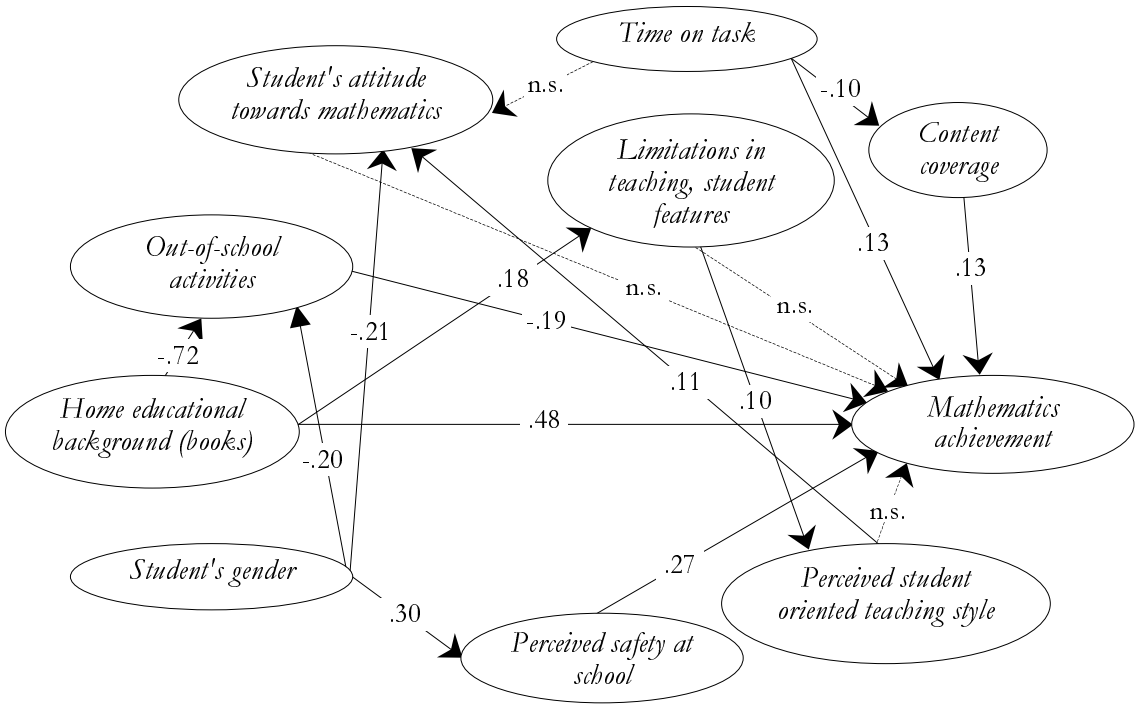


Figure 4-1d (The Netherlands)

Final recursive between-classroom path model (including aggregated student factors)