MATHEMATICS ACHIEVEMENT OF SLOVENE STUDENTS AT THE END OF COMPULSORY EDUCATION

Mojca Štraus

DOCTORAL COMMITTEE

Chairman:	Prof. dr. B.E. van Vucht Tijssen ▪ University of Twente
Promotor:	Prof. dr. Tj. Plomp ▪ University of Twente
Assistant promotor:	Dr. W.J. Pelgrum ▪ University of Twente
Members:	Prof. dr. J.J.H. van den Akker ▪ University of Twente
	Prof. dr. J. Scheerens ▪ University of Twente
	Dr. W.A.J.M. Kuiper ▪ University of Twente
	Prof. dr. J. Justin ▪ University of Ljubljana, Slovenia
	Prof. dr. A. van Streun ▪ University of Groningen

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Štraus, M.

Mathematics achievement of Slovene students at the end of compulsory education
Thesis University of Twente, Enschede – With refs. – With Dutch summary.
ISBN 90 365 2032 0

Layout: Sandra Schele
Press: PrintPartners Ipskamp - Enschede

MATHEMATICS ACHIEVEMENT
OF SLOVENE STUDENTS AT THE END
OF COMPULSORY EDUCATION

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 vrijdag 23 april 2004 om 15.00 uur

door

Mojca Štraus

geboren op 18 mei 1968

te Ljubljana, Slovenia

Promotor: Prof. dr. T. Plomp

Assistant promotor: Dr. W.J. Pelgrum

PREFACE

While writing this thesis I was privileged to enjoy encouragement and support from many people.
I am particularly grateful to my mentor Tjeerd Plomp and co-mentor Hans Pelgrum, who guided me through this project and generously shared their wisdom with me. I have enjoyed our cooperation on this thesis as well as in several IEA projects and always experienced it as inspirational and stimulating. They closely reviewed the preliminary versions of this manuscript and gave suggestions that helped me make progress. Their experienced supervision enabled that my work and ideas resulted in this thesis.

I want to thank the Department of Curriculum and its chair, Jan van den Akker, of the Faculty of Behavioural Sciences for support and hospitality during my visits to Enschede to work on the thesis. Sandra Schele took care of numerous arrangements during these visits and set up the layout of the final manuscript for which I want to thank her. Furthermore, Klaas Bos, Pauline Vos, Sarah Howie, and Rien Steen helped me in my work through conversations and/or by kindly sending me their theses. There have also been many people in other countries who always took time for discussion and answering my questions. I am especially grateful to Eugene Gonzalez and Ray Adams.

My colleagues at the Educational Research Institute Ljubljana continuously expressed their support and encouragement. In particular I want to thank Barbara Japelj Pavešić, Mojca Čuček, Alenka Gril, Marjeta Doupona Horvat, Barbara Neža Brečko and Maša Repež for their support, fruitful conversations and for providing such a pleasant working environment. To Jennifer Toth I want to thank for her suggestions and comments for improving English in the text.

I thank the Ministry of Education, Science and Sports of the Republic of Slovenia for financially supporting this research. I thank the Rubato Foundation, Enschede, for facilitating my visits to the University of Twente.

Finally, I am grateful to my parents for always believing in me, and to Boris and Nika for their understanding, support and patience.

Mojca Štraus
March 2004

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
The context of the study

In the last few decades, student achievement has played a central role among the indicators used to evaluate the quality of education systems. Accordingly, recent education reform in Slovenia included the achievement of international standards of knowledge and skills as an important goal. In order to provide additional information for the purposes of evaluating the quality of mathematics education in Slovenia, the present study examines mathematics achievement of students in the final grade of the non-reformed compulsory education in Slovenia in an international context. Data from the Third International Mathematics and Science Study (TIMSS) will be used to provide information about mathematics achievement of Slovene students. TIMSS data will also be used to provide information about achievements of students from other European countries, which are taken as a point of reference for describing the achievement of Slovene students. Another point of reference is derived from the attainment targets in the reformed mathematics curriculum. The goal of this study is to provide information that would support the efforts to successfully implement the reforms.

In this chapter, the origin and purpose of the study are first explained (1.1). This is followed by an account of the studies on Slovene mathematics achievement in an international context (1.2). Section 1.3 presents characteristics of national and international assessments; while in section 1.4, ways of utilizing assessment information are discussed. The problem statement and research questions are formulated in section 1.5. In the final section of this chapter (1.6), an overview of the following chapters is given.

1.1 INTRODUCTION

After the secession from Yugoslavia in 1991, the political, social, and economic changes urged Slovenia to reform its education system. These reforms encompassed the structure of the school system as well as the curricula of all school subjects. The goals for reformed education explicitly stated that Slovene education "*makes possible the achievement of internationally comparable standards of knowledge after the completion of the primary education*" (*White Paper*, 1996, p.92)[1]. The reform is currently being implemented into the school system through a 10-year process of stepwise transformations and it is expected to be completed in the 2008/2009 school year.

When developing the new system of education, the policy makers, educators, and subject experts were faced with the question of which areas actually needed reforming. To properly address this question, additional information is required, ranging from the information on the general structure and effectiveness of the school system, to the specific information on what is really going on in schools, and what are the outcomes of these processes. Throughout the political debate surrounding the reform in Slovenia, comparisons with other countries, and especially with those from the European Union, were emphasized. In addition to the theoretical premises for the reform, the *White Paper* (1996) gives a number of comparisons with other European countries as support for the introduction of changes in the school system. Given the contemporary, worldwide attention to achievement, among the important questions to be answered were: *How well do Slovene students perform in comparison with students from other countries? Do they reach expected levels of achievement? What should be expected of them?* The answers to these questions where sought through opinions of experts, experiences of teachers and others involved in education and, where available, data from empirical studies.

The present study deals with mathematics achievement of Slovene students at the end of compulsory education and the reform of mathematics curriculum. Based on more recent data the above questions will be addressed in an international comparative context. Furthermore, developments in Slovene students' achievements will be examined to provide insight into their progress over time. These issues will be analyzed in terms of underlying concepts and the method will be described how to tackle them. Through this, possible areas of improvement in Slovene mathematics achievement will be indicated.

The reason to focus on the subject of mathematics is two-fold. The first is that mathematics is one of the core subjects in the Slovene school curriculum. The second is that mathematics achievement has been the focus of several international

---

1 The authors used the term *primary education* to name the compulsory part of the Slovene education system. Sometimes also terms *basic* or *elementary education* are used. In this study, simply the term *compulsory education* will be used. For a description of the structure of this part of the Slovene education system see Chapter 2.

surveys, starting with the First International Mathematics Study (FIMS) in 1964 (Husén, 1967) and followed by the Second International Mathematics Study (SIMS) in the early 1980s (Robitaille & Garden, 1989). Most recently, the Third International Mathematics and Science Study (TIMSS) in 1995 (Beaton et al., 1996) and its successors in 1999 (Mullis et al., 2000) and 2003 (the study was renamed into the Trends in International Mathematics and Science Study) were conducted. All of these studies were carried out under the auspices of the International Association for Evaluation of Educational Achievement (IEA), an international cooperative of research institutes conducting research in education. Similar studies have been conducted by other organizations.

1.2 Research in Slovene mathematics achievement in an international context

At the time of the development of the new mathematics curriculum in Slovenia, international comparative data providing information on the performance of Slovene students as compared to other relevant countries were only available from a limited number of assessments. The first such assessment was based on the replication of the IEA SIMS, carried out in Slovenia in 1989 (Šetinc, 1991). However, the information provided by this study was only available for the final grade of non-compulsory upper-secondary education and is therefore not relevant in the present study.

In compulsory education, the first internationally comparative assessment of mathematics achievement in Slovenia was carried out through the second study of the International Assessment of Educational Progress (IAEP II) in 1991, a study based on the United States' National Assessment of Educational Progress (NAEP, Beaton, 1987). In the IAEP II study 9- and 13-year-old students from Slovenia and 19 other countries or, more specifically, (parts of their) educational systems, participated (Lapointe, Mead, & Askew, 1992). The results for Slovenia were seen by the country's educators as a matter of concern and were used for seeking further information on the areas in Slovene mathematics education in which reforms were needed (Japelj, 1993).
Four years later, in 1995, Slovenia participated in IEA's TIMSS (Šetinc, Japelj, & Trobec, 1997). In Slovene compulsory education, as in most other participating educational systems, TIMSS assessed achievements of students in the third and fourth grades and in the seventh and eighth grades. The results published in the TIMSS international reports seemed to indicate a somewhat better state of mathematics education in Slovenia than the IAEP II results suggested. Such an

interpretation was constructed by the media, based on the comparisons of Slovenia's position on the overall constructed scale in relation to other European countries (Matos, 1997).

Conclusions about the country's education system based on rankings of countries are problematic for several reasons. First, as is the case of Slovenia in IAEP II and TIMSS, differences in the particular country's rankings may have resulted from the differences in the studies' designs, such as the definition of the target populations, instrument development, or the scaling methodology. There may be many additional sources of differences in the implementation of the studies and approaches to reporting results. Second, although overall achievement is an important measure of the outcomes of the country's education system, questions *whether* and *which* changes are needed in particular areas of the curriculum can not be adequately addressed by using this measure.

The TIMSS databases, therefore, with internationally comparative data provide an opportunity to analyze mathematics achievements of Slovene students to address the information needs of curriculum developers and teachers as implementers of the curriculum. At the time of the development of the reformed curricula, only the TIMSS 1995 results published in international reports were available, while the databases for public use were constructed only after the reform of the curriculum was completed. Although the data could not be used to provide information that would be useful in the curriculum development process itself, several studies in Slovenia have since utilized the TIMSS databases to investigate particular features of the Slovene mathematics education. Most of these studies focus on the relationship of student achievement to the contextual variables, such as approaches to teaching and students' attitudes, and do not provide information on the contents of knowledge and skills of Slovene students in mathematics.
This was to some extent provided by the so-called TIMSS scale anchoring studies (Kelly, 1999; Mullis et al., 2000). However, these studies do not focus on comparisons between Slovenia and other countries. Specifically for Slovenia, Magajna (2000) examined TIMSS 1995 data to reveal the areas in which Slovene students performed relatively higher or lower than students from other TIMSS 1995 countries in the light of the reforms in the curriculum, Štraus, Čuček, Gril, Doupona Horvat, and Japelj (2003) employed the scale anchoring procedures on student samples in Slovenia and some other European countries to provide descriptions of knowledge and skills of Slovene students, and Štraus (2003) compared algebra achievement of Slovene students to achievements in other European countries on the basis of the TIMSS 1999 data. While these studies

provide useful information for policy making and curriculum development and implementation in Slovenia, given that the curricula have only recently been reformed and that they are only being implemented, there are further information needs in this process such as, a detailed analysis of trends in Slovene achievement and of the link between the intended and the attained curriculum.

The present study will utilize the TIMSS data to describe how well Slovene students performed in mathematics and what developments were in this performance in the late 1990s from two perspectives that were considered important in the policy documents. One perspective are the achievements in other European countries and the other perspective are the attainment targets in the intended mathematics curriculum. The purpose is to reveal areas in Slovene achievement in which improvements might be desirable. From this, additional in-depth information on Slovene mathematics achievement in a national and international context will be provided. Although the reformed mathematics curriculum in Slovenia has already been developed, such information might be useful in the process of its implementation.

1.3 CHARACTERISTICS OF NATIONAL AND INTERNATIONAL ASSESSMENTS

In broader terms, the issue in this study, as described in the previous section, deals with the quality of education. In the last few decades, the quality of education and within it, achievements of students, have been an important part of the political debate. This has been a debate in countries with well-developed education systems in the industrialized world, as well as in developing countries.

Concern about educational quality is not new. For some time, governments in many countries routinely collected and published data that indicated how their education systems functioned and developed. Data were usually provided on the numbers of schools, teachers, and students, and on efficiency indices such as student-teacher ratios and repetition rates (e.g. Husén & Tuijnman, 1994; Kellaghan & Greaney). Over the years, it became evident that merely providing wide access to education does not ensure that the goals of education are achieved. Kellaghan and Greaney (2001) describe that the competencies of human resources in today's information society are being stressed as critically important for success in global economic competition. This shifted the focus of political attention from issues in managing the quantitative growth of the system, to the monitoring the functioning of education systems in terms of their outcomes. Student achievement, school effectiveness, and accountability all became key criteria for judging the quality of educational systems. The educational systems, schools and individual students are all under increasing pressure to perform.

To address the information needs in the process of the educational reform therefore assessment data on student achievement are needed. In this study, the term *assessment* is used for the process of obtaining information on student achievement from a sample of students or a whole population with the intention to draw inferences at the system level. Another use of the term (but not applied in this study) refers to the assessment of individual students for the purpose of certification and/or providing individual feedback on the learning progress.
International and national assessments have been initiated to serve information needs in many countries. Reasons for appeal of educational assessments in policy making are described below, followed by main characteristics, motivations and questions to be answered in national and international assessments.

1.3.1 Reasons for appeal of educational assessments in policy making

The history of educational assessment only dates back to the early 1960s, and has increased in importance in the minds of policy makers since the 1980s (Husén & Tuijnman, 1994). It emerged from the growing need to include the information on educational outcomes into a knowledge base for policy decisions, for example, about the allocation of resources, for judging how well education measures up to certain stated goals, or the success or failure of an educational reform. Not only did increasing numbers of students in the educational system result in increasing costs, but costs per student were also rising, prompting concern about cost-effectiveness and accountability (Husén & Tuijnman, 1994; Kellaghan, 1996). Another reason for the growing importance of educational assessment was that a comparative information base can be built from which the hypotheses about stability and change in education can be tested.

There are additional reasons for the great appeal of educational assessment to policy makers (Linn, 2000). First, assessments are relatively inexpensive compared with other aspects of managing education, such as implementing curriculum changes that involve substantial professional development for teachers. Second, assessments can be externally mandated. It is easier to mandate assessment requirements at the system level than it is to take actions that involve actual change in what happens inside the classroom. Third, changes in assessment can be implemented rapidly. Forth, results can be made visible through the media, including poor results, which are not necessarily undesired. In the beginning of a policy maker's term, poor results might be desirable for policy makers who want to show they have had an effect. At the same time good results can always be used as support for policy making aimed at little change.

Within the framework of this study, it is important to distinguish between national and international assessments. One may wonder why international and not national data are used in this study. The main characteristics and differences between the two types of assessments are discussed below.

1.3.2 National assessments

Kellaghan and Greaney (2001) distinguish a number of motivations for carrying out a national assessment. It may be carried out with the intention of raising standards, to help maintain the standards, to provide information that can be used to aid decision making about resource allocations, and to assign accountability for student performance. National assessments may be carried out to alter the balance of control in the education system, for example to ensure that what is taught in schools is less dependent on the professional judgment of teachers, and more dependent on central authorities that mandated the assessment. They may also be carried out to compensate for the poor assessment practices of teachers.

In national assessments, answers are sought for questions such as: How well are students learning in the education system (with reference to general expectations, the aims of the curriculum, or preparation for life)? Is there evidence of particular strengths and weaknesses in students' knowledge and skills? How do particular sub-groups within the population perform? What factors are associated with student achievement? Do the achievements of students change over time? The information need addressed in a national assessment, and the procedures to carry it out, depend on the interests of key stakeholders.

The United States' National Assessment of Educational Progress (NAEP), which began in the 1980s, was the first periodic national assessment for evaluation of educational achievement (Beaton, 1987). This assessment is designed to measure student achievement in mathematics, science, and reading at specified ages and grades. The National Institute for Educational Measurement (CITO) in the Netherlands has been conducting assessment surveys in different school subjects in primary and lower secondary education every four years since 1987 (Bokhove, Van der Schoot, & Eggen, 1996; from Bos, 2002).

In Slovenia, there are no national assessments carried out particularly for the purpose of serving policy needs. There are however, national examinations which have been introduced in the reformed system at the end of the third and sixth grades, and at the end of compulsory education, i.e., the ninth grade. In addition to providing feedback to individual students and certification for completion of compulsory education, they are also expected to provide useful information for policy purposes.

1.3.3 International assessments

International assessments of student cognitive achievements arose from the interest of several countries to obtain additional information on the quality of their educational systems through between-country comparisons of the performance of their students. The general steps in conducting an international assessment are that representatives from participating countries agree on the population of students, the curriculum domain to be assessed, and on an instrument to assess achievement in the chosen domain. The instrument is administered to a representative sample of students, usually simultaneously in all participating countries, and comparative analyses of the data are carried out.

The International Association for Evaluation of Educational Achievement (IEA) has been carrying out studies of school achievement in a variety of countries since 1959. It is an independent, non-governmental and international cooperative of research centers of different education systems, with currently more than 50 members. IEA's international comparative studies have two main goals (Plomp, 1998; Plomp, Howie, & McGaw, 2003). First, they are intended to provide policy makers and educational practitioners with information about the quality of their education in relation to relevant reference countries. This is carried out through comparisons of scores or sub-scores on an international achievement test. Second, they are intended to assist participating countries in understanding the reasons for observed differences. Within the second goal, an important part of the attention in the IEA studies is given to the extent to which a country's intended curriculum (what should be taught in a particular grade) is implemented in schools and attained by students. This means that most IEA studies are curriculum based, which is their significant feature distinguishing them from the international studies conducted by other agencies. An analysis of the curricula of participating countries and the development of achievement instruments based on these curricula are the starting points of the design of an IEA study. This link between the achievement instruments and the curricula of the participating countries is also the reason why the IEA's TIMSS study is considered most appropriate to provide data for addressing the problem in the present study.

While there is no doubt about the usefulness of national assessments carried out in the context of the educational system and curriculum, without problems with translation and adoption of the instrumentation and equivalence of samples of students, there are also important benefits from participation in the international comparative assessments. The main advantage of international assessments over national assessments is that they provide between-country comparative data. Since there are no absolute standards for educational achievement, comparative studies are essential to provide policy-makers and educators with descriptive information about the range of educational quality in relation to other national systems (Plomp, 1998). In this way, international assessments contribute to setting realistic standards for educational systems, as well as to monitoring educational quality.
Comparative studies may also be helpful in understanding the relationships of other variables of the school system to the observed differences in student performance, by exploring cross-nationally relations between school achievement and factors, such as the curricula, amount of time spent on school work, and many other possible explanatory measures. Furthermore, Beaton et al. (1999) emphasize that the benefit of participation in international assessments of educational achievement is also that the countries are obliged to scrutinize their curricula more closely. A curriculum in a particular subject area may become outdated in some respects and therefore fail to include topics or approaches of emerging importance. The international studies that involve close analysis and comparison of the curricula of participating countries impel a thorough review by each country of its own curriculum. In this way ministries and curriculum experts can be alerted to differences between their country's curriculum and that of other countries in terms of the emphasis given to, and the content covered by, different subject areas. Such differences do not necessarily imply any deficiency because a country may well have a good reason for adopting a curriculum involving variations from the international pattern, for example, reasons associated with the state of development of its education system.
There are some additional benefits in conducting international assessments that pertain more to the developing, or less-developed countries (Plomp, Howie, & McGaw, 2003). First, the demand of international assessments to follow rigorous survey procedures may be of considerable benefit to building the research capacity in a country in which traditions of empirical educational research and associated technologies are not strong. Second, through participation in an international assessment, some countries collect baseline data in certain subject areas where

previously there were no data available. Third, mutual participation with more experienced countries heightens awareness of the actions taken in these countries for improvement of the education system, which provides opportunities to draw lessons from their experiences. And finally, as a consequence of media attention given to the international assessments, education can gain in priority among the areas that need policy makers' attention.

Other agencies, in addition to IEA, have initialized international assessments in the last 15 years. Following a preparation period of several years, the Organization for Economic Co-operation and Development (OECD) launched a Programme for International Student Assessment (PISA) in 2000 (see e.g., OECD, 2002; Schleicher, 2000). The assessment is designed to collect data at regular, three-year intervals on samples of 15-year olds in OECD member countries, as well as in a number of other countries. A significant feature of PISA is that the assessment framework for each domain tested (reading literacy, mathematics, and science) is structured to reflect important knowledge and skills that students will need in adult life, and does not rely on common components of the curricula of participating countries as the focus for assessment (Plomp, Howie, & McGaw 2003). This is the fundamental difference between IEA and OECD studies. Since Slovenia did not participate in PISA, and since the possible identified areas in student mathematics achievement in which improvement might be desired would lack the link to the Slovene school curricula, PISA database is not suitable for the present study.

Another example of an international assessment study is the International Assessment of Educational Progress (IAEP), conducted in 1988 and 1991 under the direction of Educational Testing Service, under contract to the U.S. Department of Education. As already mentioned, this study represented an extension of the U.S. 1986 NAEP assessment in mathematics and science. In 1991, in addition to the collection of data on mathematics and science achievements in twenty countries including Slovenia, data on contextual variables, including time given to homework, availability of books at home, and teacher characteristics was assessed. The IAEP data are also not considered suitable for addressing the problem in this study for several reasons. The most important reason is that the study was based on the United States' NAEP design, using instruments developed for the purposes of their national assessment. Due to considerable differences between the United States' and Slovene mathematics curriculum (e.g., Schmidt et al., 1997), the NAEP instruments are less appropriate as measures of student achievement in Slovenia.

1.4 UTILIZING ASSESSMENT INFORMATION

From the perspective of an educational system, the benefits of conducting national and international assessments are that information potentially useful for improving the educational system is obtained. Usefulness of this information for policy needs is considered highly important although, in most cases, it can not be expected that clear causal link from the finding of an assessment to the policy decision will be established (e.g., Bryk & Hermanson, 1994). The international assessments, especially the IEA TIMSS study in the late 1990s, have been given many accounts for their impact on the policy (e.g., Dossey & Lindquist, 2002; Kellaghan, 1996; Kellaghan & Greaney, 2001; Robitaille, Beaton, & Plomp, 2000).

As previously mentioned, data from an international comparative assessment will be used in this study. Functions that international assessments may serve and the general ways in which the data from these assessments may be presented are discussed below to provide a context for formulating the problem statement and research questions for this study.

1.4.1 Functions of international assessments

The international assessments may serve a number of functions for national and international educational policymaking and practice. The general ways in which the results of international assessments might be used were described by Plomp, Howie, and McGaw (2003, with references to Kellaghan, 1996; and Plomp, 1998). First, descriptive comparisons with other countries might serve to identify particular aspects of a national system, such as, the achievement levels of students, or the content of the curriculum that could be considered problematic because they differ from the practice in other countries. These comparisons may be of particular value when they are made with countries of special interest, such as 'cultural' neighbors or economic competitors. They could lead to further investigations of the aspects of educational system and may lead to actions to remedy the deficiencies.
Second, assessments may be conducted on a regular basis with the purpose of making informed decisions about change when and where it is needed. In this case, the information provided by an assessment is used for monitoring the educational system. The TIMSS study, with the first data collection in 1995, was repeated in 1999 and 2003 to provide trend data to serve this function.
Third, the findings of an international assessment can contribute to the understanding of differences between, as well as, within educational systems. This provides a base for informed decision-making about the aspects of educational system.

Fourth, international assessments reveal variations between educational systems that may be taken as a starting point for research leading to a better understanding of the factors that contribute to the effectiveness of education.
And finally, international comparative achievement studies serve to promote general "enlightenment" (Kellaghan, 1996, with reference to Weiss, 1981). The use of findings of international studies can enrich public discussions by providing information about other educational systems. This might serve to clarify, for example, assumptions about what schools try to achieve, what they actually achieve, and what is possible to achieve. In this case, findings are not directly related to individual decisions that might be taken, but rather contribute to a gradual diffusion of ideas into the sphere of organizational decision-making.

Most of these functions are not confined to the data from the international assessments but can be served when data are obtained from a national assessment. For example, national assessment data can serve the functions of "monitoring", "understanding", "research", and "enlightenment". The function of "descriptive comparisons" can be served if sub-regions or schools within the country are compared. While the data from the international assessments are used mostly at the system level, for example, for curriculum revision or pedagogical innovations, the results of national assessments, in which all schools participated, may also be used at the school level for accountability purposes (e.g., Fitz, 1996).

1.4.2 Presenting assessment information

As the results of international assessments can serve different functions, they need to be presented and interpreted using different approaches. In order to best serve the information needs of policy makers, the results of assessment studies should be presented summarily. For many other users however, information that is summarized at the level of average country scores may not be sufficient. For some users, overall scores may mask more than they reveal (Mislevy, 1995; Schmidt et al., 1998). Curriculum analysts, textbook writers, and teachers may want to know how students performed in particular areas of the curriculum, how different groups of students performed, or what is academically reasonable to expect from students. It is therefore important that the information derived from the assessment is presented in a way that enables meaningful interpretations to its users.
Following its aim, the results of the analyses of international comparative data in the present study need to be presented in a way that will be useful in implementation of the curriculum reform. Several approaches to such presentations and how they can be interpreted are described below.

The results of international assessments are generally presented as average achievement scores of participating countries on a common numeric scale. On such scales, average score in a particular country can be compared to average scores in other countries. Such comparisons are called norm-referenced interpretations (Hambleton & Sireci, 1997) and they mainly serve in a function of "descriptive comparisons".
Although such interpretations of the assessment results can be useful, they have often been used rather uncritically by the media and policy makers. Simple lists of education systems ranked by the mean achievement score on a test are unlikely to produce valuable guidance for development and implementation of improvement measures. Furthermore, such information may often focus attention on the league tables of countries in the international assessments or schools in the national assessments. This may lead to the overshadowing of more interesting aspects of performance, such as variance in achievement across different areas of the domain, or gender or ethnic-group differences. In consequence, it may disguise important characteristics of student achievement and mislead the inferences about the quality of educational system.

Further, to provide information on what students actually know and can do, assessment results need to be presented, not with reference to other students but, with reference to knowledge and skills that are expected to be mastered. These are usually based on objectives specified in the curriculum. Such interpretations of the results are called criterion-referenced interpretations (Hambleton & Sireci, 1997). They can be used to determine the extent of mastery of curriculum objectives, the achievement of certain attainment targets, or the achievement of performance standards, and through this, they can provide diagnostic information that is needed for improving the performance.
When mastery of curriculum objectives is investigated the achievement scores are usually presented in terms of the proportion of students that master each objective. For example, a student may be assumed to have achieved mastery if he or she correctly answers a predetermined number of items that assess the objective. When curriculum is structured in terms of the levels of performance expected of students at particular ages or grades, the assessment results may be presented to provide information on proportions of students reaching each target level.

Criterion-referenced interpretations are also not without problems (Wiliam, 1996). In addition to criticisms of arbitrariness, because of the use of the expert judgment in setting the 'criterion', the attribution of the meanings to the achievement scores can run into problems of specificity. At the same time, more generally worded statements, which could manage collectively to exhaust a domain, can be too vague to be interpreted consistently by different readers.

Assessment information can be interpreted following both, norm-referenced and criterion-referenced approach. As argued, one approach may yield more appropriate information than the other depending on the purpose of the assessment and the needs of the users. In the present study, both, norm-referenced and criterion-referenced interpretations of Slovene students' achievement are considered important as both provide information useful to address the problem under study. While the TIMSS international reports provide norm-referenced interpretations (Beaton et al., 1996; Mullis et al., 2000), as well as criterion-referenced interpretations (Kelly, 1999; Mullis et al., 2000) of mathematics achievements of students in participating countries, they do not focus on comparisons of Slovenia to other relevant countries nor do they take into account the intentions for achievement of Slovene students embedded in the policy documents and the curriculum. The results in these reports are therefore used as a starting point for further analyses in this study.

1.5 PROBLEM STATEMENT AND RESEARCH QUESTIONS

In the previous sections, the premises for posing the problem statement in the present study were outlined. It was argued that there were limited system level data on student achievement available at the time of curriculum development in Slovenia. Furthermore, no international comparative data that would allow for examination of developments in Slovene achievement over time were available. Since then, Slovenia participated in several follow-ups of the international comparative assessment TIMSS allowing for such issues to be addressed. This lead to the formulation of the following problem statement in this study:

> *How can the mathematics performance of Slovene students at the end of compulsory education in the non-reformed system and its developments in the second half of the 1990s be described to serve the needs for comparative international information as input in the process of curriculum reform and its implementation?*

In this problem, two perspectives for describing Slovene achievement are embedded. The first is the perspective of the intended curriculum for mathematics in Slovenia or, more specifically, the attainment targets that were specified in the reformed curriculum. Analyzing mathematics achievement of students in the non-reformed system from this perspective will enable insight into 'the starting point' of the new curriculum. Future assessments that may be carried out in Slovenia may utilize this information to find out whether improvement measures introduced in the reformed curriculum had desired effects on student achievement.

The second perspective from which Slovene mathematics achievement will be examined are achievements of students from other European countries. As explained, the relevance of this perspective also emerged from the Slovene policy documents. In order to identify areas in Slovene achievement in which improvements might be desired, countries with similar or higher overall achievements than in Slovenia are considered relevant for comparisons. The rationale for the selection of these countries is described in more detail in Chapter 3 and the actual procedure is described in Chapter 5. The countries[2] that emerged from this selection were Belgium-Flemish, the Netherlands, Hungary, and the Slovak Republic. The relevance of comparisons of Slovenia with these countries is underlined by the membership of Belgium-Flemish and the Netherlands in the European Union and by participation of Hungary and the Slovak Republic, as of Slovenia, in the process of accession to the association in May 2004.

Corresponding to the two perspectives that are important for describing Slovene mathematics achievement, two research questions are posed in this study:

1. *How well did Slovene students at the end of compulsory education in the non-reformed system in the late 1990s perform in mathematics when compared to the attainment targets in the reformed mathematics curriculum and what were the developments in this performance between 1995 and 1999?*

2. *How well did Slovene students at the end of compulsory education in the non-reformed system in the late 1990s perform in mathematics when compared to the performance of students in other European countries and what were the developments in this performance between 1995 and 1999?*

As argued in the previous section, there is no single 'right' analysis for describing student achievement. In international reports usually a single number is constructed that summarizes the performance of students in a particular country

2 The term *country* is used in this thesis as a synonym for *educational system*. Certain members of IEA, for example Belgium-Flemish, are part of a country but operate from an educational perspective independently from the other part(s) of the country.

which can be compared to the numbers constructed for other countries. Additional estimates of achievement are usually shown for particular content areas in the domain assessed. In this study, analyses will be carried out for the overall mathematics domain, for content subdomains in mathematics that will reflect the structure of content in the Slovene curriculum, and for so-called cognitive categories (e.g., Bloom, 1956; Robitaille et al., 1993). Finally, contents of individual items will be examined on which particularly high or low achievement will be observed. These analyses will enable norm-referenced as well as criterion-referenced interpretations of achievement results of Slovene students.

Using two different perspectives will provide more insight into students' achievements than if a single perspective was used. Furthermore, the differences between the results of the analyses based on each perspective can be examined. In this study, this is called 'convergence' between the two perspectives. The analysis of convergence of the results from the two perspectives may highlight areas in Slovene achievement in which improvements might be desired from both perspectives. Furthermore, areas in which expectations in the intended curriculum might be considered too high or too low taking into account achievements of students from other European countries may also be highlighted. In terms used by Kellaghan (1996) for this function of international assessments, this is called "enlightenment".

As indicated previously, data on student mathematics achievement from the TIMSS study will be used to address these research questions. The TIMSS data collections have been carried out in 1995, 1999, and in 2003. TIMSS 2003 data are not yet available for analyses. To describe the performance of Slovene students at the end of compulsory education in the non-reformed system, the TIMSS 1999 data will be used. For examining the developments, the TIMSS 1999 data will be compared to the data from TIMSS 1995.

1.6 OVERVIEW OF THE FOLLOWING CHAPTERS

Chapter 2 provides an overview of the Slovene education system and the reforms that are being introduced since the late 1990s. Changes in the mathematics curriculum are outlined in more detail to provide a background for describing the results of this study. Conceptualization of the study based on a literature review is given in Chapter 3. Main concepts used in this study are discussed in order to provide a basis for presenting the design and the results in subsequent chapters.

Chapter 4 outlines background of the TIMSS study. The conceptual framework and research questions, design, sampling, instrument development, data collection and data processing in TIMSS are also summarized in this chapter.

The design of the present study is described in Chapter 5. The analytical approaches for addressing the research questions are elaborated. The operational research questions are derived that will be answered by employing several analytical procedures. Chapter 6 gives results of the analyses for the first research question. Slovene achievement in mathematics and its developments in the late 1990s are examined from the perspective of the attainment targets in the reformed intended curriculum in Slovenia. Chapter 7 gives results for the second research question through which Slovene achievement is compared to achievements in other European countries. In the final section of Chapter 7, convergence of the results obtained from the two perspectives is examined.

Finally, Chapter 8 summarizes the research and its main findings. Discussion is given on the most notable outcomes and issues emerging from them with regard to the mathematics achievement and the reformed curriculum in Slovenia. Reflections on methodological issues and findings of the study are followed by recommendations how to further support the implementation process of the reformed mathematics curriculum in Slovenia. Recommendations for future research in this area are also given.

CHAPTER 2
Education reform in Slovenia

Slovenia is a country located in the south of central Europe. It gained independence in 1991 after the disintegration of Yugoslavia. Reforms following the move to independence were fundamental as they had to reflect the political, economic and social changes, which came as the consequence of the independence. Apart from creating a legislative basis for changes, the reform efforts were focused on the development of curriculum and assessment systems and improving the effectiveness and transparency of the education system in general. The education system in Slovenia and its reforms are described in this chapter.

In section 2.1, a few basic figures about Slovenia are given and section 2.2 describes basic elements of the organization of policy making in Slovene compulsory education. The structure of the non-reformed system and the changes made to this structure by the reforms are described in section 2.3. Main characteristics of the curriculum reforms are pointed out in section 2.4, followed by a review of the changes in the curriculum for mathematics in section 2.5. The data in this chapter were assembled from Statistical Office of the Republic of Slovenia (2002ab) and the descriptions of the system are based on the *White Paper* (1996) and Plevnik (Ed., 1998).

2.1 THE COUNTRY'S CHARACTERISTICS

Slovenia has an area of 20,273 km^2 and a population of 2 million. 83% of the inhabitants are Slovenes by ethnic origin. Two indigenous ethnic minorities live in Slovenia: Hungarian (0.32%) and Italian (0.11%). There are also Romanics, Albanians, Montenegrins, Croats, Macedonians, Muslims, Serbs and other undeclared by ethnic affiliation. The main religious affiliation is Catholicism. 69 % of inhabitants are Catholic, 1.2 % Protestant, 0.6 % Orthodox, 0.6 % Islamic, while 28.2 % do not express any religious affiliation (or data were not available).

The official language is Slovenian, while in areas inhabited by members of the Italian and Hungarian minorities, Italian and Hungarian are also official languages. Slovenia is a unitary state with a republican form of government. The first multi-party elections took place in 1990. Since then a parliament has been formed by various parties from the left and right. The parliament appoints the Prime Minister at the proposal of the President. Governmental power is fully exercised by the Prime Minister.

2.2 POLICY MAKING IN COMPULSORY EDUCATION

In the decades before the independence Slovene compulsory education was characterized by centralization and minimal diversity. This uniformity was based on the idea of providing equal opportunities for everyone to learn the basics that were required for continuing education. Education in Slovenia was, and still is centrally run by the Ministry of Education and Sports (recently merged with the Ministry of Science into the Ministry of Education, Science and Sports; in the following named the Ministry of Education, or the Ministry).

The Ministry is responsible for defining national policy on education. Its prime responsibilities relate to the structuring and funding of the system, the management of schools (with exception of one privately run Waldorf school), inspection procedures and financial aid. The Ministry controls the education system through legislation. Education is funded through the state budget, and schools are directly financed by the Ministry according to norms which are set and controlled by the Ministry. Local communities provide a proportion of the resources for investments, maintenance and equipment expenses.

As support to the Ministry in carrying out educational development and counseling, there is the National Educational Institute (*Zavod Republike Slovenije za šolstvo*) which has nine regional offices that offers advice to, and supervision of schools with respect to curricula, textbooks, time-tables, etc. The National Examination Center (*Državni izpitni center*) is responsible for preparing, supervising the administration of, and providing feedback on the statewide student exams at all levels of pre-university education.

In his decisions regarding compulsory education, the Minister of Education is obliged to take into account recommendations from the Council of Experts of the Republic of Slovenia for General Education (*Strokovni svet*, further called Education Council or Council). The Council is a permanent advisory body elected by Parliament, and is comprised of 26 members who represent various institutions that are involved in some way with education (research institutions, universities, or counseling centers). The Council's role is to determine the contents of education curricula, approve curriculum guides, textbooks and other education materials, and propose criteria and standards for school equipment. To propose new policies, the Minister of Education also consults with experts, school representatives, organizations of parents and students, teachers, and other stakeholders in education.

2.3 THE STRUCTURAL REFORM OF THE EDUCATION SYSTEM

Compulsory education in Slovenia is also named *elementary education* or *basic education*. By finishing elementary education, which in the non-reformed system covers education of students from age of 7 to age of 14 years and in the reformed system from age of 6 to age of 14 years, students also complete their statutory, compulsory education. Figure 2.1 presents the structure of the education system in Slovenia. After compulsory school, most students continue their education. Except for privately run Waldorf primary school (supported only in part by the state) all elementary schools in Slovenia are public and fully funded by the state. Compulsory education in Slovenia is therefore free of charge.

In the late 1970s, the number of children in successive age cohorts entering Slovene schools was almost 30,000. Declining birth rates since 1980 have yielded contemporary cohorts of approximately 20,000. Small schools in rural areas count as school buildings but not as organizational units, because they are organizationally considered part of another, larger school. In such cases, the larger school is called the central school and smaller school is called the settlement school. In the school year 2001/2002, there were 448 central schools and 365 settlement schools.

Because the implementation of reforms in the school system is gradually taking place, there are currently two school systems in Slovenia, the non-reformed system and the reformed system. Schools first started to introduce reforms in the school year 1999/2000. At the same time, older students and students in other schools attended the classes according to the non-reformed system. In the school year 2001/2002, schools were still providing instruction according to the non-reformed system to approximately 160000 students. In schools that began with the implementation of the reforms, instruction according to the reformed system was provided for 14000 students. These students either entered education in the reformed system, or transferred from the non-reformed to the reformed system in grade 7 (of the reformed system). This transfer occurred in order to speed up the process of reform implementation. In the years following 1999/2000, the introduction of reforms was extended to an additional number of schools, until all remaining schools introduced the earlier entrance into grade 1 of the reformed system by 2003/2004. Instruction according to the reformed system will be given in all grades in all schools by 2007/2008. The basic characteristics of the structure of the non-reformed system will be described below, followed by the main changes introduced in the system by the on-going reforms.

2.3.1 The non-reformed system

Compulsory education of the non-reformed system in Slovenia is eight years comprised of students between 7 and 14 years of age. It is divided into two four year cycles; the first four years of primary education, and the second four years of lower secondary education. According to ISCED 97 (*Unescov priročnik* - Unesco manual, 2001) classification of these two cycles corresponds to ISCED 1 and ISCED 2 levels (see Figure 2.1).

In grades 1 through 4, children are taught all subjects by one teacher although specialist teachers are employed for subjects such as music and physical education. In grades 5 through 8, children are taught by subject teachers, most of whom have been qualified to teach two subjects (for example, mathematics and physics, or geography and history).

The school year lasts 38 weeks or 190 school days, and is divided into assessment cycles. Classes are held five days a week. The compulsory program includes core and optional subjects as well as days used for various field activities (cultural and science days, practical activities, sport days). There are also remedial and additional classes and extracurricular activities provided by the schools which are not compulsory for students to attend, but are compulsory for schools to provide.
The number of lesson periods per week varies depending on the grade students attend. Lower grade students (grades 1 through 4) typically have 20 lesson periods per week with an additional 2 periods of non-compulsory remedial classes, or classes for advanced students, 2 periods of extracurricular activities, and a so-called home class period (half a period per week). The home-class period is intended for students to analyze their study results, or look for the ways and methods of resolving conflicts in their everyday school life.

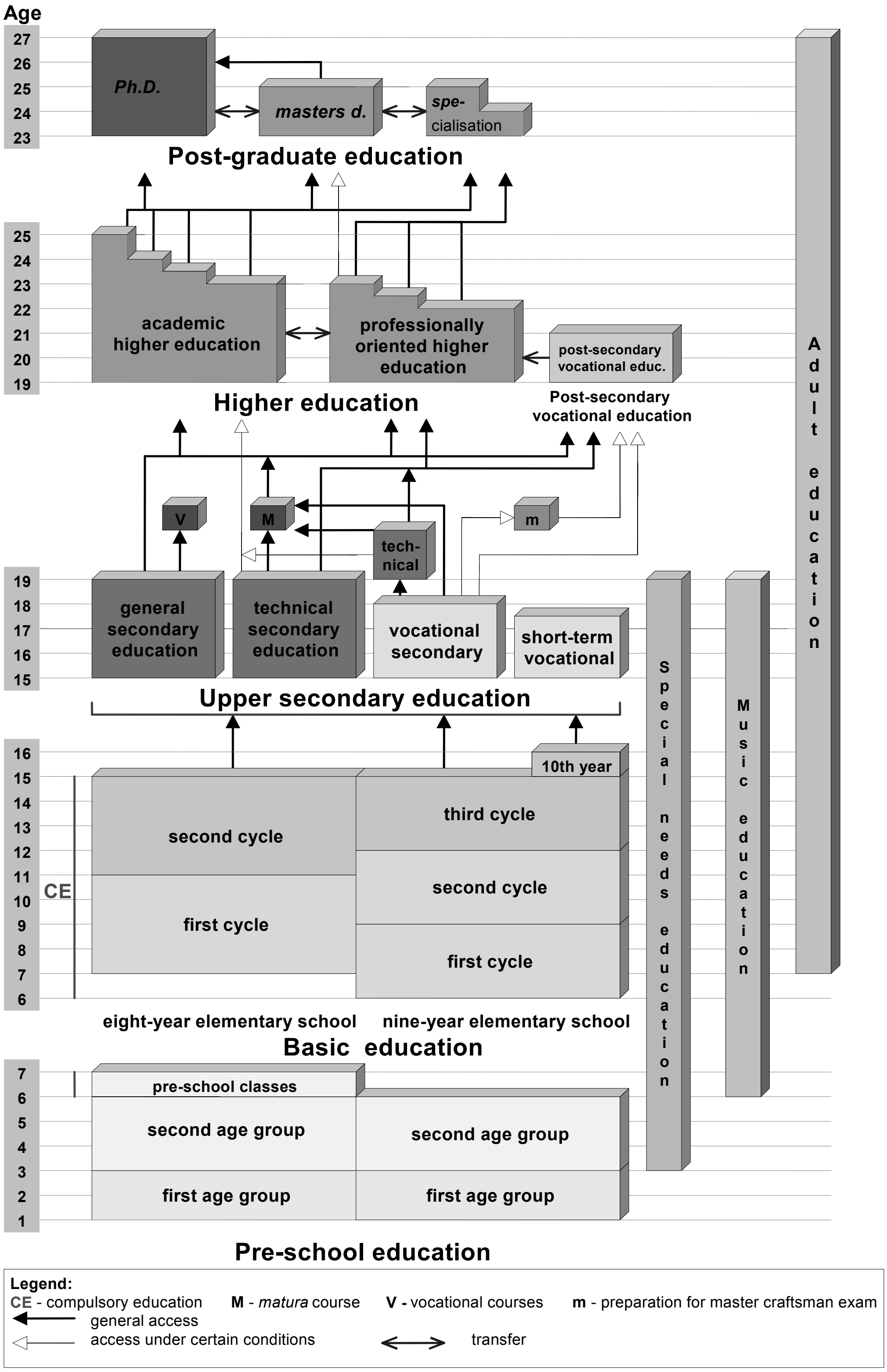


Figure 2.1 Structure of the education system in Slovenia

Note: Based on Plevnik and Žižmond (2000), from Brečko (2003).

Upper grade students (grades 5 through 8) typically have 26 lesson periods per week with an additional 2 periods of remedial classes or, classes for advanced students, 3 periods of extracurricular activities, and a home class period (half a period per week). Lessons are generally held in the morning and each period lasts 45 minutes.

The compulsory education offers the following mandatory subjects: Slovene language, foreign languages (English, German or French), mathematics, history, geography, ethics and society or civics, chemistry, biology, physics, drawing, music, design and technology, physical education, and home economics. In lower grades the subjects are combined into groups according to the prescribed syllabus (*Program življenja in dela v osnovni šoli* – The Programe of Life and Work in Elementary School, 1984, further called *Curriculum Guide*). Progression from grade to grade at all school levels is based upon marks given by teachers on a numerical scale. The exception is the first assessment cycle in the first grade of compulsory education where teachers do not grade students numerically, but they monitor their progress which they then report to their parents. In the assessment cycles in higher grades, two methods of grading are used: on numerical scale (form one to five, with five being the highest mark) in 'instructional' subjects (e.g., mathematics, mother tongue), and a three-level grading scale (very successful, successful, less successful) in 'educational' subjects (sports education, music, and arts). The marks given by teachers are aggregated at the end of the school year. Students who receive an unsatisfactory mark for any subject are given a chance to take an additional test scheduled and evaluated by the teacher before the next school year begins. With only one unsatisfactory mark, a student may be allowed to progress up to the next grade, but if a student has more than one unsatisfactory mark or fails in the same subject for two consecutive years, that student must repeat the grade.

2.3.2 The structural reforms

The following paragraphs provide a brief overview of the main structural reforms in the compulsory education system. As already mentioned, other levels of pre-university education were also modified including the education of students with special needs and teacher training.

Beginning of education and the duration of compulsory education

The reforms of the Slovene education system included an extension of the compulsory education from eight to nine years. This extension was done in a way, that the beginning of the compulsory education was prolonged by one year, which means that children enter school one year earlier, at the age of six instead of seven.

The nine years of compulsory education in the reformed system are divided into three cycles, each covering three consecutive years. According to ISCED 97 classification, the first two cycles (grades 1 to 6) correspond to primary education, ISCED 1 level, and the third cycle corresponds to lower secondary education, ISCED 2 level (see Figure 2.1).

In addition to the extension of compulsory education to nine years, a non-compulsory tenth grade of elementary education was also introduced. It is intended for students who have not successfully completed the ninth grade of compulsory education, or those who are not satisfied with their grades in the final assessment of knowledge and skills at the end of the ninth grade. In the course of this tenth grade, students may acquire knowledge and skills that are necessary to complete their compulsory education, or improve their results in the final assessment needed to enroll in secondary schools.

Organization of school time and compulsory school program

In the reformed system the requirements for students' attendance where changed to some degree in order to offer a higher degree of flexibility of instruction. Students' weekly attendance requirements comprise no more than 22 periods in the first cycle, 26 periods in the second cycle, and 30 periods in the third cycle. Within the cycles, number of periods increases from lower to the higher grades.

In the syllabus, the beginning of the first foreign language commences in the fourth grade while in the non-reformed system it was introduced at the fifth grade. Foreign languages are offered as part of the extracurricular activities at lower grades in both systems. In the first and the second cycle of the reformed system, subjects are linked and/or merged into subject areas (for example, science which includes biology, physics, chemistry, the field of arts, etc.).

Assessment and conditions for advancement

The developers of the reforms argue that the system of teacher assessment in the non-reformed education system did not maintain a balance of importance among individual subjects, since the students' attitude toward a subject was largely determined by the assessment and not by the knowledge and skills provided by the instruction. They further argue that this system paid little attention to the characteristics of a child's development (individual differences, pace of development) and learning (types and levels of knowledge, usefulness of knowledge) both at the beginning of education and in individual cycles (*White Paper*, 1996).

To avoid this lack of balance it was recommended that in the reformed system more emphasis should be put on other methods of teacher assessment. National formulae for descriptive grading were provided to teachers in the first and second cycles. In the first grade, teachers use only descriptive grading, while in other grades of the first and the second cycles, descriptive grading is combined with numerical grading so that teacher reports include numerical grades in individual subjects and a grade in overall achievement.

At the end of the first and second cycles, the knowledge and skills of students are assessed by means of a non-compulsory national test in mathematics and their mother tongue. Additionally, in the second cycle, an optional subject is tested. These tests are prepared by the National Examination Center and are administered according to a standard procedure. The results are used as feedback for schools, students and their parents, but they do not influence either the grades or whether or not a student advances from the lower to the higher cycle. At the end of the second cycle the results of the national tests are used to differentiate between a student's level in the third cycle.
In the third cycle, teachers assess the students' progress during the year by means of numerical grading. At the end of the third cycle, a compulsory national test in mathematics, mother tongue and an optional subject is administered as an addition to the certification of completion of compulsory education. The student report card lists a grade for each subject; for subjects not taken on the national test, a final grade is given by the teacher, and for subjects taken in the national test, a final grade consists of equal proportions of results from the national testing and the final grade given by the teacher.

In all cycles students and their parents receive teacher's reports at least twice in a school year and at the end of each school year. Students do not generally repeat a grade in the first two cycles, except the last grade of the second cycle (grade six) if they do not receive passing grades in all subjects. The results of national tests are not published and they can not be used as a criterion for remuneration of schools or teachers.

Differentiation at school
In the non-reformed system there was no external differentiation of students. To accommodate different needs of students, teachers used different approaches to teaching within the classes. In the reformed system, external ability differentiation is being introduced into the second and the third cycles. In the second cycle, and grade 7 of the third cycle, a flexible form of differentiation is introduced in at least two subjects, one of which is mathematics. The flexibility of the differentiation is that students spend most of their time in heterogeneous (home) classes, where they deal with the basic subject matter, and no more than a quarter of the time in

homogenous classes, where instruction is based on the ability levels. Generally, students are differentiated into three ability levels.
In grades 8 and 9 of the third cycle, formal external differentiation into classes according to ability is carried out in at most three subjects; mathematics being one of the compulsory subjects for differentiation. Criteria for the assignment of students to ability levels are based on the results of a national test given at the end of the sixth grade, achievement in a particular subject in the seventh grade, and wishes and interests of students and their parents. Therefore the ability level may be decided upon by the student. Students may transfer from one ability level to another at the end of individual assessment cycles, and as an exception, during an assessment cycle.

2.4 THE PREMISES AND GOALS OF CURRICULUM REFORM

School curriculum in Slovenia is prepared at national level. The curriculum guide used in the non-reformed system was written in 1984 and consisted of 610 pages of specific content and objectives to be taught, sequence of contents, numbers of hours to be devoted to each specific objective in each subject, in each grade (*Curriculum Guide*, 1984).

In the Slovene policy documents, five main reasons underlying the need for the curriculum reform in Slovenia are outlined (Nacionalni kurikularni svet, 1996, further named National Curriculum Council). These reasons include: ageing of the population; transition to post-industrial system in which new technologies are playing an important role; economic, social and political changes in the country; environmental issues; and the country's recent independence and the need for integration in the European and world economic and communication networks. Some of these reasons may underlie the need for education reforms in other countries as well, but this is even more important in a country like Slovenia which has seen such radical change over the past few years and which faces more in the near future (OECD, 1998).
Among the problems in Slovene education, the National Curriculum Council recognized the compartmentalization of the school subjects, overloaded programs and school syllabi, lagging behind the developed countries in the educational achievement levels, and the poor quality of knowledge and skills of students. Accordingly, the goals of the curriculum reform included "*the attainment of internationally verifiable standards of knowledge*" and "*enlargement of the quality and permanency of the knowledge and skills of students*" (National Curriculum Council, 1996).

Following these conclusions, the intention of the curriculum development in Slovenia was to reduce the curriculum content and build more integration between different school subjects, placing more emphasis on inter-disciplinary knowledge and cross-curricular topics. It was concluded that the curriculum should help students to develop the capacity to utilize the knowledge and skills they acquire in school for thoughtful or innovative purposes. One of the major components of the reformed curricula of school subjects was the specification of attainment targets for each subject in each grade. In the past no explicit framework of standards for the outcomes of Slovene compulsory education existed. The intention of setting attainment targets for student outcomes was that they would serve as a basis for further curriculum development and implementation. In addition, the existence of these attainment targets supports the autonomy of teachers and at the same time enables monitoring the functioning of the educational system.

2.5 THE REFORM OF THE MATHEMATICS CURRICULUM

As in many countries, mathematics is one of the core subjects in the Slovene school syllabus. It is introduced in grade 1 and remains in the syllabi of all higher grades. The non-reformed mathematics curriculum guide is a 22 page part of the 610 page curriculum document prepared at national level which includes guidelines for work, the syllabus and also the curricula for all other school subjects (*Curriculum guide*, 1984). In this curriculum, only the titles and subtitles of content areas that should be covered in each grade are specified. These are accompanied by numbers of school periods allocated for the particular content area and approximately ten lines of instruction for implementation. In the non-reformed system, the textbook to be used in mathematics classes in each grade was prescribed by the authorities (until the late 1990s). These textbooks were used by teachers as an additional document describing the intended curriculum for mathematics in the non-reformed system.

The reform of the mathematics curriculum in Slovenia was developed between 1996 and 1997 and was adopted by the Slovene Education Council in 1998. The following description of the reforms in the Slovene mathematics curriculum is based on Magajna (2000).

The weaknesses in mathematics knowledge and skills of Slovene students were recognized as follows: poor understanding of numbers and operations; poor abilities to solve demanding problems and tasks, especially in the area of data representation and analysis; a weak link between operating skills and

mathematical skills; the late, although efficient treatment of certain topics in algebra, proportionality, and indirect measurement tasks.
Based on the views of mathematics teachers about the non-reformed mathematics curriculum, the analysis of mathematics curricula in several other countries, results that were available from the IAEP II and TIMSS 1995 studies, and modern theories of teaching mathematics, the Mathematics Curriculum Development Panel (*Predmetna kurikularna komisija za matematiko*) concluded that major changes to the Slovene mathematics curriculum were required. However, the Panel decided that any changes should be introduced in the curriculum carefully and gradually. The Panel argued that past experiences and experiences in other countries have shown that radical changes are risky and often do not yield desired results.

The main features of reform of the mathematics curricula in Slovenia are therefore as follows. More emphasis was put on conceptual knowledge and skills especially through six lesson periods per year of "project work" in order to support conceptual learning. Certain topics were excluded from the curriculum, such as, vectors, systems of linear equations, procedures for construction of a triangle similar to a given triangle, and parts of geometrical projections. Other topics were rearranged. Decimal fractions were introduced to younger students beginning in grade 6 of the reformed system, as compared to grade 6 in the non-reformed system. This introduction was carried out through measurement topics instead of fractions. Topics in three-dimensional geometry, for example, the naming and recognition of a cube, were introduced in grade 1, to six year old students. There were also some changes made to teaching measurement topics.

The 1984 mathematics curriculum guide did not include topics from data representation and analysis. Therefore, in the reformed curriculum, a whole new sub-area of data representation and analysis was introduced comprising of simple procedures for data collection, classification, presentation and learning of, and computing central tendencies of data. It was intended that these mathematical topics were to be integrated with other topics, at least at the lower grades. It needs to be mentioned however, that most mathematics teachers and educators in Slovenia considered the 1984 curriculum guide to be long out-dated. By the late 1990s, there were several additional textbooks approved by the Education Council that included these topics and were used by teachers in their classes.
With the introduction of the data representation and analysis of the reformed curriculum, the use of computer technology was considered important, with emphasis on working with spreadsheets. Calculator use was introduced a year earlier than in the non-reformed system, now in grade 6 of the reformed system.

Other mathematical topics that were not reallocated or excluded were to be introduced at the same age as in the non-reformed system. Due to an earlier beginning of the compulsory education in the reformed system these topics are introduced at the sequential grade as compared to the non-reformed system.
As mentioned previously, the differentiation in ability of students in eighth and ninth grades was introduced in order to better organize teaching and learning of mathematics for students with different abilities. At the same time, the learning objectives from the reformed mathematics curriculum remained unchanged between ability levels. Each student is supposed to be offered the whole range of mathematical topics in which they can learn by following different paths best suited to their abilities, and consequently, achieve different results.

The reformed mathematics curriculum guide for compulsory education in Slovenia (*Učni načrt*, 2002, further called *Curriculum Guide*) contains 86 pages of learning objectives for mathematics in grades 1 to 9. As in the non-reformed curriculum guide, it is structured by grades and, within grades, by content areas. The numbers of school periods to be allocated to each content area in each grade are specified within which teachers are free to allocate time to topics included. In contrast to the non-reformed guide, the objectives are given in the form of contents and processes that students are desired to master as a result of instruction. They are accompanied with didactical recommendations and concrete examples of items to guide teachers in their implementation. Often, cross-curricular linkages are indicated.

The operationalized objectives for instruction are given at the end of each grade in the form of performance objectives or attainment targets. As mentioned in Chapter 1, these attainment targets are the main new feature in the reformed curriculum. They represent a departure from a traditional approach of only specifying the contents to be taught in each subject, in each grade. These attainment targets are set at two levels in most grades and at three levels in the two final grades (grades eight and nine). The lowest level is called "minimal standards" and they represent what nearly all students should attain[1]. In other words, when a student masters the set of contents and skills that are represented in the minimal standards he or she receives a passing grade. A student may receive a passing grade even when he or she does not master a whole set of minimum standards; the decision to what extent students should attain the minimum standards in order to successfully pass the subject is left to teachers.

[1] In the curriculum guide the term *standards* is used. In this thesis, they are called also *attainment targets* as they represent to knowledge and skills students should attain. The meanings of the concept of standards are discussed in more detail in Chapter 3.

The second level is called "fundamental standards". They are defined as the knowledge and skills expected of the average achieving student and what teachers should strive for their students to learn. The "higher standards" represent the knowledge and skills to be attained by higher achieving students and, as mentioned, are set only in grades eight and nine.
There are up to 17 attainment targets listed for each grade. They are in the form of short sentences. For example, in grade 6, one of the "fundamental" attainment targets states that "*a student is able to compare different decimal fractions*" (*Curriculum Guide*, 2002, p. 79; translation MŠ). Attainment targets at Level 3 in grades eight and nine include specialized knowledge and skills that usually only higher achieving students can master.

Operational guidelines on how the three ability levels should be determined and what percentage of students should master these knowledge and skills in order to say that the standards are met are not given. As explained in Chapter 1, these attainment targets (also called the standards) will be used in the present study as one of the reference points for describing mathematics achievement of Slovene students.

CHAPTER 3
Conceptualization of the study

The problem in this study concerns mathematics achievement of Slovene students. From the discussion in Chapter 1, it follows that several concepts need to be considered in order to address this problem. For example, an assessment needs to be carried out and correspondence of achievement measures with pre-selected reference points needs to be analyzed in order to determine whether this achievement reaches the desired levels. In this chapter, these and other concepts on which this study is based will be discussed to provide the research context for this study. The concepts are conceived as components of the monitoring process of the functioning of education system. Based on this conceptual framework the design and the results of this study will be presented in subsequent chapters.

This chapter begins with section 3.1 in which the main concepts used in this study are defined. Their relationships are presented in a conceptual framework. In the following sections each of the concepts from this framework are discussed. Educational goals and standards are discussed in section 3.2, and monitoring in section 3.3. Conceptualizations of the curriculum and, within it, of student achievement are discussed in section 3.4. Section 3.5 elaborates on issues pertaining assessment of student achievement. The final section (3.6) discusses measurement of correspondence of achievement with predefined reference points and measurement of trends.

3.1 CONCEPTUAL FRAMEWORK

The research problem in this study is to describe how well Slovene students at the end of compulsory education in the late 1990s performed in mathematics. As argued in Chapter 1, this question refers to a broader concept of the quality of education. Quality may be related to many aspects of education, for example, school climate, equipment, or teaching approaches. As also discussed in Chapter 1, *student achievement* is an important aspect of educational quality. Increasingly, in many countries student achievement is *monitored* through an *assessment* with the intention to identify areas in the *curriculum* in which improvement might be desired. To evaluate information obtained from assessments, appropriate *reference points* are needed. A reference point may be constructed on the basis of *the goals of education*, the *standards,* or on the basis of achievements in other countries. Assessment information is then interpreted by measuring the *correspondence* of achievement with the selected reference points. This can be followed by identification of *strengths* and *weaknesses* in achievement, *diagnosis* of possible causes and development and implementation of *improvement measures* in the curriculum. These concepts are conceived as components of the monitoring cycle as shown in Figure 3.1. Whether or not improvement measures have taken place, the next cycle of monitoring can be carried out to examine *trends* in student achievement. In case improvement measures have been introduced, trends can be used to indicate whether the desired effects in student achievement have occurred.

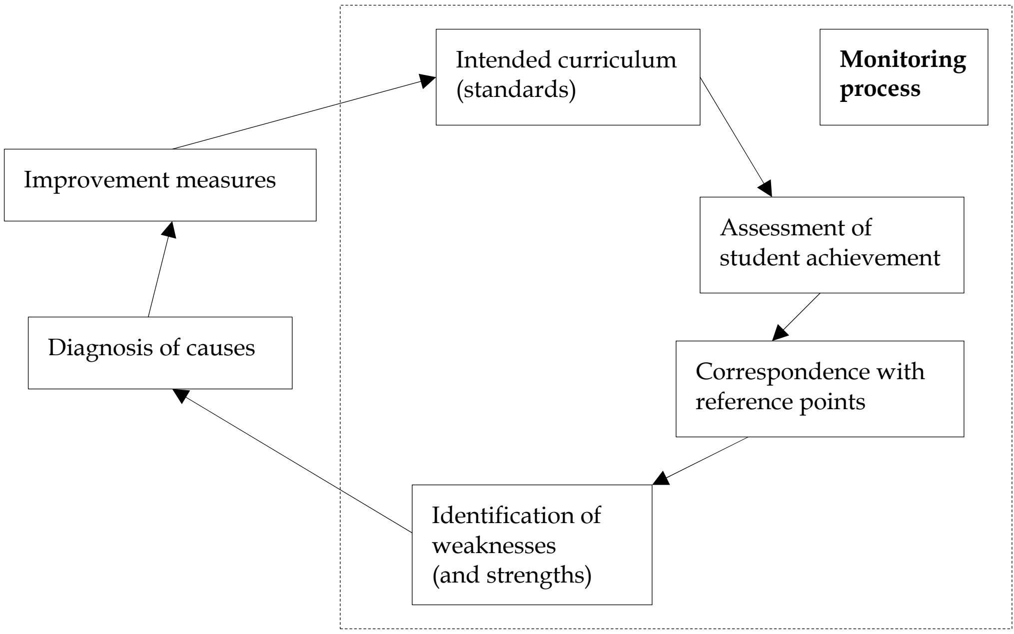


Figure 3.1 Conceptual framework for this study

Note: Based on Pelgrum (1998, p.7)

In this thesis, the focus lies on the concepts that are a part of the *monitoring process* in Figure 3.1. Diagnosis of causes of observed weaknesses (and strengths) and development of improvement measures, also presented in Figure 3.1, are beyond the scope of this study. In the following, descriptions of the concepts in the monitoring process will be provided and based on a literature review.

3.2 EDUCATIONAL GOALS AND STANDARDS

Many nations have explicit or implicit *goals* for education. Specifically defined goals may be written in policy documents. Implicit goals of education may be what is expected of the education system but not explicitly written. They are often set at the national level by the central government and specified in very general terms. In order to measure the functioning of the education system, goals need to be translated into more tangible objectives, formulated for example as *education standards*. In the subsections below, first a discussion on the meanings of standards is presented, followed by a description of two approaches to standards setting that are relevant in this study.

3.2.1 The meanings of standards

The US Joint Committee on Standards of Educational Evaluation (1994) defines that "*a standard is a principle mutually agreed to by people engaged in a professional practice, that, if met, will enhance the quality and fairness of that professional practice, for example, evaluation?*" (p. 2).

Education standards can be set for various elements of the education system. The elements of education system are often classified into three categories: inputs into schools, processes in schools, and outcomes of schooling (Postlethwaite, 1994). Inputs into schools are, for example, school buildings, school supplies, and teachers. Standards for inputs, often called norms are, for example, the minimum specifications of what a school should possess or the minimum student-teacher ratio.
School processes are, for example, teacher workload, curriculum (which in the case of national curricula can also be classified as the input), or the organization of the school. The standards for these processes could be the specification of the number of hours each teacher spends in a classroom, the extent to which the curriculum should be covered, or the specification of the required contacts of school with parents and community.
The educational outcomes are generally concerned with students' cognitive achievement and attitudes toward subject areas of interest at selected age or grade levels. Since the focus in this study is on student achievement, the following discussion concerns achievement standards.

A review of the literature reveals that the term *standard for student achievement* is used in a variety of meanings. In addition to differences in terminology, the definitions of standards differ in the specificity of operationalization of this term.

For example, Gonzalez and Beaton (1994) describe that a *standard* can be conceived as a conceptual definition that refers to the desired levels of performance that is to be reached or surpassed by most, if not all the students; or, they may refer to a benchmark against which the achievement of an individual or group can be compared. Gonzalez and Beaton name the operationalization of the standard onto a numeric scale a *cut score*. The performance is seen as a trait or internal disposition, an individual's score is used as an *indicator* of this trait, and the cut off level compared to the score is an indicator of whether the individual has met the standard. Similarly, Wiliam (1996) defines standard setting as the identification of certain points on a numeric scale with particular standards, with the intention to enhance the inferences that are warranted from the test scores.
In other cases, the term achievement standard is used to refer both to the *kind* of knowledge or skills students should learn and to *how well* they are expected to master them (Thomas, 1994). According to Thomas, in addition to defining the content and levels, applying standards also means specifying how achievement will be assessed and therefore how the achievement standards are operationalized.

As described in Chapter 2, the main new feature in the reformed curriculum in Slovenia are the attainment targets for student achievement. The terminology most applicable for these targets, and therefore adopted in this study, is given by Phillips (1994) and also by Husén and Tuijnman (1994). They distinguish between *content standards* and *performance standards*. A content standard represents what students are expected to learn, involving knowledge and skills essential to a domain under consideration. The performance standards describe what students must do in order to demonstrate that acceptable levels of learning have occurred; in other words, to indicate to what extent the content standards have been reached. Performance standards specify "how good is good enough" and they are often given as a cut score on a test, but, in general, they can be any demonstration that students have mastered a pre-defined level of learning.
The attainment targets in the reformed Slovene mathematics curriculum represent the content standards. The three levels of these standards could be seen as the performance standards at system level although they are not given on a numerical scale. The procedure through which they will be operationalized onto such a scale will be described in Chapter 5.

Standards are usually set by a group of judges or other individuals that are considered to be either experts in the field or to have certain stakes in the outcomes being measured. This "arbitrariness" of standards setting has often been criticized, especially when used in high-stakes accountability systems. Wiliam (1996) argues that the selection of both the content standards and the cut score with which they are represented are arbitrary and are driven by (often implicit) a set of values. To address such criticisms, Gonzalez and Beaton (1994) wrote that what should be sought is an informed procedure of standard setting with numerous pieces of evidence that will support the selection of a given cut score.

Given that the attainment targets and the performance levels are a novelty in the Slovene intended curriculum and that there are no explicit guidelines on the procedures to be used for the selection of the cut scores, this study will use expert judgment on the meaning of these attainment targets for measuring student achievement at system level.

3.2.2 Approaches to standard setting

Of the variety of approaches to standard setting (e.g., Gonzalez & Beaton, 1994; Thomas, 1994), two are relevant for this study. Their description is based on Postlethwaite (1994). The first approach is where a ministry or a body of expert judges sets a standard before the data are collected. This is called the *a priori* approach. An example is when a panel of experts identifies a set of items or performances, which students in a given grade would need to answer or carry out correctly to conclude that they have met the standard. Such a standard might be used for example to decide when students are ready to proceed to the next grade. At system level, a priori standards could be set in the form of percentages of students that need to answer correctly a set of items or a given proportion of these items.

The second approach is the derivation of standards after the assessment data have been collected. This is called *a posteriori* approach. This approach may involve identifying one or more cut-off scores on the achievement scale based on criteria of relevance, such as, for example, that the selected cut-off points best distinguish between "poor", "average", and "advanced" performers. The cut-off scores divide the continuum of student achievement into levels. The achievements of students with scores at the particular level determine the standard at this level.

In the present study, the two approaches to standard setting will be used to provide (operationalized) sets of *reference points* (also called 'perspectives' in Chapter 1) for describing mathematics achievement of Slovene students. As indicated in the research questions, two sets of reference points will be used, the attainment targets from the reformed curriculum and achievements of students from other European countries. The first reference point is constructed a priori, while the second is constructed a posteriori. The derivation of these sets of reference points from the Slovene curriculum and from the TIMSS data will be described in Chapter 5.

3.3 MONITORING

To obtain information about the progress of the educational system towards the desired goals or standards, the educational system needs to be monitored. *Monitoring* refers to systematic and regular collection of data about important aspects of education in order to identify changes occurring in the system. The purpose of monitoring is to provide information on the basis of which it can be determined whether the system goals have been met or whether adjustments need to be made in the system in order to increase the correspondence with these goals (Pelgrum, 1990, 1995).

In this thesis, monitoring is focused on student achievement. Postlethwaite (1994) describes the aim of measuring and continuously monitoring student achievement to be twofold: first, to identify those aspects of each subject matter that are being well achieved and poorly achieved; and second, to identify if achievement levels are remaining constant over time or are improving or deteriorating. If achievement is found to be deteriorating in important aspects of a subject matter, action can be taken through curriculum development or through teacher training to improve them.
Monitoring implies not only the measurement of achievement as the output of a system, but also the evaluation of the data obtained (Pelgrum, 1990). Consequently, feedback adjustments can be made at the appropriate level of the education system with the intention to improve student achievement. Monitoring helps in this process by identifying discrepancies between system goals and system outputs. This definition of monitoring also applies to this study.

3.4 CURRICULUM

This study focuses on curriculum as a context for explaining the results of the observed phenomena in student achievement. The areas in which improvement in Slovene achievement might be desired will be sought within the mathematics curriculum. The curriculum is therefore seen as one of the most important vehicles through which improvement in student achievement can be initiated. The terminology and models used in the literature to describe the curriculum and its elements, varies. For the present study, those conceptualizations of the curriculum are important in which student achievement is seen as their integral part.

3.4.1 Conceptualizations of curriculum

The conceptualization of the curriculum, given by Goodlad, Klein and Tye (1979) (from Pelgrum, 1990; Vos, 2002) focuses on decisions at the societal, institutional, instructional, and individual levels. They identify the following appearances of the curriculum: (1) the ideal curriculum as were the intentions of curriculum developers, (2) the formal curriculum as is in the official documents, (3) the perceived curriculum as is seen by teachers, (4) the operational curriculum in the instructional process, and (5) the experienced curriculum as experienced by students. Van den Akker (1988, 1998; from Vos, 2002) adds to this system another appearance at student level: (6) the attained curriculum (the learning results of the students).

This curricular definition can be seen as corresponding with the conceptualization of the curriculum in the IEA studies at three levels controlled by different stakeholders in the education system: macro level – the intended curriculum (what the community values and what students should learn), meso – the implemented curriculum (what and how schools and teachers teach), and micro – the attained curriculum (what students learn) (Plomp, 1998). The first IEA study to highlight this distinction was The Second International Mathematics Study (SIMS, Travers & Westbury, 1989).

The *intended* curriculum is designed by the educational authorities and is reflected in formal publications adapted by the educational system, such as curriculum guidelines, syllabi or course activities and textbooks. The intended curriculum is designed to reflect the goals of education and at the same time to be taught in classrooms. Its delivery, or what is actually taught in the classroom, is the *implemented* curriculum. Although it is based on the intended curriculum, it may be influenced by local contextual variables, including the school's administration and climate, classroom characteristics, resources, and local or community interests and involvement.

The *attained* curriculum is generally described by the knowledge, skills and attitudes students have acquired during the course of their instruction. In the case of The Third International Mathematics and Science Study (TIMSS), the attained curriculum was defined as "*… the mathematics and science that students have learned and their attitudes towards the two subjects*" (Robitaille & Maxwell, 1996, p.37). Although student achievement is greatly affected by both, the implemented curriculum and societal context, it is also a function of factors under the control of the individual student, such as the student's effort, attitude, and personal interest.

The present study will adopt the conceptualization of the curriculum from the IEA studies. It is concerned with the intended curriculum for mathematics in Slovenia and with the attained curriculum, or more specifically, with the knowledge and skills students have learnt in mathematics. According to the above

conceptualizations, textbooks are an integral part of the formal curriculum and, thus, of the intended curriculum. However, as described in Chapter 2, in the reformed system in Slovenia teachers can choose from several approved textbooks. For practical reasons, textbooks are not considered in this study. The main focus therefore is on the curriculum guide that was prepared during the curriculum development process, which will serve as a source of information about the goals for mathematics achievement of Slovene students.
Although the implemented curriculum can be seen as an important mediating factor between the intended and the attained curriculum (e.g., Pelgrum, 1990), it is beyond the scope of this study to examine this appearance of the curriculum and its link to the intended and implemented curriculum.
Because of the importance of the concept of student achievement for this study, its conceptualization is discussed in more detail in the next section.

3.4.2 Student achievement

Central in the conceptual framework of this study is student achievement. The definition seems relatively straightforward: *Student achievement is what students know and can do*. Sometimes a more precise definition is useful: *Student achievement is what students know and can do at a certain age or in a certain grade in a certain subject area.* Although it may be affected by many other factors than participation in the educational system such as, the experiences in everyday life (Resnick, 1987; Saxe, 1988), there is no doubt that student achievement is seen as an important part of the outcomes of education (e.g., Bottani & Tuijnman, 1994; Kellaghan & Greaney, 2001).

Although this definition may seem straightforward, Cole (1990) gives insight into the characteristics and implications of two major general conceptions of educational achievement that have influenced thinking in education in past decades. The first is the conception of educational achievement as *basic skills and facts*, and the second achievement as *higher order skills and advanced knowledge*. Cole (1990) wrote that the concept of achievement as basic skills and facts emerged from the field of behavioral psychology which was closely connected with the learning of specific, discrete skills described as precise, well-delimited behaviors (see also Popham, 1993). The consequences of such a conception, according to Cole (1990), were that the views that important school skills can and should be listed as discrete pieces of desired competence, have immediate behavioral outcomes that can be segmented and individually tested, and be clearly linked to a specific school curriculum were promoted.

It should be noted that the name *basic skills and facts* is a simplification of what Cole (1990) describes under this name. For example, basic mathematics skills may be seen as the ability to add, subtract, multiply, and divide in response to explicit instruction to do so. However, being able to use those operations in solving problems including recognizing the operations needed, may also be seen as part of this basic skills and facts conceptualization. By Cole's (1990) characterization, the predominant conception of achievement as basic skills and facts is one "*in which achievement is represented by recall of separate facts or simple demonstrations of discrete skills that can be taught, learned and assessed in direct forms and in short periods of time*" (p.3).

The concept of achievement is therefore closely related to the way it is measured (see e.g., Dochy & Moerkerke, 1997). Here, these issues are touched upon only as part of the discussion on the conceptualization of student achievement. Measurement issues are further discussed in more detail later in this chapter. Cole (1990) claims that the above conception of achievement was promoted by, and in turn promoted, developments in educational measurement, which turned toward testing smaller, more curriculum-specific skills. Similarly, Popham (1993) describes that, in the 1960s and 1970s in the United States, the attention was put on measurable instructional objectives which had become known as "behavioral objectives".

In earlier national assessments conducted mostly in the United States and in earlier international assessments, the attainment of these objectives was measured mostly using multiple-choice item format (Hambleton & Sireci, 1997; Mislevy, 1995). Hambleton and Sireci wrote that the predominant use of multiple-choice item format was influenced by its cost-effectiveness, which is shown in the facts that it enabled assessment of a wide array of content in a relatively short amount of time, that it can easily be objectively scored and, probably not unimportant, that in the 1950s an optical scanner was invented. Such a scanner could mechanically score items where examinees darkened ovals on an answer sheet that corresponded to their choice from a list of options.

In addition to measurement issues, there are other advantages of such a conception of achievement. Clearly explicated instructional objectives help clarify the nature of one's instructional aspirations (Popham, 1993). The conception of achievement by focusing on the desired outcome suggests specific instructional actions that should be followed in order to produce desired learning. In this way, the link between the desired achievement and curriculum is made transparent.

In the 1990s, the conception of achievement as basic skills and facts has received much criticism (Broadfoot 1994, 2001; Herman, 1997; Ivic, 1994). First of all, the limitations of multiple-choice item format have become to receive more attention than its advantages. Multiple-choice items have been criticized as placing stringent limits on the types of proficiencies that can be assessed (Hambleton & Sireci, 1997). Popham (1993) claims that, in cognitive achievement tests, many multiple-choice items are easier than comparable constructed response items, even after correction for guessing. He also observed that the shortcoming of specifying desired outcomes as behavioral objectives was that they resulted in focusing on increasingly smaller and more specific segments of learner postinstruction behavior, while the most important goals of education do not lend themselves readily to a behavioral formulation.

At the more conceptual level, the main criticisms concerned the view of achievement as a set of 'discrete' competencies from which knowledge is formed in a linear-additive process (Ivic, 1994). Such a conception does not allow for generalization on students' problem solving abilities, critical and creative thinking, expertise, associative and interpretative capabilities, as well as the formation of complex social competencies such as cooperation, communication and team spirit.

These criticisms are addressed by the second conception of student achievement described by Cole (1990) as higher order skills and advanced knowledge, which includes terms as those mentioned above, namely complex knowledge, critical thinking, problem solving, understanding and expertise. Explicating these higher order skills and knowledge is difficult because the conception is more diffuse and arising out of several different contexts. However, in the past decades main research stems from cognitive psychology and cognitive science approach to the study of learning and instruction (Walberg & Haertel, 1994). With reference to Glaser (1984, 1988) and Rabinowitz and Glaser (1985), Cole (1990) explains that expert knowledge demonstrates a coherence of what is known (relatedness), knowledge of domain-specific patterns or principles, use of patterns and principles in problem solving, recognition of situations and conditions for using knowledge, highly efficient performance, and use of self-regulating skills (metacognitive strategies, such as forward reasoning).

Similarly, Herman (1997, with reference to Glaser & Silver, 1994) describes that to know something is not just to have received information, but also, to have interpreted it and related it to other knowledge one already has. It is important to know not just how to perform, but also when to perform and how to adapt that performance to new situations. Thus, the presence or absence of discrete bits of information is not of primary importance in the assessment of meaningful learning. Rather what is highly valued are the so-called metacognitive skills and learning, namely how and whether students organize, structure, and use that information in context to solve problems. In addition to acquisition of knowledge and skills, the dispositions to use the skills and strategies, the knowledge of when to apply them, and the ability to learn from their experiences need to be acquired.

An example of higher order skills and advanced knowledge conception of achievement can be presented in contrast to the former, basic skills and facts conception. While students may solve problems correctly using arithmetic rules they, when probed, often show lack of understanding of operations they used or principles involved in using them. They often fail to use the skills and facts they acquired in new problem situations in which they would be relevant. Walberg and Haertel (1994, with reference to Larkin & Rief, 1979) give examples of differences in strategies used by novice and expert problem solvers in physics. Novices quickly resorted to converting the physics problem into a series of equations and immediately tried to manipulate the equations, while experts examined the terms of the problem more thoroughly and clarified that their position was sound before introducing the equations. More examples of individual manifestations of this conception of achievement are given by Cole (1990).

Both conceptions have been emphasized in describing what students should learn. But they contrast markedly in implications for practice and research and, consequently, for the measurement of achievement. Broadfoot (1994, with reference to Resnick & Resnick, 1989) argues that complex competencies such as "thinking" can not be assessed in the decomposed and decontextualized approach assumed in standardized testing since the "organic whole" of such a task is, like teaching, more than a sum of its various elements. Broadfoot further argues that short and superficial questions, used for example in reading comprehension can not test, or give a student an opportunity to demonstrate, higher order thinking skills.
Herman (1997) claims that measurement of higher order skills and advanced knowledge has been incorporated in the alternative assessments, also called authentic assessment or performance assessment, as an alternative to conventional multiple-choice testing that was popular in many assessments in the United States in past decades. Alternative assessments range from portfolios of student work or extended projects that may consume an entire school year, to open ended questions also called constructed response questions which resemble multiple-choice test items where the response options have been omitted. By Herman's definition, alternative assessment typically requires students to actively accomplish complex and significant tasks, while bringing to bear knowledge, recent learning, and relevant skills to solve realistic or authentic problems.
In the same line, Walberg and Haertel (1994) emphasize that different types of items and test format beyond conventional paper-pencil tests relying on

unconstructed responses are desirable for determining how readily students transfer learning of novel tasks and what types of solution strategies they employ. In addition, diverse methods of scoring, beyond number correct and response latency, are needed to diagnose learning difficulties.

In response to the criticisms of the basic skills and fact conception of achievement measured mainly by multiple-choice items and in an attempt to include higher order skills and advanced knowledge into the conceptualization of achievement, many national (e.g., NAEP) and international assessments (e.g., IEA and PISA) have included in the measurement instruments various types of constructed response item formats. Examples range from items requiring short-answer responses through items that need to be answered by several sentences, called extended response items. Such item formats have been in use in instructional settings for a long time, but according to Hambleton and Sireci (1997), it is only recently that they are being used as part of large-scale assessments. They are seen as facilitating a more comprehensive and realistic assessment of knowledge and skills.
Although the progress in introducing more authentic forms of assessment in the international studies is recognized, many authors still see a number of obstacles in measuring higher order skills and advanced knowledge in these studies. In addition to the problem of the cost of such assessments, Broadfoot (1994) and Walberg and Haertel (1994) argue that the limitations of existing assessment techniques together with political expedience can (still) lead to employing undesirable conceptions of achievement. At the same time they caution against simply applying the new forms of assessment because their statistical characteristics are not yet well understood. The use of many different formats of assessment might lead to the need of setting new standards and rules of evidence required to establish the quality of measurement. So far, it is difficult to see how the more complex learning outcomes might be incorporated into large-scale assessments.

In 1990, Cole concluded that there is a lack of an overarching framework for understanding the two separate conceptions and their relations to each other. Little in the skills and facts conception points to anything beyond. At the same time, the higher order skills and advanced knowledge views often acknowledge the existence of more elementary skills and knowledge, but generally offer little sense of their importance or how they are used in the higher order activities.

The synthesis of the two conceptions can be seen from the conceptualization constructed by the committee of the United States Academy of Science, in 1998 (Kilpatrick, 2001). Following a previously conducted study on reading, the committee started a study of mathematics learning that would synthesize research on pre-kindergarten through eighth-grade mathematics learning, provide research-based recommendations for teaching, teacher education and curriculum for improving student learning and to identify areas where research is needed, and give advice and guidance to educators, researchers, publishers, policy makers, and parents. In its work, the committee also needed to address the definition of "successful mathematics learning". Among the terms "mathematical literacy", "numeracy", "mastery of mathematics", and "mathematical competence" the committee settled on the term "mathematical proficiency", defining it in the terms of the following five strands 'to be interwoven in concert' (p.106): (a) *conceptual understanding*, which refers to student's comprehension of mathematical concepts, operations, and relations; (b) *procedural fluency*, or the student's skill in carrying out mathematical procedures flexibly, accurately, efficiently, and appropriately; (c) *strategic competence*, the student's ability to formulate, represent, and solve mathematical problems; (d) *adaptive reasoning*, the capacity for logical thought and for reflection on, explanation of , and justification of mathematical arguments; and (e) *productive disposition*, which includes the student's habitual inclination to see mathematics as a sensible, useful, and worthwhile subject to be learned, coupled with a belief in the value of diligent work and in one's own efficacy as a doer of mathematics (p.107).

What is, therefore, between all these views, the conceptualization of achievement used in this study? The approach to settle the concept of achievement is guided by the purposes of this study. As explained in Chapter 1, the purpose of investigations of student achievement in the present study is to yield information that is useful for mathematics curriculum development and implementation in Slovenia based on conclusions from comparisons in an international context. As also previously explained, the major new feature in the reformed mathematics curriculum in Slovenia are attainment targets for each grade. These attainment targets are written in terms of specific performance behaviors that can each be linked to specific curriculum objectives. Although competencies, such as critical thinking and communication skills are recognized to be an important part of the outcomes of education in Slovenia and elsewhere, the conception of achievement as purely higher-order skills and advanced knowledge without a clear link to the curriculum and its basic skills and facts elements is considered too complex for this study. Furthermore, the specificity of behavioral objectives and attainment targets in the Slovene curriculum and the transparency of their link to the individual items in the instruments for measuring achievement in this study is considered highly important for addressing the research problem. This led to the decision to adopt the conception of student achievement as the sum of 'discrete' behavioral elements for this study.

3.5 ASSESSMENT

The collection and presentation of information about educational outcomes is called *assessment* (Husén & Tuijnman, 1994). There are different approaches to assessment, each serving particular purposes and implying a certain style of measurement. In Chapter 1, characteristics of national and international assessments of importance to present the problem in this study were discussed. In an assessment, an *indicator* of student achievement is constructed. The link between this indicator and the concept of student achievement was discussed in the previous section. Several other issues in measuring student achievement are discussed below.

3.5.1 Coverage

In education, a measurement instrument for student outcomes is generally a test, which consists of one or more test items relevant to a certain domain. A test usually contains a limited number of items from the domain. Therefore *coverage of the curriculum by the test* is a relevant issue. It is related to the concept of validity (e.g., Crooks et al., 1996; Mislevy et al., 2003; Wolf, 1994, 1998). Most often, the question to be answered is whether scores on an assessment test can be used to make inferences about achievement of students at a particular age or grade level in a country. This question can be addressed at the level of the intended, as well as the implemented curriculum.

At the level of the intended curriculum, coverage by the test means to what extent the objectives in the curriculum are covered by test items. This was termed 'test-coverage index' by Wolf (1998, with reference to Rosier & Keeves, 1991). It can be measured using judgments of experts. This then provides information to make judgments about the validity of the test. It shows to what extent the items in the test could be interpreted as representing the intended curriculum and, thus, enabling the link between the intended and the attained curriculum.

At the level of the implemented curriculum the concept of coverage by the test represents the extent to which test items cover what teachers have actually taught. It can be measured by collecting data from teachers. In this case, the coverage indicates to what degree the items in the test could be interpreted as representing the implementing curriculum and, thus, enabling the link between this appearance of the curriculum and the attained curriculum.

In the present study, the coverage of the intended curriculum by the test will be examined. The procedure that will be used to estimate this coverage is described in Chapter 5.

There is also a 'reversed' issue of the coverage. In the case of international assessments, achievement tests may not suit the curriculum of a particular country. A test may include knowledge and skills not included in the curricula of several participating countries. In such cases, *the coverage of the test by the curriculum* is relevant. For example in TIMSS, the estimate of the coverage of the test by the intended curriculum was termed the 'test curriculum matching index' (Beaton et al., 1996; Mullis et al., 2000). It was measured as the percentage of items judged by country's experts whether, according to the curriculum, students should master the skills tested by each test item. The coverage of the test by the implemented curriculum also received attention in international assessments. It is generally associated with the term 'opportunity to learn' (OTL) and is measured through instruments in which teachers indicate whether a certain topic has been taught prior to test administration. These instruments were either topic based (Mullis et al., 2000) or item based (Beaton et al., 1996; Travers & Westbury, 1989). Studies examining the appropriateness of TIMSS achievement tests in the Netherlands were conducted by Kuiper et al. (1998) and Vos (2002).
Using the international achievement test in this study, the concept of coverage of the test by the intended curriculum is important. Procedures for estimating this coverage in Slovenia will be described in Chapter 5.

3.5.2 Achievement scales

In educational assessments, the individual student's achievement in the domain is estimated on the basis of his or her responses to the items in the test. This estimation is based on an assumption that achievement can be represented on a continuous dimension and on a unidimensional scale. The construction of a scale enables representation of unobservable characteristics, such as achievement in particular domain, in a quantitative manner. The international studies that have tried to measure educational performance across the world, as TIMSS, have reported their results in some type of scaling metric particular to the assessment instruments used. As already mentioned, important considerations in construction of these scales are reliability and validity of inferences that can be made from them. Mislevy et al. (2003) presented a general framework of the assessment design models relating what is observed to what is inferred about student achievement which helps to address issues of reliability and validity, as well as others. However, these concepts will not be discussed any further at this point. Validity and reliability in TIMSS will be discussed in Chapter 4 and in this study in Chapter 5.

In general, two models can be distinguished for constructing achievement scales, the classical test theory model and item response theory (IRT) model. Gonzalez and Beaton (1994) discuss that under the assumption of classical test theory the test items may be considered a random sample from a theoretical domain. The overall score on the test or the proportion correct is considered as an estimate of student's performance if all of the items in the domain had been administered. The test score is assumed to be the best available estimate of the true score of the student. The standard error of this measure represents the variation of these scores obtained by the student if numerous random samples of the same number of test items from the domain had been administered.
Under the assumptions of IRT, the interpretation of an outcome measure is different. The items are assumed to be located along a difficulty scale, and the score of the student on the set of items administered is considered to be the best indicator of where the student is located along the scale. This location is called the ability of the student in the domain of interest. It is assumed to be the same regardless of the difficulty of the items responded to by the student. The error of measurement is the variation in the estimated students' ability if numerous sets of items were administered to the same student and ability estimated were to be obtained with each set of items.

As explained in Chapter 1, there may be several purposes of constructing an achievement scale. When an achievement scale is constructed with a purpose to describe the performance of individual students in terms of how they do in relation to other students, the scale is called a normative scale. The construction of a normative scale is relatively straightforward. Regardless of how a test was constructed or what it measures, it is possible to determine a student's relative standing along the achievement scale. As described in Chapter 1, these are called norm-referenced interpretations (Hambleton & Sireci, 1997). There are many normative metrics commonly used to describe student test performance; for example, percentile scores, grade-equivalent scores, and standard scores. In international studies countries are usually ordered according to their mean scores on a common scale. The particular country's performance is then interpreted as better than, worse than or the same as that of other countries.

However, unless looking at individual items or very small groups of items measuring similar knowledge and skills, normative scales generally are not used to convey concrete information about what students know and can do. In international assessment, information that one country performed better than another on a test covering a wide rage of content does not reveal what kind of tasks students from the higher performing country can successfully complete in contrast to the students from the lower performing country. As also discussed in Chapter 1, interpretations of the results providing information on what students know and can do with respect to the content and processes assessed are called criterion-referenced interpretations (Hambleton & Sireci, 1997). They can be based on a priori or on a posteriori criteria (see subsection 3.2.2). A priori criteria are usually set by a panel of experts. TIMSS scale anchoring studies (Kelly, 1999; Mullis et al., 2000) are examples of criterion-referenced interpretations of assessment results based on a posteriori criteria.

Both models for constructing achievement scales mentioned previously, classical test theory model and IRT model, can be used for the purposes of norm-referenced and criterion-referenced interpretations. However, these interpretations differ between the two models. Because the resulting scale is independent of the sample of items and the sample of students, Gonzalez and Beaton (1994) favor the use of IRT when constructing an achievement scale (see also Hambleton & Rogers, 1989). In TIMSS, IRT scales have been used to summarize the performance on a test and compare achievements between countries. They were also used in the TIMSS scale-anchoring studies (Kelly, 1999; Kelly et al., 2000; Mullis et al., 2000). However, for practical reasons it was decided to use the percent correct scales from the classical test theory model in this study. In support to this decision is the fact, that the percent correct scales are easier to understand for the lay public (Kelly, 1999).

3.5.3 Content areas and cognitive levels

Intended curricula are usually structured in *content areas* (Schmidt et al., 1997), as is done in the Slovene curriculum (*Curriculum guide*, 1984, 2002). Content areas are named using generic terms such as 'measurement' and 'probability'. In reporting assessment information it is relevant to examine variability of achievement across these areas. As explained in Chapter 1, providing such information is also important for this study. The content areas that will be used will be defined according to the Slovene mathematics curriculum. Further description of these content areas will be given in Chapter 5.

In the process of instrument development, a second dimension is often considered, termed 'objectives' (Wolf, 1998) or 'performance expectations' (Robitaille et al., 1993). This dimension describes the kinds of performances students are expected to demonstrate in the various content areas of the domain. These classifications are based on work by Bloom (1956). They are usually considered to be hierarchical,

from less complex to more complex cognitive processes (Mullis et al., 2003). For example, in TIMSS 1995 and TIMSS 1999 performance expectation categories were defined as 'knowing', 'using routine procedures', 'using complex procedures', 'investigating and solving problems', and 'communicating and reasoning'.
Cognitive complexity should not be confused with item difficulty. For nearly all cognitive skills, it is possible to create relatively easy items, as well as very challenging items (Mullis et al., 2003).

Although in the Slovene reformed curriculum cognitive levels are not explicitly addressed, their importance in the process of development of national examinations is emphasized (Rutar Ilc, 2003). These levels will also be considered in this study. Variability of achievement of Slovene student across the cognitive categories will be examined using the TIMSS structure. To avoid possible confusion with expected achievement on an item, instead of 'performance expectation' the terms *cognitive levels, levels of cognitive requirements* or *cognitive categories* will be used.

3.6 CORRESPONDENCE OF ACHIEVEMENT AND TRENDS

Within this section, concepts will be introduced that are important for utilization of assessment information in the present study. How the correspondence of achievement with predefined reference points can be defined and measured, and how strengths and weaknesses in student achievement can be identified will be discussed. Finally, the concept of trends and, within it, the concepts of identical and cloned items will also be addressed.

3.6.1 Measuring correspondence of achievement with predefined reference points

As previously discussed, in order to determine whether improvement measures in the education system are needed, assessment information must be interpreted in terms of the *correspondence* of achievement with predefined reference points. Measurement of correspondence implies that there are two or several comparable objects between which this correspondence is measured. As was also discussed before, assessment information on student achievement is usually given in the form of scores on a numeric scale. In order to determine the correspondence of achievement to the selected reference points, the reference points need to be in the form that enable these comparisons.

In the case of this study, by taking an international achievement test administered in Slovenia and in the reference countries as a measurement instrument for achievement, the second reference point in this study (achievements of students in these countries) is in the same metric as the target measure of achievement in Slovenia. However, the reference point on the basis of attainment targets is not as straightforwardly comparable to the achievement measure. A procedure is needed to transform the attainment targets onto a numeric scale comparable to the target measure. In terms of discussion in section 3.2, there is a distinction between the standard (or the attainment target) and its cut score (Gonzalez & Beaton, 1994). In this study, the cut scores for the attainment targets in the Slovene curriculum need to be determined. The procedure used to obtain these cut scores will be described in Chapter 5.

As previously mentioned, the interpretations of assessment data may serve different information needs of different users and are generally distinguished as norm-referenced and criterion-referenced interpretations. They are usually obtained by employing different procedures for analyzing assessment data. For this study, the distinction between R-techniques and Q-techniques is applicable (Nunnally, 1967). In R-techniques, one considers how well respondents are doing (R for respondent, for example student) while Q-techniques analyze how well certain content elements are doing (Q for question or item through which this content is operationalized).
These techniques will be called *student based* and *item based analyses* in this study. Student based analysis is most commonly used to describe student achievement in a domain or in several subdomains, such as content areas and cognitive levels. Item based analysis can be used to reveal characteristics of the test or individual items such as item difficulty (item percent correct) or item discrimination (see e.g., Crocker & Algina, 1986). They can also be used to reveal the amount of *compensation* between low scores and high scores on items in the test. For example, while student achievement in a domain or in subdomains may be satisfactory, this may be due to a relatively few high achieving items while on most items achievement may be lower than desired. Detailed studies of assessment information have been emphasized for example by Mislevy (1995) and Schmidt et al. (1998). Pelgrum et al. (1986) used item based approach on SIMS data to examine similarities and differences between countries taking into account OTL information. Such analyses (and even analyses of selection of wrong answers in multiple-choice items) may reveal important information for curriculum developers and didactical experts.

In the present study, both approaches to analysis of assessment data will be used. Student based analyses will be used to describe correspondence of Slovene achievement to the reference points and item based analyses will be used to describe correspondence of observed difficulties of individual items with the reference points. Through this, item based analysis will reveal the numbers of items on which the correspondence or non-correspondence of student achievement with the reference points occurred.

3.6.2 Identification of strengths and weaknesses

Conceptually, a *strength* is something that is above the 'expected' level or the level that 'should' have been attained. A *weakness* is something that is below that level. The level that should have been attained is termed the reference point in this study and, as explained, two reference points will be selected. The notions of strengths and weaknesses imply that there is a hierarchy in the target and reference measures from which they are to be determined. In this study, this hierarchy will be provided by having numerical measures for both. Procedures that will be used for identification of strengths and weaknesses are described in Chapter 5.

3.6.3 Measuring trends in student achievement

Trends generally mean comparison of measures over a period of time. In a longitudinal study, trends are measured for the same units of analysis (Keeves, 1992). Trends can also be measured at the level of a whole population. In this case the population definition and major parts of the instruments remain the same, while the units of analysis in this population change. The latter definition of trends applies to the TIMSS data and also to this study.

When assessment instruments for measuring trends are not exactly equal in the successive measurements, an issue of comparability of these instruments for making inferences about trends in student achievement emerges. This issue is particularly important when using item percent correct technology, as in this study. For example, if items in the second measurement are easier, they yield an impression that student achievement has increased. If items in the second measurement are more difficult, student achievement appears to decrease. This issue will be addressed in this study by defining sets of *identical* and *cloned* (Vos, 2002) items in achievement instruments and carrying out analyses separately on these two sets.

An item is defined as cloned from an original item, if it is, although not identical, very similar to the original item in terms of content, format, and difficulty. In a cloned item, only a detail is altered. It is assumed that most students being able to solve the first item would also be able to solve the second item, and vice versa. In this way, scores on cloned items in the first measurement could be considered comparable to scores in the second measurement.

However, changes introduced in items may influence students' percent correct scores on these items even though student achievement has not changed. Comparability of these items therefore needs to be checked through scores on these items as well. If scores on both, identical and cloned items do not change, the cloned items can be considered comparable. In all other cases, additional detailed analyses of item content may be needed to distinguish to what extent the changes in scores have been caused by the changes in items.

In the present study both sets of items will be used in case the results show no change in students' scores on both sets of items. This will increase the coverage of the domain. In all other cases, only identical items will be used. More detailed procedures will be described in Chapter 5.

CHAPTER 4
TIMSS

The present study of mathematics achievement of Slovene students will employ the international data from the Third International Mathematics and Science Study (TIMSS). Up to date, three TIMSS surveys have been carried out, in 1995, 1999, and 2003[1]. The TIMSS 1995 and TIMSS 1999 assessments were conducted following the same sampling procedures and measuring identical constructs. The elements of the TIMSS design and methodology that are described in this chapter include the conceptual framework, research questions, target populations and sampling, instrument development, and data collection and scaling procedures. The most important feature of TIMSS for the present study is that it is based on the curricula of participating countries. This enabled a link between student achievement (the attained curriculum) and the curriculum as is prescribed in the official documents (the intended curriculum). Through this link, areas in student achievement in which improvements might be desired can be identified and possible remedial actions developed.

In section 4.1, the TIMSS conceptual framework and research questions are presented. Target populations and sampling design are outlined in section 4.2, followed in section 4.3 by a description of how the three levels of the curriculum were measured in TIMSS. The procedures employed in the TIMSS data collection are presented in section 4.4, while section 4.5 briefly describes the data analysis and scaling methodology used in the study. In section 4.6 the validity and reliability of TIMSS achievement instruments are discussed and section 4.7 provides an account of the TIMSS data collection procedures in Slovenia.

1 TIMSS 2003 was renamed into the Trends in International Mathematics and Science Study. These data are not yet publicly available and will not be used in this study.

4.1 Generic conceptual framework and research questions for TIMSS

TIMSS is an assessment of student achievement in two core school subjects with an intention to learn more about the nature and the extent of student achievement in mathematics and science and the context in which it occurs (Robitaille & Robeck, 1996). The conceptual framework for TIMSS was derived from previous studies conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA). The earlier IEA studies recognized the centrality of the notion of curriculum in any examination of the teaching and learning of subject matter in schools (IEA, 1998). These studies tried to explain the observed differences in achievement of students between and within countries in the light of the prescribed curricula, processes in schools and other, predominantly background variables.

TIMSS also determined the curriculum as the focus of investigation of the sources of differences in student achievement between, as well as, within countries. A generic conceptual framework for TIMSS was developed over a series of IEA studies, and specifically from the development of the conceptual framework for Second International Mathematics Study (SIMS, Travers & Westbury, 1989). As described in Chapter 3, this framework incorporates three appearances of the curriculum: the curriculum as mandated at system level (the intended curriculum), the curriculum as taught by teachers in classrooms (the implemented curriculum), and the curriculum as learned by students (the attained curriculum). It is summarized in Figure 4.1. Conceptualizing the curriculum in this way made it possible for TIMSS to provide information that is relevant to in-depth discussions of reasons for varied student achievement, as well as, other issues in education such as attitudes of students and satisfaction of teachers.

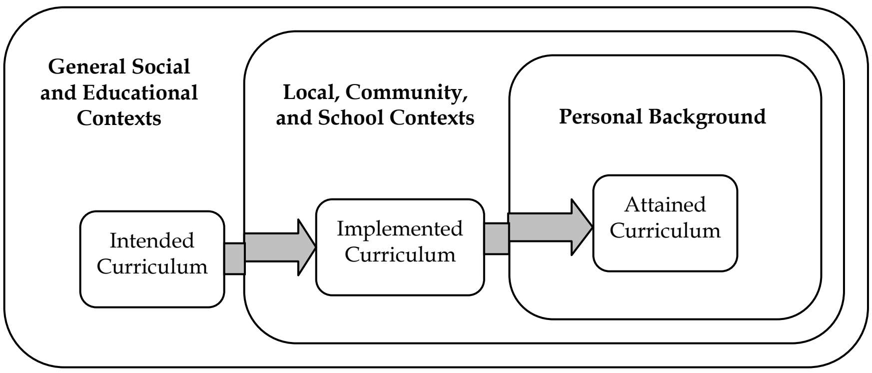


Figure 4.1. Conceptual framework for IEA studies

Note: Adapted from Robitaille and Maxwell (1996, p.37)

There were four general research questions based on the conceptual framework presented in Figure 3.1 (Robitaille & Garden, 1996*)*:
Research question 1 (the intended curriculum): How do countries vary in the intended learning goals for mathematics and science; and what characteristics of educational systems, schools, and students influence the development of these goals?
Research question 2 (the 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?
Research question 3 (the attained curriculum): What mathematics and science concepts, processes, and attitudes have student learned; and what factors are linked to students' opportunity to learn?
Research question 4 (relationships between curricula and social educational contexts): How are the intended, the implemented, and the attained curriculum related with respect to the contexts of educational, the arrangements for teaching and learning, and the outcomes of the educational process?

The first three research questions for TIMSS can be addressed by providing descriptive information on the variables measuring corresponding components of the curriculum. The fourth research question focuses on the information that could help researchers and educators to understand how their educational systems function and to explain the differences in student achievement across and within countries.

4.2 TARGET POPULATIONS AND SAMPLING

The research design of TIMSS focused on the issues that arose from the postulated research questions. Basically two research designs were considered; the cross-sectional design where students are assessed at one point in time, and the longitudinal design in which the same students are assessed at two or more points in time (Robitaille & Robeck, 1996). Since any study needs to provide a balance between the complexity required to respond to specific questions and the need for simplicity, timeliness, and cost-effectiveness, most studies are cross-sectional (Beaton et al., 1999). Because of their complexity, international comparative longitudinal studies are rare. It is not only difficult to retain contact with students over time, but expected high attrition rates have to be reflected in the design, and this adds to the cost of the studies.

Therefore, the basic design of TIMSS was cross-sectional. In addition to the characterization of the achievement of students at each sampled grade level, such a design allowed for comparisons between variables across populations. However, the TIMSS 1995 survey also included a compromise between a cross-sectional and a longitudinal survey design, as it assessed adjacent grades for the populations of primary and lower secondary education (see below). This allowed more information about achievement to be obtained (across several grades) and an analysis of variation in cumulative schooling effects.

4.2.1 Populations

The design of TIMSS was intended to facilitate the investigation of mathematics and science curricula from a variety of perspectives. In TIMSS 1995, three populations were sampled, and these corresponded roughly to the end of primary education, the end of compulsory education and non-departmentalized education in mathematics or science in many countries, and the end of secondary schooling (Robitaille & Garden, 1996). Subsequent TIMSS surveys assessed subsets of the original set of populations.

In TIMSS 1995, the intended target populations of students assessed were:

- *Population 1 consisted of students in the pair of adjacent grades that contained the most students who were 9 years old at the time of testing;*
- *Population 2 consisted of students in the pair of adjacent grades that contained the most students who were 13 years old at the time of testing; and*
- *Population 3 consisted of students in the last year of secondary school, regardless of the type of program in which they were enrolled.*

In the majority of the countries, the two grades selected for Population 1 were the equivalent of grades three and four and for Population 2 the modal grade equivalents were grades seven and eight (IEA, 1997; Gonzalez & Smith, Eds., 1997). All countries were required to participate in the Population 2 study; participation at the other levels was optional. The replication in 1999, also named TIMSS-R, focused on the upper grade of Population 2 (IEA, 2001; Gonzalez & Miles, 2001). This was grade eight in almost all countries. Of the 38 countries that participated in TIMSS-R, 26 had also participated in the 1995 study.

The present study focuses on Population 2. There are several reasons for this. One is that in Slovenia as well as in some other European countries grade eight currently is the final year of compulsory education (in the non-reformed system in Slovenia). Student achievements in this grade therefore represent an important source of information on the outcomes of schooling and the effectiveness of the education system. Another reason to focus on this population is that the existence of two data collections enables a research design of trend studies. Student achievements of eighth-grade students can be studied in a longitudinal perspective in order to obtain an understanding of developments over time.

4.2.2 Sampling

IEA studies have traditionally employed a research design based on the use of intact classes of students. This means that the studies focus on classrooms at a particular grade level, as opposed to focusing on students of a particular age. This is important for studies investigating the linkages among the intended, implemented, and attained curricula, as is also the case in the present study. Such studies are more likely to contribute to understanding of what kinds of curricula and instructional practices are associated with the highest levels of student attainment (Robitaille & Garden, 1996).

Because random sampling of students in a country is practically impossible, samples in large-scale studies commonly involve two or more stages. The following description is based on Foy et al. (1996). The basic sample design used in TIMSS was a two-stage stratified cluster design. The first stage consisted of a sample of schools; the second stage consisted of a sample of intact mathematics classrooms from each eligible target grade in the sampled schools. The design required schools to be sampled using a probability proportional to size; while classrooms were sampled with equal probabilities. In some larger countries, the initial sampling units were counties or metropolitan areas, and schools were selected from these. In TIMSS, the simple equivalent sample size (i.e., if students had been sampled randomly from the total population rather than from schools) was required not to be less than 400 students. TIMSS sampling approach was designed to yield 150 schools for each of Populations 1 and 2, and one classroom per each grade. Replacement schools were also drawn in a sample to enable schools to be substituted in the event that a school in the primary sample declined to participate. Two replacement schools were drawn for each sampled school.

The national sampling designs were reviewed by the sampling referee. The participating countries were allowed to adapt the TIMSS sample design for their educational system, using more sophisticated designs and procedures than the base design provided, however, these solutions had to be approved and monitored by the international project management. To be acceptable for TIMSS, national sample designs had to result in probability samples, which gave unbiased weighted estimates of population parameters, and for which estimates of sampling variance could be computed.

4.3 MEASURING THE THREE APPEARANCES OF THE CURRICULUM IN TIMSS

In section 4.1, the IEA conceptual framework for measuring outcomes of education was described. This conceptual framework implies that the outcomes in TIMSS were measured at three levels of the curriculum; the intended, the implemented and the attained curriculum. These three appearances of the curriculum were measured using the questionnaires for experts, school principles, teachers, and students. Student achievement was measured through internationally designed achievement tests. These instruments are further described below with emphasis given on the development of achievement tests.

4.3.1 TIMSS questionnaires

In TIMSS, the data on the intended curriculum were gathered through questionnaires that were administered at national level focusing on organizational structure, courses, demographics, and teacher credentials, as well as questionnaires gathering information about national-level curriculum plans, reforms, issues, and policies with respect to mathematics curricula.

To study the implemented curriculum, TIMSS school and teacher questionnaires captured several types of information about schools and teachers which can be expected to influence, or account for how each teacher implements the curriculum. Key questions sought information specifically related to the selection and the use of teaching resources by the teachers for their classes, their perceptions of the characteristics of these classes, and the perceived impediments to their being able to teach more effectively. The most direct measures of the curriculum students were exposed to in the sampled classrooms were derived from a series of questions relating to the extent of content coverage, teaching practices, and methods of assessment.

The attained curriculum in TIMSS consisted of the mathematics and science that students have acquired in the course of their studies. The attained curriculum as the outcome of education was measured by achievement tests and questionnaires presented to students. The student background questionnaires were designed to examine students' perceptions of the importance of success in mathematics, and how successful they were relative to other students and whether they liked these subjects. Development of achievement instruments is described in more detail in the following subsection.

4.3.2 Curriculum frameworks and development of achievement instruments

To define the range of mathematics topics to be addressed in TIMSS, the TIMSS curriculum frameworks (Robitaille et al., 1993) were developed for TIMSS 1995 by groups of mathematics educators with input from the participating countries. For subsequent TIMSS surveys, these frameworks were adapted to reflect the curricular developments in the participating countries and in the world in general (Mullis et al., 2003). The TIMSS 1995 frameworks included the entire span of curricula from the beginning of schooling through the completion of secondary school. They were organized in a system of categories by which the contents of textbooks and curriculum guides were coded and analyzed, and on which the assessment instruments were based.

For the purpose of measuring student achievement two dimensions, *content* and *performance expectations*, were defined in the TIMSS frameworks as also described in Chapter 3. The purpose of performance expectations is to describe the many kinds of performance or behavior that a given test item or curriculum unit might elicit from students; such as, understanding, theorizing, analyzing, and problems solving (Robitaille et al., 1993). Another dimension of *perspectives* was added to refer to the nature of the presentation of the content in the curriculum materials and was used for the TIMSS curriculum analysis (Schmidt et al., 1997). This analysis led to a blueprint for the tests that covered most of the curricula of participating countries. In the process of curriculum development each item was assigned a content parameter and a parameter for performance expectations. For reporting TIMSS 1999 results, categories 'fractions and number sense', 'measurement', 'geometry', 'data representation, analysis and probability', and 'algebra' were used for content dimension, and 'knowing', 'using routine procedures', 'using complex procedures', 'investigating and problem solving', and 'communicating and reasoning' for the dimension of performance expectations.[2]

Tables 4.1 and 4.2 present the distribution of mathematics TIMSS 1995 and TIMSS 1999 Population 2 items and score points by five mathematics content reporting categories and by five performance categories used in TIMSS international reports.

[2] TIMSS 1995 reporting categories were different to some extent and were revised in TIMSS 1999. Because the present study uses TIMSS 1999 data for its main part, the categories from this study are presented.

The largest number of items in both tests covered the contents of 'fractions and number sense', while the items were approximately evenly distributed across the other four categories.
About one-third of the TIMSS 1995 mathematics items were kept secure to measure trends over time. Using terminology from Chapter 3, these are called 'identical' items. A considerable number of items that were released were replaced in TIMSS 1999 mainly by items of similar content, format, and difficulty (Mullis et al., 2000). These are called 'cloned' items in the present study. It was assumed that students being able to answer the original item would also be able to answer the clone, and vice versa. In total, the mathematics test for TIMSS 1995 Population 2 included 155 items and the mathematics test for TIMSS 1999 Population 2 included 162 items. There were 152 compatible items (cloned and identical) between the two tests with 157 score points in total. Additionally, some items that constituted the same tasks were merged into one item with several score levels per task in the process of construction of the international database.

In both data collections, items were allocated to mutually exclusive clusters, labeled A to Z (Adams & Gonzalez, 1996). The clusters were systematically assigned to eight test booklets so that one cluster appeared in all test booklets, some clusters appeared in several test booklets, and some clusters appeared in one test booklet. The test booklets were systematically distributed to students in sampled classes and each student was presented one booklet. The existence of a core cluster made it possible to test the equivalence of the samples of students that had been assigned to each booklet. This so-called rotated test design reduced testing time per student while allowing data to be collected on a large number of items covering major aspects of curricula. The time allocated to each booklet for Population 2 was 90 minutes.

Table 4.1 Distribution of TIMSS items across content categories

	TIMSS 1999			TIMSS 1995			Identical Items			Cloned Items		
	% of items	Number of items	Number of score points	% of items	Number of items	Number of score points	% of items	Number of items	Number of score points	% of items	Number of items	Number of score points
Fractions and number sense	38	61	62	39	60	61	35	17	17	39	59	60
Measurement	15	24	26	14	22	24	13	6	6	14	22	24
Data representation, analysis and probability	13	21	22	14	21	22	17	8	8	14	21	22
Geometry	13	21	21	13	20	20	13	6	6	13	20	20
Algebra	22	35	38	21	32	33	23	11	11	20	30	31
Total	100	162	169	100	155	160	100	48	48	100	152	157

Note: Adapted from Mullis et al. (2000, pp. 319-320).

Table 4.2 Distribution of TIMSS items across performance categories

	TIMSS 1999			TIMSS 1995			Identical Items			Cloned Items		
	% of items	Number of items	Number of score points	% of items	Number of items	Number of score points	% of items	Number of items	Number of score points	% of items	Number of items	Number of score points
Knowing	19	30	30	21	32	32	33	16	16	20	30	30
Using routine procedures	23	38	39	25	38	39	17	8	8	25	38	38
Using complex procedures	24	39	40	24	37	39	27	13	13	24	37	37
Investigating and solving problems	31	51	53	31	48	49	23	11	11	31	47	48
Communicating and reasoning	2	4	7	-	-	-	-	-	-	-	-	-
Total	100	162	169	100	155	160	100	48	48	100	152	157

Note: Adapted from Mullis et al. (2000, pp. 319-320).

As discussed in Chapter 3, large-scale surveys of student achievement have traditionally used, either exclusively or mainly, multiple-choice items. To also measure the important achievement outcomes that are either impossible to measure, or difficult to measure well, using multiple-choice items, a variety of item types were employed in TIMSS. Each test booklet contained items in three different item formats: multiple-choice, short answer, and extended-response in approximately the same proportions. About three-quarters of the items in each test were multiple-choice items, each with four or five response options. About one-quarter of the items were open-ended items requiring students to construct and write their own answers. Some of these items required students to supply a short answer to a problem or question and were called "short-answer" items, while others required an explanation or extended response to a question or problem, and were called "extended-response" items.

The TIMSS mathematics achievement tests were developed through an international consensus involving input from experts in mathematics and measurement specialists (Garden, 1996). The aim of TIMSS instrument development was to have items that had maximum validity across participating countries and to test as wide range of school mathematics curriculum as possible. The items underwent an iterative development and review process, including the pilot testing. Every effort was made to help ensure that the tests represented the curricula of the participating countries and that the items did not exhibit any bias towards or against particular countries.

4.3.3 Translation

In many countries that participate in international comparative studies the translation of the instruments into national language is necessary before the actual data collection. The TIMSS data collection instruments were prepared in (American) English and each participating country was responsible to translate the instruments into the national language. Although there is no way to be absolutely sure that all translations had exactly the same meaning and the same level of language difficulty, the translation process in TIMSS was designed to ensure standard instruments across countries (O'Connor & Malak, 2000). National Research Coordinators received guidelines for translating the testing instruments into their national languages and cultural context. Translators were permitted to adapt the test as necessary to make unfamiliar contextual terms culturally appropriate (e.g., the names used in stems of the items). It was important, however, that their changes did not affect the meaning of the question, the reading level of the test, the difficulty level of item, or the likelihood of another possible or correct answer for the test item.

After the translation was completed, the translated instruments were checked by an international translation company against the TIMSS international version to assess faithfulness of translation. The National Research Coordinators then received feedback from the translation company and the International Study Center suggesting additional revisions. After these had been made the final version was checked by the International Study Center after test was administered using statistical procedures (Mullis & Martin, 2000).

4.4 DATA COLLECTION

In large-scale international studies such as TIMSS, data collection requirements need to be specified carefully to ensure the comparability of the procedures that are followed. Data collection for TIMSS 1995 populations in the northern hemisphere took place in the first half of 1995. In the southern hemisphere, data collection for populations 1 and 2 was carried out in the latter part of 1994. For TIMSS 1999, data were collected in the northern hemisphere in the first half of 1999, while in the southern hemisphere, they were collected in the second part of 1998.

In order for the achievement results to be comparable, the TIMSS tests had to be administered under standardized conditions. Manuals were generally written for the national centers and coordinators of data collection in schools. In many countries it was necessary to train the people who would administer the tests. The main way of assuring standard conditions of data collection in TIMSS was through a number of training sessions that were held for National Research Coordinators who in turn trained the actual personnel who administered the tests in their respective countries.

4.4.1 Quality assurance

Quality assurance efforts are an important part of any study. Any study needs to include measures to ensure that the findings of the study will not be compromised or called into doubt because of a failure to implement high standards of quality control to every aspect of the study (Beaton et al., 1999). One of the major strengths of TIMSS is the extent to which high standards of quality assurance were proposed and implemented (O'Connor & Stemler, 2000).

Each country was asked to identify a person to serve as a quality control monitor for their national study. The monitors attended an international training session during which they were briefed on the kinds of activities they were expected to perform. These included instructions for visiting their national center, interviewing the national research coordinators, visiting a number of schools participating in the study, and observing the data collection process in a number of classes. Each national quality control monitor provided an independent report to the international study center.

4.4.2 Scoring of constructed response items

As already mentioned, approximately a quarter of items required a constructed response. Some required the students to provide only an answer (short answer items), while others required students to show all of their work (extended response items). The multiple-choice items were each scored one point as well as the short-answer items. To obtain more detailed information about answering the extended response items, partial responses to these items were given scores. Students on such items received scores ranging from 0 to 3 points depending on the item and the degree of correctness of their response to the item.

To ensure reliable scoring procedures in each country, scoring guides were developed. For each of the constructed response items scoring guides contained rubrics and explanations of how to implement them together with example student responses for various rubric categories (Garden & Smith, 2000). These guides and the training packets containing examples of student responses for practice in applying rubrics were used in training seminars for those responsible for coordinating the scoring of those items in their own country. National centers were then responsible for training personnel in their countries to apply the scoring rubrics reliably.

4.5 MEASURES OF ACHIEVEMENT

The TIMSS mathematics and science achievement results were scaled to summarize the data for publication and further research. The principal method by which student achievements were reported in TIMSS was through scale scores derived by using the item response theory (IRT) model. With this procedure, the performance of a sample of students can be summarized on a common scale even when different students have been administered different items. In addition to providing a basis for estimating mean achievement, IRT scale scores permit estimates of how students within countries vary and provide information on percentiles of performance. Detailed descriptions of scaling procedures for TIMSS 1995 are given in Adams, Wu, and Macaskill (1997), and for TIMSS 1999 in Yamamoto and Kulick (2000).

A single scale for mathematics and another for science were produced in TIMSS 1995 and in TIMSS 1999. The TIMSS 1995 data were rescaled to enable comparisons with the TIMSS 1999 results. Average overall achievements in mathematics in the participating countries were compared on these scales (Beaton et al., 1996; Mullis et al., 2000). As mentioned in Chapter 1, in addition to normative comparisons of mean achievement, so-called scale anchoring studies were conducted in TIMSS to investigate students' content knowledge at different levels of the achievement scales (Kelly, 1999; Mullis et al., 2000).
Special concern in TIMSS was addressed to estimating the standard errors of all published statistics. The usual formulas for standard errors are appropriate for simple random samples, whereas the TIMSS sampling procedure was multistage with intact classes sampled. Therefore, the jackknife method was used to compute error estimates (Gonzalez & Foy, 2000).

4.6 VALIDITY AND RELIABILITY OF THE TIMSS INSTRUMENTS

As previously explained, in TIMSS, curriculum was seen as a broad explanatory factor underlying student achievement. For the validity of the TIMSS achievement tests, the concepts of curriculum coverage of the test was an important concern (see Chapter 3). To investigate the appropriateness of the TIMSS mathematics tests for the defined populations of students to be assessed, and to show how student performance for individual countries varied when based only on the test questions that were judged to be relevant to their own country, the Test-Curriculum Matching Analysis (TCMA) (Beaton et al., 1996, Mullis et al., 2000) was developed and conducted in TIMSS. The national research coordinator of each country was asked to consult with a person or persons who were very familiar with the curricula at the grades being tested in order to provide information on whether or not each item was in the country's intended curriculum at each of the appropriate grades being tested.

Beaton et al. (1996) and Mullis et al. (2000) report that most TIMSS countries indicated that some items were not included in their intended curricula at the grades being tested, however, the majority of countries indicated that most items were appropriate for their students. According to Beaton et al. and Mullis et al. the TCMA results provide evidence that the TIMSS mathematics tests represent a reasonable basis for comparing achievement for the participating countries in the two data collections. Insofar as countries rejected items that would be difficult for their own students, these items tended to be difficult for students in other countries as well. The analysis showed that omitting such items improved the results for that country, but also tended to improve the results of all other countries, so that the overall pattern of results was largely unaffected.
To assess the reliability of TIMSS achievement tests and procedures, reliability across the eight test booklets was computed. In the countries that were included in the present study, Belgium Flemish, the Netherlands, Hungary, and the Slovak Republic, the median reliabilities ranged from 0.89 to 0.91 in both data collections (Beaton et al., 1996; Mullis et al., 2000).

4.7 TIMSS IN SLOVENIA

Slovenia participated in all TIMSS data collections; in 1995, 1999, and 2003. In all these data collections, the Slovene national TIMSS team followed the international design of the survey as closely as possible. To increase the quality of the data collected, specially trained test administrators visited schools on the agreed upon date and administered achievement instruments to students in the selected classes. However, a decision was taken to deviate from the international design with regard to the target population definition and was approved by the International Study Center. These deviations in Population 2 and their underlying reasons are explained below.

According to the international definition of the Population 2, students in a pair of adjacent grades that contained the most students who were 13 years old at the time of testing were supposed to be sampled. In Slovenia, the grade with the majority of

13 year old students is grade seven. This grade was sampled in TIMSS 1995. The adjacent grade with the second largest proportion of 13 year old students in Slovenia was grade six. However, the TIMSS curriculum analysis (Schmidt et al. 1997), revealed that, by grade six, most of the contents included in the international achievement instruments were not covered in the Slovene mathematics and science curricula. For example, by grade 6, Slovene students did not learn any of the chemistry or physics topics included in the TIMSS science achievement tests. Similarly, sixth-grade students did not learn about functions, equations and several other algebra and geometry topics included in the TIMSS achievement instruments. On the other hand, Slovene grade eight mathematics and science curricula covered most of these topics.

For this reason Slovenia (as well as Germany, Columbia and Romania; Beaton et al., 1996) decided to sample students in grade eight as the adjacent grade. Consequently, by TIMSS design, grade eight was assessed as the upper grade of Population 2 in TIMSS 1995 and also in TIMSS 1999.

Slovene students that participated in the TIMSS data collections in 1995 and 1999 were on average half a year older than their counterparts in many other countries (Beaton et al., 1996; Mullis et al., 2001). If grade six was sampled as the adjacent grade, Slovene students would be on average half a year younger than students in other countries. While higher age of Slovene students could have increased the country's rank, it was considered in Slovenia that the results would be less informative if grade seven were taken as the upper grade. Furthermore, since grade 8 was the final grade in the compulsory education in Slovenia the data were collected on the outcomes of the compulsory education.

For Population 2 there were 121 schools sampled in grade seven and 122 in grade eight in TIMSS 1995, and 149 schools in grade 8 in TIMSS 1999. In total, 2708 grade 8 students were assessed in TIMSS 1995, and 3109 grade 8 students in TIMSS 1999. The data were collected at the end of the school years 1994/1995 and 1998/1999.

CHAPTER 5
Research design

This chapter describes procedures that were employed to address the research questions in this study, outlined in Chapter 1. The underlying concepts that are important in addressing these questions were discussed in Chapter 3. In this chapter, these concepts will be elaborated in terms of measurements that will be carried out in order to provide information for addressing the research questions and to transform the general research questions into an operationalized form. TIMSS 1995 and 1999 data were used for this purpose. More specifically, items in the TIMSS achievement tests were used to describe the intended and the attained curricula in Slovenia as well as the attained curricula in the reference countries. Responses of students to these items were used to assess correspondence of Slovene achievement with both, the standards and achievements of students in the reference countries. These comparisons were carried out by first taking students and next items as units of analysis.

The first section (5.1) examines how the reference points for describing student achievement will be constructed in this study. The next section (5.2) elaborates on the procedure for measuring the attainment targets in the intended curriculum. Section 5.3 describes how the correspondence between the attained curriculum and the selected reference points will be measured. The procedure for operationalization of the attainment targets is described in this section followed by procedures for assessing the correspondences between achievement and the reference points and analyzing trends in these correspondences. In the final section (5.4) the generic structure for formulating operationalized research questions with examples of these questions is presented.

5.1 CONSTRUCTION OF THE REFERENCE POINTS

The research questions, outlined in Chapter 1, refer to the concepts of strengths and weaknesses in student achievement that were discussed in Chapter 3. In this study, mathematics achievements of students and their developments over time will be analyzed using two reference points, the intended curriculum for mathematics in Slovenia and the achievements of students from several other European countries. First, it will be described how the attainment targets in the Slovene mathematics curriculum were measured.

5.1.1 Measuring the attainment targets in the intended curriculum

As discussed in Chapter 3, the intended curriculum is the appearance of a curriculum at system level. In this study, the Slovene curriculum for mathematics in the reformed system was used. As described in Chapter 2, it is structured by grades (1 through 9) and, within each grade, by content areas and objectives that should be covered. Additionally, attainment targets are specified for most grades at two levels, and at three levels for the two final grades (grades 8 and 9). Contents and objectives for mathematics in all grades are considered in this study. As also discussed in Chapter 2, textbooks to be used for instruction are not prescribed in the reformed system and teachers are free to choose among the available textbooks. Although textbooks can be seen as part of the intended curriculum (see Chapter 3), this study focused on the curriculum guide prepared at system level and did not consider the textbooks used by teachers.

The procedure used to describe the intended curriculum in this study employs items from the TIMSS international achievement tests. Items are used in the Slovene intended curriculum to help illustrate intentions embedded in the objectives (see Chapter 2). At the same time, items can be used to describe the intended curriculum more quantitatively. In this approach a large set of items is taken and persons knowledgeable about the curriculum are asked to indicate the matching between these items and the curriculum (see Chapter 3). Mutual coverage is desired to be as large as possible. While test can contain items which do not match the intended curriculum, there might also be parts of the intended curriculum which are not represented by the items in the test.

In the present study, the TIMSS achievement tests were not designed specifically to describe the Slovene intended curriculum for mathematics. For this reason, the extent of the mutual coverage between the intended curriculum and the TIMSS items needed to be examined in order to provide evidence on the appropriateness of the test for measuring and linking the intended and the attained curricula in Slovenia.
This coverage was addressed by first asking a Slovene mathematics curriculum specialist to estimate the coverage of the attainment targets by the TIMSS items. For each attainment target in the curriculum, the specialist indicated whether it is

covered by the items in the TIMSS tests. There might have been several TIMSS items covering the same attainment target or several attainment targets covered by the same TIMSS item. The latter occurred, for example, when an attainment target in the lower grade was 'upgraded' in the higher grade. For example, the attainment target for grade 3 "a student can order whole numbers up to 1000" (*Curriculum Guide*, 2002, p.76, translation MŠ) was upgraded in grade 4 into "a student can order whole numbers up to 10000" (p.77). In such cases it was considered that an item that covered the attainment target for the upper grade also covered the attainment target for the lower grade. The same occurred between the two levels of these targets. Most attainment targets at Level 1, except two, can be considered included in attainment targets at Level 2 in the similar manner as above. In such cases, if an item covered a Level 2 attainment target it was also considered to cover the corresponding Level 1 attainment target. This however, does not hold for the attainment targets at Level 3, which include more specialized knowledge and skills intended only for higher achieving students to attain.

The percentage of attainment targets at Levels 1 and 2 that were covered in this way was 77 % for the TIMSS 1995 as well as for the TIMSS 1999 tests. The percentage for both tests is equal due to similarities between the two tests (see Chapter 4). The main topics that were not covered in the TIMSS tests were geometry topics about circles, triangles, constructing angles and triangles, parts of three-dimensional geometry including the Pythagorean theorem, and simplifying symbolic expressions using properties of operations. These topics are mainly covered in the final grades of compulsory education. Also, very few attainment targets at Level 3 were covered. Therefore the TIMSS items can be seen as reasonably covering the "general" part of the intended curriculum, while they do not cover its "specialized" part.

Second, the 'reversed' question was addressed. For each TIMSS item, the curriculum specialist indicated whether it is covered by the attainment targets in the curriculum. The coverage of the TIMSS tests by the intended curriculum in Slovenia was examined already in the Test Curriculum Matching Analysis (TCMA, see Chapter 4) where it was found that this coverage was considerable, 88 % for TIMSS 1995 (Beaton et al., 1996) and 96 % for TIMSS 1999 (Mullis et al., 2000). The increase in the coverage from 1995 to 1999 was mostly due to data representation items (see Chapter 2).
However, these data were based on the non-reformed curriculum. In the present study, the coverage of the TIMSS tests by the reformed curriculum needed to be examined. For each item, the mathematics curriculum specialists indicated whether

it is covered by the attainment targets in the reformed curriculum. The percentage of TIMSS 1995 items that were covered was 96 %, and of TIMSS 1999 items 97 %. Again, the similarity in these percentages is due to the similarities between the two achievement tests. The above percentages for the mutual coverage were seen by the mathematics expert as sufficiently large to enable meaningful examination of Slovene achievement on the basis of TIMSS items. As mentioned, this examination is meaningful for the attainment targets at Levels 1 and 2, while the "specialized" knowledge and skills at Level 3 in grades eight and nine are not sufficiently represented in the TIMSS tests.

In the following, procedures that were used for measuring the standards in the intended curriculum are described. To provide data for the first reference point in this study, the measurements of the intended curriculum consisted of allocating each TIMSS item to one of the levels of the attainment targets. Accordingly, the measures were given values "Level 1", "Level 2", and "Level 3".
While the coverage of the attainment targets by the TIMSS items was assessed by a single mathematics expert in Slovenia (a mathematics curriculum expert and school advisor), the coverage of the TIMSS items by the attainment targets and allocation to the levels were carried out by two additional experts (a member of the national mathematics curriculum development panel and a mathematics teacher). The experts were asked to consider all possible item characteristics that might influence the level of the intended standards, such as familiarity of the context to Slovene students, complexity of wording, and item format.
For each item, the median of the three measures was taken as the measure that was used in this study. The reliability of these expert judgments was 0.72 which was deemed acceptable (e.g., Wolf, 1994) considering that the attainment targets are a novelty in the Slovene mathematics education.

In the process of allocation of items to the levels of the standards, it was found that for six items in the TIMSS 1995 and for five items in the TIMSS 1999 achievement tests, the standards in the intended curriculum could not be determined. Also, two TIMSS 1999 items had problems with translation. These items were excluded from further analysis in this study. As already indicated, it was found that only six items in the two tests were allocated at Level 3. These items were also excluded from further analysis since they do not provide a sufficient coverage of the attainment targets at this level.
The TIMSS test items that are included, therefore, represent the attainment targets at Level 1 and 2 in the intended curriculum. Since, as described in Chapter 2, the attainment targets at Level 3 are set only in the two final grades and are meant to

be achieved only by a smaller proportion of students, the inferences from the results of this study are considered valid for the "general" part of the curriculum. Of the 161 TIMSS items, 147 items were included in the analysis.

5.1.2 Selection of countries for a reference point

As explained in Chapter 1, in addition to the attainment targets, a reference point called 'European dimension' (White paper, 1996) emerges from the Slovene educational policy documents. This reference point was based in this study on achievements of students from several other European countries. In order to identify possible areas for improvements in Slovene mathematics education, countries with similar or higher overall achievement than Slovenia were selected. Furthermore, since developments in Slovene achievement in the late 1990s are also addressed in the research problem, countries were selected that had similar or higher overall achievement in both measurements in this period. The countries were selected based on the IRT scores for average achievements reported in Mullis et al. (2000). Four countries[1] were selected in this way: Belgium-Flemish, the Netherlands, Hungary, and the Slovak Republic. The average achievement in Belgium-Flemish was significantly higher than in Slovenia on this scale in both TIMSS data collections, while the average achievements in the other three countries were similar to Slovenia.

For brevity, from here on the reference point constructed on the basis of the attainment targets in the Slovene intended curriculum will be called also 'the standards', while the second reference point will be called 'the achievements of students from the reference countries'.

5.2 PROCESSING OF MEASURES OF THE ATTAINED CURRICULUM

As mentioned on several occasions, the achievement data that used in this study were collected in the TIMSS 1995 (Gonzalez & Smith, 1997; IEA, 1997) and the TIMSS 1999 surveys (Gonzalez & Miles, 2001; IEA, 2001). As was discussed in Chapter 3, measurements in this study were processed using the percent correct technology (also called p-values). As also indicated in that chapter, two approaches to analyses of the data were used, student based and item based analyses. When students were taken as units of analysis, the average percent correct on a set of items (e.g., on the whole test or in the subdomains) in a particular country was computed

[1] As explained, country is used synonymously for education system.

in order to estimate achievement of students on this set of items. When units of analysis were items, the item percent correct was computed as the percentage of students responding correctly to a particular item. The usual term for this is *item difficulty* although some authors prefer to use the term 'item easiness' because with higher percent correct an item is relatively easier. In this study the former is used.

In the TIMSS scoring procedures, described in Chapter 4, most of the TIMSS items were scored dichotomously, indicating whether the item was answered correctly or incorrectly. As also explained, some of the items required a student to construct an extended response. For such items partially correct responses were scored. The students were assigned a score ranging from 0 to 3 points in TIMSS 1995, and from 0 to 2 points in TIMSS 1999, depending on the item and the degree of correctness of their responses to this item.

When an item response can have only two values, 0 for incorrect and 1 for correct, the average score on the item for a sample of students is also the proportion correct. However, this does not hold for an item where responses can score more than 1 point. For such items it was necessary to find a way to use the proportion correct to represent the responses. In this study the TIMSS procedure for computing proportion correct on such items was used (Beaton & Gonzalez, 1997). This procedure includes transforming graded responses into a series of dichotomous variables as described below.

If for example, an item had possible score points 0, 1, and 2, two variables were created: V_1 equaled 1 if the student received 1 or 2, and 0 otherwise, and V_2 equaled 1 if the student received 2, and 0 otherwise. For such an item two types of percent correct were computed; the percentage of students receiving at least one point (the percentage of students receiving 1 on V_1), and the percentage of students receiving full credit (the percentage of students receiving 1 on V_2). As already mentioned, the average percentages correct in this study were computed using these partial credit variables. There were 153 such variables included in the analysis of the TIMSS 1999 data and 93 such variables in the analysis of trends.

Percentages correct as indicators of students achievement and item difficulties are not precise. They are estimates of the true values in the whole population obtained from measurements on a sample of students. For this reason, sampling errors need to be taken into account when reporting results. With each estimate a standard error (which is equal to sampling error in this analysis since other measurement errors are not considered) will be presented. Since in TIMSS a two-stage clustered sampling design was used (see Chapter 4), special procedures need to be employed for estimation of these standard errors. The procedure employed in this study is the jackknife procedure (Brick et al., 1997). These standard errors enabled the construction of confidence intervals for the estimates and an assessment of the statistical significance of the correspondence of Slovene achievement with the reference points. The procedures for assessing this correspondence are described in the following section.

5.3 CORRESPONDENCE OF THE ACHIEVEMENT IN SLOVENIA WITH THE SELECTED REFERENCE POINTS

In this section, procedures that were needed for assessing the correspondence of mathematics achievement in Slovenia with selected reference points are described. The first subsection describes the operationalization of the attainment targets onto a numeric scale that enabled comparisons with the achievement scores. The remaining subsections describe the procedures used in student based and in item based analyses, how trends were examined and how analyses were carried out considering different content areas and cognitive categories of the TIMSS items.

5.3.1 Operationalization of the levels of the attainment targets

As previously explained, in order to describe the correspondence of mathematics achievement of Slovene students with the standards in the intended curriculum, the two appearances of the curriculum were linked in this study by means of individual items from the TIMSS achievement tests. However, the measurements on the intended curriculum described in the previous section were carried out in the non-numeric form of the level of the attainment target for individual item, while the measurements on the attained curriculum were provided as the percentage of Slovene students answering a particular item correctly. These two measurements were linked in this study by operationalizing the intended level for each item into the *intended percentage correct*. This operationalization was carried out on the basis of general descriptions of the differences between the three levels of the attainment targets and was as follows:

- For items allocated at *Level 1*, the intended percentage correct was set to 75 %.
- For items allocated at *Level 2*, the intended percentage correct was set to 50 %.

This operationalization is, of course, a very simple model for the intended percent correct (or difficulty) for individual items. Items that are used to measure student achievement may and should vary in their intended difficulty as well as in other characteristics. However, in absence of explicit guidelines of how these levels of the attainment targets should be operationalized in order to assess whether they have been achieved, this operationalization was deemed sufficient. It was also considered plausible by the mathematics curriculum experts that allocated the TIMSS items to the levels of the attainment targets.

Once the target and reference measures were calculated in the form of comparable scores, correspondence between the two was assessed. There may be different procedures used for this depending on the nature of the scales. In this study, both the target and the reference measures were on a continuous numeric scale, implying straightforward numeric procedures for assessing correspondence between them. Below, procedures for analysis of this correspondence in the student based and item based analyses are described.

5.3.2 Correspondence when students are taken as units of analysis

The procedures used for analysis of the correspondence between Slovene achievement and the selected reference points will be described first for the reference point that was constructed on the basis of the attainment targets in the curriculum. The elements of the procedures for assessing the correspondence between Slovene achievement and the achievements in the reference countries that differ from those for the first reference point will be described thereafter.

The correspondence between Slovene achievement and the attainment targets in the curriculum was examined as follows. First, average percent correct estimates of achievement of Slovene students on the whole test and their standard errors were computed. Second, the level of the standard for the test was determined as the average of the intended percents correct for items in the test. Average percent correct estimates of achievement were then compared to the level of the standard. When the estimate for Slovene achievement was significantly higher than the reference point, this was taken as an indication of a strength in Slovene achievement.[2] In case it was significantly lower, this was taken as an indication of a weakness. In the third case, Slovene achievement was described as corresponding with the standards. Correspondence of achievement at respective levels of the standards was also examined to indicate whether there were differences in these correspondences between the two levels.

---

[2] Generally, the level of significance was taken at 0.05. When Slovene achievement was compared to the standards a 'conservative' critical value of 3 was taken to avoid capitalizing on small differences from the standards. When Slovene achievement was compared to the achievements in the reference countries the critical value was determined using Bonferroni adjustment for multiple comparisons (Gonzalez & Gregory, 2000). It was at 2.235. For comparisons between individual items, the critical value for differences in item percents correct was also set to 3 because of large numbers of possible simultaneous comparisons.

In the case when the reference point was constructed on the basis of the achievements of students from the four reference countries, there were four reference measures computed as the average percent correct estimates of achievement in these countries and their standard errors. In comparison with each of these countries it was indicated whether Slovene achievement was significantly higher or lower. For this a usual test of significance was used (t-test). In this way, comparisons with each of the four reference countries were examined. However, since no particular country was taken as 'the most important' for comparisons with Slovenia, it was said that there are indications of strengths or weaknesses in Slovene achievement if significantly higher or lower achievement in Slovenia was observed in at least two comparisons.

5.3.3 Correspondence when items are taken as units of analysis

As for student-based analyses, procedures in item based analyses will first be explained for the reference point of the attainment targets in the curriculum. The elements that are different for the second reference point are described thereafter.

As the first step in this analysis, estimates of item difficulties and their standard errors were computed for each item from the responses of Slovene students. They were called the 'Slovene item difficulties'. The intended item percents correct for the TIMSS items were taken as the 'intended item difficulties'. Items for which the Slovene item difficulties were significantly higher than their intended difficulties were classified as 'strong'. Items for which Slovene item difficulties were significantly lower than their intended difficulties were classified as 'weak'. Other items were classified as 'neutral'.

As argued in Chapter 3, it is important to examine the numbers of strong and weak items in order to reveal whether the correspondence or non-correspondence with the standards observed for student achievement occurred over a number of items or whether there were, for example, a few high achieving items that masked the non-correspondence of achievement on the remaining majority of items. At the same time the contents of individual strong and weak items serve the criterion-referenced interpretations of Slovene achievement. However, inferences on the basis of individual items may be sometimes (undesirably) sensitive to particular item characteristics, such as, wording, translation differences, format, graphical representations or misprinting. Inferences about strengths and weaknesses in Slovene achievement were therefore made if relatively high or low numbers of items were classified as weak or strong. For example, if only strong or neutral items were found, this was taken as an indication of a strength in Slovene achievement.

Similarly, if only weak items were found this was taken as an indication of a weakness.
In these and other cases, the sign test for significance was used. In some cases, correlations were also computed. However, correlation is a measure of the relative correspondence, and could be high even though all item difficulties in one group (of the two groups that are correlated) were significantly lower than the difficulties in the other group. Thus, correlations were used only to support the observations about the absolute changes in difficulties. Analyses of correspondence were also carried out for the respective levels of the standards.

When the reference point was constructed on the basis of achievements of students from the four reference countries, sets of strong, weak and neutral items were identified for each of these countries. Since, as previously noted, no country is taken as the most important, these sets were additionally clustered into sets of strong, of weak and of neutral items in the following way. An item was classified as strong in comparison with the reference countries if its difficulty for Slovenia was significantly higher than its difficulties in at least two reference countries. Similarly, an item was classified as weak if its difficulty for Slovenia was significantly lower than its difficulties in at least two reference countries. Other items were classified as neutral. In cases when the Slovene item difficulty was significantly higher than the difficulties in two reference countries and at the same time significantly lower that the difficulties in the remaining two countries, such items were also classified as neutral.

5.3.4 Trends

As previously explained, mathematics achievement of Slovene students was described based on the most recently available data, TIMSS 1999, while trends were examined comparing the results from TIMSS 1995 and TIMSS 1999. However, as explained in Chapter 4, the TIMSS 1995 and TIMSS 1999 achievement tests were not exactly equal. The two tests overlapped in a set of items, while most of the remaining items were 'clones', i.e., not identical but similar in content, format and difficulty (see Chapter 3). It therefore first needed to be investigated as to whether the cloned items can be considered equivalent for describing strengths and weaknesses in mathematics achievement of Slovene students; and, in consequence, for examining the trends over this period. The rationale for using cloned items is that a wider coverage of the domain is provided.
As explained in Chapter 3, this issue was addressed by carrying out separate analyses for the identical and the cloned items. Only in the case that no significant

change in student achievement for Slovenia had been observed on both sets of items, further investigations of trends would use also the cloned items. In all other cases, only identical items were used for analyses of the trends.
Once the appropriate set of items was determined, trends were examined on this set of items carrying out the procedures for student based and item based analyses described in the previous sections on the 1995 and 1999 data and comparing the results. Changes in achievement in this period were determined using a usual test for significance (t-test).

5.3.5 Location of strengths and weaknesses in content areas and in cognitive categories

The international achievement tests are usually designed to cover a wide range of mathematics topics. In TIMSS it was observed that there are differences among countries in the grade levels at which particular topics are introduced and in the teaching emphases given to some topics (Schmidt et al., 1997). As explained in Chapter 4, assessment frameworks were prepared in TIMSS on the basis of which items were grouped into content categories. Achievements in these categories were compared to provide information about the possible effects of the curricular variation on average achievement.

As described in Chapter 1, it is also of interest in the present study to know whether strengths in mathematics achievement of Slovene students can be observed primarily in one particular content area (or several content areas), or weaknesses in another content area (or several content areas). However, the structure of the Slovene intended curriculum to some extent differs from the TIMSS content categories. This is understandable since the TIMSS frameworks were not developed to suit the curriculum of any country in particular, but rather to serve as a base for construction of the international achievement test. The differences between Slovene curriculum and the TIMSS content categories are, for example, on items that were classified in TIMSS as 'fractions and numbers sense' items while in Slovenia the knowledge and skills that these items require are taught as proportionality topics within a broader area of functions. There are also differences for items that required measurement of areas of shapes ('measurement' in TIMSS) for which the required knowledge and skills are taught in Slovenia as geometry topics. The non-correspondence of the TIMSS content areas with the structure of the Slovene mathematics curriculum was observed also by Magajna (2000). Classification of TIMSS items was therefore re-examined by a Slovene mathematics curriculum expert to derive content areas that better reflected the Slovene intended curriculum.

As also described in Chapter 4, the TIMSS assessment frameworks included the second dimension of 'performance expectations'. During the development of the test, each item was assigned to a category of these performance expectations. In the Slovene intended curriculum cognitive levels are not explicitly addressed (the three levels of the attainment targets are not linked to cognitive levels). However, as explained in Chapters 1 and 2, it is relevant to examine Slovene achievement by these levels (Rutar Ilc, 2003). It was therefore decided to use TIMSS cognitive categories in this study to address the research questions in Chapter 1. As for the content categories, the mathematics expert was asked to review the TIMSS classification into cognitive categories in the light of the (implicit) intentions of the curriculum. In general, the TIMSS cognitive categories were considered suitable by the mathematics expert. However, since only four items were assigned to the category 'communicating and reasoning' in TIMSS (see Table 4.2), they were merged into the previous category of 'investigating and solving problems'.

The Slovene content areas and cognitive categories used in this study are presented in Table 5.1. In each cell of the cross table in Table 5.1 the numbers of TIMSS 1999 items with corresponding values for the two dimensions are given. The distribution of the TIMSS 1995 items across these cells was similar.

Table 5.1 Distribution of the TIMSS 1999 items across the Slovene content areas and cognitive categories used in this study

	Knowing	Using routine procedures	Using complex procedures	Investigating and solving problems	Total
Natural numbers	1	3	3	2	9
Meaning of rational numbers	9	2	8	1	20
Operations with rational numbers	-	8	2	9	19
Algebraic expressions	6	8	1	10	25
Functions and Proportionality	2	6	3	10	21
Geometrical shapes	3	4	6	10	23
Measurement	5	4	4	5	18
Data representation	2	2	9	2	15
Probability	1	-	1	1	3
Total	29	37	37	50	153

5.3.6 Reviewing most notable strengths and weaknesses

As was discussed on several occasions, analyses at the levels of the overall domain and at several subdomains may not provide sufficient information for some purposes. It was argued in Chapters 1 and 3 that individual items may reveal additional information useful for curriculum development and implementation. In the present study, analyses of the content and characteristics of individual items might reveal information which may not have been obtained by examining average scores on sets of items or by examining the numbers of items that deviated from the reference points. For this reason, individual items for which item difficulties deviated to a large extent from the reference points were also examined. They were called *particularly strong* and *particularly weak* items. When the reference point was constructed on the basis of the attainment targets, the items for which the observed difficulties significantly differed by at least 10 percentage points from the standards were selected. The range of 10 percentage points was determined after the variability of item difficulties around the levels of the standards was examined. When the reference point was the achievements of students from the reference countries, the items that were strong or weak in comparison with at least two countries were selected. These items were also labeled particularly strong and particularly weak for this analysis. Finally, the sets of particularly strong and of particularly weak items were compared between the two reference points.

The analysis of the content and characteristics of particularly strong and particularly weak items were used to discuss strengths and weaknesses in Slovene achievement as compared to the respective reference points. However, in such interpretations caution is needed because, as already mentioned, students' scores on an item may be (undesirably) sensitive to translation inconsistencies in items, format, graphical representations, misprinting, etc. Nonetheless, through this analysis, information on areas in which most notable strengths and most notable weaknesses in Slovene mathematics education as reflected in students' scores on these items was provided.

5.4 OPERATIONAL RESEARCH QUESTIONS

In Chapter 1, two broad research questions were posed for this study. Based on the description of analytical procedures in this chapter, these research questions can be transformed into operational research questions. These operational research questions are presented in the form of a list of questions due to repetition that would occur in this list. Rather a structure is presented from which each operational research question can be easily constructed. This structure is shown in Table 5.2.

Table 5.2 Generic structure for the operational research questions

	Standards in the curriculum								**Achievements in the reference countries**			
	Student based analyses				Item based analyses				Student based analyses		Item based analyses	
	TIMSS 1999		Trends 1995 - 1999		TIMSS 1999		Trends 1995 - 1999		TIMSS 1999	Trends 1995 - 1999	TIMSS 1999	Trends 1995 - 1999
	Overall	Levels of standards	Overall	Levels of standards	Overall	Levels of standards	Overall	Levels of standards				
Overall test	A	B	C		D							
Content areas					E							
Cognitive categories												
Item level					F							

Based on this structure, specific operational research questions can be formulated. As indicated in Table 5.2, in the analysis student based and item based approaches were employed at the levels of the overall achievement test, within the respective content areas, cognitive categories, and at item level. Within each of these steps, trends in correspondence of achievement with the reference points were also investigated. Examples of operationalized research questions that can be drawn from the structure in Table 5.2 are as follows:

A. *To what extent did achievements of Slovene students at the end of compulsory education in 1999 correspond with the standards?*
B. *Were the correspondences similar when considering different levels of the standards?*
C. *To what extent did correspondences in 1999 deviate from those in 1995?*

In an item based analysis at the overall level, a possible question is:
D. *To what extent did item difficulties for Slovenia in 1999 correspond to the standards?*

As discussed previously, on the bases of these correspondences, items were classified as strong, neutral or weak. An example of the operationalized research question addressing the location of these items in the content areas of the curriculum is
E. *In which content areas were strong and weak items located?*

Reviews of contents of individual items can be seen as student based or as item based approach. Although the structure in this analysis is different from that of the overall test, content areas, and cognitive categories, this analysis is also indicated as part of the generic structure for the operational research questions in Table 5.2 under item based analysis. Item reviews were carried out by addressing questions such as

F. *What are the contents and characteristics of the particularly strong and particularly weak items?*

In Table 5.2, all possible analyses for addressing the general research questions are indicated. However, for some of these analyses there may be insufficient numbers of items and those were not carried out. It is deemed that for each analysis at least 5 items are needed. These requirements are to some extent arbitrary, however they are based on the examination of the data from which it was determined what requirements are reasonable.

The explicit operational research questions addressed by the analyses that were described in this chapter will be outlined in the following chapters where the results of these analyses are also presented. The presentation of the results is structured according to the two main research questions. In the following chapter, Chapter 6, mathematics achievement of Slovene students will be analyzed from the perspective of the attainment targets in the intended curriculum. In Chapter 7, the results of comparisons with the reference countries will be presented.

CHAPTER 6
Correspondence of mathematics achievement in Slovenia with the attainment targets

As explained in the previous chapters, the research questions in this study focus on the strengths and weaknesses of mathematics achievement of Slovene students. Points of reference for addressing these questions are the attainment targets that were specified in the reformed intended curriculum for mathematics education in Slovenia and the achievements of students from four other European countries. In this chapter, the strengths and weaknesses of mathematics achievement of Slovene students will be described from the perspective of the attainment targets, also called the standards. Once a description of strengths and weaknesses was available at the level of the overall domain, more detailed analyses were carried out by 'zooming in' on the question of where in the curriculum these strengths and weaknesses are located. As explained in Chapter 5, this was done in terms of curriculum content areas and levels of cognitive requirements. Within these research questions, developments in student achievement and in its correspondence with the standards over the period between 1995 and 1999 were also examined.

In section 6.1, correspondence between the standards in the intended curriculum and the observed achievement of Slovene students is investigated at the level of the overall test. In this section, as in other sections in this chapter, first the results of student based analysis are presented, followed by the results of item based analysis. In section 6.2, strengths and weaknesses are located in the content areas of the Slovene curriculum, and, in section 6.3, within levels of cognitive requirements. In section 6.4, the most notable strengths and weaknesses in mathematics achievement of Slovene students are reviewed through particularly strong and particularly weak items, respectively. In the final section (6.5), main results from the analyses for the first research question are summarized.

6.1 CORRESPONDENCE AT THE LEVEL OF THE OVERALL MATHEMATICS DOMAIN

As explained in Chapter 5, the method that was used for assessing correspondence of Slovene achievement with the attainment targets in the intended curriculum consisted of classifying TIMSS items into three levels and defining the intended percentage correct for these levels. At Level 1, the expected percentage correct was set to 75 percent, at Level 2 it was set to 50 percent, and at Level 3 to 25 percent. As it appeared that there were only a few items that were classified at Level 3, it was decided that this level would not be considered in the analysis.

Following the design of this study, achievement of Slovene student and its correspondence with the standards are described using the data from the TIMSS 1999 study. Developments in this achievement and its correspondence with the standards are examined using the TIMSS 1995 and TIMSS 1999 databases. In order to take into account the adaptations that were made in the TIMSS achievement tests between the two data collections, the analyses were carried out separately for identical and cloned items (see Chapters 4 and 5).
Furthermore, in the design of this study two approaches to analyses were described. In the first approach, students are taken as units of analysis, while in the second approach the units of analysis are items. These two approaches were carried out taking all items in the achievement test together, as well as considering different content areas, levels of cognitive requirements, and finally individual items. In Chapter 5, the generic structure of the operational research questions was developed for which the analyses in this chapter were carried out. The specific questions addressed and the results of the analyses for these questions are presented in the following sections.

6.1.1 Students as units of analysis

In this subsection, the following questions are addressed:

- *To what extent did achievement of Slovene students at the end of compulsory education in 1999 correspond with the standards?*
- *Were the correspondences similar when considering different levels of the standards?*

To examine the appropriateness of the cloned items for analysis of trends in this study, the following question is also addressed:

- *Were the correspondences similar when considering the sets of identical and cloned items in 1995 and 1999 separately?*

After the appropriate set of items for trend analysis is determined, the following two questions will also be addressed:

- *To what extent did the correspondences in 1999 deviate from those in 1995?*
- *To what extent did the correspondences in 1999 deviate from those in 1995 when considering different levels of the standards?*

The results of analyses for answering these questions are presented in Tables 6.1 and 6.2.

Table 6.1 Mathematics achievements of Slovene students in 1999[1]

	Overall		Level 1 standard			Level 2 standard		
	Number of items	Average percent correct	Number of items	Average percent correct		Number of items	Average percent correct	
TIMSS 1999	153	61 (0.7)	69	70 (0.7)	▼	84	53 (0.8)	▲

▼ Average percent correct significantly lower than the standard
▲ Average percent correct significantly higher than the standard
( ) Standard errors are presented in parentheses

The results in Table 6.1 show that the average score of Slovene students on all items in the TIMSS 1999 achievement test was 61 percent which was found as corresponding with the standards. This could be interpreted as showing that, in overall, achievement of Slovene students in mathematics is satisfactory when compared to the attainment targets in the curriculum.

When considering the different levels of the standards, it can be observed that the average score at Level 2 was lower than at Level 1 ($p<0.05$[2]) as was expected from the hierarchy of these levels (see Chapter 2). However, the correspondences of scores with the standards were different for the two levels. At Level 1, the score of 70 percent correct was significantly lower than the intended level of 75 %, while at Level 2, the score of 53 percent correct was significantly higher than the intended 50 %. This indicates that while scores in Slovenia corresponded to the standards at the level of the overall test, there seem to have been weaknesses at Level 1 and strengths at Level 2.

1 In all tables in this thesis, rounded estimates are presented while tests of significance were carried out using non-rounded values. This may cause some inconsistencies in the tables.

2 The significance of this difference is not indicated in Table 6.1 to avoid confusion with indications whether the scores correspond with the standards. This significance can be determined using t-test with the standard error that is presented in Table 6.1. The same holds for similar comparisons in other tables in the remainder of this thesis.

Table 6.2 Trends in mathematics achievements of Slovene students between 1995 and 1999

	Overall			Level 1 standard			Level 2 standard		
		TIMSS 1995	TIMSS 1999		TIMSS 1995	TIMSS 1999		TIMSS 1995	TIMSS 1999
	Number of items	Average percent correct	Average percent correct	Number of items	Average percent correct	Average percent correct	Number of items	Average percent correct	Average percent correct
Identical items	48	69 (0.7)	70 (0.7)	30	72 (0.7) ▼	73 (0.7)	18	65 (0.7) ▲	65 (0.7) ▲
Cloned items	92	60 (0.7)	59 (0.7)	38	73 (0.7)	69 (0.7) ▼	54	51 (0.7)	52 (0.7)

▼ Average percent correct significantly lower than the standard

▲ Average percent correct significantly higher than the standard

( ) Standard errors are presented in parentheses

Table 6.2 shows results of analyses of trends that were carried out separately for identical and cloned items in TIMSS 1995 and TIMSS 1999. At the overall level, scores of Slovene students on identical items were similar between the two measurements (69 and 70 percent correct) and they corresponded with the standards in both occasions. Although scores on cloned items were lower (60 and 59 percent correct, $p<0.05$ for both pairs of estimates), they were also similar between the two measurements and corresponded with the standards.
When these scores are broken down by the levels of the standards, differences in the test scores between 1995 and 1999, as well as in their correspondences with the standards can be observed. While at Level 2, the scores on identical items (65 percent correct on both occasions) as well as on cloned items (51 and 52 percent correct) were similar between 1995 and 1999, at Level 1, the scores on identical items were similar (72 and 73 percent correct) and the scores on cloned items changed (73 and 69 percent correct, $p<0.05$).

The question regarding the appropriate set of items for analysis of the trends can now be answered as well. Since the change in scores is observed only on the set of cloned items, while the scores remained stable on the set of identical items, it seems plausible to argue that mathematics achievement of Slovene students did not change in the period between 1995 and 1999. In addition, it seems plausible to hypothesize that the changes in the observed test scores of students on the set of cloned items could have been caused by the changes introduced in the items. This shows that the cloned items *can not* be used to examine trends in the strengths and weaknesses in Slovene achievement in the present study. From here on, trends in student achievement will therefore be examined using the set of identical items only.

The results of trend analysis on identical items indicate no significant changes in scores of Slovene students between 1995 and 1999. This stability in mathematics achievement of Slovene students was also observed by Mullis et al. (2000) through comparisons of IRT scores. The above results based on the breakdown of the item percent correct scores on identical items by the levels of the standards additionally support the observations of stability in Slovene achievement in this period.
Regarding the correspondence of achievement on the identical items with the standards, differences can be observed between the two levels. On identical items at Level 1, the scores of Slovene students were lower than intended in 1995 and they corresponded with the standards in 1999. However, it can also be observed that this change in correspondence is due only to a small (and non-significant) change in actual scores. It could be said that there was also only 'a slight change' in correspondence.

6.1.2 Items as units of analysis

In this subsection, items are taken as units of analysis. As described in Chapter 5, an item was classified as strong if the observed difficulty (in the form of the item percent correct) was significantly higher than the level of the standard for this item. Similarly, an item for which the observed difficulty was significantly lower than the level of the standard was classified as weak.

The following questions are addressed in this subsection:

- *To what extent did item difficulties for Slovenia in 1999 correspond with the standards?*
- *Were the correspondences similar when considering different levels of the standards?*
- *To what extent did the correspondences in 1999 deviate from those in 1995?*
- *To what extent did the correspondences in 1999 deviate from those in 1995 when considering different levels of the standards?*

The results of the analysis for these questions are presented in Figure 6.1 and Tables 6.3 and 6.4. In Figure 6.1, the correspondences of item difficulties with the standards in the Slovene curriculum are presented. There are two areas in the figure: the items that were allocated at Level 1 and those allocated at Level 2. Within these areas, items are ordered by their decreasing difficulties. For both levels, the intended percents correct are also indicated as a straight line (75 percent correct for Level 1 and 50 percent correct for Level 2).

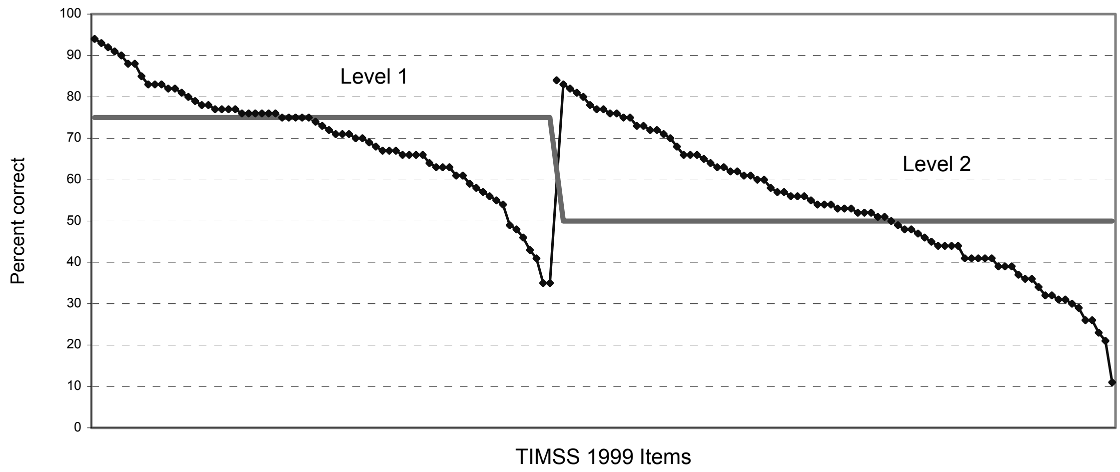


Figure 6.1 Correspondences of item difficulties with the standards in 1999

It can be observed that, on average, items that were allocated at Level 2 had lower difficulties than items that were allocated at Level 1, as was also seen in Table 6.1. However, there is considerable variability in item difficulties at both levels of the standards and, consequently, items also vary in their correspondences with the levels of the standards. At Level 1, item difficulties ranged from 94 to 35 percent correct. At Level 2, this range was from 84 to 11 percent correct. The numbers of items for which the observed item difficulties were higher (strong items), lower (weak items), or did not differ significantly from the level of the standard (neutral items) are presented in Table 6.3 (see below). In addition, it can also be observed from Figure 6.1 that for some items at Level 1 their difficulties were even lower than intended for Level 2 (3 items; $p<0.05$ for all items), and for some items at Level 2, difficulties were even higher than intended for Level 1 (4 items; $p<0.05$ for all).

It seems plausible to argue that these items indicate considerable discrepancies between the intended curriculum in the reformed system and the attained curriculum in the non-reformed system. While this does not mean that the reformed curriculum cannot be realized as intended (this cannot yet be measured due to the ongoing process of introduction of the new curriculum), it seems plausible to argue that there were areas in which student achievement was considerably lower than the 'new' intentions and at the same time areas in which it was considerably higher than these intentions. This lack of correspondence can, of course, be a consequence of the differences between the two curricula. However, the analysis of the differences between the non-reformed and the reformed curriculum that might have caused these discrepancies are beyond the scope of this study. An overview of these differences was given in Chapter 2. In this study, the standards in the reformed mathematics curriculum are taken as the current vision of the knowledge and skills that should be mastered by Slovene students and the curriculum is the vehicle towards this vision. The above results show the outcomes of the non-reformed curriculum with regard to this vision.

Table 6.3 Numbers (and percentages) of strong, neutral and weak items in 1999

	Overall		Level 1 standard		Level 2 standard	
Strong	47	(31 %)	13	(19 %)	34	(41 %)
Neutral	55	(36 %)	28	(41 %)	27	(32 %)
Weak	51	(33 %)	28	(41 %)	23	(27 %)
Total	153	(100 %)	69	(100 %)	84	(100 %)

As indicated above, Table 6.3 presents the numbers and percentages of strong, weak, and neutral items as compared to the standards. At the level of the overall test, similar proportions of items were identified as strong, weak, or neutral. This shows that the correspondence of the test scores with the standards observed in Table 6.1 is a consequence of a balance in item difficulties; approximately as many as there were items for which difficulties were significantly lower than the standards, there were items for which difficulties were significantly higher than the standards.

When numbers of strong, weak, and neutral items are broken down by the levels of the standards, there seem to be differences between the two levels similarly as in student based analysis. Of the 69 items at Level 1, approximately a fifth were classified as strong. The remaining items were similarly distributed across the categories of neutral and weak items. At Level 2, somewhat less than a half of items were strong (41 %), approximately a third neutral (32 %), and approximately a quarter (27 %) weak. These observations coincide with observations from the student based analysis that average scores of Slovene students were lower than intended at Level 1 and higher than intended at Level 2. It further shows that the compensation between the strong and the weak items did not occur to an extent that large numbers of weak items would be masked by a few strong items indicating correspondence of average scores with the standards or even higher achievement.

More specifically, there were less strong items and more weak items at Level 1, than could be expected based on the results at the overall level ($p<0.05$[3]). As seen in Table 6.1, this resulted in achievement that was lower than intended. Although significantly higher average scores than intended were observed at Level 2 (Table 6.1), the difference between numbers of strong and weak items was not significant.

3 $Bin(41,p=0.5)$, $P(X<=13)<0.05$.

Table 6.4 Trends in numbers of strong, neutral and weak items between 1995 and 1999

Overall test

	Strong 1999	Neutral 1999	Weak 1999
Strong 1995	20	2	-
Neutral 1995	-	10	-
Weak 1995	-	1	15

Level 1 standard

	Strong 1999	Neutral 1999	Weak 1999
Strong 1995	7	1	-
Neutral 1995	-	8	-
Weak 1995	-	-	14

Level 2 standard

	Strong 1999	Neutral 1999	Weak 1999
Strong 1995	13	1	-
Neutral 1995	-	2	-
Weak 1995	-	1	1

In Table 6.4, results of item based analysis of trends are presented. Three cross-tabulations are presented, each showing the numbers of (identical) items that remained in, or changed 'status' of a strong, a weak or a neutral item between 1995 and 1999. Because the numbers of items are small, direct numbers are presented instead of percentages. In the cross-tabulation for the overall test, it can be seen that only three identical items changed from a strong or a weak item in 1995 to a neutral item in 1999. This shows that not only was there stability in the correspondence of the Slovene students' average scores with the standards between 1995 and 1999 (Table 6.2) but that this stability can also be observed at item level.
Table 6.4 shows that among the identical items in the two tests, the largest percentage of items (42 %) was strong, approximately a third was weak and the remaining fifth was neutral. Looking at the breakdown by the levels of the standards, at Level 1, approximately a half of items was weak, a quarter strong, and a quarter neutral. Of the three items that changed status, one was at Level 1, changing from a strong item in 1995 to a neutral item in 1999. At Level 2, nearly three quarters of items were strong. At this level, one strong and one weak item in 1995 became neutral items in 1999. The remaining identical items were strong, neutral or weak in both measurements.

These trend results therefore indicate stability in correspondence of item difficulties with the standards in the second half of the 1990s. Also, the correlation between item difficulties in 1995 and 1999 is 0.96. The results of item based analysis reveal similar differences in the correspondence of scores between the two levels of the standards as were observed in student based analysis. Using the results on the set of identical items, it seems plausible to conclude that Slovene overall achievement, as well as its correspondence with the standards, remained stable in the period between 1995 and 1999 (Table 6.2) and that stability was also observed in correspondences of individual item difficulties (Table 6.4). This was observed for the overall test, as well as for the respective levels of the standards.

In general, it seems that achievement of Slovene students corresponded with the standards at Level 2 while it was lower than the standards at Level 1, on average, as well as across individual items and that these strengths and weaknesses remained stable between 1995 and 1999.

6.2 LOCATION OF STRENGTHS AND WEAKNESSES IN CONTENT AREAS

In order to determine in which content areas strengths and weaknesses were located, the TIMSS items were classified according to the content areas in the Slovene mathematics curriculum (see Chapter 5). Following the design of the study, first students and then items are taken as units of analysis.

6.2.1 Students as units of analysis

In this subsection the following questions are addressed:

- *To what extent did achievement of Slovene students correspond with the standards when considering different content areas?*
- *Were the correspondences similar when considering different levels of the standards?*
- *To what extent did the correspondences in 1999 deviate from those in 1995?*

Trends with respect to the levels of the standards are not examined due to insufficient numbers of items in most categories. The results of analyses for the above questions are presented in Tables 6.5 and 6.6.

As shown in Table 6.5, average scores of Slovene students varied across content areas and across levels of the standards. In one cell in Table 6.5 the results are not shown and the content area 'probability' is not presented altogether due to insufficient number of items. It can be observed from Table 6.5 that correspondences of scores with the standards varied across content areas. At the overall level, a satisfying correspondence was observed for most content areas. On items about 'natural numbers', 'algebraic expressions', and 'data representation', average scores of Slovene students were even significantly higher than the standards. Lower than the standards were average scores in content areas 'meaning of rational numbers' and 'geometrical shapes'.

Table 6.5 Mathematics achievement of Slovene students in 1999 by content areas

	Overall			Level 1 standard			Level 2 standard		
	Number of items	Average percent correct		Number of items	Average percent correct		Number of items	Average percent correct	
Natural numbers	9	75 (0.8)	▲	8	77 (0.8)		1	-	
Meaning of rational numbers	20	64 (0.8)	▼	15	66 (0.8)	▼	5	59 (1.4)	▲
Operations with rational numbers	19	59 (1.0)		8	70 (1.2)	▼	11	52 (1.0)	
Algebraic expressions	25	61 (0.9)	▲	6	69 (1.1)	▼	19	58 (0.9)	▲
Functions and proportionality	21	54 (0.8)		5	69 (0.9)	▼	16	50 (0.9)	
Geometrical shapes	23	55 (0.9)	▼	8	55 (1.3)	▼	15	55 (0.9)	▲
Measurement	18	59 (0.8)		7	78 (0.9)	▲	11	48 (0.9)	
Data representation	15	71 (0.8)	▲	9	82 (0.6)	▲	6	56 (1.3)	▲

▼ Average percent correct significantly lower than the standard
▲ Average percent correct significantly higher than the standard
( ) Standard errors are presented in parentheses
\- Data are not shown due to insufficient number of items

These correspondences differed between the two levels of the standards for most content areas. As expected, scores were higher at Level 1 than at Level 2 in most content areas ($p<0.05$). An exception was the group of items about 'geometrical shapes' on which scores were identical at both levels of the standards.
Except for the content areas 'natural numbers', 'measurement', and 'data representation', students' scores at Level 1 were lower than the standards. In 'geometrical shapes', these scores were even 20 percentage points lower than the standards. In many of these content areas, average scores of Slovene students at Level 2 were higher than the standards and in no content area at this level they were lower.

This shows that the difference between the two levels observed in the overall scores was reflected also in most of the content areas of the curriculum. Although at the overall level indications of weaknesses were found for two content areas only, in quite a few content areas weaknesses at Level 1 seem to be compensated by strengths at Level 2. There also seem to be strong areas in overall, though. In 'data representation', significantly higher scores than intended were observed at both levels. Also, in 'measurement' significantly higher scores than the standards were observed at Level 1 and not at Level 2. In this content areas there seems to be a different pattern of achievement than in other areas.

Table 6.6 Trends in mathematics achievement of Slovene students between 1995 and 1999 by content areas

	Number of items	TIMSS 1995 Average percent correct		TIMSS 1999 Average percent correct	
Natural numbers	3	-		-	
Meaning of rational numbers	11	67 (0.9)	▼	67 (0.7)	▼
Operations with rational numbers	1	-		-	
Algebraic expressions	6	71 (0.8)	▲	72 (0.7)	▲
Functions and proportionality	8	67 (0.8)	▲	66 (0.9)	▲
Geometrical shapes	7	61 (0.9)	▲	61 (1.0)	▲
Measurement	5	73 (0.9)	▲	72 (0.8)	▲
Data representation	6	77 (0.7)	▲	78 (0.7)	▲

▼ Average percent correct significantly lower than the standard
▲ Average percent correct significantly higher than the standard
( ) Standard errors are presented in parentheses
- Data are not shown due to insufficient number of items

Results of analyses of trends in Slovene students' scores by content areas are presented in Table 6.6. For content areas 'natural numbers' and 'operations with rational numbers' there were insufficient numbers of items to present the data. It can be seen that the stability in overall scores observed in Table 6.2, is preserved across the content areas. For none of the content areas the differences in scores between 1995 and 1999 were significant. Furthermore, the correspondences with the standards also did not change although they were, for most content areas, different from those in Table 6.5 because of lower numbers of items. As mentioned, trends within the levels of the standards were not examined due to insufficient numbers of items.

6.2.2 Items as units of analysis

In this section, the distributions of strong and weak items across content areas will be examined. As described in the design, an item was classified as strong or weak, respectively, if the observed item difficulty was significantly higher or lower than the level of the standard for this item. The analysis in this section will be carried out only at the level of the overall test, since the breakdown by the levels of the standards and further into categories of strong, weak, and neutral items yielded insufficient numbers of items in most cases. Also, trends will not be examined in this subsection since basically no changes in correspondences of item difficulties were observed already in the previous section. The following question therefore remains to be addressed in this subsection:

- *In which content areas were strong and weak items located?*

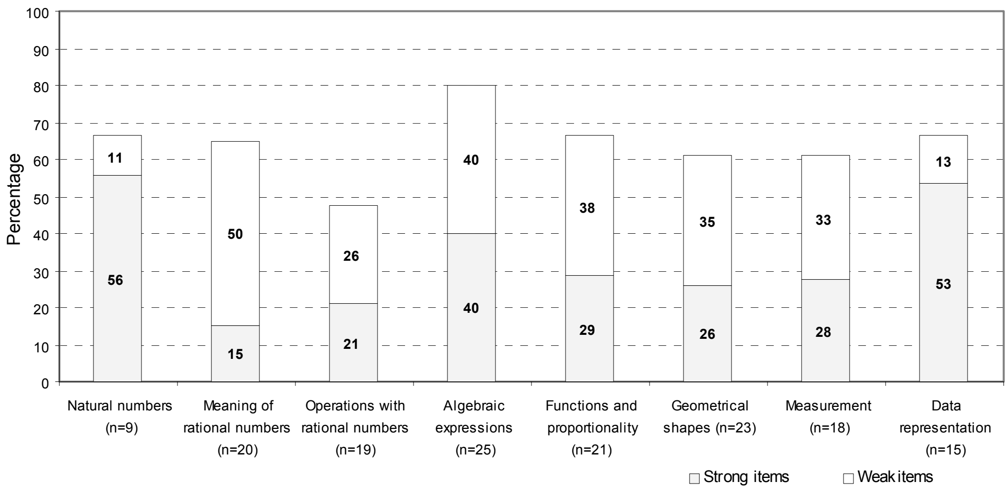


Figure 6.2 Percentages of strong and weak items in 1999 by content areas

The results of analyses for this question are presented in Figure 6.2. Because content areas were represented by different numbers of items, in Figure 6.2, the percentages of strong and weak items in individual content area are shown. It can be observed that in none of the content areas there were only strong or only weak items, but rather that the percentages of strong and weak items in content areas vary. In 'natural numbers', a single item (of total 9 items) was identified as weak and more than half of items were identified as strong. Similarly, more than half of items in 'data representation' was strong and less than a seventh was weak. If these observations are combined with observations from the student based analysis, it can be seen that the correspondences of students' average scores with the standards in these areas (Table 6.1) were a consequence of the correspondences that occurred over a number of items.

In contrast, a relatively small percentage of items about 'meaning of rational numbers' were strong, and a half of these items were weak. In this area, mostly weaknesses in Slovene achievement seem to be reflected in the TIMSS 1999 items. In the remaining content areas, there were similar percentages between strong and weak items. Among these, item difficulties in 'algebraic expressions' deviated from the standards for the highest percentage of items. While 40 % of items in this area were strong, 40 % of them were also weak. In contrast to 'natural numbers' and 'data representation', it seems that in this area, in spite of the average scores that corresponded to the standards (Table 6.1), a number of weak items could be found. At the same time, while average scores in 'geometrical shapes' were lower than the standards (Table 6.1), approximately a quarter strong items were found in this area. This shows that it is difficult to describe these content areas as strong or weak in general.

The above results of the student based and item based analyses indicate, that while in most content areas students' scores corresponded with the standards, in most a number of weak items were also found. This shows that general descriptions of these content areas as strong or weak are not warranted. As exception, indications of strengths were found in two content areas, 'natural numbers' and 'data representation's, while in 'meaning of rational numbers' it seems plausible to argue that weaknesses might have existed.

6.3 LOCATION OF STRENGTHS AND WEAKNESSES AT LEVELS OF COGNITIVE REQUIREMENTS

In this section, strengths and weaknesses in Slovene achievement will be located at levels of cognitive requirements. Each item was assigned to a category of cognitive requirements as explained in Chapters 4 and 5. As in the previous sections, strengths and weaknesses will be examined by first taking students and, second, items as units of analysis. The results of each of these two approaches are presented in the following two subsections.

6.3.1 Students as units of analysis

The results of the student based approach to the analysis at which levels of cognitive requirements the strengths and weaknesses can be located and of trends at the overall level of the standards are presented in Tables 6.7 and 6.8 below. Trends by the individual levels of the standards are, as for content areas, not examined due to insufficient data. The following questions are addressed by this analysis:

- *To what extent did achievement of Slovene students correspond with the standards when considering different cognitive categories?*
- *Were the correspondences similar when considering different levels of the standards?*
- *To what extent did the correspondences in 1999 deviate from those in 1995?*

Table 6.7 Mathematics achievement of Slovene students in 1999 by cognitive categories

	Overall			Level 1 standard			Level 2 standard		
	Number of items	Average percent correct		Number of items	Average percent correct		Number of items	Average percent correct	
Knowing	28	66 (0.7)	▲	16	69 (0.8)	▼	12	63 (0.8)	▲
Using routine procedures	37	65 (0.8)		20	73 (0.8)		17	55 (0.9)	▲
Using complex procedures	37	63 (0.7)	▲	16	74 (0.7)		21	55 (0.8)	▲
Investigating and solving problems	51	53 (0.8)	▼	17	64 (0.9)	▼	34	47 (0.8)	▼

▼ Average percent correct significantly lower than the standard
▲ Average percent correct significantly higher than the standard
( ) Standard errors are presented in parentheses

As can be observed from Table 6.7, attainment of Slovene students in 1999 was the highest on items that required 'knowing', lower on items that required 'using routine procedures' and 'using complex procedures', and the lowest on items that required 'investigating and solving problems'. Moreover, average scores in 'investigating and solving problems' were significantly lower than in the other three categories ($p<0.05$ for all comparisons).

At the level of the overall test, students' scores corresponded with the standards on items that required 'knowing', 'using routine procedures', and 'using complex procedures'. Their average scores were lower than the standards on the group of items that required 'investigating and solving problems'.
When looking at breakdown by the levels of the standards, at Level 1, correspondence with the standards was observed for 'using routine procedures' and 'using complex procedures'. It is interesting to observe that at this level, the average score for 'using routine procedures' was significantly higher than the average score for 'knowing' ($p<0.05$). At this level therefore, weaknesses in Slovene achievement seem to also be reflected in items with low cognitive requirements, although the scores at the overall level in this category were higher than the standards.
At Level 2, average scores of Slovene students were higher than the standards for most categories, except for 'investigating and solving problems' where they were significantly lower. However, this occurred by only three percentage points difference which perhaps could be seen as not critical. Compared to other categories, though, the scores in this category differed from the pattern.

On the basis of these results it seems plausible to argue that more general weaknesses in Slovene achievement as compared to the standards can be observed and that they are mostly reflected in items that were allocated at Level 1 at the highest and the lowest cognitive level. Scores in the categories 'using routine procedures' and 'using complex procedures' could be described as indicating strengths since they corresponded with the standards at both levels. Arguably, this contrast between the categories itself may not be desired in the educational outcomes in Slovenia.

Table 6.8 Trends in mathematics achievement of Slovene students between 1995 and 1999 by cognitive categories

	Number of items	TIMSS 1995 Average percent correct		TIMSS 1999 Average percent correct	
Knowing	16	72 (0.7)	▲	72 (0.6)	▲
Using routine procedures	8	69 (0.8)	▲	70 (0.8)	▲
Using complex procedures	13	68 (0.8)		69 (0.7)	▲
Investigating and solving problems	11	68 (0.7)	▲	67 (0.8)	▲

▼ Average percent correct significantly lower than the standa
▲ Average percent correct significantly higher than the standa
( ) Standard errors are presented in parentheses

In Table 6.8, trends in Slovene scores by cognitive categories are presented. As in previous analyses, stability can be observed in students' scores within respective cognitive categories between 1995 and 1999. A change in correspondence in 'using complex procedures' occurred due only to a slight change in scores. These results seem to indicate that there were no changes in student achievement in Slovenia with respect to the different cognitive levels. It seems plausible to argue that the strengths and weaknesses observed above also existed in 1995.

6.3.2 Items as units of analysis

The following questions are addressed in this subsection:

- *In which cognitive categories were the strong and weak items located?*
- *Were the locations similar when considering different levels of the standards?*

As for content areas, trends are not addressed in item based analysis for cognitive categories since basically no changes in strong, weak and neutral items were observed already at the level of the overall test.

Percentages of items that were identified as strong or weak in different cognitive categories are presented in Figure 6.3. It can be observed from Figure 6.3 that at the level of the overall test as well as at Levels 1 and 2 separately strong and weak items can be found in all cognitive categories. Furthermore, in most cognitive categories at the overall level the percentages of strong and weak items within categories are relatively similar, although in 'investigating and solving problems' somewhat more weak than strong items seem to be found (not significant). It seems plausible to argue that these results coincide with the observations in student based analysis where scores in 'investigating and solving problems' were lower than intended. It can be seen that such scores occurred over a number of items in this category.

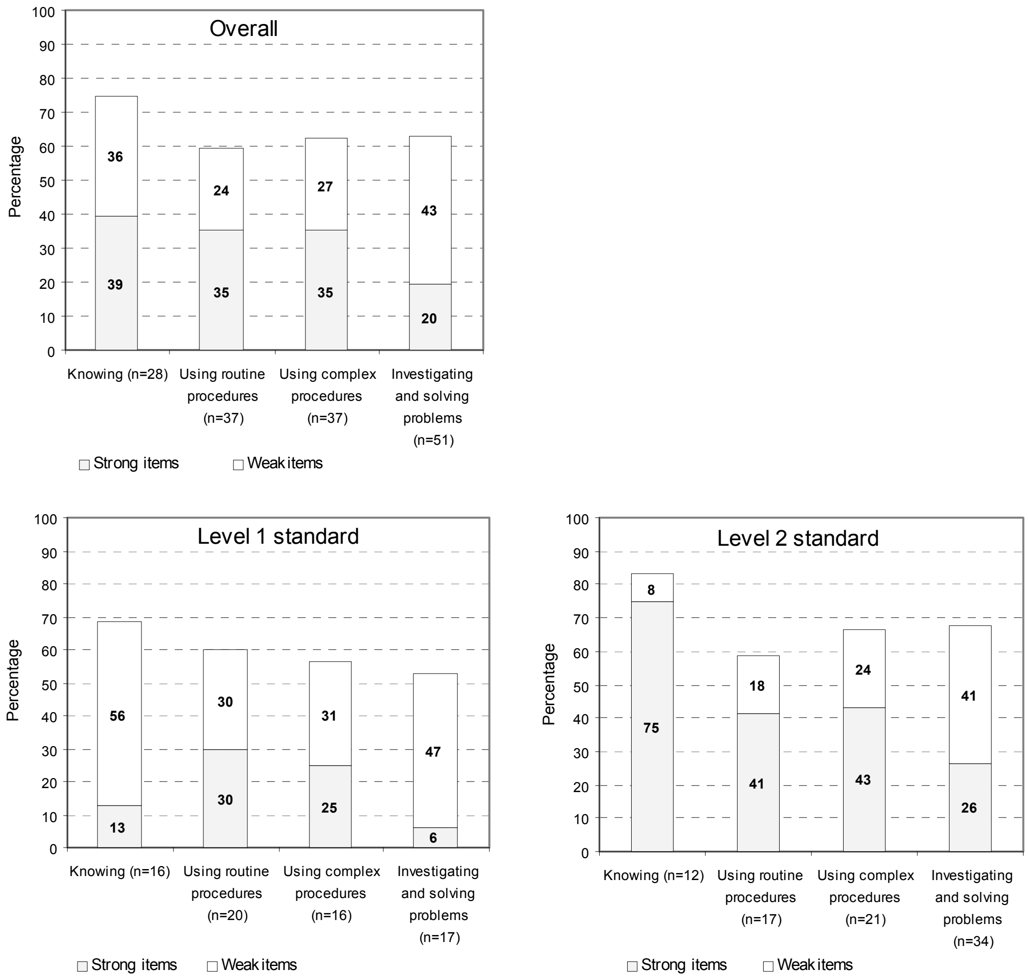


Figure 6.3 Percentages of strong and weak items in 1999 by cognitive categories and by the levels of the standards

When broken down by the levels of the standards, differences between the two levels observed in the previous analyses can also be observed within cognitive categories. At Level 1, in categories 'knowing' and 'investigating and solving problems', there seem to be larger percentages of weak than strong items ($p<0.05$ only for 'investigating and solving problems'[4]). At Level 2, there seem to be larger

4 Bin(9,p=0.5), P(X<=1)<0.05

percentages of strong than weak items in 'knowing' ($p<0.05$[5]). For this cognitive category the difference between Level 1 and Level 2 seems most striking with the percentages of strong and weak items at Level 1 seemingly reversed at Level 2.

On the basis of these observations it seems plausible to hypothesize that strengths in Slovene achievement were reflected in items that required 'using routine procedures' and 'using complex procedures' and in items that required 'knowing' at Level 2, while weaknesses were reflected in items that required 'investigating and solving problems' and in items that required 'knowing' at Level 1. This hypothesis could perhaps be tested in more detail if larger numbers of items measuring attainment of individual levels of the standards were available.

6.4 MOST NOTABLE STRENGTHS AND WEAKNESSES AS COMPARED TO THE ATTAINMENT TARGETS

In the previous sections it was observed that although average students' scores were higher or lower than the standards in some content areas, strong and weak items were found in all content areas (excluding 'probability' in which there is an insufficient number of items). Similarly, locating strengths and weaknesses in cognitive categories revealed higher average scores than intended in the lower cognitive categories and lower average scores than intended at the highest cognitive category, however, strong and weak items were found in all categories. This indicates that there may be other characteristics of items (as well as other background variables) influencing students' scores. Following the design of this study, to shed more light on the strengths and weaknesses observed in Slovene achievement, individual items will be examined. Items for which Slovene achievement was at least 10 percentage points above or below the operationalized intended attainment will be reviewed (this difference was significant for all items). These items are called *particularly strong* and *particularly weak*, respectively. They are presented and discussed in the following two subsections.

6.4.1 Most notable strengths

Following the generic structure of the operational research questions in Chapter 5, the following question is addressed in this subsection:

- *What are the contents and characteristics of the particularly strong items when compared to the standards?*

[5] $Bin(10,p=0.5)$, $P(X<=1)<0.05$.

In Table 6.9, particularly strong items for Slovenia when compared to the standards are presented. There are 37 items in this table ordered by content area and cognitive requirements (24 % of items in the analysis). As observed from Table 6.9, approximately a fifth of particularly strong items wes allocated at Level 1 (7 items) and the remaining items were allocated at Level 2. This coincides with observations in the previous sections that strengths were mostly reflected in items at Level 2.
All content areas are represented in Table 6.9. The largest number of particularly strong items was about 'algebraic expressions' (9 items) and the lowest number of these items was found in 'meaning of rational numbers' (2 items). Of particularly strong items at Level 1, three are 'data representation' items and two are about 'natural numbers', the content areas in which indications of strengths at this level were found in the previous sections.
Cognitive categories also are approximately equally represented by particularly strong items. Except for four items (M08, L17, V04B, and V02_1), other items in Table 6.9 were in multiple choice format. For illustration, some of these items are presented in Figure 6.4.

The item with the highest difficulty in Table 6.9 required rounding a whole number to the nearest hundred (N11, natural numbers). Nearly all Slovene students (94 %) answered this item correctly. Similarly, at least 90 % of Slovene students correctly answered items that required identifying a sum that is closest to a given number expression (H09, natural numbers), subtract two decimal numbers (R07, operations with rational numbers), identifying best unit to measure the mass of an egg (D11, measurement), and read a table to find required information (P16, data representation). Slovene students were also able to perform high on the remaining five 'data representation' items in Table 6.9, which also required interpreting data in charts in order to find required information. These results are interesting since there were no 'data representation' topics in the Slovene intended curriculum until late 1990s (see Chapter 2). In 'measurement' in addition to item D11, particularly strong items required determining measurement accuracy of ruler (F10), identifying angle closest to 45 degrees in a circle (N15), and estimate actual length of building compared to length of car (L09).

Table 6.9 Particularly strong items when compared to the standards in 1999

Intended standard	Content area	Cognitive category	Item content	Item label	Item format	Average percent correct
Level 1	Natural numbers	Using routine procedures	Round 17175 to nearest hundred.	N11	MC	94
Level 2	Natural numbers	Using routine procedures	Approximate number of magazines sold/year from sold/week.	Q06	MC	64
Level 1	Natural numbers	Using complex procedures	Identify sum which is closest to 691 + 208.	H09	MC	92
Level 2	Meaning of rational numbers	Knowing	Identify list of equivalent fractions.	N14	MC	65
Level 2	Meaning of rational numbers	Using complex procedures	Identify the smallest fraction.	M04	MC	78
Level 1	Operations with rational numbers	Using routine procedures	Subtract two decimal numbers to 0.001.	R07	MC	90
Level 2	Operations with rational numbers	Using routine procedures	Multiply two decimals.	M08	SA	62
Level 2	Operations with rational numbers	Using complex procedures	Find height of stack from paper thickness	T04	MC	62
Level 2	Algebraic expressions	Knowing	Identify algebraic equation representing relationship: 7n+6=41.	B12	MC	83
Level 2	Algebraic expressions	Knowing	Select symbolic linear equation which fits word problem about magazines.	H12	MC	81
Level 2	Algebraic expressions	Knowing	Identify expression equivalent to n x n x n.	P09	MC	80
Level 2	Algebraic expressions	Knowing	Identify equivalent algebraic expression for k+k+k+k+k.	P11	MC	72
Level 2	Algebraic expressions	Using routine procedures	Simplify and solve linear equation for X.	L17	SA	76
Level 2	Algebraic expressions	Using routine procedures	Solve linear equation for X.	O07	MC	73
Level 2	Algebraic expressions	Investigating and solving problems	Solve algebraic word problem equating masses on a scale.	A02	MC	84
Level 2	Algebraic expressions	Investigating and solving problems	Determine number of matchsticks which continues pattern based on figures.	C05	MC	61
Level 2	Algebraic expressions	Investigating and solving problems	Extend circle pattern from 4 figures to 7th.	V04B	ER	66
Level 2	Functions and Proportionality	Using routine procedures	Solve ratio problem for x: 7/13 = x/52.	D08	MC	77
Level 2	Functions and Proportionality	Using routine procedures	Identify missing number in table showing xy relationship.	J17	MC	73
Level 2	Functions and Proportionality	Using complex procedures	Approximate tons of fertilizer sold increased by 15%.	K06	MC	75
Level 2	Functions and Proportionality	Investigating and solving problems	Predict number of total defective bulbs based on random sample.	H11	MC	71
Level 2	Functions and Proportionality	Investigating and solving problems	Solve word problem involving ratios: find true statement about number of men/women at meeting.	Q05	MC	72
Level 2	Geometrical shapes	Knowing	Identify similar triangles of different size.	J15	MC	66
Level 2	Geometrical shapes	Using routine procedures	Find center point of rotation for symmetry-related rectangles.	O08	MC	77
Level 2	Geometrical shapes	Using complex procedures	Identify cube made by folding 2-dimensional net.	B11	MC	68
Level 2	Geometrical shapes	Investigating and solving problems	Identify relationship of angles in symmetric polygon.	D07	MC	66
Level 2	Geometrical shapes	Investigating and solving problems	Find angle based on supplementary pairs.	M07	MC	70
Level 1	Measurement	Knowing	Identify best unit to measure the mass of an egg.	D11	MC	93
Level 2	Measurement	Knowing	Determine measurement accuracy of ruler.	F10	MC	75
Level 2	Measurement	Knowing	Identify angle closest to 45 degrees in a circle.	N15	MC	61
Level 2	Measurement	Using complex procedures	Estimate actual length of building compared to length of car.	L09	MC	82

MC multiple choice item
SA short answer item
ER extended response item

Table 6.9 Particularly strong items when compared to the standards in 1999 (Continued)

Intended standard	Content area	Cognitive category	Item content	Item label	Item format	Average percent correct
Level 1	Data representation	Using routine procedures	Interpret distance/time graph to determine intersection point of two plots.	E01	MC	88
Level 1	Data representation	Using complex procedures	Identify day/time in table at shown temperature.	P16	MC	91
Level 1	Data representation	Using complex procedures	Interpret data in bar graph: number of pencils sold.	Q04	MC	88
Level 2	Data representation	Using complex procedures	Interpret line graph to find region of largest increase.	B07	MC	76
Level 2	Data representation	Using complex procedures	Interpret time/length of string line graph to determine time for pendulum to swing 20 times.	R09	MC	63
Level 2	Data representation	Investigating and solving problems	Solve word problem: determine cheaper magazine subscription by unit comparison.	V02_1	ER	63
Total	37 items					

MC multiple choice item
SA short answer item
ER extended response item

Three of the particularly strong items regarding 'algebraic expressions' required 'investigating and solving problems' however, it could be said that they were not generally difficult. Item that required solving algebraic word problem equating masses on a scale (A02) could have been solved either using a formal approach by constructing an algebraic equation or by an informal approach that students in Slovenia practice already in the lower grades. It might be hypothesized that because the item included a graphical representation of the equation that was to be constructed, it was not difficult even for students choosing the formal approach. Also, the other two items required knowledge of patterns and they could have been answered using 'common knowledge'.

All the remaining particularly strong items in 'algebraic expressions', although they were allocated at Level 2, were answered as high as required at Level 1. More specifically, difficulties for these items did not differ significantly from the intended 75 % at Level 1. This can also be observed for five particularly strong items in 'functions and proportionality' for which it seems plausible to hypothesize that they, except Q05, could also have been answered using algebra skills. These results may therefore be interpreted as showing that Slovene students in the non-reformed system mastered algebra skills required by these items to a higher extent than intended in the reformed curriculum.

Two among the particularly strong items in Table 6.9 could be described as not very familiar to Slovene students through their intended curriculum for mathematics, although they were considered covered by the attainment targets. These are items that required predicting number of total defective bulbs based on a random sample (H11, functions and proportionality), and identifying cube made by folding a two-dimensional net (B11, geometrical shapes). It seems plausible to argue that Slovene students answered these two items by using 'common knowledge' in addition to what they learned in school.

L17. Find the value of x if $12x - 10 = 6x + 32$.

Answer: ________________________

N11. A company produced 17 175 cars in 1996. For a report, this number was rounded to the nearest hundred. Which was the number of cars given in the report?

A. 17 000

B. 17 100

C. 17 200

D. 17 270

H11. From a batch of 3000 light bulbs, 100 were selected at random and tested. If 5 of the light bulbs in the sample were found to be defective, about how many defective light bulbs would be expected in the entire batch?

A. 15

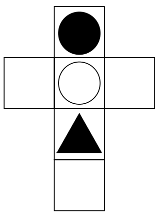
B. 60

C. 150

D. 300

E. 600

B11.



Which of these cubes could be made by folding the figure above?

A. B. C. D.

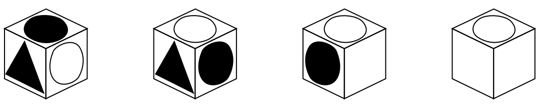


Figure 6.4 Examples of particularly strong items as compared to the attainment targets

6.4.2 Most notable weaknesses

In this subsection, the following question is addressed:

- *What are the contents and characteristics of the particularly weak items when compared to the standards?*

Similarly as for strengths, particularly weak items as compared to the standards were sought among the TIMSS 1999 items. They are presented in Table 6.10. As described in Chapter 4, for some of the TIMSS items partial responses were also scored. Such items had two score levels (only one TIMSS 1995 item had three score levels) and they were generally more difficult than other items. As explained in Chapter 5, an adaptation of percent correct approach was used to compute percents correct on such items. The labels of such items indicate the score level for which the percents correct were computed. For example, the percentage of students that received at least 1 point on item U01 (the difficulty of the score level 1) is presented as U01_1. The percentage of students that received 2 points on this item (the difficulty of the score level 2) is presented as U01_2. As explained in Chapter 5, such items were included in analysis in the form of these partial correct variables. There were six such items in TIMSS 1999, however for reasons mentioned in Chapter 5, not all were included in this analysis.

In Table 6.10, several items with two score levels can be observed. Some of these items were found as particularly weak only at the highest score level (e.g., T01_2, 'algebraic expressions') and some at both score levels (e.g., U01_1 and U01_2, 'measurement').

The number of particularly weak items when compared to the attainment targets equals the number of particularly strong items (37 items). The particularly weak items are somewhat differently distributed across content areas and cognitive categories, still there are all content areas and all cognitive levels represented in Table 6.10. There was a larger number of particularly weak items at Level 1 in Table 6.10 than was the number of particularly strong items at this level in Table 6.9. This coincides with observations from the previous sections that weaknesses were found mostly at Level 1. The largest numbers of particularly weak items were found in content areas 'geometrical shapes' (8 items) and 'functions and proportionality' (7 items). There were also six items about 'algebraic expressions' and six about 'meaning of rational numbers'. A half of particularly weak items required 'investigating and solving problems' (18 items) and the largest percentage of the remaining items required 'using complex procedures' (8 items). Some of these items are presented in Figure 6.5.

Table 6.10 Particularly weak items when compared to the standards in 1999

Intended standard	Content area	Cognitive category	Item content	Item label	Item format	Average percent correct
Level 1	Natural numbers	Using complex procedures	Write two possibilities for actual height from rounded value.	V01	SA	41
Level 1	Meaning of rational numbers	Knowing	Identify smallest of decimal fractions.	B10	MC	58
Level 1	Meaning of rational numbers	Knowing	Identify fraction of circle which is shaded.	F12	MC	59
Level 1	Meaning of rational numbers	Knowing	Identify two hundred six and nine-tenths.	L10	MC	61
Level 1	Meaning of rational numbers	Knowing	Shade in 3/8 of squares in grid.	N19	SA	55
Level 1	Meaning of rational numbers	Using routine procedures	Reduce decimal to fraction in lowest terms.	P17	SA	48
Level 1	Meaning of rational numbers	Using complex procedures	Determine actual length of box which is rounded to 9 cm.	A03	MC	61
Level 1	Operations with rational numbers	Using routine procedures	Subtract three fractions.	L18	MC	63
Level 1	Operations with rational numbers	Investigating and solving problems	Determine how much money left after spending 5/8.	R14	SA	35
Level 2	Operations with rational numbers	Investigating and solving problems	Solve multi-step word problem with decimal fractions: amount of paint left.	N17	MC	36
Level 2	Operations with rational numbers	Investigating and solving problems	Solve multi-step word problem with decimals: amount of fuel left.	S03	MC	31
Level 2	Algebraic expressions	Using complex procedures	Identify linear expression based on word problem: number of books.	Q01	MC	39
Level 1	Algebraic expressions	Investigating and solving problems	Complete chart by observing sequence of triangles in each of 3 figures.	S01A	SA	56
Level 2	Algebraic expressions	Investigating and solving problems	Determine next number which is common to two number sequences.	I04	SA	34
Level 2	Algebraic expressions	Investigating and solving problems	Solve algebra word problem with fractions: number of marbles in bag to start with.	N16	MC	39
Level 2	Algebraic expressions	Investigating and solving problems	Extend circle pattern from 3 figures to 7th.	S01B	ER	30
Level 2	Algebraic expressions	Investigating and solving problems	Solve multi-step algebra word problem of simultaneous equations: boy and girls from total club members.	T01_2	ER	37
Level 1	Functions and Proportionality	Using routine procedures	Find relationship between numbers in a set of ordered pairs.	E05	MC	63
Level 2	Functions and Proportionality	Using complex procedures	Calculate ratio of rectangle areas given relationship between sides.	U02B_1	ER	32
Level 2	Functions and Proportionality	Using complex procedures	Calculate ratio of rectangle areas given relationship between sides.	U02B_2	ER	29
Level 1	Functions and Proportionality	Investigating and solving problems	Find 1/3 of number from relationship	F11	MC	54
Level 2	Functions and Proportionality	Investigating and solving problems	Find ratio between side and perimeter for given rectangle.	P08	MC	21
Level 2	Functions and Proportionality	Investigating and solving problems	Solve algebra word problem involving ratios: number of boxes with 12 books.	T02A	ER	31
Level 2	Functions and Proportionality	Investigating and solving problems	Solve multi-step word problem computing fractions: fraction of smaller boxes.	T02B	ER	11

MC multiple choice item
SA short answer item
ER extended response item

Table 6.10 Particularly weak items when compared to the standards in 1999 (Continued)

Intended standard	Content area	Cognitive category	Item content	Item label	Item format	Average percent correct
Level 1	Geometrical shapes	Knowing	False statement of congruent triangles in a rectangle	A05	MC	63
Level 1	Geometrical shapes	Using routine procedures	Find perimeter of whole figure containing 5 squares of known size.	S02C	ER	35
Level 1	Geometrical shapes	Using complex procedures	Determine into how many right triangles a rectangle can be divided.	R11	MC	57
Level 2	Geometrical shapes	Using complex procedures	Determine unknown angle in rotated congruent triangles.	K08	MC	39
Level 2	Geometrical shapes	Using complex procedures	Find measure of angle by calculation involving addition/subtraction of adjacent angles.	Q10	SA	32
Level 1	Geometrical shapes	Investigating and solving problems	Find area of paved walkway around pool	I07	MC	49
Level 1	Geometrical shapes	Investigating and solving problems	Determine area of path around garden.	J10	MC	46
Level 1	Geometrical shapes	Investigating and solving problems	Find area of triangle inside a square	U03	SA	43
Level 2	Measurement	Using routine procedures	Draw new rectangle based length/width ratios of another rectangle.	U02A_2	ER	26
Level 2	Measurement	Investigating and solving problems	Estimate the time for water to cool.	U01_1	ER	26
Level 2	Measurement	Investigating and solving problems	Explain method of estimation (Time for water to cool).	U01_2	ER	23
Level 1	Data representation	Knowing	Interpret test score frequency table to find number above a given level.	G01	MC	64
Level 2	Data representation	Investigating and solving problems	Solve word problem: determine cheaper magazine subscription by unit comparison.	V02_2	ER	36
Total	37 items					

MC multiple choice item
SA short answer item
ER extended response item

Item difficulties of particularly weak items ranged from 64 percent correct for interpreting test score frequency table to find number above given level (G01, data representation) to 11 percent correct on a multi-step problem (T02B, functions and proportionality, see Figure 6.5). The non-correspondence with the standards for item G01 may be explained by the presence of a distracter in the available responses that gave a number of at least given level instead of above given level (24 % of Slovene students chose this answer). The non-correspondence with the standards for item T02B and, at the same time, for the low score on this item may be explained by the fact that students had to first find the answer to the first part of the task (T02A) in order to answer the second part (T02B). Item T02A however, was also found as particularly weak with 31 percent correct.

Identifying the smallest fraction (M04, meaning of rational numbers) was observed as particularly strong, in Table 6.9. However, identifying the smallest decimal fraction (B10, meaning of rational numbers, see Figure 6.5) was observed as particularly weak, in Table 6.10. This difference can also be explained by the choices of answers that were presented to students. While in M04 'easily comparable' fractions were given as choices, there was a strong distracter in B10 (the number 0.5 might have been identified as smaller than 0.125, 15 % of Slovene students chose this distracter as the correct answer).

B10. Which of these is the smallest number?

A. 0.625

B. 0.25

C. 0.375

D. 0.5

E. 0.125

N17. A painter had 25 L of paint. He used 2.5 L of paint every hour. He finished the job in 5.5 hours. How much paint did he have left?

A. 10.25 L

B. 11.25 L

C. 12.75 L

D. 13.75 L

T02. A book publisher sent 140 copies of a certain book to a bookstore. The publisher packed the books in two types of boxes. One type of box held 8 copies of the book, and the other type of box held 12 copies of the book. The boxes were all full, and there were equal numbers of both types of boxes.

a) How many boxes holding 12 books were sent to the bookstore?

Answer: ________________

b) What fraction of the books sent to the bookstore were packed in the smaller boxes?

Answer: ________________

P08. The rectangle below is twice as long as it is wide.

What is the ratio of the ctangle to its perimeter?

A. $\frac{1}{2}$

B. $\frac{1}{3}$

C. $\frac{1}{4}$

D. $\frac{1}{6}$

Figure 6.5 Examples of particularly weak items as compared to the attainment targets

In 'operations with rational numbers', particularly weak items were mostly multi-step word problems (R14, N17, S03) in addition to an item that required subtracting three fractions (L18). Again, it seems that many students simply forgot to perform the second subtraction (28 % of students). Similarly, solving word problems or problems in unusual settings such as finding ratio between side and perimeter of a rectangle (P08) was required by particularly weak items in 'functions and proportionality'. In P08, 53 % of students chose distracter A (see Figure 6.5).

In general, in contrast to particularly strong items in Table 6.9, most of particularly weak items in Table 6.10 were word problems that were presented in an unusual setting (e.g., L11, A05, I04), related several mathematical concepts within the same task (e.g., U02B, P08), or required multi-step procedures (e.g., R14, J10, N17, N16). It seems plausible to hypothesize that these characteristics influenced lower student scores on these items.

6.5 ANSWERS TO THE FIRST RESEARCH QUESTION

In this chapter, the first research question was addressed. This research question examined correspondence of Slovene achievement in mathematics with the attainment targets in the reformed curriculum. The approaches to the analyses in this chapter were based on students as units of analysis and followed by analysis in which items were taken as units. The correspondences of Slovene achievement with the attainment targets were examined at the level of the overall mathematics domain, in content areas, cognitive categories, and at the level of individual items. Correspondences were examined also for different levels of the attainment targets as well as in the period between 1995 and 1999.

It was observed that, at the overall level, Slovene achievement corresponded with the standards. However, within this correspondence weaknesses at Level 1 were masked by strengths at Level 2. This was observed also in most content areas. Average scores of Slovene students corresponded with the standards at both levels in content areas 'measurement' and 'data representation', and in 'natural numbers' at Level 1 (there were no items at Level 2 in this area). These areas can be described as corresponding with the standards in the curriculum in general. The result for 'data representation' is especially interesting since there were no data representation topics in the Slovene curriculum until the late 1990s.

When achievement in cognitive categories was examined, it was observed that scores in the middle categories 'using routine procedures' and 'using complex procedures' corresponded with the standards at both levels and scores in the lowest category 'knowing' corresponded with the standards at Level 2. In the remaining category 'investigating and solving problems', and in 'knowing' at Level 1, average scores of Slovene students were lower than the standards.

The review of items that were particularly strong or particularly weak when compared to the standards indicated some possible explanations for the observed results. In all analyses no significant changes in student achievement and therefore in its correspondence with the standards were observed between 1995 and 1999.

CHAPTER 7
Correspondence of mathematics achievement in Slovenia with achievements in the reference countries

In this chapter, mathematics achievement of Slovene students will be described from the perspective of achievements of students from other European countries. As explained in Chapter 5, for this purpose four other European countries were selected as reference: Belgium Flemish, the Netherlands, Hungary, and the Slovak Republic. This group includes two educational systems from the European Union and two from the countries that are, as Slovenia, in the process of accession to the association in May 2004. The structure of the analyses of achievement from this perspective is similar as in the previous chapter. After the strengths and weaknesses are described at the level of the overall test, their locations within the content areas in the Slovene mathematics curriculum and at the levels of cognitive requirements are sought. Within these steps, students are taken as units of analysis followed by analysis in which items are taken as units. Particularly strong and particularly weak items when compared to the reference countries are also reviewed. Finally, sets of strengths and weaknesses in Slovene achievement that were identified from the two perspectives are compared.

In the first section (7.1), the results of student and item based comparisons of Slovenia to the reference countries at the level of the overall achievement test are presented. In section 7.2, the two approaches are employed for locating the strengths and weaknesses in content areas; and in cognitive categories in section 7.3. Section 7.4 reviews the most notable strengths and weaknesses identified in mathematics achievement of Slovene students as compared to the reference countries. In the final section (7.5), the convergence in the descriptions of Slovene mathematics achievement obtained from the two perspectives is examined.

7.1 CORRESPONDENCE AT THE LEVEL OF THE OVERALL MATHEMATICS DOMAIN

Following the decisions made in the previous chapter, the TIMSS 1999 database will be used for most analyses conducted in this chapter. In the previous chapter, no changes in student achievement, as well as no changes in strong and weak items were observed in Slovenia between 1995 and 1999. However, it is of interest to see how this trend in Slovenia compares to trends in other countries. There may have been changes in achievements of students from the countries that were selected as reference. In consequence, the descriptions of strengths and weaknesses in Slovene achievement from the perspective of these countries might change between 1995 and 1999. For this reason, trends will be examined in this chapter also in item based analyses employing a set of identical items in the two measurements.

7.1.1 Students as units of analysis

The following questions are addressed in this subsection:

- *To what extent did achievement of Slovene students at the end of compulsory education in 1999 correspond to achievements of students from the reference countries?*
- *To what extent did the correspondences in 1999 deviate from those in 1995?*

The results of the analyses for these questions are presented in Tables 7.1 and 7.2.

Table 7.1 Mathematics achievement of Slovene students in 1999 compared to the reference countries

	Number of items	Slovenia	Belgium Flemish	Netherlands	Hungary	Slovak Republic
TIMSS 1999	153	61 (0.7)	67 (0.8) ▲	61 (1.8)	62 (0.9)	63 (1.1)

In table, estimates of average percent correct are reported
▼ Average percent correct significantly lower than in Slovenia
▲ Average percent correct significantly higher than in Slovenia
( ) Standard errors are presented in parentheses

As shown in Table 7.1, achievements in Slovenia and in the four reference countries ranged from 61 percent correct in Slovenia and in the Netherlands, to 67 percent correct in Belgium Flemish. The overall scores of students from these countries were similar to Slovenia. These similarities were expected and can be explained as a consequence of the selection process of the reference countries (see Chapter 5). Only in Belgium Flemish was achievement of students in 1999 significantly higher than in Slovenia. As mentioned in Chapter 5, this was observed also by Mullis et al. (2000) using IRT scores.

Table 7.2 Trends in mathematics achievement of Slovene students between 1995 and 1999 compared to the reference countries

	Number of items	Slovenia	Belgium Flemish		Netherlands	Hungary		Slovak Republic
TIMSS 1995	48	69 (0.7)	73 (1.3)	▲	70 (1.6)	67 (0.8)	▼	69 (0.7)
TIMSS 1999	48	70 (0.6)	76 (0.7)	▲	74 (1.6)	68 (0.8)		69 (0.9)

In table, estimates of average percent correct are reported
▼ Average percent correct significantly lower than in Slovenia
▲ Average percent correct significantly higher than in Slovenia
( ) Standard errors are presented in parentheses

Trends in students' scores in Slovenia and the reference countries are presented in Table 7.2. Again, similarities with Slovenia can be observed for most countries and, in most cases, these similarities remained stable between 1995 and 1999. In no country were changes between 1995 and 1999 significant. For Slovenia, this was already observed in Chapter 6. Although the average score in 1999 in the Netherlands was 4 percentage points higher than in 1995, this difference was not significant and also did not influence the correspondence between the average score in Slovenia and in the Netherlands (considering relatively large standard errors of estimates for the Netherlands). The average score in Slovenia was also similar to the average score in the Slovak Republic in 1995, as well as in 1999. The average scores in Belgium Flemish were significantly higher than in Slovenia in both measurements. The change in correspondence of the scores with Hungary occurred due only to a slight change in achievement[1].

In general, it can be argued that average scores in Slovenia remained similar to the scores in the reference countries in the period between 1995 and 1999. Future assessments may show whether slightly unfavorable, but non-significant, shifts in average scores for Slovenia in comparisons with Belgium Flemish and the Netherlands are indications of a lack of progress, or are merely a consequence of chance fluctuations.

7.1.2 Items as units of analysis

Following the procedure for item based analyses in this study as described in Chapter 5, an item was identified as strong in comparison with a particular reference country if the observed item difficulty for Slovenia was significantly higher than in the reference country. Similarly, an item was identified as weak in comparison with a particular country, if the observed item difficulty in Slovenia was significantly lower than in the reference country. The following questions are addressed in this subsection:

[1] As mentioned in Chapter 6, inconsistencies in tables may occur because rounded estimates are presented.

- *To what extent did item difficulties for Slovenia correspond to those in the reference countries?*
- *To what extent did the correspondences in 1999 deviate from those in 1995?*

The results of analysis for the first question are presented in Figure 7.1 and Table 7.3, and for the second question in Table 7.4.

In Figure 7.1, correspondences of the Slovene item difficulties with the difficulties in the reference countries are presented. The 'middle' thick line presents the Slovene item difficulties by which items are also ordered in this figure. The darker thin line generally above the Slovene difficulties presents for each item the highest difficulty in the four reference countries. The lighter thin line generally below the Slovene difficulties, presents the lowest item difficulty in the four reference countries.

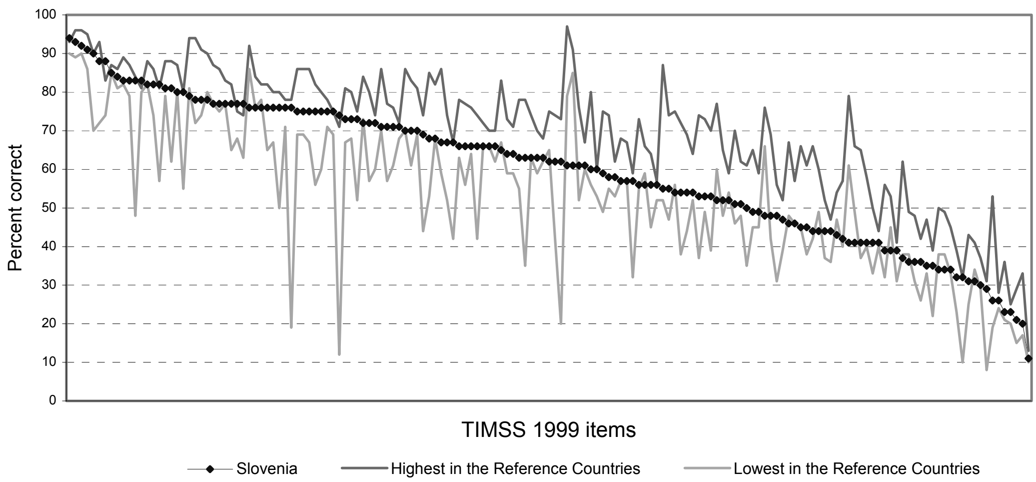


Figure 7.1 Correspondences of item difficulties for Slovenia with those for the reference countries in 1999

Generally, Slovene difficulties are in between the highest and the lowest item difficulty. However, the orderings of the highest and the lowest difficulties in the reference countries considerably deviate from the ordering of the Slovene difficulties. For several items that had high difficulty in Slovenia, there were considerably lower difficulties in at least one of the reference countries. For several items that had low difficulty in Slovenia, there were considerably higher difficulties in at least one of the reference countries. It can also be observed that, for some items, the Slovene difficulty was the lowest and, for some items, it was the highest, although the latter did not occur as often. This is understandable since a significantly higher achieving country is included in comparisons (Belgium Flemish, see Table 7.1) while no significantly lower achieving countries are included.

However, it can be observed that there are a few items for which Slovene difficulties were considerably lower than the lowest difficulty in the reference countries. It seems plausible to hypothesize that weaknesses in Slovene achievement are reflected in these items. They will be reviewed later in this chapter.

Table 7.3 Numbers (and percentages) of strong, neutral and weak items for Slovenia when compared to the reference countries in 1999

	Strong	Neutral	Weak	Total
Belgium Flemish	8 (5 %)	92 (60 %)	53 (35 %)	153 (100 %)
Netherlands	22 (14 %)	101 (66 %)	30 (20 %)	153 (100 %)
Hungary	14 (9 %)	121 (79 %)	18 (12 %)	153 (100 %)
Slovak Republic	13 (8 %)	123 (80 %)	17 (11 %)	153 (100 %)
In at least two comparisons	13 (8 %)	106 (69 %)	34 (22 %)	153 (100 %)

Table 7.3 shows numbers of strong, weak, and neutral items in comparison with the reference countries. In parentheses, percentages of these items relative to the total number of items are given. As shown in Table 7.3, in comparison with all four reference countries, the observed item difficulties for Slovenia were similar to those in the reference countries for at least half of items. In comparison with Belgium Flemish, eight items (5 %) were identified as strong and 53 items (38 %) were identified as weak for Slovenia. That the number of strong items is relatively low and the number of weak items is relatively high is again plausible due to the significantly higher overall achievement in Belgium Flemish (see Table 7.1).
In comparison with the Netherlands, 22 items (14 %) were identified as strong and 30 items (20 %) as weak for Slovenia. In comparison with the two candidate countries, Hungary and the Slovak Republic, lower numbers of strong and weak items can be observed. Also in comparison with these two countries, numbers of weak items are only slightly larger than the numbers of strong items. It can be said that there is a balance in these numbers and therefore, a balance in items for which the Slovene difficulties significantly deviated from the difficulties in each of these two countries. It is therefore interesting to observe that there are larger discrepancies in these numbers between Slovenia and the Netherlands even though the average score in the Netherlands was also similar to the average score in Slovenia.

For all countries, however, the numbers of weak items were (at least slightly) larger than the numbers of strong items. Although this did not affect the comparisons of average scores with these countries, except for Belgium Flemish, it might be interesting for the perspective of curriculum development and implementation.

As an additional indicator of correspondences between item difficulties in Slovenia with those in the reference countries, correlations between these difficulties were computed. Correlations for Belgium Flemish, Hungary and the Slovak Republic were all above 0.85, while the correlation for the Netherlands seems to be lower (r=0.72), showing less alignment between items that had high and low difficulties in Slovenia and those that had high and low difficulties, respectively, in the Netherlands.

From Table 7.3 it can be seen that each reference country provides 'its own' set of strong and weak items for Slovenia. Following the design of this study, additional criterion for classifying an item strong or weak was used. By this criterion, an item was classified as strong if it appeared strong in comparison with at least two reference countries. Similarly, an item was classified as weak if it appeared weak in at least two comparisons. In this way, 34 items in the TIMSS 1999 achievement test were classified as weak and 13 items strong for Slovenia as also shown in Table 7.3. One item, F11, appeared strong in comparison with two countries (the Netherlands and the Slovak Republic) and at the same time weak in the other two comparisons (with Belgium Flemish and Hungary). This item was classified as neutral. In this way, approximately a fifth of items was classified as weak for Slovenia and approximately a tenth as strong. The weak items may be used to investigate in which areas improvements are possible in Slovene achievement, while the strong items indicate where Slovene achievement is already relatively high.

As indicated in the questions to be addressed in this subsection, trends in correspondences of item difficulties for Slovenia with those in the reference countries were also investigated. The results are presented in Table 7.4. For each reference country, a cross table of the numbers of identical items that were classified as strong, neutral, or weak in each of the two measurements are presented. Because of small numbers of items, direct numbers instead of percentages are presented.

The first general observation from Table 7.4 is that, in comparisons with all reference countries, the majority of items were neutral in both measurements. Furthermore, in no comparison were there items that changed from strong to weak, or from weak to strong. From this it seems plausible to argue that no major changes in comparisons between Slovenia and the reference countries occurred. This statement can be further supported by examining the correlation between item difficulties within each country at the two occasions. These correlations were all above 0.93, and, as already mentioned in Chapter 6, this correlation for Slovenia was 0.96.

Table 7.4 Trends in numbers of strong, neutral and weak items for Slovenia between 1995 and 1999 when compared to the reference countries

Belgium Flemish

	Strong 1999	Neutral 1999	Weak 1999
Strong 1995	1	4	-
Neutral 1995	-	19	9
Weak 1995	-	2	13

Netherlands

	Strong 1999	Neutral 1999	Weak 1999
Strong 1995	4	5	-
Neutral 1995	1	20	8
Weak 1995	-	1	9

Hungary

	Strong 1999	Neutral 1999	Weak 1999
Strong 1995	7	6	-
Neutral 1995	2	26	1
Weak 1995	-	2	4

Slovak Republic

	Strong 1999	Neutral 1999	Weak 1999
Strong 1995	6	5	-
Neutral 1995	5	24	2
Weak 1995	-	1	6

In at least two comparisons:

	Strong 1999	Neutral 1999	Weak 1999
Strong 1995	3	6	-
Neutral 1995	2	19	9
Weak 1995	-	1	8

However, several 'smaller' changes in strong and weak items can be observed in Table 7.3. For example, in comparison with Belgium Flemish, nine items that were neutral in 1995 became weak in 1999 and four items that were strong in 1995 became neutral in 1999. At the same time, two items that were weak in 1995 became neutral in 1999. The number of items that decreased in status when compared to a particular reference country can be obtained as the sum of the above diagonal elements in a cross table. Similarly, the numbers of items that increased in status can be obtained as the sum of the below diagonal elements. In comparison with Belgium Flemish, 13 items decreased and two items increased in their status, which was found to be a significant difference[2].

When Slovene difficulties are compared with those for the Netherlands, similar observations emerge. Although the average score on the overall TIMSS 1999 test in the Netherlands was not significantly higher than in Slovenia (Table 7.1), of those items that changed status between 1995 and 1999, more items became weak than strong.[3] In comparison with Hungary and the Slovak Republic no significant differences were observed.

From these results of item based analysis of trends, it seems plausible to argue that Slovenia shows a slight lack of progress when compared to the two EU countries while it remains comparable to the two candidate countries. However, some caution is needed in interpreting the results of comparisons with the Netherlands. While item based analysis of trends shows an increase in differences between Slovenia and this country, the overall scores on the TIMSS 1999 test in the

2 $B(15, p=0.5)$, $P(X>=13)<0.05$.

3 $B(15, p=0.5)$, $P(X>=13)<0.05$.

Netherlands were similar to Slovenia (Table 7.1). This shows that average scores in the Netherlands on the remaining items in the TIMSS 1999 test were not as high as on identical items. It remains an open question as to what trends would have been observed if a larger set of items were used for their analysis. Therefore the hypothesis of a lack of progress remains to be tested in future.

When items are classified as strong or weak using the additional criterion that they appear strong or weak in comparison with at least two reference countries, similar trends as in comparison with the EU countries can be observed. The largest proportion of items (40 %) was neutral in both measurements. While three items increased in status, significantly more items decreased.[4] On the basis of this, it also seems plausible to hypothesize that, although Slovene achievement remained stable between 1995 and 1999, there was a slight lack of progress at item level as compared to the reference countries in this period. Due to relatively low number of items in this analysis, this hypothesis needs also to be checked in the future.

7.2 LOCATION OF STRENGTHS AND WEAKNESSES IN CONTENT AREAS

As discussed in Chapter 6, to 'zoom in' on the strengths and weaknesses in Slovene achievement, in this section they will be located in the content areas of the Slovene mathematics curriculum. In the first subsection, students are considered as units of analysis, and in the second, the units of analysis are items.

7.2.1 Students as units of analysis

The following questions are addressed in this subsection:

- *To what extent did achievements of Slovene students correspond to achievements of students in the reference countries when considering different content areas?*
- *To what extent did the correspondences in 1999 deviate from those in 1995?*

The results of analyses for these questions are given in Tables 7.5 and 7.6. As in Chapter 6, due to insufficient number of items, content area 'probability' is not shown in these results.

[4] $B(20,p=0.5)$, $P(X>=17)<0.05$.

Table 7.5 Mathematics achievement of Slovene students in 1999 by content areas compared to the reference countries

	Number of items	Slovenia	Belgium Flemish		Netherlands		Hungary		Slovak Republic	
Natural numbers	9	75 (0.8)	80 (0.9)	▲	75 (2.0)		76 (0.9)		80 (1.0)	▲
Meaning of rational numbers	20	64 (0.8)	77 (0.7)	▲	74 (1.8)	▲	68 (1.0)	▲	66 (1.1)	
Operations with rational numbers	19	59 (1.0)	62 (1.1)		53 (2.3)		61 (1.0)		61 (1.3)	
Algebraic expressions	25	61 (0.9)	63 (1.0)		53 (2.0)	▼	63 (1.0)		63 (1.2)	
Functions and proportionality	21	54 (0.8)	58 (0.8)	▲	57 (1.8)		57 (0.9)		53 (1.2)	
Geometrical shapes	23	55 (0.9)	62 (1.0)	▲	54 (2.0)		50 (1.1)	▼	59 (1.3)	▲
Measurement	18	59 (0.8)	67 (0.6)	▲	63 (1.4)		64 (0.8)	▲	61 (1.1)	
Data representation	15	71 (0.8)	75 (0.8)	▲	73 (1.8)		68 (1.0)		69 (1.0)	

In table, estimates of average percent correct are reported
▼ Average percent correct significantly lower than in Slovenia
▲ Average percent correct significantly higher than in Slovenia
( ) Standard errors are presented in parentheses

In general, Slovene scores corresponded with the scores in the reference countries in most content areas. As expected, the largest numbers of differences can be found in comparison with Belgium Flemish. When this country is taken as reference, Slovene scores were significantly lower in most content areas, except in 'operations with rational numbers' and 'algebraic expressions', where they were similar to scores in this country.
Considering that Belgium Flemish was the highest achieving European country in TIMSS 1999 (Mullis et al., 2000), the results in Table 7.5 can be interpreted as showing the areas in Slovene achievement where further improvements might be possible. From this perspective, Slovene students achieve to a sufficient level in the two areas mentioned above, while improvements might be possible in other content areas of the curriculum.

When the other three reference countries are considered, it can be observed that Slovene scores in most content areas corresponded with the scores in these countries. Differences between Slovenia and anyone of these three reference countries were observed in at most three content areas. However, most of these differences were in favor of the reference country. Slovene students' scores were significantly lower than the scores of Hungarian students in 'meaning of rational numbers', and 'measurement', and they were significantly higher in 'geometrical shapes'. In comparison with the Slovak Republic, scores in Slovenia were significantly lower in 'natural numbers' and in 'geometrical shapes'. In comparison with the Netherlands, they were lower also in 'meaning of rational numbers' and they were higher in 'algebraic expressions'. The average score in 'operations with rational numbers' corresponded with the scores in all reference countries.

An overview of Table 7.4 across countries and across content areas reveals that Slovene achievement was significantly lower than in at least two reference countries in content areas 'natural numbers', 'meaning of rational numbers', 'measurement', and 'geometrical shapes'. It seems plausible to argue that, especially in 'meaning of rational numbers' in which Slovene scores were significantly lower than in three reference countries, more general weaknesses in Slovene achievement existed in these content areas. Considering the 'European dimension' from the policy documents, further improvements might be focused on these areas.

The results of analysis of trends in students' scores in content areas between 1995 and 1999 are presented in Table 7.6. For each content area two rows are presented, the first showing the scores in 1995 and the second in 1999. Some cells in Table 7.6 do not show data due to insufficient numbers of items for analysis.

Table 7.6 Trends in mathematics achievement of Slovene students between 1995 and 1999 by content areas compared to the reference countries

	Number of items	Trends	Slovenia	Belgium Flemish	Netherlands	Hungary	Slovak Republic
Natural numbers	3	1995	-	-	-	-	-
		1999	-	-	-	-	-
Meaning of rational numbers	11	1995	67 (0.9)	77 (1.2) ▲	74 (1.5) ▲	66 (0.9)	67 (0.9)
		1999	67 (0.7)	79 (0.6) ▲	79 (1.8) ▲	68 (0.9)	66 (1.1)
Operations with rational numbers	1	1995	-	-	-	-	-
		1999	-	-	-	-	-
Algebraic expressions	6	1995	71 (0.8)	73 (1.3)	67 (1.8)	70 (0.9)	70 (0.9)
		1999	72 (0.7)	75 (0.9)	72 (1.8)	71 (0.8)	71 (1.0)
Functions and proportionality	8	1995	67 (0.8)	71 (1.8)	66 (2.0)	67 (0.9)	66 (1.0)
		1999	66 (0.9)	74 (0.8) ▲	70 (2.1)	71 (0.9) ▲	64 (1.1)
Geometrical shapes	7	1995	61 (0.9)	64 (1.5)	60 (1.8)	53 (1.1) ▼	67 (0.9) ▲
		1999	61 (1.0)	68 (1.1) ▲	63 (1.7)	52 (1.0) ▼	67 (1.2) ▲
Measurement	5	1995	73 (0.9)	74 (1.7)	72 (1.5)	72 (0.9)	73 (0.8)
		1999	72 (0.8)	74 (1.0)	73 (1.7)	71 (0.8)	71 (1.0)
Data representation	6	1995	77 (0.7)	74 (1.3)	78 (1.6)	73 (0.7) ▼	72 (0.8) ▼
		1999	78 (0.7)	77 (1.1)	80 (1.5)	73 (0.9) ▼	75 (1.0) ▼

In table, estimates of average percent correct are reported
▼ Average percent correct significantly lower than in Slovenia
▲ Average percent correct significantly higher than in Slovenia
( ) Standard errors are presented in parentheses

A general observation from Table 7.6 is that there are not many changes, especially in the correspondences between Slovene scores and scores in the reference countries (these scores may differ from those in Table 7.5 due to lower numbers of items). A few changes can be observed though. While in 1995, no significant differences on eight 'functions and proportionality' items between Slovenia and any of the reference countries were observed, in 1999, scores on these items in Belgium Flemish and Hungary were significantly higher. This occurred also on seven items in content area 'geometrical shapes' in comparison with Belgium Flemish. These changes were due to at least three percentage point increases in the reference countries, although these increases themselves were not significant. It seems plausible to hypothesize that a slight lack of progress observed in item based analysis in the previous section may be due to lack of progress on items from these two content areas.

7.2.2 Items as units of analysis

The question addressed in this subsection is:

- *To what extent did item difficulties for Slovenia correspond to those for the reference countries when different content areas are considered?*

Trends in these correspondences between 1995 and 1999 will not be investigated since the breakdown of the set of identical items into the content areas and further classification of items into categories of strong, weak and neutral yielded insufficient numbers of items. As previously noted, items will further be classified as strong or weak in a particular content area if they appeared strong or weak in comparison with at least two reference countries (excluding item F11, see previous section).

Results for answering the above question are presented in Figure 7.2. Since content areas contained different numbers of items, percentages of strong and weak items within content areas are presented in Figure 7.2.

It can be observed that in most content areas, except in 'meaning of rational numbers', less than 40 % of item difficulties deviated from at least two reference difficulties. In 'meaning of rational numbers' however, more than half of items were weak in comparison with at least two countries. Another observation worth noting is that in 'measurement' no strong items were found, while there were more than a quarter weak items. Combined with the results of student based analysis, it seems plausible to describe these two areas as weak for Slovenia.

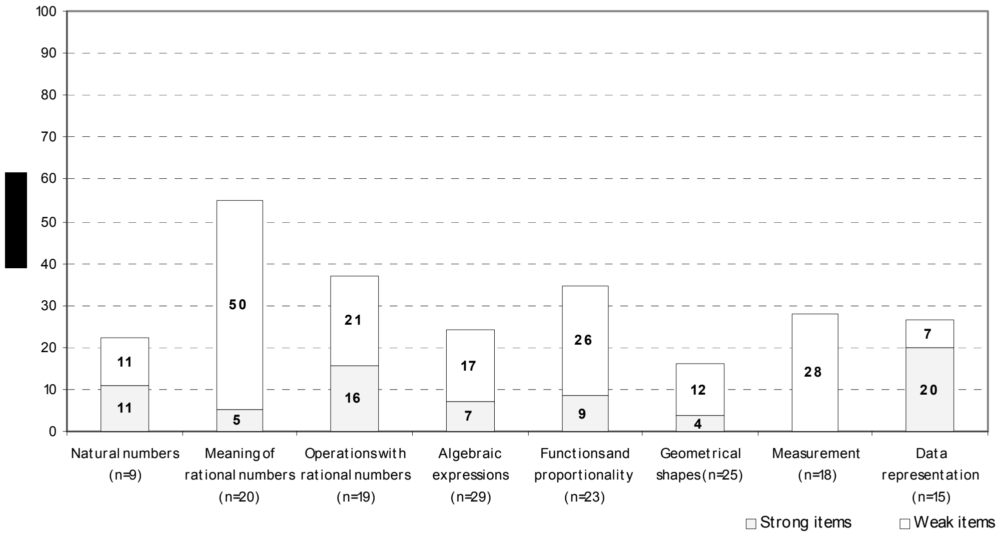


Figure 7.2 Percentages of strong and weak items for Slovenia by content areas when compared to the reference countries in 1999

Approximately a quarter of weak items were also found in 'functions and proportionality'. However, due probably to some strong items (9 %), this area was not found weak in student based analysis. In contrast, in 'data representation' there were a fifth strong items. This is interesting since, as explained in Chapter 2, the non-reformed curriculum guide did not include these topics until they were introduced in the textbooks in the late 1990s. In student based analysis, the scores in this area corresponded with the scores in most countries, except with the scores in Belgium Flemish. This area therefore does not emerge as strong in general, although some interesting findings may be observed from reviewing individual items. As mentioned, this will be done later in this chapter.

7.3 LOCATION OF STRENGTHS AND WEAKNESSES AT LEVELS OF COGNITIVE REQUIREMENTS

In this section, strengths and weaknesses in Slovene achievement when compared to the reference countries are located at the levels of cognitive requirements of items. As in other sections, the first subsection presents results of students-based analysis and the second of item-based analysis.

7.3.1 Students as units of analysis

The following questions are addressed in this subsection:

- *To what extent did achievement of Slovene students correspond to achievements of students from the reference countries when considering different cognitive categories?*

- *To what extent did the correspondences in 1999 deviate from those in 1995?*

The answers are based on the results in Table 7.7 and 7.8.

Table 7.7 Mathematics achievement of Slovene students in 1999 by cognitive categories compared to the reference countries

	Number of items	Slovenia	Belgium Flemish	Netherlands	Hungary	Slovak Republic
Knowing	28	66 (0.7)	74 (0.9) ▲	71 (1.6) ▲	70 (0.9) ▲	68 (1.0)
Using routine procedures	37	65 (0.8)	67 (0.8)	56 (2.1) ▼	67 (0.9)	66 (1.1)
Using complex procedures	37	63 (0.7)	70 (0.6) ▲	67 (1.5)	64 (0.9)	65 (1.1)
Investigating and solving problems	51	53 (0.8)	59 (0.9) ▲	55 (2.0)	53 (0.9)	54 (1.2)

In table, estimates of average percent correct are reported
▼ Average percent correct significantly lower than in Slovenia
▲ Average percent correct significantly higher than in Slovenia
( ) Standard errors are presented in parentheses

An overall impression from Table 7.7 is, except in comparison with Belgium Flemish, that the Slovene scores corresponded with the scores in the reference countries in most cognitive categories. It is interesting to note that differences in the scores by cognitive levels were observed mostly in comparisons with the two EU reference countries while they 'nearly perfectly' corresponded with the scores in the other two candidate countries. Although some strong and some weak items for Slovenia when compared to these countries were identified (Table 7.3), their higher and lower difficulties seem to have been compensated within each cognitive category.

In the two EU countries, Flemish students' scores were significantly higher than those of Slovene students in most cognitive categories except in 'using routine procedures'. At the same time at this level, the Dutch students achieved significantly lower than Slovene students. In comparison with these two countries, 'using routine procedures' could be described as strong in Slovenia. In contrast in the category 'knowing', the average score in Slovenia was significantly lower than the score in the Netherlands, while in the two highest cognitive categories the scores were similar. This might be interpreted as showing that the main differences in achievements of the Slovene students and the Dutch students can be found in the lowest two cognitive categories.

In comparison with the two candidate countries, the score in 'knowing' was significantly lower in Slovenia than in Hungary and similar in other categories. In all cognitive categories the Slovene scores were similar to those in the Slovak Republic.

From these observations, it seems plausible to argue that weaknesses in Slovene achievement are reflected in items that required 'knowing' and strengths in items

that required 'using routine procedures'. In the two highest cognitive categories, only Flemish students scored higher than Slovene students, therefore showing that improvements are possible but also that the current achievement might be described as satisfactory with regard to achievements in the remaining three reference countries.
An interesting pattern regarding ordering of cognitive levels with regard to scores occurred in the five European countries. In Slovenia, Hungary and the Slovak Republic scores in these categories are decreasing by increasing complexity of the category. Moreover, in these three countries, scores in the highest (most complex) cognitive category, 'investigating and solving problems', were significantly lower than achievements in the next highest category, 'using complex procedures' and then also lower than in other categories ($p<0.05$ for all comparisons).

In Belgium Flemish and the Netherlands these patterns are different. In both measurements, the scores in 'using routine procedures' were *significantly lower* ($p<0.05$) than the scores in 'using complex procedures'. Furthermore, in the Netherlands, the score in 'using routine procedures' was similar to the score in 'investigating and solving problems' while the score in 'using complex procedures' was similar to the score in 'knowing'. The scores in categories 'using routine procedures' and 'using complex procedures' therefore were reversed in the two EU reference countries.

Table 7.8 Trends in mathematics achievement of Slovene students between 1995 and 1999 by cognitive categories compared to the reference countries

	Number of items	Trends	Slovenia	Belgium Flemish	Netherlands	Hungary	Slovak Republic
Knowing	16	1995	72 (0.7)	76 (1.1) ▲	75 (1.5)	71 (0.8)	69 (0.8) ▼
		1999	72 (0.6)	78 (0.6) ▲	79 (1.6) ▲	71 (0.8)	70 (0.9)
Using routine procedures	8	1995	69 (0.8)	72 (1.4)	65 (1.9)	65 (0.9) ▼	69 (0.9)
		1999	70 (0.8)	73 (0.8)	68 (1.9)	68 (0.9)	69 (1)
Using complex procedures	13	1995	68 (0.8)	74 (1.4) ▲	74 (1.6) ▲	68 (0.8)	71 (0.8)
		1999	69 (0.7)	78 (0.7) ▲	77 (1.6) ▲	69 (0.9)	71 (1)
Investigating and solving problems	11	1995	68 (0.7)	70 (1.4)	62 (1.5) ▼	62 (0.9) ▼	66 (0.8)
		1999	67 (0.8)	72 (0.9) ▲	67 (1.8)	63 (0.9) ▼	66 (1.1)

In table, estimates of average percent correct are reported
▼ Average percent correct significantly lower than in Slovenia
▲ Average percent correct significantly higher than in Slovenia
( ) Standard errors are presented in parentheses

In Table 7.8, the results of analysis of trends in scores in different cognitive categories are presented. In general, there is stability in correspondences of students' scores in Slovenia with those in the reference countries. A few 'small' changes can be observed though. The general observation is that most of these changes are in favor of the reference countries. In comparison with Belgium Flemish, while scores remained significantly higher on identical items in 'knowing' and 'using complex procedures', they became significantly higher in 1999 in 'using routine procedures' and also 'investigating and solving problems'. Since these changes were due to only slight increases in scores of Flemish students and even due to smaller estimates of standard errors, future studies might reveal whether these differences will develop into a stronger trend. In comparison with the Netherlands, the Dutch score in 'investigating and solving problems' changed from a significantly lower score in 1995 to a similar score to Slovenia in 1999. This also was due to small absolute changes in scores and remains to be examined in future studies. Similarly, all other changes in Table 7.8 are small and remain to be examined in the future. In general, it seems plausible to hypothesize that there was stability in the correspondence of Slovene scores with the scores in the reference countries also across the cognitive categories.

7.3.2 Items as units of analysis

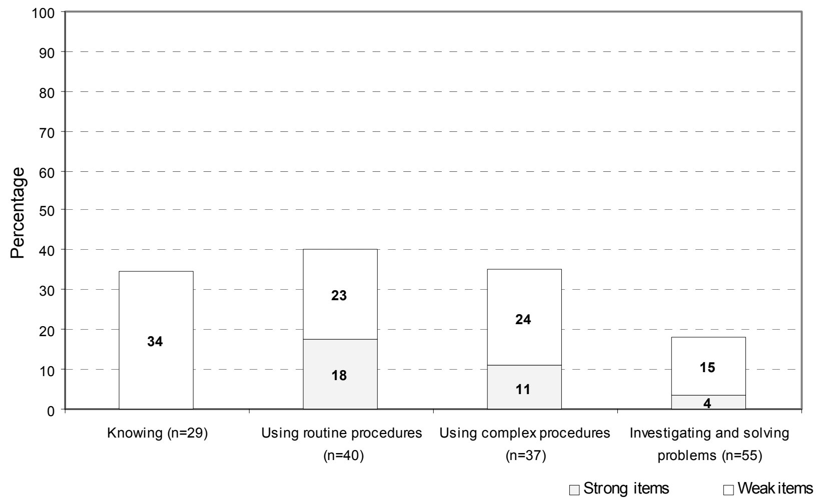
In this subsection, the question that is addressed is:

- *To what extent did item difficulties for Slovenia correspond to those for the reference countries when different cognitive levels are considered?*

As for content areas, trends are not examined in item based analysis due to the insufficient numbers of items in the breakdowns of categories. The results of the analysis for the above question are given in Figure 7.3. As in previous sections, the items were classified as strong or weak if they appeared as strong or weak in comparison with at least two reference countries (excluding item F11, see previous sections).

It can be seen from Figure 7.3 that in all cognitive categories, the majorities of items were classified as neutral (at least 60 %). This is understandable since most countries had similar overall scores to Slovenia. However, the foremost observation is that in 'knowing', no items were classified as strong while approximately a third of items were classified as weak. This lack of strength was also revealed in student based analysis (Table 7.7) and can be interpreted as reflecting more general weakness in Slovene mathematics achievement in this area. In other cognitive categories at least a few strong items were found and in none of these categories the differences between numbers of strong and weak items were significant.

Figure 7.3 Percentages of strong and weak items for Slovenia by cognitive categories when compared to the reference countries in 1999



Furthermore, it is interesting to observe that Slovene item difficulties corresponded with the difficulties in the reference countries for the highest percentage of items in 'investigating and solving problems' (more than 80 %). In student based analysis, the Slovene score in this category differed only from the score in Belgium Flemish and therefore it is plausible to argue that the correspondence observed in average scores occurred over a number of items. While the Slovene scores were lower in this category than in other categories (Table 7.7), scores in the reference countries also tended to be lower on approximately the same items. In 'using routine procedures' and 'using complex procedures' there seem to have been somewhat larger compensation between high and low achievement at item level. The higher percentages of strong and weak items observed in Figure 7.3, did not influence correspondences of the average scores in these categories with the scores in the reference countries (Table 7.7).

7.4 MOST NOTABLE STRENGTHS AND WEAKNESSES AS COMPARED TO ACHIEVEMENTS IN THE REFERENCE EUROPEAN COUNTRIES

Following the design of this study, particularly strong and particularly weak items for Slovenia as compared to the reference countries will be reviewed in this section. As in item based analyses in the previous sections, an item was classified as strong if it appeared strong in comparison with at least two reference countries. Similarly, an item was classified as weak if it appeared weak in comparison with at least two reference countries. In section 7.1 in this way 13 strong and 34 weak items were found (see Table 7.3). Following the terminology of the corresponding section in Chapter 6, these items will be called particularly strong and particularly weak in this section. The contents and other characteristics of these items will be reviewed in the following two subsections.

7.4.1 Most notable strengths

Particularly strong items for Slovenia when compared to the reference countries are presented in Table 7.9 in order to address the following question:

- *What are the contents and characteristics of the particularly strong items when compared to the reference countries?*

Items in Table 7.9 are ordered by their content areas and cognitive requirements. As shown in Table 7.9, most content areas are represented by at least one particularly strong item, except 'measurement'. That in 'measurement' no strong items could be found was already observed in Figure 7.2. There are also no strong items in the cognitive category 'knowing', as observed also in Figure 7.3. It can be seen, that the majority of the particularly strong items required 'using routine procedures' (7 items), half of the remaining items required 'using complex procedures' (3 items), and half 'investigating and solving problems' (3 items). Two of these items are presented in Figure 7.4.

It is interesting to see that there is one item for which the Slovene item difficulty was significantly higher than in all reference countries. This item required interpreting distance/time graph to determine the intersection point of two plots (E01, data representation). Although, as mentioned, data representation topics were not included in the textbooks until late 1990s, students assessed in TIMSS 1999 already had some experience with these topics in school. It can also be hypothesized that Slovene students could have acquired knowledge and skills to solve this item by studying other topics in mathematics (e.g., functions) or in other school subjects (e.g., physics). It is not immediately evident though, what could be the cause for the significantly lower difficulties on this item in all reference countries. An inspection of the curricula of these countries might offer some explanations, however this is not done here.

Table 7.9 Particularly strong items for Slovenia in 1999 when compared to the reference countries

Content area	Cognitive category	Item content	Item label	Slovenia	Belgium Flemish	Netherlands	Hungary	Slovak Republic
Natural numbers	Investigating and solving problems	Solve multi-step word problem with 4-digit whole numbers.	E04	82	81	57 ▼	63 ▼	74 ▼
Meaning of rational numbers	Using complex procedures	Identify shaded figure which shows 2/5 = 4/10.	B09	67	74 ▲	69	53 ▼	52 ▼
Operations with rational numbers	Using routine procedures	Divide fractions.	J12	74	58 ▼	12 ▼	66	71
Operations with rational numbers	Using routine procedures	Multiply two decimals.	M08	62	39 ▼	20 ▼	65	73
Operations with rational numbers	Using routine procedures	Subtract two decimal numbers to 0.001.	R07	90	73 ▼	70 ▼	90	87
Algebraic expressions	Using routine procedures	Simplify and solve linear equation for X.	L17	76	58 ▼	19 ▼	74	78
Algebraic expressions	Using routine procedures	Calculate (5X+3)/(4X-3) for X=3.	N13	76	67 ▼	50 ▼	80	72
Functions and Proportionality	Using complex procedures	Calculate ratio of rectangle areas given relationship between sides.	U02B_1	32	23 ▼	10 ▼	30	32
Functions and Proportionality	Using complex procedures	Calculate ratio of rectangle areas given relationship between sides.	U02B_2	29	20 ▼	8 ▼	22	31
Geometrical shapes	Investigating and solving problems	Find fourth quadrilateral angle, given the other three.	L16	53	70 ▲	39 ▼	39 ▼	49
Data representation	Using routine procedures	Find difference between the means of two sets of scores.	A06	82	78	74 ▼	76 ▼	86 ▲
Data representation	Using routine procedures	Interpret distance/time graph to determine intersection point of two plots.	E01	88	74 ▼	80 ▼	83 ▼	82 ▼
Data representation	Using complex procedures	Interpret barchart histogram of travel time to find number above 10 min.	H07	77	82	82	69 ▼	65 ▼
Total	13 items							

In table, item difficulties (item percents correct) are presented
▲ Item difficulty significantly higher than in Slovenia
▼ Item difficulty significantly lower than in Slovenia

Slovene students also scored high compared to the reference countries on a multi-step word problem with four-digit whole numbers (E04, natural numbers). The average score in Slovenia on this item was significantly higher than in three reference countries, except in Belgium Flemish. Slovene students therefore seem to have mastered well the knowledge and skills required by these items.

N13. If $x = 3$, what is the value of $\frac{5x+3}{4x-3}$?

Answer: ________________

J12. Divide: $\frac{6}{55} \div \frac{3}{25} =$

Answer: ________________

Figure 7.4 Examples of particularly strong items for Slovenia when compared to the reference countries

It is interesting to observe that on approximately half of particularly strong items scores in Slovenia were simultaneously higher than the scores in Belgium Flemish and in the Netherlands (7 items), while in the other pairs of reference countries this did not occur as often. Six of these seven items required 'using routine procedures'. Although these are small numbers of items and the results should be interpreted with caution due to previously mentioned potential problems, it seems plausible to hypothesize, based also on observations from the previous sections, that there are strengths in Slovene achievement in this category, relative to the two EU countries. At the same time this could also be interpreted as pointing out that larger differences in student achievement at item level existed between Slovenia and these two countries, than between Slovenia and the other two candidate countries.

While most of particularly strong items in Table 7.9 show high difficulties for Slovenia, two items had lower difficulties in absolute terms (U02B_1 and U02B_2, functions and proportionality, 32 % and 29 %). These two items (actually one item with 2 score points for full answer) required calculation of a ratio of areas of two rectangles given relationship between sides. However, as can be seen from Table 7.9, in the reference countries, the difficulties for this item were even lower. Therefore, although Slovene scores on these items may not be seen as satisfactory since these items were particularly weak when compared to the standards (Table 6.10), they can be described as relatively high when compared to the reference countries. Such information is important in order to understand and use the standards in the curriculum.

There are also three particularly strong items in Table 7.9 that were identified as weak in comparison with one of the reference countries (B09, meaning of rational numbers; L16, geometrical shapes; and A06, data representation). Whether these items indeed could be described as reflecting strengths in Slovene achievement remains to be answered. It is possible that other criteria should be considered in order to answer this question, for example the importance of the particular knowledge and skills that these items required in the light of the Slovene curriculum. This, of course, is important also when examining all other items.

7.4.2 Most notable weaknesses

In this subsection, the final question from the structure of operational research questions in Chapter 5 is addressed:

- *What are the contents and characteristics of the particularly weak items when compared to the reference countries?*

Table 7.10 presents particularly weak items for Slovenia when compared to the reference countries. There are 34 such items. As previously mentioned, items are ordered by content areas and cognitive categories. It can be observed that all content areas and all cognitive categories are represented in Table 7.10. Three of these items are presented in Figure 7.5.

There are five items that had significantly lower difficulties in Slovenia than in all four reference countries. These items required writing two possibilities for actual height from rounded value (V01, natural numbers), identifying two hundred six and nine-tenths (L10, meaning of rational numbers), determining actual length of box which is rounded to a given whole number value (A03, meaning of rational numbers), calculating time to finish homework after ¾ hour (O06, measurement), and using data in a chart to identify number of times when humidity was equal to a given value (L11, data representation). Of these, two items, V01 and A03, were also identified as particularly weak with respect to the standards in the curriculum (Table 6.10). As previously mentioned, the convergence between the results for the two reference points will be examined in the last section. Items V01 and A03 can be described as requiring 'indirect thinking' in the procedure of rounding. It can be hypothesized that this characteristic influenced lower scores of the Slovene students on these items. Although no information is available why in other countries the difficulties of these items were higher than in Slovenia, this fact in itself could be regarded as valuable information for the Slovene curriculum developers. Weaknesses in Slovene achievement with regard to understanding of decimal fractions, including rounding, relative to other TIMSS 1995 countries were observed already by Magajna (2000). For the process of identifying possible causes for these deficiencies and development of improvement measures (see Chapter 3), curricula of these countries were an important source of information (National Curriculum Council, 1996).

Table 7.10 Particularly weak items for Slovenia in 1999 when compared to the reference countries

Content area	Cognitive category	Item content	Item label	Slovenia	Belgium Flemish	Netherlands	Hungary	Slovak Republic
Natural numbers	Using complex procedures	Write two possibilities for actual height from rounded value.	V01	41	61 ▲	64 ▲	66 ▲	79 ▲
Meaning of rational numbers	Knowing	Identify smallest of decimal fractions.	B10	58	67 ▲	74 ▲	55	60
Meaning of rational numbers	Knowing	Estimate rational number from point P on number line.	D12	78	84 ▲	90 ▲	80	80
Meaning of rational numbers	Knowing	Identify fraction of circle which is shaded.	F12	59	74 ▲	75 ▲	56	49 ▼
Meaning of rational numbers	Knowing	Identify figure showing 2/3 shaded.	H08	78	94 ▲	92 ▲	78	72
Meaning of rational numbers	Knowing	Identify two hundred six and nine-tenths.	L10	61	87 ▲	79 ▲	97 ▲	90 ▲
Meaning of rational numbers	Using routine procedures	Identify number rounded to the nearest 100th.	O04	52	64	60	74 ▲	77 ▲
Meaning of rational numbers	Using complex procedures	Determine actual length of box which is rounded to 9 cm.	A03	61	85 ▲	89 ▲	91 ▲	89 ▲
Meaning of rational numbers	Using complex procedures	Select correct pair of numbers: # 1 < 2.25 < # 2.	C04	56	66 ▲	59	59	64 ▲
Meaning of rational numbers	Using complex procedures	Identify smallest fraction.	D09	75	82 ▲	86 ▲	71	67
Meaning of rational numbers	Using complex procedures	Identify decimal fraction between 0.07 and 0.08.	F09	75	84 ▲	86 ▲	69	74
Operations with rational numbers	Using routine procedures	Identify point on number line from distance of other points.	N12	46	65 ▲	67 ▲	51	48
Operations with rational numbers	Using complex procedures	Find height of stack from paper thickness	T04	62	75 ▲	72 ▲	65	69
Operations with rational numbers	Investigating and solving problems	Identify fraction represented in word problem: birthdays in first half of year.	G05	68	76 ▲	85 ▲	53 ▼	61
Operations with rational numbers	Investigating and solving problems	Solve division word problem with common fractions: scoops of flour needed to fill bag.	O09	56	69 ▲	73 ▲	57	55
Algebraic expressions	Knowing	Identify true expression when a,b,c, are different real numbers.	R10	41	56 ▲	40	58 ▲	48
Algebraic expressions	Investigating and solving problems	Solve algebra word problem with fractions: number of marbles in bag to start with.	N16	39	52 ▲	53 ▲	45	45
Algebraic expressions	Investigating and solving problems	Solve multi-step algebra word problem of simultaneous equations: boy and girls from total club members.	T01_1	45	66 ▲	45	54 ▲	60 ▲
Algebraic expressions	Investigating and solving problems	Solve multi-step algebra word problem of simultaneous equations: boy and girls from total club members.	T01_2	37	62 ▲	38	46 ▲	53 ▲

In table, item difficulties (item percents correct) are presented
▲ Item difficulty significantly higher than in Slovenia
▼ Item difficulty significantly lower than in Slovenia

Table 7.10 Particularly weak items for Slovenia in 1999 when compared to the reference countries (Continued)

Content area	Cognitive category	Item content	Item label	Slovenia	Belgium Flemish	Netherlands	Hungary	Slovak Republic
Functions and Proportionality	Knowing	Identify equation expressing xy relationship from table of values.	H10	60	60	71 ▲	80 ▲	56
Functions and Proportionality	Using routine procedures	Laps run by Carol and Alice from ratio 4/3 = 12/?	A04	53	73 ▲	71 ▲	49	65 ▲
Functions and Proportionality	Using routine procedures	Find relationship between numbers in a set of ordered pairs.	E05	63	67	74 ▲	73 ▲	63
Functions and Proportionality	Investigating and solving problems	Predict number of total defective bulbs based on random sample.	H11	71	79 ▲	86 ▲	76	70
Functions and Proportionality	Knowing	Identify equation representing relation in x/y table	L14	54	67 ▲	64	75 ▲	56
Geometrical shapes	Knowing	False statement of congruent triangles in a rectangle	A05	63	74 ▲	74 ▲	35 ▼	78 ▲
Geometrical shapes	Using complex procedures	Identify cube made by folding 2-dimensional net.	B11	68	82 ▲	82 ▲	68	74
Geometrical shapes	Using complex procedures	Determine angle in overlapping congruent triangles.	E02	48	69 ▲	42	49	64 ▲
Measurement	Knowing	Identify angle closest to 45 degrees in a circle.	N15	61	70	52	72 ▲	76 ▲
Measurement	Using routine procedures	Find who needs most paces to walk to end of hallway from chart showing lengths of paces.	L13	79	94 ▲	90 ▲	91 ▲	81
Measurement	Using routine procedures	Calculate time to finish homework after 3/4 hour.	O06	76	90 ▲	90 ▲	92 ▲	86 ▲
Measurement	Using routine procedures	Draw new rectangle based length/width ratios of another rectangle.	U02A_1	41	66 ▲	52	58 ▲	49
Measurement	Using routine procedures	Draw new rectangle based length/width ratios of another rectangle.	U02A_2	26	53 ▲	38 ▲	43 ▲	19
Data representation	Using complex procedures	Use data in a chart to identify number of times when humidity was exactly 20 percent.	L11	48	72 ▲	76 ▲	69 ▲	66 ▲
Probability	Knowing	Predict probability of occurrence for fifth independent coin toss.	F08	70	74	79 ▲	86 ▲	70
Total	34 items							

In table, item difficulties (item percents correct) are presented
▲ Item difficulty significantly higher than in Slovenia
▼ Item difficulty significantly lower than in Slovenia

Some explanations and interpretations of low scores in Slovenia as compared to the reference countries may also be offered for other items. For example, it may be that students in the non-reformed system in Slovenia are rarely presented with non-linear graphs, such as in item L11, or such questions about these graphs. As data representation topics were included in the textbooks only in the late 1990s, it may be that 'unusual' types of items were not included.

V01. The height of a boy was reported as 140 cm. The height had been rounded to the nearest 10 cm. What are two possiblities for the boy's actual height?

Answer: ______________cm and ______________ cm.

L10. Which number is two hundred six and nine-tenths?

A. 206.09

B. 206.9

C. 206.910

D. 2006.9

T04. A sheet of paper is 0.012 cm thick. Of the following, which would be the height of a stack of 400 sheets of this paper?

A. 0.048 cm

B. 0.48 cm

C. 4.8 cm

D. 48 cm

Figure 7.5 Examples of particularly weak items for Slovenia when compared to the reference countries

There are also five items in Table 7.10 for which item difficulties in Slovenia were significantly lower than in three of the reference countries. These items required solving a multi-step algebra word problem of simultaneous equations (T01_1 and T01_2, algebraic expressions), finding a value from a given ratio (A04, functions and proportionality), finding the case in which the largest value of paces for a given length is needed (L13, measurement), and drawing a rectangle based on ratios of sides with a given rectangle (U01_2, measurement). On item L13 for example, the Slovene score of 79 % could not be described as low in absolute terms, however, in the reference countries, the scores on this item were generally higher. While the characteristics of other items may point to possible explanations for lower absolute scores (e.g., ratios being covered only in the final grade), it is not immediately evident why in the reference countries scores are higher.

As for the particularly strong items (Table 7.9), for the particularly weak items the pair of two countries, Belgium Flemish and the Netherlands, can also be observed as 'causing' most particularly weak items for Slovenia. On more than 60 % of these items (21 items), the Slovene difficulties were lower than the difficulties in these two countries. Furthermore, on 14 items the Slovene difficulties were lower only than in these two countries. On the basis of this it seems plausible to hypothesize that the differences in these comparisons may also be due to differences in mathematics curricula, which may be larger between Slovenia and the two EU countries than, between Slovenia and the other two candidate countries. To provide more detailed information on possible sources of these differences, further studies would be needed. The database from the TIMSS curriculum analysis (see Schmidt et al., 1997) might be useful for such studies.

7.5 CONVERGENCE OF THE RESULTS WHEN COMPARED TO THE TWO REFERENCE POINTS

By using two reference points, two descriptions of Slovene achievement in mathematics were obtained. In this section, it will be analyzed to what extent the strengths and weaknesses observed in Slovene achievement in comparison with one reference point can also be observed in comparison with the other reference point. This convergence in strengths and weaknesses will be examined for content areas and cognitive categories as well as for individual items.

7.5.1 Convergence in content areas and cognitive categories

To assess this convergence, the results from Chapter 6 and from this chapter will be compared. Table 7.11 presents a summary of the results from these two chapters for locating strengths and weaknesses in Slovene achievement in content areas of the Slovene curriculum.

As shown in Table 7.11, there were similarities and differences in the descriptions of Slovene achievement between the two perspectives. From both perspectives weaknesses in Slovene achievement were observed in content area 'meaning of rational numbers' and, partially, in cognitive category 'knowing' (compared to the standards in the curriculum, weaknesses were observed at Level 1 only). Content areas 'algebraic expressions', 'functions and proportionality', 'operations with rational numbers', and 'data representation' can be described as satisfactory from both perspectives.

Table 7.11 Summary of the results of student based and item based analysis of correspondence of achievement with the two reference points

	Correspondence with the attainment targets		Correspondence with the reference countries	
	Student based analysis	Item based analysis	Student based analysis	Item based analysis
Natural numbers	▲	▲	▼	○
Meaning of rational numbers	▼	▼	▼	▼
Operations with rational numbers	○	○	○	○
Algebraic expressions	▲	○	○	○
Functions and proportionality	○	○	○	○
Geometrical shapes	▼	○	▼	○
Measurement	○	○	▼	▼
Data representation	▲	▲	○	○
Knowing	▲	○	▼	▼
Using routine procedures	○	○	○	○
Using complex procedures	▲	○	○	○
Investigating and solving problems	▼	○	○	○

▲ Area could be described as strong compared to the reference point
○ Area could be described as corresponding with the reference point
▼ Area could be described as weak compared to the reference point

There were also similarities between the results using the two reference points in observations about content area 'geometrical shapes'. In this content area, average scores of Slovene students were significantly lower than the standards and also lower than in two reference countries (Belgium Flemish and the Slovak Republic, Table 7.5). However, in item based analysis it was found that there is a relatively large proportion of items on which Slovene student achieved comparably or even higher than the level of the corresponding reference point. For this reason, this content area was not emphasized as weak in the results of this study.
Satisfactory with respect to both reference points were also achievements in the cognitive categories 'using routine procedures' and 'using complex procedures'.

The differences between the two perspectives emerged in the content area 'measurement', in which Slovene achievement seemed to correspond with the standards but it was lower than in two of the reference countries (Belgium Flemish and Hungary, Table 7.5). Furthermore, in this content area, no strong items when compared to the reference countries were found. Further investigations in this area might be needed to reveal possible sources of these differences.

Similarly, average scores in 'natural numbers' were higher than the standards, but were lower than in two of the reference countries (Belgium Flemish and the Slovak Republic, Table 7.5). However, since for the majority of items it was observed that student achievement on these items corresponded also with achievements in the reference countries, this content area may not be generally described as weak. When scores in different cognitive categories were examined, the differences were that, while in 'knowing', Slovene scores were in overall significantly higher than the standards, they were significantly lower than the average scores in three reference countries, except the Slovak Republic. In contrast, while average scores of Slovene students were significantly lower than the standards in 'investigating and solving problems', they were nonetheless similar to the scores in the reference countries. These results reveal that the standards in the curriculum need to be understood in the light of the achievements of students from other countries.

7.5.2 Convergence in particularly strong and particularly weak items

Convergence of the results for the two research questions can be examined also by comparing sets of particularly strong and particularly weak items identified in comparisons with the two reference points. More specifically, the lists of particularly strong and particularly weak items in Tables 6.9, 6.10, 7.9, and 7.10 were compared. These comparisons revealed the items that were identified as particularly weak in comparison with both reference points; the items that were identified as particularly strong in comparison with both reference points; items that were identified as particularly weak or particularly strong in comparison with one reference point but not with the other; and, some items that also were identified as particularly strong in comparison with one reference point and at the same time as particularly weak in comparison with the other. These items will be discussed below.

Particularly strong items in comparison with both reference points

Four items were particularly strong in comparisons with both reference points. These items all required 'using routine procedures' for which it was observed in previous sections, that they reflected the strengths in Slovene achievement were found. The items required multiplying two decimals (M08, operations with rational numbers), subtracting two decimals (R07, operations with rational numbers), simplifying and solving linear equation (L17, algebraic expressions, see Figure 6.4), and interpreting distance and time graph to determine intersection point of two plots (E01, data representation).

Particularly weak items in comparison with both reference points

Comparisons of the lists of items in Tables 6.10 and 7.10 yielded 9 items that were particularly weak when compared to the standards and also particularly weak when compared to the reference countries. These items required writing two possibilities for actual height from rounded value (V01, natural numbers, Figure 7.5), identifying two hundred six and nine-tenths (L10, meaning of rational numbers, Figure 7.5), determining the actual length of box which was rounded to a given whole number (A03, meaning of rational numbers), solving algebra word problem with fractions (N16, algebraic expressions), solving multi-step algebra word problem of simultaneous equations (T01_2, algebraic expressions), identifying false statement about congruent triangles in a rectangle (A05, geometrical shapes), and drawing a new rectangle based on the ratio of length and width with another rectangle (U02A_2).

As already mentioned, it can be hypothesized that items V01 and A03 required 'inverse' rounding. It is interesting to observe that Slovene students scored high on an item that also required rounding, however in a straightforward manner (item N11, 94 %, Table 6.9). It seems plausible to hypothesize that the two items above had low difficulties because of the inversed problem and that this points to weaknesses in students understanding of the procedure of rounding.

For item L10 (see Figure 7.5), it can be hypothesized, that students might be mislead by a distracter. Analysis of incorrect responses revealed that 14 % of students chose option C as the correct answer. At the same time even larger percentage of students (24 %) chose option A. This item is also a particularly weak from the perspective of achievements in the reference countries. It seems plausible to hypothesize that improvements might be desired in students' understanding of decimal fractions in Slovenia. This weakness in Slovene achievement was observed also by Magajna (2000). The topic of decimal fractions was one of the explicit areas that were addressed by the reformed curriculum. As explained in Chapter 2, the introduction of decimal fractions takes place in an earlier grade (grade 6 of the reformed system) and through measurement topics as opposed to introduction in grade 6 in the non-reformed system (a year older students) through the concept of fractions. Future studies might show whether weaknesses in this area observed in this study have been remedied.

Particularly weak items in comparison with the standards and at the same time particularly strong items in comparison with the achievements in the reference countries

These items were determined through comparisons of Tables 6.10 and 7.9. Two items, or more specifically, one item with two score points was found particularly weak when compared to the standards that was particularly strong when compared to the achievements in the reference country. This item (U02B_1 and U02B_2) was already observed above as the item with low difficulty in absolute terms while with still relatively high difficulty compared to the reference countries. As mentioned, while it can be hypothesized that this item had low difficulties because it was the second part of a multi-step task (comprising items U02A and U02B), there is no evident explanation for even lower difficulties in other countries (Belgium Flemish and the Netherlands). Since these items are secured for future assessments they can not be presented.

Particularly strong items in comparison with the standards and at the same time particularly weak items in comparison with achievements in the reference countries
Comparison of Tables 6.9 and 7.10 yielded four such items. These items required finding height of stack from thickness of one paper (T04, operations with rational numbers, see Figure 7.5), predict total number of defective bulbs based on random sample (H11, fractions and proportionality, see Figure 6.4), identifying cube made by folding a two-dimensional net (B11, geometrical shapes, see Figure 6.4), and identifying angle closest to 45 degrees (N15, measurement).
It is interesting to observe that, while items B11 and H11 appeared particularly strong when compared to the standards, they were particularly weak when compared to the reference countries. As discussed above, it can be hypothesized that these two items were not familiar to Slovene students. Analyses of the curricula might reveal whether the difference between Slovenia and the reference countries might be hypothesized as a consequence of differences in the curricula of these countries.

7.6 ANSWERS TO THE SECOND RESEARCH QUESTION

The results of analysis for the second research question were presented in this chapter. Slovene achievement in mathematics at the end of compulsory education was compared to four other European countries, Belgium Flemish, the Netherlands, Hungary, and the Slovak Republic. Similar approaches to analyses were used as for the first research question in Chapter 6.
In overall, correspondences in achievements between Slovenia and most of the reference countries were expected due to the criteria for selection of the reference countries, that they have similar or higher overall achievement. A more detailed analysis revealed some differences. In four content areas, 'natural numbers', 'meaning of rational numbers', 'geometrical shapes', and 'measurement', Slovene

average scores were significantly lower than in at least two reference countries. Moreover, Slovene average scores were significantly higher only in comparison with the Netherlands in 'algebraic expressions', and in comparison with Hungary in 'geometrical shapes'. In 'measurement' furthermore, no particularly strong items were found. In cognitive categories, Slovene achievement corresponded to achievement in the reference countries in higher categories and it was significantly lower than in three countries in the lowest category, named 'knowing'.

As in Chapter 6, no significant changes in student achievement in these countries between 1995 and 1999 emerged from the analyses. There was however a slight lack of progress in Slovenia observed at item level. A review of the particularly strong and particularly weak items revealed that the largest percentages of these items emerged from comparisons with Belgium Flemish and the Netherlands.

In the final section of this chapter, the convergence between the results for the two research questions was examined. This analysis revealed that there are similarities as well as differences between the two sets of results. In 'meaning of rational numbers' weaknesses in Slovene achievement were observed in comparison with both reference points.

CHAPTER 8
Summary, reflections and recommendations

The title of this thesis indicates that it focuses on mathematics achievement of students in Slovenia. Recently, major curriculum reforms were developed in Slovenia in mathematics education as well as in all other subjects. To provide supporting information in the current process of the implementation of the reforms in mathematics, this study describes student achievement in comparison with the visions and intentions embedded in the new curriculum as well as in comparison with several other European countries. This structure of comparisons was reflected in the two research questions that were posed for this study. The results for these research questions were presented in the previous chapters. This final chapter presents a brief overview of the study and its main findings. Reflections on the methodology and main conclusions are followed by recommendations for policy and future research.

In section 8.1 the research questions, conceptual framework and research design are summarized. Main findings are given in section 8.2. This is followed by a discussion on methodological issues and conclusions of this study in section 8.3. Recommendations for policy implementation and future research are given in the final section (8.4).

8.1 SUMMARY OF CONCEPTUAL FRAMEWORK AND STUDY DESIGN

In the past decade, there have been major changes in Slovenia. Country's independence in 1991 yielded reforms in nearly all aspects of social life. Within this, education was one of the earliest areas to receive attention. After the legislation for education reform was adopted in 1996, panels of curriculum experts were mandated to review the existing curricula and prepare proposals for their reforms. In addition to compulsory education, this process encompassed all other levels of pre-university education, including kindergarten. At the level of compulsory education, the major additions to the curricula included the attainment targets, also called standards, which were specified for each subject in each grade. Within this context, the present study aimed to describe mathematics achievement of Slovene students at the end of compulsory education *prior to* the implementation of the reformed curriculum. The study adds to other studies, because it sheds more light onto the areas in which improvements might be desired and possible. It was not possible to describe mathematics achievement of Slovene students as the outcome of the reformed curriculum because the introduction of the reforms into the school system were just starting to take place, but the type of information yielded in this study may help curriculum reformers and practitioners refine their conclusions from the review of the old curriculum and develop additional approaches that teachers might use in their teaching. Through this, the success of the reforms is additionally strengthened.

In describing the Slovene mathematics achievement at the end of compulsory education, reference points were needed. Two general reference points were constructed. The first was constructed on the basis of the attainment targets in the reformed curriculum. Even though students assessed studied the non-reformed curriculum, the aim of this study was to describe the knowledge and skills of Slovene students in the light of the intentions embedded in the reforms. Through this, information on what is 'the baseline' for the new curriculum was provided.
Since reaching internationally comparable achievement was emphasized in the reform, comparisons with other countries were also considered important in this study. This led to the construction of the second reference point. In the very near future, in May 2004, Slovenia will become a member of the EU along with nine other European countries. In the present study, it was of interest to compare mathematics achievements of Slovene students to achievements of students in the EU countries and in the other candidate countries. In order to provide information on areas of possible improvements, countries with similar or higher average achievements were selected for these comparisons: Belgium Flemish, the Netherlands, Hungary, and the Slovak Republic.[1]
Although the main part of the description of mathematics achievement in Slovenia was focused on one point in time, the year 1999, the study also examined the developments in this achievement with respect to the selected reference points over a period of four years in the second half of the 1990s.

1 As explained in Chapter 1, for reasons of brevity the term 'country' is used synonymously with 'educational system'.

The interests pointed out above resulted in the following two research questions for this study:

1. *How well did Slovene students at the end of compulsory education in the non-reformed system in the late 1990s perform in mathematics when compared to the attainment targets in the reformed mathematics curriculum and what were the developments in this performance between 1995 and 1999?*

2. *How well did Slovene students at the end of compulsory education in the non-reformed system in the late 1990s perform in mathematics when compared to the performance of students in other European countries and what were the developments in this performance between 1995 and 1999?*

After the introduction of the research in Chapter 1 and the description of the education system and reforms in Slovenia in Chapter 2, concepts and methods to tackle the above research questions are outlined in Chapter 3. On the basis of a literature review, concepts that were important for addressing the research questions in this study, such as goals, standards, assessment of student achievement, measuring the correspondence with the selected reference points and, from this, identifying weaknesses and strengths are described in this chapter.

As follows from the research questions, international comparative data were needed for this study, collected at several points in time to enable examination of trends. The data from the 1995 and 1999 data collections in the Third International Mathematics and Science Study (TIMSS), conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) were used. In Chapter 4, the elements of the TIMSS design and methodology important for the present study are summarized. They include the conceptual framework, research questions, target populations and sampling, instrument development, data collection and scaling procedures in TIMSS. The most important feature of TIMSS for the present study is that the design of achievement instruments was driven by the curricula of participating countries. This enabled the link between student achievement (the attained curriculum) and the curriculum as prescribed in the official documents (the intended curriculum).

The research design is presented in Chapter 5. First reference points are described in more detail. The reference point from the attainment targets in the curriculum was constructed by specifying for each item in the TIMSS achievement tests the levels of these targets represented in the item. As described in Chapter 2, these attainment targets were specified in the reformed curriculum at three levels, called "minimum standards, "fundamental standards", and "higher level standards". For reasons of brevity, they were called Level 1, Level 2, and Level 3 standards in this thesis, which means that there are 'Level 1 items', 'Level 2 items' and 'Level 3 items' in the achievement test.

In this way, the TIMSS items were used to describe the intended and the attained curricula in Slovenia as well as the attained curricula in the reference countries. The measures of the attained curricula were obtained through the students' scores on these items. The measures of the intended curricula were provided by three mathematics experts in Slovenia in the form of the levels of the attainment targets at which the items were allocated and operationalized into 'intended percent correct', being 75 % for Level 1, 50 % for Level 2 and 25 % for Level 3. Since it was found that only six TIMSS items were allocated at Level 3, they were excluded from further analysis.

The general research questions were operationalized in the design of the study. The overall structure of the operationalized research questions is presented in Table 5.2, while each individual question is further elaborated in the respective chapters with the results (Chapters 6 and 7). Two basic analysis approaches were used: a student based and an item based analysis. In the first approach, average scores of students on items and their correspondence with the reference points were examined, whilst in the second approach, the correspondences of achievements on individual items with the reference points were examined.

In student based analysis, when the estimate of Slovene achievement was significantly higher than the reference point, this was taken as an indication of strengths in Slovene achievement. In case it was significantly lower, this was taken as an indication of weaknesses. In the third case, Slovene achievement was described as corresponding with the reference point. When the reference point was constructed on the basis of achievements of students from the reference countries, only when Slovene achievement was significantly higher than in at least two reference countries this was considered an indication of strengths. If it was significantly lower than in at least two reference countries, this was considered as an indication of weaknesses.

In item based analysis, items with significantly higher difficulties (item percents correct) than the reference point were classified as strong. Items with significantly lower difficulties were classified as weak. In case of the four reference countries, items were classified as strong or weak, respectively, when their difficulties for Slovenia were significantly higher or lower than in at least two reference countries. Inferences about strengths and weaknesses in Slovene achievement were made if relatively high or low numbers of items were classified as weak or strong.

The analyses were carried out at the levels of the overall mathematics achievement scores, of scores in several subdomains, as well as at the level of individual items. The mathematics subdomains were defined as content areas which were based on the structure of content in the Slovene mathematics curriculum, and as cognitive categories which were based on the TIMSS classification (called 'performance expectations' in TIMSS). The main findings from these analyses are described in the following section.

8.2 MAIN FINDINGS

Following the structure of the two research questions, the findings of this study were presented in chapters 6 and 7. Main findings from each of these chapters are summarized in the following two subsections. In the final subsection, convergence of the results between the two research questions is examined.

8.2.1 Findings from the first research question

The most general finding in this study is that mathematics achievement of Slovene students in the non-reformed system corresponded with the attainment targets (called also the standards) in the reformed curriculum. However, this is at the level of overall mathematics achievement only. Detailed analyses revealed variation in the correspondences. As mentioned, two levels of the standards were examined, named Level 1 and Level 2. When looking at achievements at Level 1 weaknesses in student achievement were observed. These were compensated with strengths at Level 2 (see Table 6.1). There seem to be deficiencies in the knowledge and skills of Slovene students that are undesired from the perspective of the intended reformed curriculum. In order to achieve the goals of the reform, attention has to be put on the knowledge and skills at Level 1, that is on knowledge and skills that nearly all students are intended to master.

Further analyses, carried out for different content subdomains in the Slovene curriculum (called content areas in this thesis and different from TIMSS classification) revealed that the contrast between Level 1 and Level 2 occurred in most of these areas (see Table 6.5). However, in two content areas, labeled as 'natural numbers' and 'data representation', student achievement in Slovenia can be described as strong. In these content areas students' average scores corresponded with the standards at both levels and therefore also at the overall level. The result for 'data representation' is especially interesting since, as described in Chapter 2, no topics from this area were included in Slovene intended (non-reformed) curriculum until additions to this curriculum were prepared in the late 1990s. Speculatively, Slovene students might have answered the TIMSS 'data representation' items using 'common knowledge' in addition to what they learned in school. Nonetheless, the achievement in this area seems to be of similar levels than achievements in the reference countries and at the same time higher than required by the standards in the curriculum.

In contrast, in the content area 'meaning of rational numbers', Slovene achievement seems to be weaker than intended according to the attainment targets. Students' average scores were significantly higher or lower than the standards also in some other content areas. However, the analysis in which items were taken as units (see Chapter 5) revealed that nonetheless there were substantial proportions of items in each of these content areas on which achievement of Slovene students was lower or higher than the standards, respectively, and thus, impeding general conclusions about strengths or weaknesses in these content areas. For example, although the average score in 'algebraic expressions' was higher than the standards, there were equal proportions of items on which students achieved significantly higher and lower than the standards and, therefore, this area cannot be described as strong in general (see Table 6.5 and Figure 6.2).

Variation in achievement of Slovene students and its correspondence with the standards was also examined across so-called cognitive categories. This concept is based on work of Bloom (1956) and was used in TIMSS as 'performance expectations' (see Chapters 3 and 4). Cognitive categories indicate the types of performances students are expected to exhibit when engaged with the item content. Four cognitive categories from TIMSS classification of items were used in this study: 'knowing', 'using routine procedures', 'using complex procedures', and 'investigating and solving problems'. They were considered to be hierarchically ordered, from less complex to more complex performances although items of variety of difficulties can be constructed for each category.
When correspondence of mathematics achievement of Slovene students with the standards was examined with respect to these categories, it was found that overall, students' scores corresponded with the standards in the three lower categories and were significantly lower in the highest category. Nonetheless, some strength in this category was reflected in a fifth of items on which achievement was significantly higher than the standards.
In the lowest category of 'knowing', students' scores corresponded with the standards only at Level 2, while they were significantly lower than the standards at Level 1. Therefore weaknesses in student achievement also seem to be reflected in

this category. Examination at item level revealed that these weaknesses seem mostly to be reflected in items about 'meaning of rational numbers' (see Table 6.10).

8.2.2 Findings from the second research question

The second research question complemented the description of mathematics achievement of Slovene students from the perspective of the intended curriculum. The second perspective was constructed from the achievements of students from four other European countries: Belgium Flemish, the Netherlands, Hungary, and the Slovak Republic. As mentioned in the previous section, countries with similar or higher overall achievements in mathematics than Slovenia were selected as reference in order to provide information on areas in which improvements might be desirable or possible. Therefore, by definition, average achievement in Slovenia corresponded with achievements in these countries, except in Belgium Flemish, which had significantly higher overall achievement. Through the second research question, it was more of interest to compare Slovene achievement with achievements in the reference countries with respect to the content areas in the Slovene curriculum, to the categories of cognitive requirements in items, as well as at the level of individual items.

In these analyses it was revealed that Slovene scores were significantly lower in comparison with Belgium Flemish in most content areas, except in 'operations with rational numbers' and 'algebraic expressions'. In comparison with the remaining three countries, differences with Slovenia were observed in fewer content areas, however, most of these differences were in favor of the particular reference country.

Considering only the cases in which Slovene scores were significantly lower than in at least two reference countries, content areas 'natural numbers', 'meaning of rational numbers', 'geometrical shapes', and 'measurement' seem to be weak for Slovenia. However, additional analysis in which items were taken as units of analysis revealed that, in content areas 'natural numbers' and 'geometrical shapes', there were few items on which Slovene achievement was significantly lower than in at least two reference countries. Therefore these two content areas cannot be described as weak in general. In contrast, in 'meaning of rational numbers' and 'measurement' relatively large proportions of such items were found showing that indeed there may be weaknesses in Slovene achievement in these areas.

Variation of achievement across cognitive categories was also examined in the second research question. This analysis revealed that Slovene achievement in the lowest category, labeled as 'knowing', was significantly lower than in three reference countries, except in the Slovak Republic. This also may be interpreted as indicating where weaknesses in Slovene achievement in comparison with the reference countries existed.

As mentioned in Chapter 7, an interesting observation emerged from the results of comparisons with the reference countries in the cognitive categories. In general, these categories were considered hierarchical in the sense that in lower categories higher student achievements are expected. This hierarchy was roughly observed in three countries, Slovenia, Hungary, and the Slovak Republic, although significant differences in achievement were found only between the highest and the other three categories. In the two EU countries it was observed that achievements in 'using routine procedures' were significantly lower ($p<0.05$) than in 'using complex procedures'. Furthermore, in the Netherlands, achievement in the category 'using routine procedures' was similar to the achievement in 'investigating and solving problems'.

For the Netherlands, this observation occurred in the context of the mathematics reform that was carried out in early 1990s based on "Realistic Mathematics Education" (Vos, 2002). Vos describes that an important component in this reform was to enable students to make mental images. 'Bare' items, showing mostly mathematical symbols and little text were replaced with items in which the problem is stated within context which approximates the real-life of students. The solutions of such items can then be meaningfully interpreted within this context.

The observations in this study raise a question whether the lower achievement of the Dutch students' in the category using routine procedures as compared to the three candidate countries may be linked to the effects of the reform. This might also be important information for curriculum developers in other countries.

8.2.3 Convergence of the results for the two research questions

Similarities as well as differences have been observed in the descriptions of Slovene achievement from the perspectives of the attainment targets and of achievements of students from the reference countries.

Content areas 'operations with rational numbers', 'algebraic expressions', and 'functions and proportionality' can be described as generally satisfactory from both perspectives. Also, achievements in cognitive categories 'using routine procedures' and 'using complex procedures' seem to correspond with both reference points. For content area 'meaning of rational numbers' and, partially, for cognitive category 'knowing' weaknesses in Slovene achievement were observed from both perspectives. The two latter similarities can certainly be described as undesired for the outcomes of the Slovene mathematics education.

At the same time, in other areas, the descriptions of Slovene achievement differed between the two reference points. The main difference was observed for content area 'measurement'. While Slovene scores in this area corresponded with the standards, they were significantly lower than in two reference countries, Belgium Flemish and Hungary. Moreover, there were no strong items in this content area for Slovenia (items on which Slovene scores would be significantly higher than in at least two reference countries). This finding can be used for the refinement of the standards in the curriculum. Similarly, it is important to know that while Slovene students achieved significantly lower than the standards when items required skills of 'investigating and solving problems', they nonetheless achieved comparably on these items as their counterparts in the reference countries.

8.3 DISCUSSION

In this section, a reflection is given on the methodology in this study in which several choices were made. This is followed by a discussion of the conclusions following from the results of this study.

8.3.1 Methodological issues

As indicated in the second research question, it was important for this study to use international data. There are a number of advantages as well as disadvantages in this. An obvious advantage is that comparisons with other countries on the same measurement instruments are available. Furthermore, an advantage is that these data were collected using up-to-date methodology which was developed following explicit and high standards (Martin et al., 1999). Measurement instruments were carefully prepared and their statistical properties checked to ensure reliability and validity for the aims of the study. For example, in TIMSS three-quarters of the items were multiple choice items. While the 'over utilization' of these type of items can be criticized, their statistical properties are important for ensuring comparability of the results of international studies.

There are also disadvantages in the utilization of international data for describing achievement nationally. While in national assessments instruments can be prepared to reflect the national curriculum in the subject assessed, it is difficult in international assessments to expect perfect coverage of the curriculum of a particular country. In this study, it was observed that the attainment targets at higher level were considerably underrepresented in the TIMSS achievement tests. Nevertheless, at the level of "minimum" (called Level 1) and "fundamental" targets (called Level 2) coverage was sufficient to enable interpretations that are meaningful for users in Slovenia.

In addition, content areas defined in international studies may not reflect the structure of the country's curriculum. In this study, the content areas were re-defined to better suit the purposes of describing Slovene achievement in mathematics. For example, items about areas of geometrical shapes were classified in TIMSS as measurement topics, while in Slovenia they are taught as a geometry topic. TIMSS content category 'fractions and number sense' was divided for Slovenia into three categories, 'natural numbers', 'meaning of rational numbers', and 'operations with rational numbers'. By using the Slovene content areas, weaknesses were detected in meaning of rational numbers which were not revealed in the TIMSS classification. The inappropriateness of the TIMSS content areas for the purposes of describing Slovene achievement was already observed by Magajna (2000).

An important methodological issue in educational assessment is how the students' responses on individual items will be summarized. As discussed in Chapter 3, two general approaches are used in TIMSS, the item percent correct approach and the IRT approach. While the latter has many advantages, there were practical impediments for its utilization in this study. It is time consuming and generally not understood by the lay public. Therefore, in this study, percents correct were computed for estimates of item difficulties and student achievement.
This caused some disadvantages for other parts of this study. For example, estimates of student achievement in the form of average item percents correct are sensitive to particular items used and the particular students sampled. As a consequence, when estimating trends in student achievement it is recommended that the same instruments be used. In TIMSS, while there was a set of identical items in the achievement tests of 1995 and 1999, most items were changed in some details. These were called cloned items. To examine the possibility of also using cloned items for trend analysis, it was first examined to what extent trend results on the identical items differed from the results on the cloned items. Considerable differences in trend comparisons of student achievement were found between these two groups of items and therefore it was decided to only use identical items for the analysis of trends.

The re-classification of items into the Slovene content areas was carried out by a single mathematics curriculum expert while the attainment targets were determined by three experts. These persons were chosen for their expert knowledge on the mathematics curriculum in Slovenia. However, a possible threat to the validity of this study might be that items were incorrectly allocated to content areas or levels of the standards. These measurements could have been supported by additional measurements in teacher population or from other experts. This would enhance the validity of the inferences.

Another methodological choice in this study was the operationalization of the levels of the attainment targets in the intended curriculum. As explained in Chapter 2, these targets are set at three levels, but the operational guidelines for measurement of their achievement were not given. The operationalization of these levels in this study used a simple model of three values of 'intended percent correct' based on the description of the general differences between the three levels of the standards and the judgments of experts. This model has its strengths and weaknesses. The advantage is that it enabled direct measurements of the correspondence of student achievement with these levels. It also preserved the hierarchical ordering of the levels in terms of intended achievement. A possible weakness of this model is that, in reality, the underlying intentions in the curriculum to what level of achievement students should master particular contents and processes within the same levels of the standards vary and it is probably not possible to present them as a single number of 'intended percent correct'. Possible individual 'artificially' strong and weak items or strong and weak items that were 'missed' because another value of the intended percent correct for this item might have to be used, may be identified by examining contents of items and their observed difficulties. However, even this procedure may not be sufficient for revealing only the 'real' strengths and weaknesses since these two concepts are dependent on the concept of what 'should' have been attained which might, in turn, be different for different persons setting this desired level of attainment. Nevertheless, for the purposes of summarizing the achievements in this study as compared to the standards this model was deemed sufficient.

In carrying out an educational assessment for the purposes of monitoring and developing remedial actions, measures of the implemented curriculum are important (Pelgrum, 1990). This study did not include this appearance of the curriculum for the following reason. Since the attained curriculum and the intended curriculum in this study were compared between two different systems, the weaknesses observed in achievement when compared to the standards were not necessarily weaknesses from the perspective of the non-reformed intended curriculum. It would be meaningful to use measurements of the implemented curriculum for trying to explain the discrepancies between the intended and the attained curriculum within the same system or to try to develop new approaches to teaching in the new system. Since this study described the 'starting point' of the new curriculum and did not embark on finding the causes of the observed discrepancies, the implemented curriculum was not considered. But it needs to be considered in possible future studies that will examine whether the attained curriculum has progressed from the starting point in the desired direction.

It was argued that the content areas defined in the international reports do not sufficiently reflect the structure of the Slovene mathematics curriculum (see also Magajna, 2000). The international content categories were redefined in this study to better reflect the structure of the national curriculum. The importance for doing this was shown by the fact, that in the content area 'meaning of rational numbers' weaknesses in student achievement were observed in this study from the perspective of the attainment targets as well as in comparisons with the reference countries. This weakness did not emerge in comparisons provided by the international reports since they were masked by strengths in other parts of the large content area 'fractions and number sense'.

Finally, although TIMSS achievement tests were designed to cover a wide range of content area and special data collection procedures were used to gather comparable data on each item (e.g., rotation of the eight different test booklets), the numbers of items for some of the analyses in this study were too low. This occurred in the content area probability. Fortunately, this area is 'small' in the Slovene curriculum and the inferences about student achievement by excluding this area were not seriously affected. Low numbers of items in content areas and cognitive categories also impeded some analyses of trends at item level as well as at the levels of the standards.

8.3.2 Conclusions regarding mathematics achievement in Slovenia

Pelgrum (1990) argued that the main goal of educational assessments is to improve education by monitoring the output of the education system. Information about the curricular appearances (the intended, the implemented, and the attained curriculum) and, in particular, comparisons between them constitutes important feedback for educational actors at different levels of education. The discrepancies between the curricular appearances could provide important information, which can be used when considering at which levels in the system improvement measures should be introduced.

As discussed in Chapter 1, there is in many countries a growing interest in national and international assessments in order to provide such information. International assessments have especially increased in importance considering the high numbers of participating countries in IEA studies and PISA. In Slovenia, there are no national assessments that would be carried out specifically with the purpose to serve policy needs. The country, however, participated in a number of international assessments. Being a smaller European country it needs to develop and continuously nurture ties with other countries. One of the aspects in this is comparing the country's education system with other education systems in Europe.

As also described in Chapter 1, there may be many benefits of participation in international assessments. It is not unimportant for Slovenia that a research capacity was built and studies were carried out following high standards of quality. However, the results of these assessments presented in international reports and in a few national reports are not sufficient for the country to benefit sufficiently from the time consuming and expensive data collection. There is no statistical procedure that could describe the education system in a single number and there are many different users of assessment information whose needs have to be addressed using different approaches.

In the process of the curriculum reform and its implementation, information specifically addressing the needs of curriculum developers and teachers needs to be provided. In Slovenia international assessment data have not been used sufficiently for this purpose. Kellaghan and Greaney (2001) observed that utilization of assessment data and getting results in a meaningful form to teachers pose particular challenges. This can also be observed in Slovenia. Slovene teachers generally express great willingness to participate in system-level assessments and recognize their importance for the monitoring purposes. However, the feedback they receive often lacks usefulness for their classroom teaching.

In effort to provide this feedback, this study 'zoomed in' on Slovene results in a large international comparative study to go beyond norm-referenced comparisons in the international reports. It tried to address the information needs of actors in the current reform of mathematics education. Although this new curriculum has already been developed, the information resulting from this study is expected to be useful in its implementation.

This study was basically a descriptive study. It provided information on the strengths and weaknesses observed in achievement of students when compared to the relevant reference points. By linking student achievement in the non-reformed system to the reformed curriculum it provides information on the point from which the reformed curriculum is starting. This information may be also used in possible future studies to examine whether the desired effects of the reforms can be observed.

The most general finding in this study was that from the perspective of the standards in the intended curriculum as well as from the international perspective, it seems that Slovene education system yielded satisfactory results. In depth investigations pointed out weaknesses from the two perspectives. There were also strengths observed. However, they were mostly from the perspective of the standards while there were little strengths observed from the perspective of the reference countries. Since only higher or similarly achieving countries were selected for comparisons, this does not necessarily mean that the standards in the curriculum are too low, but it may mean that they are 'overachieved' also in the reference countries.

The results of this study pointed out the importance of in-depth analyses of the national results in an international context. For example, the international TIMSS reports (Beaton et al., 1996; Mullis et al., 2000) showed that Slovenia is among the higher achieving European countries. However in more detailed comparisons it was found that there is room for improvement. In comparisons with the reference countries in this study it was observed that although Slovenia and the reference countries, except Belgium Flemish, have similar overall achievements, there were some 'individual' differences, most of which were in favor of the particular reference country.

This study also showed the importance of describing student achievement from several perspectives. While the intended curriculum is an important perspective for student achievement, it was shown, that there might be differences between the intentions in the national curriculum and what is possible to achieve as observed in other countries. This function of international assessments is termed "enlightenment" (Kellaghan, 1996). It may be that the intentions in the curriculum might be too high, however, it may also be that they are too low. Furthermore, this "enlightenment" is important in national examinations that are being carried out in the reformed system in Slovenia. Broadfoot (2002) emphasizes that international as well as national data should be used to form "dynamic standards" used in support for enhancing the quality of education as opposed to "arbitrary standards" that may be imposed on students and teachers resulting in 'teaching to the test'.

Using the results of this study, Slovene educators are in a position to understand better student achievement and at the same time the attainment targets that were set in the reformed curriculum. This study was not an evaluation of these attainment targets but it aimed at providing information on what they mean and how can they be used. Through measurements of student achievements and analyses of their correspondence with the standards as well as with other reference points, these standards may be made more refined in terms of wording and content as well as in terms of intended levels of achievement. Although the perspectives used in this study do certainly not reflect all possible views from which achievement of students could be described, they provide wealth of information that could help Slovene educators in their efforts to improve Slovene mathematics education and its outcomes.

8.4 RECOMMENDATIONS

8.4.1 Recommendations for policy and practice

As argued, the substantial changes in the education system in Slovenia should be supported with all means possible. Apart from setting up the legislation and the new curricula, it is important that teachers are constantly supported in their role of the implementers of these curricula. One way of support is the monitoring of education in which information on student achievement as compared to relevant reference points is provided. While teachers themselves carry out parts of this monitoring process in their everyday school life, information on the achievement of students between classes, schools and countries is also important.

It is therefore recommended that the results of this study be communicated to teachers as well as other curriculum experts. However, this information should be accompanied with possible proposals and suggestions for improvement that teachers can introduce in their teaching methods. It is also recommended that additional studies on the part of Slovene educators are carried out to identify possible causes of the weaknesses observed. Even though many improvement measures were already embedded in the reformed curriculum, detailed information on the areas where weaknesses in the non-reformed system existed is important for alerting teachers to known deficiencies, in particular by illustrating them through individual items.

But teachers should not only be given this information. It is recommended that they are also involved in the process of diagnosing causes and developing remedial actions. Only in this way it will be ensured that actions will be developed that could be implemented in classrooms. Monitoring can then be used to examine whether the desired progress occurred.

8.4.2 Recommendations for future research

The fact that the Slovene education system is currently being reformed implies the recommendation that the outputs as well as inputs and processes in the system be continuously monitored. This study provided information about the status of mathematics achievement in Slovenia at the end of the 1990s. Given the availability of the TIMSS 2003 data for Slovenia in the near future, these data can be used to provide further information on the current status of mathematics education in a way similar as in this study. Through this first indications of effects of the reforms might be provided.

Although this study did not use measures of the implemented curriculum, it is important to include them in future studies. Data on implemented curriculum might provide information for identifying possible causes of the concurrent weaknesses or even undesired effects of the reforms.
In this study, student achievement was described at the level of the overall population. Future research could be carried out describing achievement for different subgroups of students, for example ability levels, or school environment. A study describing student achievement in different ability groups of students might be of special interest to the Slovene teachers in the two final grades where students are taught mathematics in differentiated classes.

Limitations of paper and pencil test for covering certain areas in student achievement were also recognized in this study. For example, there were very few items that would cover the higher level attainment targets in the Slovene curriculum. Also, while in some of the content areas students' average scores were higher than the standards, it was observed that there were also weak items in these areas. Larger numbers of items in these areas might reveal the strengths and weaknesses more clearly. A wider and at the same time more refined coverage of the domain might be achieved through other types of testing, such as tailored testing and internet testing. They might be more suitable for obtaining detailed information on student achievement but also avoiding presenting large numbers of items to students.

And last but not the least, educational studies have recognized the importance of the background variables in trying to explain variations in student achievement between and within countries (IEA, 1998). Through the three curriculum appearances student achievement is linked to the contextual factors, such as the school environment, teacher characteristics, and students' own attitudes towards the subject. In studies of possible causes of weaknesses in student achievement and development of improvement measures, these factors should also be considered.

References

Adams, R.J., & Gonzalez, E.J. (1996). Design of the TIMSS achievement instruments. In D.F. Robitaille, & R.A. Garden (Eds.), *TIMSS monograph no.2: research questions and study design* (pp. 81-102). Vancouver, Canada: Pacific Educational Press.

Adams, R.J., Wu, M.L., & Macaskill, G. (1997). Scaling methodology and procedures for the mathematics and science scales. In M.O. Martin, & D.L. Kelly (Eds.), *TIMSS technical report, volume II: implementation and analysis* (pp. 111-145). Chestnut Hill, MA: Boston College.

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

Akker, J.J.H., van den, (1998). *De uitbeelding van heat curriculum* [Representing the curriculum]. [Inaugural lecture]. Enschede, the Netherlands: University of Twente.

Beaton, A.E. (1987). *Implementing the new design. The NAEP 1983-84 technical report.* Princeton, NJ: Educational Testing Service.

Beaton, A.E., & Gonzalez, E.J. (1997). Reporting achievement in mathematics and science content areas. In M.O. Martin, & D.L. Kelly (Eds.), *TIMSS technical report, volume II: implementation, and analysis* (pp. 175-185). Chestnut Hill, MA: Boston College.

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

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

Bloom, B.S. (1956). *Taxonomy of educational objectives; the classification of educational goals.* London: Longman

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.T. (2002). *The benefits and limitations of large-scale international comparative achievement studies. The case of IEA's TIMSS study.* Doctoral dissertation. Enschede, the Netherlands: University of Twente.

Bottani, N., & Tuijnman, A. (1994). International Education Indicators: framework, development and interpretation. In OECD, *Making education count: developing and using international indicators* (pp. 21-35). Paris: OECD.

Brečko, B. (2003). National policies and practices on ICT in education : Slovenia. In T. Plomp (Ed.), *Cross-national information and communication technology polices and practices in education* (pp. 525-539). Greenwich: Information Age.

Brick, J.M., Broene, P., & Severynse, J. (1997). *A user's guide to WesVarPC. Version 2.1.* Rockville, MD: Westat.

Broadfoot, P. (1994). Achievements of learning. In OECD, *Making education count: developing and using international indicators* (pp. 237-263). Paris: OECD.

Broadfoot, P. (2001). Editorial: new wine in old bottles? The challenge of change for educational assessment. *Assessment in education, 8*, 109-112.

Broadfoot, B. (2002). Editorial. Dynamic versus arbitrary standards: recognizing the human factor in assessment. *Assessment in education, 9,* 157-159.

Bryk, A., & Hermanson, K. (1994). Observations on the structure, interpretation and use of education indicators systems. In OECD, *Making education count: developing and using international indicators* (pp. 37-53). Paris: OECD.

Cole, N.S. (1990). Conceptions of educational achievement. *Educational researcher*, *19*, 2-7.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Hartcourt Brace Jovanovich College.

Crooks, T.J., Kane, M.T., & Cohen, A.S. (1996). Threats to the valid use of assessments. *Assessment in education, 3,* 265-285.

Curriculum guide (1984). See *Program življenja in dela v osnovni šoli.*

Curriculum guide (2002). See *Učni načrt: program osnovnošolskega izobraževanja: matematika.*

Dochy, F.J., & Moerkerke, G. (1997). Assessment as a major influence on learning and instruction. *International Journal of Educational Research, 27*, 415-432.

Dossey, J.A., & Lindquist, M.M. (2002). The impact of TIMSS on mathematics standards movement in the United States. In D.F. Robitaille, & A.E. Beaton, *Secondary analysis of the TIMSS data* (pp. 63-80). Dordrecht, the Netherlands: Kluwer Academic.

Fitz, C.T. (1996). Official indicator systems in the U.K.: examinations and inspections. *International Journal on Educational Research*, *25*, 239-247.

Foy, P., Rust, K., & Schleicher, A. (1996). Sample design. In M.O. Martin, & D.L. Kelly (Eds.), *TIMSS technical report, volume I: design and development* (pp. 4-1 – 4-17). Chestnut Hill, MA: Boston College.

Garden, R.A. (1996). Development of the TIMSS achievement items. In D.F. Robitaille, & R.A. Garden (Eds.), *TIMSS monograph no.2: research questions and study design* (pp. 69-80). Vancouver, Canada: Pacific Educational Press.

Garden, R.A., & Smith, T.A. (2000). TIMSS test development. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS technical report, IEA's repeat of the Third International Mathematics and Science Study at the eighth grade* (pp. 49-67). Chestnut Hill, MA: Boston College.

Glaser, R. (1984). Education and thinking: the role of knowledge. *American psychologist, 39*, 93-104.

Glaser, R. (1988). Cognitive and environmental perspectives on assessing achievement. In *Assessment in the service of learning: proceedings of the 1987 ETS Invitational Conference*. Princeton, NJ: Educational Testing Service.

Glaser, R., & Silver, E. (1994). Assessment, testing and instruction. *Annual review of psychology, 40*, 631-666.

Gonzalez, E.J., & Beaton, A.E. (1994). The determination of cut scores for standards. In A.C. Tuijnman, & T.N. Postlethwaite (Eds.), *Monitoring the standards of education: papers in honor of John P. Keeves* (pp. 171-190). Oxford: Pergamon Press.

Gonzalez, E.J., & Foy, P. (2000). Estimation of sampling variance. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS technical report, IEA's repeat of the Third International Mathematics and Science Study at the eighth grade* (pp. 203-222). Chestnut Hill, MA: Boston College.

Gonzalez, E.J., & Gregory, K.D. (2000). Reporting student achievement in mathematics and science. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS technical report, IEA's repeat of the Third International Mathematics and Science Study at the eighth grade* (pp. 279-295). Chestnut Hill, MA: Boston College.

Gonzalez, E.J., & Miles, J.A. (Eds.). (2001). *User guide for the TIMSS 1999 international database.* Chestnut Hill, MA: Boston College.

Gonzalez, E.J., & Smith, T.A. (Eds.). (1997). *User guide for the TIMSS international database: primary and middle school years assessment*. Chestnut Hill, MA: Boston College.

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.

Hambleton, R.K., & Rogers, H.J. (1989). Solving criterion-referenced measurement problems with item response models. *International journal of educational research, 13*, 145-173.

Hambleton, R.K., & Sireci, S.G. (1997). Future directions for norm-referenced and criterion-referenced achievement testing. *International journal of educational research, 27*, 379-393.

Herman, J.L. (1997). Large-scale assessment in support of school reform: lesson in the search for alternative measures. *International journal of educational research, 27*, 395-413.

Husén, T. (Ed.) (1967). *International study of achievement in mathematics: a comparison of twelve countries (vols. I and II).* New York: Wiley.

Husén, T., & Tuijnman, A. (1994). Monitoring standards in education: Why and how it came about. In A.C. Tuijnman, & T.N. Postlethwaite (Eds.), *Monitoring the standards of education: papers in honor of John P. Keeves* (pp. 1-21). Oxford: Pergamon Press.

IEA (International Association for Evaluation of Educational Achievement) (1997). *TIMSS international database for the primary and middle school years*. Chestnut Hill, MA: Boston College.

IEA (1998). *IEA guidebook 1998: activities, institutions and people*. Amsterdam: Author.

IEA (2001). *TIMSS 1999 international database*. Chestnut Hill, MA: Boston College.

Ivic, I. (1994). Theories of mental development and assessing educational outcomes. In OECD, *Making education count: developing and using international indicators* (pp. 197-218). Paris: OECD.

Japelj, B. (1993). Rezultati raziskave IAEP [The results of the IAEP study]. *Matematika v šoli, 2*, 13-22.

Joint Committee on Standards of Educational Evaluation (1994). *The program evaluation standards: how to assess evaluation of educational programs (*2nd ed.). Thousand Oaks: Sage.

Keeves, J.P. (1992). Longitudinal research methods. In J.P. Keeves (Ed.), *The IEA technical handbook* (pp. 15-38). The Hague, The Netherlands: IEA.

Kellaghan, T. (1996). IEA studies and educational policy. *Assessment in education, 3,* 143-160.

Kellaghan, T., & Greaney, V. (2001) *Using assessment to improve quality of education.* Paris: UNESCO International Institute for Educational Planning.

Kelly, D.L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) achievement scales using scale anchoring*. Doctoral disseration. Chestnut Hill, MA: Boston College.

Kelly, D., Mullis, I.V.S., & Martin, O.M. (2000). *Profiles of student achievement in mathematics at the TIMSS international banchmarks: U.S. performance and standards in an international context.* Chestnut Hill, MA: Boston College.

Kilpatrick, J. (2001). Understanding mathematical literacy: the contribution of research. *Educational studies in mathematics, 47*, 101-116.

Kuiper, W., Bos, K., & Plomp, T. (1999). Mathematics achievement in the Netherlands and appropriateness of the TIMSS mathematics test. *Educational research and evaluation, 5*, 79-84.

Lapointe A.E., Mead, N.A., & Askew, G.M. (1992). *Learning mathematics*. Princeton, NJ: Educational Testing Service.

Larkin, J.H., & Rief, F. (1979). Understanding and teaching problem solving in physics. *European journal of science education, 1*, 191-203.

Linn, R.L. (2000). Assessments and accountability. *Educational researcher, 29,* 4-16.

Magajna, Z. (2000). Obravnava primanjkljajev matematičnega znanja slovenskih učencev v učnem načrtu slovenske osnovnošolske matematike [Addressing Slovene students' lack of mathematical skills in the elementary school mathematics syllabus]. *Sodobna pedagogika*, *2*, 162-180.

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

Matos, U. (1997). Slovenski šolarji v mednarodni konkurenci [Slovene students in international competition]. *Mladina*, *14*, 18-19.

Mislevy, R.J. (1995). What can we learn form international assessments? *Educational evaluation and policy analysis, 17*, 419-437.

Mislevy, R.J., Wilson, M.R., Erickan, K., & Chudowsky, N. (2003). Psychometric principles in student assessment. In T. Kellaghan, D.L. Stufflebeam, & L.A. Wingate (Eds.), *International handbook of educational evaluation* (pp. 489-532). Dordrecht, the Netherlands: Kluwer Academic.

Mullis, I.V.S, & Martin, M.O. (2000). Item analysis and review. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS technical report, IEA's repeat of the Third International Mathematics and Science Study at the eighth grade* (pp. 225-234). Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Gregory, K.D., Garden, R.A., O'Connor, K.M., Chrostowski, S.J., & Smith, T.A. (2000). *TIMSS 1999 international mathematics report: findings from IEA's repeat of the Third International Mathematics Science Study at the eighth grade.* Chestnut Hill, MA: Boston College.

Mullis, I.V.S., Martin, M.O., Smith, T.A., Garden, R.A., Gregory, K.D., Gonzalez, E.J. Crostowski, S.J., & O'Connor, K.M. (2003). *TIMSS assessment frameworks and specifications* (2nd ed.). Chestnut Hill: Boston College.

National Curriculum Council (1996). See Nacionalni Kurikularni Svet.

Nacionalni Kurikularni Svet (1996). *Izhodišča kurikularne prenove* [Premises for the Curriculum Reform]. Ljubljana, Slovenia: Ministry of Education and Sports.

Nunnally, J.C. (1967). *Psychometric theory*. New York: McGraw Hill.

O'Connor, K.M., & Malak, B. (2000). Translation and cultural adaptation of the TIMSS instruments. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS technical report, IEA's repeat of the Third International Mathematics and Science Study at the eighth grade* (pp. 89-100). Chestnut Hill, MA: Boston College.

O'Connor, K.M., & Stemler, S.E. (2000). Quality control in the TIMSS data collection. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS technical report, IEA's repeat of the Third International Mathematics and Science Study at the eighth grade* (pp. 135-154). Chestnut Hill, MA: Boston College.

OECD Organisation for Economic Cooperation and Development (1998). *Reviews of national policies for education – Slovenia.* Examiners' report. Paris: Author.

OECD (2002). *The OECD Programme for International Student Assessment*. Paris: Author. Retreived February 2004 from http://pisa.oecd.org/

Pelgrum, W.J., (1990). *Educational assessment: monitoring, evaluation and the curriculum*. De Lier, the Netherlands: Academisch Boekencentrum.

Pelgrum, W.J. (1995). *Monitoring education indicators in the EU Member states*. Enschede, the Netherlands: University of Twente.

Pelgrum, W.J. (1998). *European network for educational research assessment, effectiveness and innovation.* Brussels: European Commission.

Pelgrum, W.J., Eggen, T.J.H.M., & Plomp, T. (1986). *The implemented and attained mathematics curriculum: a comparison of eighteen countries.* Enschede, the Netherlands: University of Twente.

Phillips, G.W. (1994). Methods and issues in setting performance standards. In A.C. Tuijnman, & T.N. Postlethwaite (Eds.), *Monitoring the standards of education: papers in honor of John P. Keeves* (pp. 191-212). Oxford: Pergamon Press.

Plevnik, T., Žižmond, A. (Eds.) (2000). *The Education System in Slovenia 2000*. Ljubljana, Slovenia: Ministry of Education and Sports, Eurydice Unit .

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

Plomp, T., Howie, S., & McGaw, B. (2003). International studies of educational achievement. In T. Kellaghan, D.L. Stufflebeam, & L.A. Wingate (Eds.), *International handbook of educational evaluation* (pp.951-978). Dordrecht, the Netherlands: Kluwer Academic.

Popham, W.J., (1993). *Educational evaluation.* (3rd ed.). Boston: Allyn and Bacon.

Postlethwaite, T.N. (1994). Monitoring and evaluation in different education systems. In A.C. Tuijnman, & T.N. Postlethwaite (Eds.), *Monitoring the standards of education: papers in honor of John P. Keeves* (pp. 23-46). Oxford: Pergamon Press.

Program življenja in dela v osnovni šoli [The programme of life and work in elementary school]. (1984). Ljubljana, Slovenia: Ministry of Education and Sports.

Rabinowitz, M., & Glaser, R. (1985). Cognitive structure and process in highly competent performance. In F.D. Horowitz, & M. O'Brien (Eds.), *The gifted and talented: developmental perspectives* (pp.75-98). Washington, DC: American Psychological Association.

Resnick, L.B. (1987). Learning in school and out. *Educational researcher, 16,* 13-20.

Resnick, L.B., & Resnick, D.P. (1989). Assessing the thinking curriculum: new tools for educational reform. In B.R. Gifford, & M.C. O'Connor (Eds.), *Future assessment: changing viewvs of aptitude, achievement and instruction.* Boston: Kluwer.

Robitaille, D.F., Beaton, A.E., & Plomp T. (Eds.). (2000). *The impact of TIMSS on the teaching and learning of mathematics and science.* Vancouver, Canada: Pacific Educational Press.

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

Robitaille, D.F., & Garden, R.A. (1996). Design of the study. In D.F. Robitaille, & R.A. Garden (Eds.), *TIMSS monograph no.2: research questions and study design* (pp. 44-68). Vancouver, Canada: Pacific Educational Press.

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

Robitaille, D.F., & Robeck, E.C. (1996). The characteristics and the context of TIMSS. In D.F. Robitaille, & R.A. Garden (Eds.), *TIMSS monograph no.2: research questions and study design* (pp. 15-33). Vancouver, Canada: Pacific Educational Press.

Robitaille, D.F., Schmidt, W.H., Raizen, S.A., McKnight, C.C., Britton, E., & Nicol, C. (1993). *TIMSS monograph no.1: curriculum frameworks for mathematics and science*. Vancouver, Canada: Pacific Educational Press.

Rosier, M., & Keeves, J. (1991). *The IEA study of science I: science education and curricula in twenty-three countries.* Oxford: Pergamon Press.

Rutar Ilc, Z. (2003). Pristopi k poučevanju, preverjanju and ocenjevanju [Approaches to teaching, assessment and grading]. Ljubljana, Slovenia: National Educational Institute.

Saxe, G.B. (1988). Candy selling and math learning. *Educational researcher, 17,* 14-21.

Schleicher, A. (2000). Monitoring student knowledge and skills: The OECD Programme for International Student Assessment. In D. Schorrocks-Taylor, & E.W. Jenkins (Eds.), *Learning from others* (pp. 63-78). Boston: Kluwer Academic.

Schmidt, W.H., Jakwerth, P.M., & McKnight, C.C. (1998). Curriculum sensitive assessment: Content does make a difference. *International journal of educational research, 29*, 503-527.

Schmidt, W.H., McKnight, C.C., Valverde, G.A., Houang, R.T., & Wiley, D.E. (1997). *Many visions, many aims volume 1: A cross-national investigation of curricular intentions in school mathematics.* Dordrecht, The Netherlands: Kluwer Academic.

Statistical Office of the Republic of Slovenia. (2002a). *Popis 2002* [Census 2002]. Ljubljana, Slovenia: Author. Retrieved, February 2004, from http://www.stat.si/popis2002/si/rezultati/rezultati_red.asp?ter=SLO&st=7

Statistical Office of the Republic of Slovenia. (2002b). *Popis 2002* [Census 2002]. Ljubljana, Slovenia: Author. Retrieved, February 2004, from http://www.stat.si/popis2002/si/rezultati/rezultati_red.asp?ter=SLO&st=8

Šetinc, M. (1991). *The average achieving curriculum : comparative assessment of pre-university mathematics in Slovenia.* Ljubljana, Slovenia: The Slovene Society of Researchers in The School Field.

Šetinc, M., Japelj, B., & Trobec, M. (1997). *Znanje sedmošolcev in osmošolcev za vstop v 21. stoletje.* [The knowledge and skills of the seventh- and eighth-graders for entering the 21st century]. Ljubljana, Slovenia: Educational Research Institute.

Štraus, M. (2003). Some strengths and weaknesses in Algebra achievement of Slovene eighth-grade students. *Studies in educational evaluation*, *29*, 207-226.

Štraus, M., Čuček, M., Gril, A., Doupona Horvat, M., & Japelj, B. (2003). *Doseganje in primerljivost standardov znanja iz slovenskega kurikula s standardi znanja iz mednarodnih primerjalnih raziskav: evalvacijska študija 2000-2003: poročilo o raziskavi* [Attainment and comparability of the standards from the Slovene curriculum with the standards of knowledge and skills from the international comparative studies: evaluation study 2000-2003: a research report]. Ljubljana, Slovenia: Educational research institue.

Thomas, R.M. (1994). Approaches to setting and selecting achievement standards. In A.C. Tuijnman, & T.N. Postlethwaite (Eds.), *Monitoring the standards of education: papers in honor of John P. Keeves* (pp. 101-120). Oxford: Pergamon Press.

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

Učni načrt: program osnovnošolskega izobraževanja: matematika. [Curriculum guide: the program of elementary education: mathematics] (2002). Ljubljana, Slovenia: Ministry of Education and Sports.

Unescov priročnik za uporabo mednarodne standardne klasifikacije izobraževanja Isced 1997: problemi pri uporabi Isced 1997 v Sloveniji [Unesco manual for applying international standards for classification of education Isced 1997. Problems in applying Isced 1997 in Slovenia] (2001). Ljubljana, Slovenia: Ministry of Education and Sports.

Vos, P. (2002). *Like an ocean liner changing course: the grade 8 mathematics curriculum in the Netherlands, 1995-2000.* Doctoral dissertation. Enschede, the Netherlands: University of Twente.

Walberg, H.J., & Haertel, G.D. (1994). The implications of cognitive psychology for measuring student achievement. In OECD, *Making education count: developing and using international indicators* (pp. 219-236). Paris: OECD.

Weiss, C.H. (1981). Measuring the use of evaluation. In J.A. Ciarlo (Ed.), *Utilizing evaluation: concepts and measurement techniques* (pp. 17-33). London: Sage.

White paper on education in the Republic of Slovenia (1996). Ljubljana, Slovenia: Ministry of Education and Sports.

Wiliam, D. (1996). Meanings and consequences in standard setting. *Assessment in education, 3*, 287-307.

Wolf, R.M. (1994). The validity and reliability of outcome measures. In A.C. Tuijnman, & T.N. Postlethwaite (Eds.), *Monitoring the standards of education: papers in honor of John P. Keeves* (pp. 121-132). Oxford: Pergamon Press.

Wolf, R.M. (1998). Validity issues in international assessments. *International journal of educational research, 29*, 491-501.

Yamamoto, K., & Kulick, E. (2000). Scaling methodology and procedures for the TIMSS mathematics and science scales. In M.O. Martin, K.D. Gregory, & S.E. Stemler (Eds.), *TIMSS technical report, IEA's repeat of the Third International Mathematics and Science Study at the eighth grade* (pp. 237-263). Chestnut Hill, MA: Boston College.

ENGLISH SUMMARY
Mathematics Achievement of Slovene Students at the End of Compulsory Education

BACKGROUND OF THE STUDY

In the past decade, there have been major changes in Slovenia. Country's independence in 1991 yielded reforms in nearly all aspects of social life. Within this, education was one of the earliest areas to receive attention. After the legislation for education reform was adopted in 1996, panels of curriculum experts were mandated to review the existing curricula and prepare proposals for their reforms. In addition to compulsory education, this process encompassed all other levels of pre-university education, including kindergarten. At the level of compulsory education, the major additions to the curricula included the attainment targets, also called standards, which were specified for each subject in each grade.

Within this context, the present study describes mathematics achievement of Slovene students at the end of compulsory education *prior to* the implementation of the reformed curriculum from two perspectives. Furthermore, developments in Slovene students' achievements were also examined over time by comparing achievement data from 1995 to 1999. Through this, possible areas of improvement in Slovene mathematics education were indicated.

RESEARCH QUESTIONS AND STUDY DESIGN

Two general reference points were constructed for describing Slovene achievement in this study. The first reference point was constructed on the basis of the attainment targets in the reformed curriculum. Even though the students assessed studied the non-reformed curriculum, the aim of this study was to describe the knowledge and skills of Slovene students in the light of the intentions embedded in

the reforms. Through this, information on what can be called 'the baseline' for the new curriculum was provided.
Since reaching internationally comparable achievement was emphasized in the reform, comparisons with other countries were also considered important in this study. This led to the construction of the second reference point. In the very near future, in May 2004, Slovenia will become a member of the EU along with nine other European countries. In the present study, it was of interest to compare mathematics achievements of Slovene students to achievements of students in the EU countries and in the other candidate countries. In order to provide information on areas of possible improvements, countries with similar or higher average achievements were selected for these comparisons: Belgium Flemish, the Netherlands, Hungary, and the Slovak Republic.[1]

These two reference points were reflected in the following two research questions for this study:

1. *How well did Slovene students at the end of compulsory education in the non-reformed system in the late 1990s perform in mathematics when compared to the attainment targets in the reformed mathematics curriculum and what were the developments in this performance between 1995 and 1999?*

2. *How well did Slovene students at the end of compulsory education in the non-reformed system in the late 1990s perform in mathematics when compared to the performance of students in other European countries and what were the developments in this performance between 1995 and 1999?*

To address these research questions, the data from the 1995 and 1999 data collections in the Third International Mathematics and Science Study (TIMSS), conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) were used. In the design of this study, the TIMSS items were used to describe the intended and the attained curricula in Slovenia as well as the attained curricula in the reference countries. The measures of the attained curricula were obtained through the students' scores on these items. The measures of the intended curricula were provided by three mathematics experts in Slovenia in the form of the levels of the attainment targets at which the items were allocated and operationalized into 'intended percent correct' for these items.

[1] For reasons of brevity the term 'country' is used synonymously with 'educational system'.

Two basic analysis approaches were used: a student based and an item based analysis. In the first approach, average scores of students on items and their correspondence with the reference points were examined, whilst in the second approach, the correspondences of achievements on individual items with the reference points were examined. In student based analysis, when the estimate of Slovene achievement was significantly[2] higher than the reference point, this was taken as an indication of strengths in Slovene achievement. In case it was significantly lower, this was taken as an indication of weaknesses. In the third case, Slovene achievement was described as corresponding with the reference point. When the reference point was constructed on the basis of achievements of students from the reference countries, only when Slovene achievement was significantly higher than in at least two reference countries this was considered an indication of strengths. If it was significantly lower than in at least two reference countries, this was considered as an indication of weaknesses.
In the item based analysis, items with significantly higher difficulties (item percents correct) than the reference point were classified as strong. Items with significantly lower difficulties were classified as weak. In case of the four reference countries, items were classified as strong or weak, respectively, when their difficulties for Slovenia were significantly higher or lower than in at least two reference countries. Inferences about strengths and weaknesses in Slovene achievement were made if significantly high or low numbers of items were classified as weak or strong.

The analyses were carried out at the levels of the overall mathematics achievement scores, of scores in several subdomains, as well as at the level of individual items. The mathematics subdomains were defined as content areas based on the structure of content in the Slovene mathematics curriculum, and as cognitive categories based on the TIMSS classification. Cognitive categories indicate the types of performances students are expected to exhibit when engaged with the item content. Four cognitive categories from TIMSS classification of items were used in this study: 'knowing', 'using routine procedures', 'using complex procedures', and 'investigating and solving problems'. They were considered to be hierarchically ordered, from less complex to more complex performances although items of variety of difficulties can be constructed for each category.

[2] All significance tests were reported at 0.05 level.

MAIN FINDINGS

The most general finding in this study is that mathematics achievement of Slovene students in the non-reformed system corresponded with the attainment targets in the reformed curriculum. However, this is at the level of overall mathematics achievement only. Detailed analyses revealed variation in the correspondences. In two content areas of the Slovene mathematics curriculum, 'natural numbers' and 'data representation', student achievement in Slovenia can be described as strong. In contrast, in the content area 'meaning of rational numbers', Slovene achievement seems to be weaker than intended according to the attainment targets.

When correspondence of mathematics achievement of Slovene students with the standards was examined with respect to cognitive categories, it was found that overall, students' scores corresponded with the standards in the three lower categories and were significantly lower in the highest category. Nonetheless, some strength in this category was reflected in a fifth of items on which achievement was significantly higher than the standards.

The second research question complemented the description of mathematics achievement of Slovene students from the perspective of the intended curriculum. It was revealed that Slovene scores were significantly lower in comparison with Belgium Flemish in most content areas, except in 'operations with rational numbers' and 'algebraic expressions'. In comparison with the remaining three countries, differences with Slovenia were observed in fewer content areas, however, most of these differences were in favor of the particular reference country. In general, achievement of Slovene students in content areas 'meaning of rational numbers' and 'measurement' can be described as relatively weak in comparison with the reference countries.

Slovene achievement in the lowest category, labeled as 'knowing', was significantly lower than in three reference countries, except in the Slovak Republic, also indicating that there may be weaknesses in Slovene achievement in this area. At the same time in other categories, including 'investigating and solving problems', Slovene achievement seems to correspond with achievements in the reference countries.

The analyses for both research questions showed that Slovene achievement in content areas 'operations with rational numbers', 'algebraic expressions', and 'functions and proportionality' can be described as generally satisfactory from both perspectives. Also, achievements in cognitive categories 'using routine procedures' and 'using complex procedures' seem to correspond with intentions in the curriculum and at the same time with achievement in the reference countries. However, in content area 'meaning of rational numbers' weaknesses in Slovene achievement were observed from both perspectives.

At the same time, while Slovene scores in 'measurement' corresponded with the standards, they were significantly lower than in two reference countries, Belgium Flemish and Hungary. This finding can be used for the refinement of the standards in the curriculum. Similarly, it is important to know that while Slovene students achieved significantly lower than the standards when items required skills of 'investigating and solving problems', they nonetheless achieved comparably on these items as their counterparts in the reference countries.

CONCLUSION

This study was basically a descriptive study. It provided information on the strengths and weaknesses observed in achievement of students when compared to the relevant reference points. By linking student achievement in the non-reformed system to the reformed curriculum it provides information on the point from which the reformed curriculum is starting. This information may be also used in possible future studies to examine whether the desired effects of the reforms will be realized. The study adds to other studies, because it sheds more light onto the areas in which improvements might be desired and possible. The type of information yielded in this study may help curriculum reformers and practitioners refine their conclusions from the review of the old curriculum and develop additional approaches that teachers might use in their teaching. Through this, the success of the reforms is additionally strengthened.

Using the results of this study, Slovene educators are in a position to understand better student achievement and at the same time the attainment targets that were set in the reformed curriculum. This study was not an evaluation of these attainment targets but it aimed at providing information on what they mean and how can they be used. Through measurements of student achievements and analyses of their correspondence with the attainment targets as well as with the reference countries, the attainment targets may be made more refined in terms of wording and content as well as in terms of intended levels of achievement. Although the perspectives used in this study do certainly not reflect all possible views from which achievement of students could be described, they provide wealth of information that could help Slovene educators in their efforts to improve Slovene mathematics education and its outcomes.

NEDERLANDSE SAMENVATTING
Wiskundeprestaties van Sloveense leerlingen aan het einde van de leerplichtige leeftijd

---

ACHTERGROND VAN DE STUDIE

Slovenië heeft de afgelopen tien jaar grote veranderingen ondergaan. De onafhankelijkheid van het land in 1991 heeft geleid tot hervormingen op bijna alle terreinen van de samenleving. Daarbij was de onderwijssector één van de eerste gebieden waarin hervormingen werden voorbereid. Nadat deze hervormingen in 1996 hun wettelijke basis hadden gekregen, is een aantal panels van onderwijsexperts ingesteld met de taak het toenmalige onderwijs door te lichten en voorstellen voor hervormingen voor te bereiden. De beoogde onderwijshervorming betrof niet alleen het onderwijs aan leerplichtige leerlingen, maar ook alle andere sectoren van het primair en secundair onderwijs, inclusief het kleuteronderwijs. Eén van de belangrijkste veranderingen betrof de invoering van eindtermen, ook wel standaarden genoemd, voor alle schoolvakken en voor alle leerjaren. Standaarden zijn geformuleerd op twee niveaus voor de leerjaren 1-7, en op drie niveaus voor de leerjaren 8 en 9. Standaarden op het laagste niveau, aangeduid met 'minimum standaarden', geven aan wat (bijna) alle leerlingen geacht worden te beheersen. Op het tweede niveau wordt gesproken van 'fundamentele standaarden', die verwijzen naar de kennis en vaardigheden die 'gemiddelde' leerlingen moeten kunnen verwerven; leerkrachten worden geacht zich in hun onderwijs op deze standaarden te richten. Op het hoogste niveau (en alleen voor leerjaren 8 en 9) zijn er de 'hogere standaarden' waarvan wordt verwacht dat goede leerlingen deze zullen bereiken.

Het is in deze context dat de voorliggende studie vanuit een aantal invalshoeken de wiskundeprestaties van Sloveense leerlingen aan het einde van de leerplichtige leeftijd beschrijft en analyseert en dat op een tijdstip juist voorafgaande aan de implementatie van de onderwijshervormingen. Daarnaast zijn ook ontwikkelingen in wiskundeprestaties onderzocht door gegevens uit 1995 en 1999 te vergelijken. Hierdoor kon dit onderzoek vaststellen op welke gebieden verbeteringen in het wiskundeonderwijs gewenst of nodig zijn.

ONDERZOEKSVRAGEN EN ONDERZOEKSOPZET

De wiskundeprestaties van Sloveense leerlingen zijn geanalyseerd vanuit twee invalshoeken of perspectieven. De eindtermen van het nieuwe curriculum vormden de eerste invalshoek. Ook al richt deze studie zich op de prestaties van leerlingen in het oude curriculum, het doel was om wiskundekennis en –vaardigheden te beschrijven vanuit de eindtermen van het nieuwe curriculum om zo een 'baseline' te krijgen voor het nieuwe curriculum.
Omdat in de curriculumhervorming nadrukkelijk werd gesteld dat het Sloveense onderwijs zich met dat van andere landen moet kunnen meten, was internationale vergelijking een tweede invalshoek van waaruit naar het Sloveense wiskundeonderwijs werd gekeken. Het belang van dit perspectief wordt onderstreept door het feit dat Slovenië, samen met een negental andere landen, per 1 mei 2004 lid wordt van de Europese Unie (EU). Dit feit was aanleiding de wiskundeprestaties van Sloveense leerlingen te vergelijken met die van leerlingen uit een aantal andere EU-landen, alsook met enkele andere landen die tegelijk met Slovenië tot de EU zouden toetreden. Teneinde na te gaan of er vanuit internationaal vergelijkend perspectief deelgebieden zijn aan te wijzen waarop het Sloveense wiskundeonderwijs kan worden verbeterd, zijn de prestaties van Sloveense leerlingen vergeleken met die van leerlingen uit een aantal referentielanden: België-Vlaanderen, Nederland, Hongarije en Slowakije.

Deze twee invalshoeken leidden tot de volgende onderzoeksvragen voor deze studie:

1. *Hoe goed presteren Sloveense leerlingen in wiskunde aan het einde van het leerplichtige onderwijs in 1999 in vergelijking met de eindtermen die zijn geformuleerd voor het herziene curriculum, en wat zijn de ontwikkelingen in de wiskundeprestaties tussen 1995 and 1999?*
2. *Hoe goed presteren Sloveense leerlingen in wiskunde aan het einde van het leerplichtige onderwijs in 1999 in vergelijking met een aantal andere Europese landen, en wat zijn hierin de ontwikkelingen tussen 1995 en 1999?*

Om een antwoord op deze onderzoeksvragen te vinden is gebruik gemaakt van de gegevens uit de 'Third International Mathematics and Science Studies' (TIMSS) die in 1995 and 1999 hebben plaatsgevonden onder auspiciën van de International Association for the Evaluation of Educational Achievement (IEA). Toetsen en toetsopgaven uit TIMSS zijn gebruikt om het beoogde ('intended') en gerealiseerde ('attained') wiskundecurriculum van Slovenië te beschrijven, alsook het gerealiseerde curriculum in de referentielanden. Maten voor de gerealiseerde curricula werden gebaseerd op leerlingresultaten op toetsitems. Maten voor het beoogde curriculum werden verkregen via oordelen van een drietal Sloveense wiskundeonderwijsexperts: voor elk toetsitem gaven zij de bijhorende eindterm of standaard van het nieuwe curriculum aan en daarmee dus ook het niveau waarop leerlingen worden geacht te presteren. Dit niveau is geoperationaliseerd in een 'intended percent correct': 75% voor de minimum standaarden, 50% voor de fundamentele standaarden en 25% correct voor de hogere standaarden.

In het analyseren van de TIMSS data zijn benaderingen toegepast, namelijk één uitgaande van prestaties van leerlingen en één gebaseerd op scores op toetsitems (percentage correct per item). In de eerste benadering werden gemiddelde scores van leerlingen op een aantal items (dwz op de toets of een subtoets) geanalyseerd vanuit de hierboven genoemde invalshoeken: de eindtermen van het nieuwe curriculum en de vergelijking met een aantal referentielanden. Als in de op leerlingprestaties gebaseerde benadering de prestatie van Sloveense leerlingen op een aantal items hoger was dan het referentiepunt voor deze items, dan werd dit gezien als een sterk punt van het Sloveense wiskundeonderwijs en, omgekeerd, prestaties lager dan een referentiepunt verwijzen naar een zwak aspect van het onderwijs. Bij het vergelijken van Sloveense prestaties met die van referentielanden werd alleen van een 'betere prestatie' gesproken als de Sloveense resultaten (statistisch) significant beter waren dan die van tenminste twee andere landen. En omgekeerd als de Sloveense resultaten voor een bepaald onderwerp significant lager waren dan die van tenminste twee andere landen, werd dit gezien als een indicatie van zwakte in het Sloveense wiskundeonderwijs.

In de op itemscores gebaseerde benadering, werden items met een percentage correct dat significant hoger is dan aangegeven door het referentiepunt gezien als 'sterk', en omgekeerd items met een percentage correct significant lager dan het referentiepunt werden 'zwakke' items genoemd. In de vergelijking met referentielanden, was er sprake van 'sterke' items voor Slovenië als het percentage correct op die items significant hoger was dan in tenminste twee andere landen, en omgekeerd werden op analoge wijze ook 'zwakke' items gedefinieerd. Conclusies over sterktes en zwaktes in het Sloveense wiskundeonderwijs werden gebaseerd op de aantallen van sterke, resp. zwakke items.

De analyses zijn uitgevoerd op verschillende niveaus: die van de resultaten op de TIMSS wiskundetoets als geheel, op een aantal deeltoetsen en ook op het niveau van items. Deeltoetsen representeerden niet alleen deelgebieden van de wiskunde zoals die in het Sloveense wiskundecurriculum worden onderscheiden, maar ook cognitieve categorieën zoals die in TIMSS zijn gebruikt. De cognitieve categorieën verwijzen naar het type gedrag dat leerlingen moeten vertonen om een opgave op te lossen. In deze studie zijn vier cognitieve categorieën uit TIMSS gebruikt: 'kennis', 'toepassen van routines', 'gebruik van complexe procedures' en 'onderzoeken en probleem oplossen'. Deze categorieën worden binnen TIMSS gezien als een hiërarchie, namelijk van minder complex naar meer complex cognitief gedrag, hoewel in elke categorie opgaven van verschillende moeilijkheidsgraad kunnen worden ontwikkeld.

BELANGRIJKSTE RESULTATEN

Een belangrijk algemeen resultaat van het onderzoek is dat de wiskundeprestaties van Sloveense leerlingen in het 'oude', niet-vernieuwde curriculum sporen met de eindtermen van het herziene curriculum. Echter, dit resultaat geldt alleen voor TIMSS-toets als geheel. Uit meer gedetailleerde analyses blijkt dat er nuanceringen bestaan. In twee deelgebieden van het Sloveense wiskundecurriculum, namelijk 'natuurlijke getallen' en 'data representatie', kunnen de prestaties van de Sloveense leerlingen als sterk worden beschouwd. Echter voor het deelgebied 'betekenis van rationale getallen' geldt dat de prestaties van de Sloveense leerlingen zwakker zijn dan beoogd in het vernieuwde curriculum.

De analyse van de mate van correspondentie van de wiskundeprestaties met de referentiepunten op de cognitieve categorieën (op basis van expert-oordelen) leidde tot de conclusie dat deze prestaties in de drie lagere categorieën correspondeerden met de beoogde resultaten, maar dat voor de hoogste categorie ('onderzoeken en probleem oplossen') de prestaties significant lager zijn dan gewenst volgens het herziene curriculum. Niettemin werden voor deze hoogste categorie op item-niveau een aantal sterke punten gevonden, aangezien de leerlingscores voor 20% van de items in deze categorie significant hoger zijn dan aangegeven in de standaarden.

Voor wat betreft de tweede onderzoeksvraag, bleek dat voor de meeste wiskundeonderwerpen de prestaties van de Sloveense leerlingen significant lager uitvielen dan die van Vlaamse leerlingen; uitzonderingen vormden 'bewerkingen met rationale getallen' en 'algebraïsche uitdrukkingen'. In de vergelijkingen met de andere landen bleek dat verschillen tussen Slovenië en andere landen weliswaar

optraden, maar in een kleiner aantal wiskundeonderwerpen, terwijl de verschillen meestal in het voordeel van de Sloveense leerlingen uitvielen. Een uitzondering vormen de onderwerpen 'betekenis van rationale getallen' en 'meten', die in vergelijking met de referentielanden relatief zwak blijken te zijn.
De prestaties van Sloveense leerlingen voor de laagste cognitieve categorie 'kennis' waren significant lager dan in drie van de vier referentielanden (de uitzondering is Slowakije), wat verwijst naar een aspect waar verbetering mogelijk is in het Sloveense wiskundeonderwijs. Anderzijds kon worden vastgesteld dat de prestaties van Sloveense leerlingen in de andere cognitieve categorieën, waaronder 'onderzoeken en probleem oplossen' correspondeert met die van de referentielanden.
Uit de analyses van de data voor de wiskundedeelgebieden (voor beide onderzoeksvragen) blijkt dat de prestaties van Sloveense leerlingen voor 'bewerkingen met rationale getallen', 'algebraïsche uitdrukkingen'en 'functies en verhoudingen' als bevredigend kunnen worden beschreven, zowel in vergelijking met de nieuwe eindtermen als met referentielanden. Hetzelfde geldt voor de resultaten op de cognitieve categorieën 'toepassen van routines' en 'gebruik van complexe procedures'.
Anderzijds, waar de prestaties van Sloveense leerlingen voor 'meten' corresponderen met de nieuwe eindtermen, waren deze significant lager dan in twee referentielanden, namelijk België-Vlaanderen en Hongarije. Dit resultaat kan worden gebruikt om de eindtermen voor dit deelgebied kritisch te bezien en eventueel aan te scherpen. Een andere belangrijke uitkomst is dat, hoewel Sloveense leerlingen significant lager presteerden op de subtoets 'onderzoeken en probleemoplossen' dan de nieuwe eindtermen of standaarden aangeven, hun prestaties wel op hetzelfde niveau lagen als van leerlingen in de referentielanden.

CONCLUSIES

Het onderzoek dat gerapporteerd is in dit proefschrift is een beschrijvend onderzoek. Het verschaft informatie over sterkte en zwakke punten in de wiskundeprestaties van Sloveense leerlingen vanuit twee perspectieven. Het vergelijken van leerlingprestaties in het 'oude', niet vernieuwde curriculum met de eindtermen van het nieuwe curriculum verschaft informatie over de beginsituatie van het nieuwe curriculum en vormt derhalve een 'baseline' voor dat curriculum. Deze informatie kan dan ook op een later tijdstip worden gebruikt om vast te stellen of de beoogde uitkomsten van de curriculumvernieuwing ook inderdaad zijn gerealiseerd. Dit onderzoek levert een unieke toevoeging aan bestaand onderzoek van wiskundeonderwijs in Slovenië, omdat het inzicht geeft op welke inhoudelijke deelgebieden en cognitieve aspecten verbeteringen wenselijk en mogelijk zijn. De informatie uit het onderzoek kan hen die verantwoordelijk zijn voor de curriculumvernieuwing en -implementatie in de praktijk helpen hun conclusies over het 'oude' curriculum aan te scherpen op basis waarvan aanvullende (betere) didactische benaderingen en materialen voor leerkrachten kunnen worden ontwikkeld. Op deze wijze kan dit onderzoek bijdragen aan de implementatie van het nieuwe curriculum.

De resultaten van dit onderzoek geven Sloveense leerkrachten en curriculumontwikkelaars ook de mogelijkheid tot een beter inzicht in de wiskundeleerprestaties van Sloveense leerlingen, alsook in de haalbaarheid van de eindtermen van het nieuwe curriculum. Hoewel het onderzoek geen evaluatie was van het bereiken van deze eindtermen, draagt het wel bij aan het inzicht in hoe de eindtermen in de praktijk kunnen worden geoperationaliseerd en gebruikt. Het meten van de leerprestaties en het nagaan in welke mate deze corresponderen met de eindtermen, alsook met die in referentielanden, kunnen niet alleen bijdragen aan het verfijnen van de inhouden en formuleringen van de eindtermen, maar ook aan hun operationalisering in termen van beoogde leerprestaties.

Hoewel de invalshoeken die zijn gekozen in dit onderzoek zeker niet de enige zijn vanwaar uit naar leerprestaties van leerlingen kan worden gekeken, verschaffen ze wel een schat aan informatie die leerkrachten en curriculumontwikkelaars kunnen benutten bij hun inspanningen het Sloveense wiskundeonderwijs te verbeteren.