Motivational Beliefs in the TIMSS 2003 Context: Theory, Measurement and Relation to Test Performance

Hanna Eklöf

Department of Educational Measurement
Umeå University
No. 2
Abstract

The main objective of this thesis was to explore issues related to student achievement motivation in the Swedish TIMSS 2003 (Trend in International Mathematics and Science Study) context. The thesis comprises of five empirical papers and a summary. The expectancy-value theory of achievement motivation was used as the general theoretical framework in all empirical papers, and all papers are concerned with construct validation in one form or another. Aspects of student achievement motivation were measured on a task-specific level (motivation to do well on the TIMSS test) and on a domain-specific level (self-concept in and valuing of mathematics and science) and regressed on test performance.

The first paper reports the development and validation of scores from an instrument measuring aspects related to student test-taking motivation. It was shown that a number of items in the instrument could be interpreted as a measure of test-taking motivation, and that the test-taking motivation construct was distinct from other related constructs.

The second paper related the Swedish students’ ratings of mathematics test-taking motivation to mathematics performance in TIMSS 2003. The students in the sample on average reported that they were well motivated to do their best on the TIMSS mathematics test and their ratings of test-taking motivation were positively but rather weakly related to achievement.

In the third and the fourth papers, the internal structure and relation to performance of the mathematics and science self-concept and task value scales used in TIMSS internationally was investigated for the Swedish TIMSS 2003 sample. For mathematics, it was shown that the internationally derived scales were suitable also for the Swedish sample. It was further shown that ratings of self-concept were rather strongly related to mathematics achievement while ratings of mathematics value were basically unrelated to mathematics achievement. For the science subjects, the internal structure of the scales was less simple, and ratings of self-concept and valuing of science were not very strongly related to science achievement.

The study presented in the fifth paper used interviews and an open-ended questionnaire item to further investigate student test-taking motivation and perceptions of the TIMSS test. The results mainly corroborated the results from study II.

In the introductory part of the thesis, the empirical studies are summarized, contextualized, and discussed. The discussion relates obtained results to theoretical assumptions, applied implications, and to issues of validity in the TIMSS context.

Keywords: test-taking motivation; TIMSS 2003; validity; construct validation; measurement; expectancy-value theory; self-concept; task value; factor analysis; eighth-grade students
Acknowledgements

Obviously, I would like to thank the world on a day like this but as the world consists of billions of people, I will restrict my acknowledgements in this section to a few people.

First, I would like to thank the Department of Educational Measurement for admitting me to their recently launched doctoral study programme. It has been a challenge and I have learned a lot! I would also like to thank all members of the staff for creating a very pleasant atmosphere to work in.

As a newly admitted and somewhat confused graduate student, I was seemingly free to choose to do my research within any of the department’s major projects. I quickly decided to align myself with the group working with TIMSS 2003, an international study in which student proficiency in mathematics and science is measured. As neither mathematics nor science are my specialities, I early on chose to focus on something closer to my heart: issues related to student motivation. I appreciate my supervisors and the department’s TIMSS staff for letting me do this, but this is not the only reason I would like to express my gratitude to these people.

My warmest thanks go to my supervisor, Widar Henriksson, and my assistant supervisor, Simon Wolming, for patiently reading and re-reading all my writings and for providing me with invaluable suggestions and advice. I would particularly like to thank Simon for being so supportive and approachable, and Widar for keeping me structured. Without Widar’s pragmatic thinking and strong emphasis on what writing a thesis is really about, his faith in time schedules and backwards counting, my work would not have been finished by now.

I would also like to thank my friends and colleagues in the TIMSS project for generously incorporating me into their group and showing interest in my work: Jan-Olof Lindström, Peter Nyström, Susanne Alger, Björn Sigurdsson and Niklas Eriksson. Special thanks go to Jan-Olof, for helping me realize my own study in connection with TIMSS 2003 in Sweden, and to Björn and Niklas, for readily helping me solve all sorts of minor and major problems. Your help has been invaluable to me and my thesis. Warm thanks also go to Susanne and David Alger as well as Gunnar Persson for proofreading my texts with short notice.

Further, I would like to thank Gunilla, my roommate, for making our study a nice place to work in and for her patience with my sometimes whining attitude. I would also like to mention my other fellow doctoral students as well
as Kent Löfgren, my corridor neighbour, for all support and advice I got from these persons throughout the years. My thanks are also due to Lotta Jarl for her help with all practical issues related to a dissertation.

There are people other than my academic colleagues who have been important to me and to whom I am very grateful. I would like to thank my mother and father who always have been there for me, who have supported me in all kinds of situations and who have taken a great interest in my present work. Special thanks go to my mother as she has invested several hours in all thinkable tasks related to my work with this thesis. My thanks also go to my brother for helping me with formatting issues and for always making me laugh. Not least, I want to direct a warm thank you to Andreas for all your support and help, and for putting up with my thesis writing and horse riding.

Finally, I would like to thank the schools that allowed me to administer questionnaires and perform interviews, although their spring semester was about to end and they suffered from the largest school strike in many years. Especially, I want to express my deep gratitude to all the eighth-graders who took part in my study, patiently completed the questionnaires, and readily answered my interview questions. These students and their responses are the prerequisite for my writings in this thesis and hence, without them, I would have nothing to present in the following.
Motivational Beliefs in the TIMSS 2003 Context: Theory, Measurement and Relation to Test Performance

This thesis is based on the following articles:


All referencing to the articles will follow the enumeration used above.
7. Summary of the Papers..................................................... 38
   PAPER I ............................................................................... 39
   PAPER II .............................................................................. 39
   PAPER III .............................................................................. 40
   PAPER IV .............................................................................. 41
   PAPER V .............................................................................. 42
8. Discussion ......................................................................... 43
   Main Findings From the Empirical Studies ......................... 43
   Beyond the Main Empirical Findings .................................. 44
   Theoretical Implications of Obtained Results ...................... 45
   Applied Implications of Obtained Results ......................... 47
   Validity and Validation in the Larger TIMSS Context .......... 48
   Limitations and Generalizations ....................................... 49
   Suggestions for Future Research ..................................... 51
References ............................................................................. 52
1. Introduction

Student motivation is a core issue in educational settings as achievement motivation is assumed to interact with achievement behavior in important ways (Pintrich & Schunk, 2002; Wigfield & Eccles, 2002). It is often claimed that a well motivated student performs better in achievement situations, has higher educational aspirations, expend more effort in learning new tasks, and persists longer at difficult tasks compared to a poorly motivated student (Pintrich & Schunk, 2002). In low-stakes testing situations a common assumption is that some students may lack situation-specific motivation to do their best on the test and that the results therefore can be an underestimation of student knowledge (Baumert & Demmrich, 2001).

But what is this invisible construct “motivation”? How could it be conceptualized and operationalized, measured and interpreted? How are motivational beliefs handled in large-scale, international studies like TIMSS 2003? Are students in fact not motivated to do their best on low-stakes tests and how do domain-specific and situation-specific aspects of motivation associate with test performance? And what difference does it make? Questions like these were the impetus to the research presented in this thesis.

On a general level, this thesis is about the measurement of latent constructs like motivational beliefs and about the quality of this measurement. More specifically, the empirical papers investigate the structure and relation to performance of situation-specific achievement motivation (test-taking motivation) and domain-specific achievement motivation, respectively.

All empirical papers attached to this thesis are concerned with construct validation in one form or another. The terms reliability and validity are often used in discussions of measurement quality. The view taken in this thesis is that validity is the overarching quality indicator (see Messick, 1989; Nyström, 2004; Wainer & Braun, 1988; Wikström, 2005; Wolming, 2001; Wolming, 1998). Reliability obviously also is a desirable feature in most measurement settings, but it is subordinated to validity, and perhaps even just a part of an all-inclusive validity concept (Nyström, 2004). Validity has been defined by Samuel Messick as “an integrative evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment” (Messick, 1989, p. 13) and this perspective on validity is adopted in the present thesis.
Having defined validity, a few other definitions might be in place. That the thesis is said to be about the measurement of latent constructs perhaps needs some clarification of terms. First, the term construct or latent construct refers to a theoretical, intangible quality or trait in which individuals differ (Messick, 1995) and is an abstract variable that is derived from theory or observation (Benson, 1998). Second, the term measurement is here broadly understood as the process of systematically assessing a trait, state, or quality, a process that involves both theoretical and empirical considerations. The terms measurement and assessment are similar in content and are often used interchangeably. Although I conceive of the term “measurement” as a more systematic and structured form of assessment, I too will use these terms interchangeably in the following.

Ideally, a measurement is not just a single administration of a test, or a single analysis of a test result, but a process, in which each step should be well thought-through and evaluated. In this thesis, the constructs measured are related to student achievement motivation. Motivation in turn is a psychological concept with no single, universally-accepted definition. Pintrich and Schunk (2002) define motivation as “the process whereby goal-directed activity is instigated and sustained” (p. 5), a general definition in line with most contemporary perspectives on motivation. This definition implies that motivation is a process rather than a product, that motivation involves goals, and that motivation is related to activity, an activity that is instigated and sustained. Further, the motivational processes are not possible to observe directly, but are inferred from verbalizations or overt behavior. Thus, might not be possible to directly measure “the process whereby goal-directed activity is instigated and sustained”, but rather aspects related to this process. This makes the role of theory all the more important to make the valid measurement of motivational constructs possible (Messick, 1995).

In my thesis, I have used the expectancy-value theory of achievement motivation as the general theoretical framework. In my empirical studies, I have used ratings of self-concept and valuing of the school subjects as indicators of expectancies and values on a domain-specific level, and ratings of self-efficacy and valuing of the TIMSS test as indicators of expectancies and values on a situation-specific level.

The empirical contribution of this thesis emanates from a large comparative study, TIMSS 2003 (Trends in International Mathematics and Science Study). Data used in the empirical studies is either drawn from the Swedish database for TIMSS 2003 (mainly papers III and IV) or collected in connection with the
study in Sweden (papers I, II and V). Swedish eighth-grade students were used as the study sample in all studies.

Disposition of the Thesis

The thesis consists of a summary and five empirical papers. After this introductory chapter, the thesis' contextual as well as theoretical framework will be addressed. Chapter 2 presents the contextual framework and also the empirical base for the five papers, TIMSS 2003. In chapter 3, the motivational framework guiding the research, the expectancy-value theory of achievement motivation, is introduced. Chapter 4 addresses the concept of test-taking motivation, together with previous research in this area. Having presented the two cornerstones for validation; the theoretical and the empirical base, the overarching measurement theory, validity theory, is presented in chapter 5. Chapter 6 contains a summary of the methodological choices made in the empirical studies. With these contextual, theoretical and methodological considerations as a background, the five papers are summarized in chapter 7. The last chapter (chapter 8) in this part of the thesis discusses the main findings of the thesis, and theoretical as well as applied implications are elaborated upon. Then, the papers follow in numerical order.

2. TIMSS 2003

The empirical data used in the five articles (papers I–V) was collected in connection with TIMSS 2003 (*Trends in International Mathematics and Science Study*), a multinational comparative study where student achievement in mathematics and science, as well as their contexts for learning and their motivation for learning in the different school subjects, is measured. TIMSS is organized by the *International Association for the Evaluation of Educational Achievement* (IEA), an international cooperative of national research institutions and government agencies that has conducted international student literacy studies since the 1960s. TIMSS 2003 was the third in a cycle of assessments, conducted every four years (see Mullis, Martin, Gonzales, & Chrostowski, 2004). Fifty countries participated in TIMSS 2003. Sweden was one of those and the empirical studies attached to this thesis use Swedish data only.
Swedish Achievement Results in TIMSS 2003

Sweden participated in TIMSS 2003 with about 4,200 eighth-grade students from 160 different schools. Sweden also participated with sixth-, seventh-, and eighth-grade students in TIMSS 1995, which enables trend studies (see Skolverket, 2004). In short, the Swedish results in TIMSS 2003 were rather discouraging, especially in mathematics. Compared to a group of 20 relevant countries (mainly OECD and EU members, see Skolverket, 2004), the Swedish students on average scored significantly below the average mathematics score for these 20 countries. Also, a comparison with the Swedish mathematics results in TIMSS 1995 showed a marked decrease in performance over time (see Skolverket, 2004 for a more detailed discussion of the achievement results). For the science part of the study, the Swedish results were less alarming. Compared to the Swedish science result in TIMSS 1995, the decrease in average science achievement was not as pronounced as in mathematics, and the Swedish students did rather well compared to the group of 20 comparable nations (Skolverket, 2004).

This thesis is not primarily concerned with achievement results, trend studies, or issues of comparability as the focus is on motivational beliefs in a Swedish TIMSS 2003 context, but a brief discussion on the characteristics and assumptions of international comparative studies might nevertheless add to the contextual understanding of the research presented in the thesis.

Comparative Studies – Characteristics and Assumptions

Large-scale, international comparative assessments of student proficiency in various school subjects have been growing in popularity and impact in recent decades. The public and political interest in the results from these studies is vast, media coverage is often extensive, especially in instances of surprisingly positive or negative results, and their impact on national educational systems have sometimes been considerable (Bechger, van Schooten, De Glopper, & Hox, 1999; Robitaille, Beaton, & Plomp, 2000; Sjöberg, 2005).

Political imperatives have become strong motivators for international comparative studies as a successful educational system is believed to be important for a nation’s economic well-being and its competitive strength on a global market (Robitaille & Robeck, 1996; Sjöberg, 2005). Results from these studies inform policy makers all over the world and studies like TIMSS can therefore be regarded as policy research (Messick, 1987; Sjöberg, 2005). Thus,
on a political level, studies like TIMSS are rather high-stakes. On the other hand, it is often argued that for the participating students, TIMSS is a low-stakes test in the sense that it does not have any consequences for the individual student. It is further argued that the low stakes of the test may affect students’ motivation to perform well. This argument inspired the research presented in three of the papers attached to this thesis (papers I, II and V).

TIMSS represents a very ambitious collaborative project where much effort is invested in the study design, sampling procedure, standardization of instruments and measurement procedures, complex scaling and analysis of achievement data, all to ensure comparability (see Martin, Mullis, & Chrostowski, 2004). On the other hand, TIMSS also represents a rather traditional measurement practice, where all students complete standardized pencil-and-paper tests translated into a variety of languages. The above implies two things that are related to theory of science. First, the mathematics and science literacy assessed must be common to everyone in the populations tested. One basic assumption then is that there actually is something we objectively can call knowledge and that this knowledge is structured in approximately the same way across countries and cultures. Second, it is assumed that this common literacy can be retrieved by using a common procedure across cultures and countries, and that outcomes can be validly compared. Related to this assumption, it is also assumed that all students in all countries and cultures will react similarly to the test battery they are designated to complete. These assumptions must be made. Otherwise, comparisons would have no meaning and international comparative studies would have no justification.

Results from studies like TIMSS are often presented in form of league tables where participating countries are ranked according to their mean level of achievement, and descriptive tables where background variables are summarized and reported. The results are mostly taken at face value and not problematized. Further, due to their magnitude and scope, large-scale studies like TIMSS are often based on very general theoretical frameworks and mostly lack substantive theory that can be used in the interpretation of data (Bechger et al, 1999).

The above validity issues that are related to large-scale, comparative studies are issues that the TIMSS administration is well aware of and much care is taken to make the tests “equally unfair” for each participating country and to make the results as reliable and comparable as possible. Studies of this magnitude obviously must make trade-offs between what is desirable and what is feasible. Nevertheless, given the impact studies like TIMSS can have on
educational systems and the public view of these systems, each extended and careful examination of the instruments used, the inferences made, and the possible consequences of inferences, could be a valuable contribution to the discussion of the validity of obtained results (Messick, 1987). The TIMSS administration strongly encourages researchers to use TIMSS data for secondary and extended analyses. However, many issues still remain unexplored.

Measures of Achievement Motivation in TIMSS 2003

TIMSS administers questionnaires to different actors in the school system, including school leaders, teachers, and students. In the student background questionnaire there are items asking for demographic characteristics, home and school environment, learning climate, time spent on homework, and so forth (see Skolverket, 2004, for a presentation of variables and results). There are also items asking for students’ perceptions of their ability as well as items asking how much the students value the school subjects assessed in TIMSS. In TIMSS internationally, two scales, in this thesis called self-concept and valuing of the school subject, are derived from these items through principal components analysis (PCA). These variables, domain-specific self-concept and task value, are the focus of paper III and IV in this thesis.

In the TIMSS student background questionnaire there are no items asking for how the students perceive the TIMSS test, or how motivated they are to do their best on the TIMSS test. However, this issue was explored in connection with the TIMSS study in Sweden, and is the focus of paper I, II and V in this thesis.

The importance of achievement motivation is acknowledged in the TIMSS context (Robitaille & Garden, 1996). Still, the number of items related to students’ motivation to learn the school subjects have decreased from the early IEA studies to TIMSS 1995 and further to TIMSS 2003. According to Robitaille and Garden (1996), this is not due to a lack of interest in these constructs and their relation to achievement and other variables, but rather due to scarce instrument administration time. Priorities have to be set, and these constructs are obviously not prioritized in TIMSS. One reason might be that between-country comparisons of motivational beliefs have been problematic. For example, while within-country studies have shown that student self-concept is positively related to achievement, between-country comparisons have revealed this relation as non-significant, or even negative (Shen, 2002; Shen & Pedulla, 2000; Artelt, 2005). According to Robitaille and Garden (1996), motivational
items are especially vulnerable to translations, and even if translations are correct, subject names and labels might mean different things to different people. Also, national response patterns vary. There seem to be cultural differences in how students treat the response scales, where students in some countries are reluctant to use the extreme ends of the scales. Also, social comparisons and frames of reference might be involved in student ratings of their self-concept and valuing of the school subjects (see Skaalvik & Rankin, 1995).

With this in mind, compared to other relevant countries, the Swedish students had a rather positive view of their own ability in mathematics and the science subjects, while they at the same time did not put much value on these subjects (Skolverket, 2004). Compared to 1995, it seems as if the Swedish students have become more confident in their abilities to perform well, while they value the school subjects less.

As motivational beliefs are assumed to be important for present and future achievement behavior, the scale structure of the motivational items used in TIMSS 2003 (asking for self-concept and valuing of the school subjects) and their relationship to theory and achievement seemed worth investigating, especially as the theoretical rationale for including these particular items is not made explicit in the TIMSS reports. Papers III and IV in this thesis explore these issues.

3. Achievement Motivation

Latent Motivational Constructs and the Role of Theory

Motivation has been defined as “the process whereby goal-directed activity is instigated and sustained” (Pintrich & Schunk, 2002, p. 5). A similar definition is presented by Phye (1997), who defines motivation as an internal state that arouses, directs, and maintains behavior. It is not possible to observe this internal state process, but it is instead inferred from verbalizations or overt behavior. Because of this invisibility, theory is necessary in the measurement of motivational beliefs. Without a theoretical framework, research problems become diffuse, operationalizations become problematic, and the validity of interpretation of results is very difficult, if not impossible, to support (or refute, for that matter) as validation is theory-driven as well as data-driven (Messick, 1995).

Due to the latency of psychological constructs, the construct motivation can be conceptualized in different ways, with different theories focusing on different psychological processes. I have chosen the Eccles and Wigfield expectancy-value
theory as the theoretical framework to guide me in how best to conceive of
achievement motivation. This theory seemed highly relevant in relation to the
content of the motivational items included in the TIMSS student background
questionnaire and the theory has been extensively validated (Eccles & Wigfield,
2002; Pintrich & Schunk, 2002; Wigfield & Eccles, 2002), but there are, of
course, many other possible views. It should also be noted that the constructs
“self-concept”, “self-efficacy” or “task value” do not equal “motivation”, but
they are constructs that have been hypothesized to be, and have been
empirically shown to be, related to motivated achievement behavior (Bong &
Skaalvik, 2003; Wigfield & Eccles, 2000).

A Note on Terminology

In this thesis, student self-concept and valuing of the school subjects are
perceived as aspects related to student motivation. I have accordingly referred to
these scales as “motivational” (with the exception of paper I where the TIMSS
terminology was adopted). However, the same scales are in TIMSS collectively
called “attitudes”. I have chosen to use the term “motivation” to define the
construct rather than “attitude” as the items in the scales ask for information
that is usually interpreted within a motivational theoretical framework.
Attitudes are here interpreted as more affective statements directed towards
some objective (like/dislike, approve/disapprove), while motivational ratings are
more cognitive-based evaluations of, in the case of the TIMSS items, ability in a
school subject and the value attached to learning the school subjects.

The Expectancy-Value Theory of Achievement Motivation

With the advent of modern psychology in the late 19th century and a growing
interest in individual differences (Anastasi & Urbina, 1997) also came an
interest in explaining the possible factors involved in human motivation. Since
then, different motivational theories with different explanatory frameworks
have been proposed. They have stressed inner needs, drives, instincts and
associations between stimuli, response and reinforcement as the cause of
motivated behavior (Pintrich & Schunk, 2002). The modern achievement
motivation paradigm is dominated by cognitive theories, which claim that
individuals’ thoughts, beliefs, and emotions together influence motivation (see
Pintrich & Schunk, 2002; Wigfield & Eccles, 2002). Most modern cognitive
theories on motivation also incorporate a sociocultural perspective (social cognitive theories), where it is acknowledged that the surrounding social context interacts with the individual and influences his or her motivational beliefs. Contemporary theories of motivation are often overlapping in content (Bong, 1996; Wigfield & Eccles, 2002). Some are comprehensive and include many motivational aspects (e.g. expectancy-value theory (Eccles & Wigfield, 2002)), while others are more specific (e.g. self-efficacy theory (Bandura, 1997)).

The general theoretical framework that has guided the research in the empirical studies attached to this thesis is the social cognitive expectancy-value model of achievement motivation. The expectancy-value perspective on motivation originates back in the first half of the twentieth century (Atkinson, 1957; Weiner, 1992). The most widely used expectancy-value model currently, and the model adopted in this thesis, comes from the work of Eccles and Wigfield and their colleagues (Eccles & Wigfield, 2002; Wigfield, 1994; Wigfield & Eccles, 2002). A contemporary version of the model is presented in Figure 1. Numerous empirical studies have been performed that support the assumptions of this theoretical framework. The theory is comprehensive in order to mirror as many as possible of the processes underlying motivated behavior and includes many contextual and psychological aspects that have been shown to interact and influence achievement choices and achievement behavior. This comprehensiveness makes it difficult to apply the entire model in a single study but as Bong (1996) has noted, a comprehensive model allows the researcher to focus on a smaller part of the model while still not losing sight of the big picture.
Although comprehensive, the model has two core components; one expectancy component that corresponds to the question “Can I do this task?”, and one value component that corresponds to the question “Do I want to do this task and why?”. The expectancy component in the model thus refers to the individual’s beliefs and judgments about his or her capabilities to do a task and succeed at it. The value component in the model refers to the various reasons individuals have for engaging in a task or not.

*Can I do this task?* The expectancy component is defined in terms of student self-concept, future short-term and long-term goals, and expectations for success. The component is to be viewed as rather future oriented and thus, expectancies for future success would be the most important aspect to measure. However, in construct validation studies using empirical data, Eccles and Wigfield have consistently found that different expectancy constructs like self-concept, self-efficacy and expectancy for future success are not differentiated
into separated factors although they are theoretically distinct. Individuals do not seem to differentiate between self-concept and performance expectations and therefore, based on the current understanding of the model, these two aspects are empirically interchangeable and can be treated as one construct.

In my empirical studies, I have used student ratings of self-concept in mathematics and the science subjects (mainly studies III and IV) and ratings of task-specific self-efficacy beliefs (mainly study I) as indicators of the expectancy component in the Wigfield and Eccles model. Issues related to self-concept and self-efficacy, respectively, are established research areas in their own right (see Bandura, 1997; Marsh & Craven, 1997; Marsh & Shavelson, 1985; Skaalvik & Skaalvik, 2004; Bong & Skaalvik, 2003), but the constructs can also be incorporated as part of a more comprehensive theoretical model, like the expectancy-value model.

Do I want to do this task and why? In the Eccles and Wigfield model, task value is defined in terms of four components. The different value components are attainment value (or importance), intrinsic value (or interest), utility value (or usefulness), and cost. Attainment value refers to the perceived importance of doing well on a task. Intrinsic value can be defined as the enjoyment the individual experiences when doing a task, or his or her subjective interest in the content of a task. Utility value refers to the perceived usefulness of the task in terms of the individual’s future goals. The fourth value component, cost, includes the perceived amount of effort required for the task. Confirmatory factor analyses have indicated that attainment value, intrinsic value and utility value are interrelated but empirically distinct from one another and from the expectancy component (Wigfield & Eccles, 2000). The cost component in the model is so far less well researched, and it’s relation to the other aspects of value is not entirely clear.

Findings from the expectancy-value research paradigm have shown that students’ expectancy beliefs, including self-concept, goals, and expectancy for success are strong predictors of actual achievement in terms of performance on standardized tests and grades in school subjects like mathematics and English, even stronger predictors than are previous grades (Eccles, Wigfield, Flanagan, Miller, Reuman, & Yee, 1989; Wigfield & Eccles, 2002). Values have also been shown to correlate positively with actual achievement, but when both expectancies and values are used to predict achievement, expectancy beliefs are significant predictors, and values are not significant predictors. On the other hand, in terms of intentions to take future courses and actual enrollment in those
courses, value beliefs are better predictors than are expectancy beliefs (see Meece, Wigfield, & Eccles, 1990). Findings have been rather consistent, although it can be noted that the value component in the model is less well researched than the expectancy component in the model (Eccles & Wigfield, 2000).

The expectancy-value theory of achievement motivation has been applied in a large number of studies investigating general and domain-specific achievement motivation. However, in this thesis the theory was also applied on a situation-specific level.

4. Test-Taking Motivation

Achievement motivation can be conceptualized and measured on different levels of generality. General measures of motivation are often too broad to contribute to the knowledge about the structure of motivational constructs and the association between achievement motivation and achievement behavior (Bong & Skaalvik, 2003). The most common type of motivational measure is domain-specific and measures achievement motivation for a particular domain (e.g., mathematics, science). Papers III and IV in this thesis are concerned with domain-specific measures of achievement motivation. However, achievement motivation can also be conceptualized and measured on a situation-specific level, i.e., motivation to perform well in a given situation, or on a given test. Papers I, II and V in this thesis are concerned with task specific motivation, or as I have called it, test-taking motivation.

Each year, an untold number of educational and psychological tests are administered to individuals around the world. A positive motivational disposition towards the test is often assumed to be a necessary though not sufficient condition for a good test performance (Cronbach, 1988; Zedner, 1993; Wainer, 1993; Robitaille & Gardner, 1996) and Messick (1988) noted that a poor test performance could be interpreted not only in terms of test content and student ability, but also in terms of lack of motivation. If different groups of students would differ systematically in level of motivation, and if less motivated students are disadvantaged in that they score below their actual proficiency level, test-taking motivation would be a possible source of bias (Zedner, 1993; Wainer, 1993; Mislevy, 1995; O’Leary, 2002; Baumert & Demmrich, 2001; O’Neil, Sugrue, Abedi, Baker, & Golan, 1997; Robitaille & Gardner, 1996) and hence a threat to the validity of score interpretation and use (Messick, 1995).
However, few scientific inquiries have been able to empirically show the structure and relation to performance of test-taking motivation. Knowledge of how individuals perceive the tests they are designated to complete, and their motivation to do their best on these tests, is scarce (Baumert & Demmrich, 2001; Nevo & Jäger, 1993), although obtained scores from a test are a function not only of the items in the test, but also of the persons responding to the test as well as the context of the measurement (Messick, 1995). Despite the scarce knowledge about test-takers’ perceptions, lack of motivation has sometimes been put forward as an explanation for results that are not as good as expected.

Test-Taking Motivation and Low-Stakes Tests

It has been hypothesized that one major reason why students would not be motivated to do their best on tests like TIMSS are the low stakes of the test for the participating students. Tests that have no personal consequences, i.e., low-stakes tests, are often assumed to cause a decrease in motivation and performance (Wolf & Smith, 1995; Wolf, Smith, & Birnbaum, 1995; Wise & DeMars, 2003). TIMSS is, in several aspects, a low-stakes test and the issue of test-taking motivation is therefore highly relevant in the TIMSS context. First, the result on the TIMSS test has no impact on student grades in mathematics or science, which otherwise is a common feature of educational achievement tests in many countries. Second, the results on TIMSS are mainly summarized at a national level and no individual results are given to the students or the schools. Thus, neither the students, their teachers, parents, nor peers will ever know the result of an individual student.

On the other hand, one may argue that the fact that the students represent their country in a world-wide comparative study is motivating for the students. One may also argue that the low stakes of the test make the students less anxious, and that they therefore achieve as well as they would on an ordinary test, although they are not maximally motivated.

Previous Research on Test-Taking Motivation

A vast amount of research has investigated various aspects of domain-specific achievement motivation. The research on situation-specific motivation or test-taking motivation is anything but vast. Studies are scattered in time and place, theoretically and methodologically. However, the expectancy-value theory of
achievement motivation has been applied to a number of studies investigating test-taking motivation (Wolf, Smith, & Birnbaum, 1995; Wolf & Smith, 1995; Baumert & Demmrich, 2001), and was the theoretical framework used in the investigation of test-taking motivation in this thesis as well.

The results from earlier studies actually focusing on test-taking motivation have been somewhat inconclusive and in many cases, the link between reported level of motivation and actual achievement has been weak. Studies have found, contrary to the low-stakes hypothesis, that the students are quite motivated even when the test is low-stakes for the students (The Center for Educational Testing and Evaluation, 2001), that raising the stakes does not always contribute to a corresponding rise in motivation and achievement (Baumert & Demmrich, 2001; O’Neil, Abedi, Miyoshi, & Mastergeorge, 2005), and that reported level of test-taking motivation is weakly associated with subsequent performance (O’Neil et. al., 2005; Zeidner, 1993). On the other hand, other studies have found that the stakes of the test indeed has an impact on motivation and performance (Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Wolf & Smith, 1995; Wolf, Smith, & Birnbaum, 1995).

In summary, it is not clear from previous empirical studies whether the validity of low-stakes tests like TIMSS is threatened by a lack of motivation among the participants because a) it is not clear if the participating students are lacking motivation at all and b) it is not clear whether rated level of test-taking motivation interacts with test performance. Studies I, II and V in this thesis explore these issues.

5. Validity Theory

Validity is a central feature in the field of measurement in the behavioral and social sciences. It has been for many years but the content and coverage of the concept has changed in past decades. Below a traditional and a modern conception of validity are summarized. These conceptions are by no means mutually exclusive, but differ in focus and scope. The general attitude towards validity and validation held in the present thesis is influenced by the writings of validity theorists like Lee Cronbach (1971; 1988; Cronbach & Meehl, 1955) and Samuel Messick (Messick, 1988; Messick, 1989; Messick 1995), and the validation effort in the empirical papers attached to this thesis is illustrated in the context of Benson’s (1998) strong program for construct validation.
A Traditional Conception of Validity

Validity as a concept emerged in the beginning of the twentieth century. To begin with it was a rather atheoretical and narrow concept, used to describe the representativeness of items chosen for a test or the correlation between a test and some measure outside the test. Validity was a property of tests and obtained validity coefficients were generalized to hold across samples and contexts. Implied in this use of the validity concept is a rather operationalist view that validity can be defined as the correlation of observed scores on a test with true scores on a criterion (Angoff, 1988). Traditionally, evidence of validity has been grouped into three distinct categories: content validity, criterion-related validity, and construct validity.

Content validity is about the relevance and representativeness of contents included in a measurement instrument. Ideally, items chosen for a test should be a representative sample from the universe of all possible items referring to the domain of interest. The typical method for evaluating content validity is expert judgment.

Criterion-related validity has been defined as the association between test scores and some criterion or criteria of interest external to the test. The purpose of the test is often predictive and the method used for validation is often correlation or regression.

Construct validity as a concept was initially introduced as an alternative to the other types of validity in cases where neither content validity nor criterion-related validity could be applied and/or evaluated. Construct validity as originally conceived refers to the extent to which the contents of a measurement instrument are able to measure a theoretical construct.

A Modern Conception of Validity

Samuel Messick (1988, 1989, 1995) is one of the most prominent modern validity theorists and his model of construct validity as an all-inclusive concept has been very influential on the discourse about validity. For Messick, a unified construct validity framework was necessary not only from a scientific point of view but also for the applied use of test scores.

According to modern conceptions of validity, validity is about the appropriateness, meaningfulness, and usefulness of score based inferences (APA, AERA, & NCME, 1999). Simply put, validity is about what a test score means (Gregory, 2004) and validation is the process by which test scores take on meaning (Benson, 1998). Within the modern validity theory framework, it is thus acknowledged
that it is the interpretation and use of test scores, not the test itself that is the proper subject of validation (Messick, 1989). This does not mean that the quality of the measurement instrument can be overlooked in the validation process. It does imply, however, that a sound measurement instrument is a necessary but not sufficient condition for the valid interpretation and/or use of test scores. Further, in the modern validity framework it is recognized that evaluations of validity are dependent on context, culture, scientific paradigm, prevailing values, and so forth. Validity is further seen as a matter of degree, validity evidence as always incomplete and validation as a continuing process (Benson, 1998).

Construct validity, the last validity “type” to be introduced (see Cronbach & Meehl, 1955) has taken over as the overarching aspect of validity and modern validity theory is basically a theory on construct validity that incorporates all other strategies (e.g., content-related, criterion-related, face-validity related) traditionally used for validation (Messick, 1995). According to Messick, there can be no validity without construct-referenced measurement, as no score interpretation is possible without construct-referencing (Angoff, 1988; Messick, 1988), and most contemporary theorists and researchers agree that there is a strong interdependence between theory and practice in the process of validation (Moss, 1995). Theory is particularly important when psychological constructs are at the focus of the measurement as they themselves are theoretical entities. Also, if content-related and criterion-related evidence of score validity are only part of the construct validity framework, it follows that theory is necessary in all efforts to validate inferences made from test scores, be they content-related, criterion-related, or construct-related.

Two of the major threats to validity are construct underrepresentation and construct-irrelevant variance. Construct underrepresentation is present when the empirical domain is defined too narrowly, and thereby fails to adequately represent the theoretical domain of the construct (Benson, 1998). More simply put; the measurement captures only part of the construct one is interested in measuring. Construct-irrelevant variance is present when the empirical domain contains reliable variance that is unrelated to the construct of interest. That is, one unintentionally measures things that are unrelated to the construct of interest. Both these sources of error can distort test interpretation and use.

Messick (1989, 1995) distinguished six aspects of construct validity as fundamental for all educational and psychological measurement. These are a content aspect, a substantive aspect, a structural aspect, a generalizability aspect, an external aspect and a consequential aspect.
I. The content aspect of construct validity refers to evidence of content relevance, representativeness, and technical quality. Thus, this aspect largely corresponds to content validity in the traditional conception of validity.

II. The substantive aspect of construct validity is concerned with specification of the theoretical domain of the construct and with operational definitions of the construct in terms of observed variables (Benson, 1998).

III. The structural aspect of construct validity involves relating items to the construct of interest by determining the extent to which the observed variables relate to one another and to the construct. This aspect involves traditional methods for evaluating internal consistency reliability and construct validity.

IV. The generalizability aspect refers to the extent to which score properties and interpretations generalize across populations, groups, settings, and tasks.

V. The external aspect includes convergent and discriminant evidence and evidence of criteria relevance and applied utility, and can be linked to traditional methods for investigating construct validity and criterion-related validity.

VI. The consequential aspect includes appraisal of the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use.

A Comment on the Modern Conception of Validity

According to Messick, the above six aspects are all part of a construct validation, and none of the aspects is very useful in isolation. The sixth aspect, the consequential aspect, has caused some controversy, and not everyone agrees that an appraisal of the actual as well as the potential consequences of test interpretation and test use is suitable for inclusion as part of the construct validity framework (see Kane, 2004; Popham, 1997). As I see it, Messick’s emphasis on the consequential basis of test interpretation and test use as part of validity is a sound reaction to the thoughtless use and widespread misuse of tests and test results throughout history. Messick’s two dimensions of validity: the evidential dimension and the consequential dimension (see Messick, 1988; 1989; 1995), could further be seen as an effort to merge two different research traditions, one
more quantitative, psychometric (the evidential basis), and one more qualitative, interpretative (the consequential basis). It could also be seen as an effort to merge two different practices: one scientific, theoretical, and one applied, socio-political. Messick’s conception of validity makes it clear that validity is not only about scientific evidence, but also about arguing for the soundness of the interpretation of this evidence. It also makes it clear that the validation process has not come to an end when the measurement outcomes have been interpreted, which was the case in traditional conceptions of validity.

In general, I believe that Messick’s model on validity is beneficial to those using tests as well as those affected by the consequences of test use, as it demands more reflection and a more integrative thinking on the part of the test developer and test user than traditional conceptions of validity did. At least I believe that this was his intention. It should be noted that Messick’s model of validity as a multidimensional concept has sometimes been accused of being difficult to understand and, above all, difficult to apply (Kane, 2004). Messick’s validity theory is a general theory, and which specific questions are asked in an actual validation effort is dependent on the purpose of the measurement. One single study cannot usually aspire to a thorough validation. In fact, this would be contrary to the modern conception of validation as always incomplete and as an ongoing process. The empirical papers attached to this thesis can hardly aspire to a complete validation effort. Rather, they are more concerned with construct validity as originally conceived as they are explorations into the structure and dynamics of psychological constructs. Nevertheless, Messick’s holistic view on validity has guided the research, from the formulation of the research problems to the interpretation of results. It is hence acknowledged that validity judgments are always value judgments, and that the consequences of a measurement are closely tied to the appropriateness, meaningfulness, and usefulness of score based inferences. In the context of my empirical studies, I found Messick’s six validity aspects a suitable framework for discussing aspects related to construct validity and construct validation. More specifically, I found Jeri Benson’s three-stage process for construct validation illustrative of my own validation effort.
Benson’s Strong Program for Construct Validation Applied to the Test-Taking Motivation Construct.

Jeri Benson, drawing on the work of Loevinger (1957), Cronbach (1971), Nunnally (1978), and Messick (1989), has presented a strong program for construct validation in which theory and the interplay between theory and empirical work plays a significant role (Benson, 1998). In accordance with most modern views on validity, her program conceives of construct validation as an ongoing and iterative process. Benson highlights three components as crucial to the validation of psychological constructs. These are a substantive component (which includes components I, II and to some extent component IV in the above descriptions of Messick’s six aspects), a structural component (which basically corresponds to component III above), and an external component (components IV and V above).

The substantive stage of construct validation is concerned with how the construct is defined, theoretically and empirically. According to Benson, all constructs are represented by two domains, one theoretical domain, which evolves from scientific theory, previous research, and the researcher’s own values and observations, and one empirical domain, which operationalizes the construct and contains all possible observed variables and the ways in which these variables can be measured. Depending on theoretical perspective, prevailing values and scientific paradigm, operationalizations will look different. The empirical domain is a reflection of the theoretical domain and it follows that the empirical domain will be easier to operationalize when the theoretical domain is well understood. As concerns the studies of test-taking motivation attached to this thesis, the theoretical domain for the test-taking motivation construct is not yet well understood or well articulated, and it follows that the empirical domain was rather tentatively formulated, drawing on general achievement motivation theory, the few previous studies exploring the construct, and the researcher’s own hypotheses (see Chapter 6.). As concerns the studies of self-concept and valuing of the school subjects attached to this thesis, I took already operationalized variables and tried to tie them to a theoretical domain post hoc, a less than optimal practice from a validity perspective, but necessary as no theoretical domain was specified for these variables.

Through accumulation of empirical studies, the theoretical domain and its reflection, the empirical domain, will be sharpened. Thus, over time, these domains will enable arguments about the generalizability of test score meaning.
**The structural stage** of construct validation contains “internal domain” (Benson, 1998, p. 13) studies, whose purpose is to investigate the internal structure of the observed variables, and how they covary with the proposed structure of the theoretical domain. In this stage, methods traditionally used in construct validation such as exploratory and confirmatory factor analysis, multitrait-multimethod procedures, item response theory, and/or studies of differential item functioning (DIF) are applied. It should be noted that positive results obtained at this stage of the validation process are a necessary though not sufficient condition for construct validity. Even if obtained results are in line with the theoretical assumptions, this does not imply that the interpretation of the test scores is valid. For example, in paper I, it was shown that a number of items assumed to measure aspects related to test-taking motivation associated with one common factor, and this factor was accordingly labeled “Test-taking motivation”. However, labeling this variable test-taking motivation might not be a valid interpretation of score meaning, but merely a reflection of the researcher’s values and his or her wish to measure a construct called test-taking motivation. Another researcher might have labeled this variable otherwise. Using this variable as a measure of test-taking motivation could then be an invalid use of scores that could in turn have unintended consequences for those affected by score use. Thus, studies of internal structure and variable names cannot be taken as indicators that the variables actually reflect the construct one is interested in. To guide the interpretation of what obtained scores actually mean and how they could be used, they have to be compared with something. The most important stage of construct validation, and the stage where scores begin to take on meaning, is therefore the external stage.

**The external stage** of construct validation relates the construct of interest to other constructs and characteristics. Assumed group differences are investigated as are relations with criteria of interest, and evidence of convergent and discriminant validity, which are fundamental principles for validation according to Messick (1995), is sampled through factor analytic procedures, multitrait-multimethod procedures, or any other method that can add to the understanding of obtained scores. In papers I, II, and V, the test-taking motivation construct was related to other related constructs, different methods were used to investigate the test-taking motivation construct, and all motivational variables were related to achievement variables in TIMSS 2003.

Further, in the external stage of construct validation, rival hypotheses should be specified and tested, and findings at this stage of the validation process can
contribute to the interpretation of the generalizability of results. At the external stage, where the construct of interest is related to other constructs, it is important to note that the validity of the measurement of these constructs also has to be evaluated.

This stage of construct validation may appear rather traditional and technical in nature but it is important to stress that the valid interpretation and use of test scores in general requires some sort of empirical study. Messick (1989) claimed that that validation is about *empirically-based* score interpretation, and that the methods used for collecting evidence to support a particular inference are the methods of science. Nevertheless, it is important also at this stage to acknowledge the values held by the researcher as well as the surrounding community. As score meaning is not given by nature, subjective judgment is inherent in all stages of construct validation.

According to Benson, a strong program of construct validation includes all these three stages, and no stage is alone sufficient. From the above description, it should be clear that modern conceptualizations of validity are not very different from developing, empirically investigating and evaluating any scientific theory.

Benson’s program explicitly touches upon all but one of Messick’s proposed six aspects of construct validity, and many of Messick’s points on a sound validation practice are included in the program (see Messick, 1989). However, Benson does not explicitly include the sixth, consequential, aspect. This obviously is a draw-back though, as I understand it, this aspect is implicit in each stage of the program. It can also be noted that the program seems more concerned with score interpretation than score use. One plausible reason for this is that the program has been applied in the context of measuring psychological constructs, and that a valid interpretation of obtained scores is a necessary though not sufficient condition for the valid use of obtained scores. Although Benson’s program is perhaps not complete, and although my empirical studies do not fulfil all requirements of a “strong program for validation”, I found Benson’s program for construct validation to be illustrative of my own validation effort. Nevertheless, I would like to emphasize the importance of considering the value implications and the consequences of test use, and the relation between score interpretation and score use, as this dimension is an important part of a sound validation effort, not least in the larger context of TIMSS 2003.
Validity Issues in TIMSS 2003

In TIMSS and other large-scale international assessments the consequential aspect of validity, in terms of value implications as well as the intended and unintended consequences of test interpretation and test use, is important to consider. One of the reasons for this is that the results in studies like TIMSS can sometimes have consequences that are reflected in policy decisions concerning national school systems, and in perceptions of a nation’s student population as knowledgeable or not so knowledgeable. Another reason is that studies like TIMSS have the power to dictate which aspects of learning are considered important, and what kind of educational system that is considered desirable.

If results from large-scale comparative studies like TIMSS are to function well as information for policy makers, teachers, and others involved in the education of students, and if they are to serve the purposes they were intended to serve, the interpretation and use of results from these studies need to be validated (Ferrini-Mundy & Schmidt, 2005; Messick, 1987). The validity in terms of value implications and social consequences need to be explicitly evaluated, together with other aspects of validity. There is an array of relevant issues related to validity in the TIMSS context that could be discussed. For example, evaluations of comparability, interpretability, generalizability and sustainable construct interpretations are all very important (Messick, 1987).

Another aspect that is important to consider is the “social psychology of the assessment setting” which, according to Messick, “requires careful attention” (Messick, 1989, p. 14). How do respondents react to the tasks that are presented to them in TIMSS 2003? Do they perceive the tests they are about to complete as valid? How is the study presented to them? What are their general attitudes towards tests and testing? Does every student in every country react in approximately the same way, and are all students in all countries and cultures equally motivated to do their best on the test? And how does the social psychology of the assessment setting interact with test performance? These are issues that are vital for the validity of interpretation and use of test scores, both in terms of evidence and consequences. This is particularly true for studies like TIMSS, which involve so many students from so many countries and cultures. Unfortunately these are also issues that mainly have been forgotten in the research on the validity of large-scale,
comparative studies. Not controlling for systematic differences in students’ reactions and behaviors in the test situation may introduce a bias in the obtained scores, invalidate the interpretation and use of these results, and thereby lead to unintended consequences for those affected by the use of these results.

6. Methodology

Measuring Motivational Beliefs

Latent constructs like motivation present a delicate measurement problem for the researcher, as it is not possible to observe motivation directly and as there is no universally accepted definition or theory that describes the construct. Rather, there are many definitions and several theories that seem to describe approximately the same psychological construct (Bong, 1996). Methodologically, the valid measurement of student achievement motivation demands a well articulated theoretical understanding of the construct of interest, a careful and theory-based instrument development, a proper study design, a suitable sampling of study objects, appropriate methods for analyzing data, and an acknowledgement of the different aspects of validity in the interpretation, communication, and use of obtained results.

There are several ways of assessing motivation. One common strategy refers to behavioral measures which are assumed to be the result of, or indices of, motivation. Examples of behavioral indices are choice of tasks, measures of effort, measures of persistence, and achievement (Pintrich & Schunk, 2002). However, none of these indices is without flaw. Choice of tasks is not always applicable, and effort, persistence, and achievement are easily confounded by skill (Pintrich & Schunk, 2002).

A second common assessment strategy refers to verbalizations. Verbal ratings of motivation are often in the form of self-reports, questionnaires, interviews, and/or ratings by others. The value of these ratings is dependent on the honesty of the respondents, that they understand the questions they are being asked and that they are capable of answering. Verbal reports also suffer from the fact that they reflect subjective perceptions. The correlational research design applied in studies I-IV contained verbal reports of motivational beliefs in the form of questionnaire items. In studies II, III, and IV, behavioral indices in the form of achievement results were used as the criterion variable. In study V, a more qualitative research method was
applied, and verbal reports of test-taking motivation were sampled through interviews and an open-ended questionnaire item.

Papers I, II and V are concerned with student test-taking motivation in the TIMSS context. No established measures of test-taking motivation were available and there were no items asking for test-taking motivation in the TIMSS student background questionnaire. Therefore, a test-taking motivation questionnaire was developed and applied in the Swedish TIMSS context. The study design of this development and application is illustrated in Figure 2.

![Figure 2. Study design for validation of items in the Test-Taking Motivation Questionnaire (TTMQ).](image)

For validation purposes, the measurement was not restricted to the test-taking motivation questionnaire that the students completed before they took the TIMSS test, but two items asking for test-taking motivation was added to the TIMSS Student Background Questionnaire (SBQ-TIMSS 2003) as national options. This questionnaire was completed by the students after the TIMSS mathematics and science test. Also, posttest interviews were performed with a smaller sample of the students ($n = 30$).

In studies III and IV, items constructed and administered by TIMSS centrally were used in the analyses and hence, I had no control over the inclusion of items, wording of items, or the study design.
Statistical Choices

All empirical studies in this thesis are concerned with construct validation in one form or another. Factor analysis is by far the most frequently used method in construct validation in the social sciences (Pedhazur & Schmelkin, 1991; Thompson, 2004). In the present thesis, I used factor analytic methods in studies I, III, and IV.

Factor analysis can be grouped into two major classes, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The former is not dependent upon theoretical assumptions and a priori model specifications while the latter is. In EFA, relations between measured variables in the data set are searched for in a rather unbiased manner. The purpose is to isolate factors with a simple and interpretable structure. In confirmatory factor analysis, the appropriateness and meaningfulness of a theoretically motivated model with specified relations between items and variables is assessed. It should be noted that factor analysis is a rather subjective enterprise. Within EFA as well as CFA, there is an array of choices for the researcher to make and the choices made can greatly influence the outcome of the analysis. Also, the interpretation of results is dependent on the researcher’s judgement, his or her theoretical standpoint, the values he or she embraces as well as the purpose of the study.

Explorative factor analysis was used to analyze data in studies I, III, and IV. Also, in studies III and IV, basic confirmatory factor models were constructed and the model-data fit was tested. In studies II, III, and IV, linear regression was also used to investigate the association between verbal reports of motivation and behavioral indices of motivated achievement behavior. However, as EFA was used in the majority of my empirical studies, as EFA often is the starting point for a construct validation effort, and as EFA is a method that requires a number of informed decisions on the part of the researcher, the EFA decision sequence will be described below, while the CFA and the regression analysis will only be briefly touched upon.

Explorative Factor Analysis

In EFA, a matrix of associations among measured variables is analyzed. The purpose is to search for meaningful relations between variables, or meaningful clusters of variables, usually called “factors” or “components” depending on which extraction method is used. When performing an EFA, the researcher has
to decide first which matrix of association coefficients that should be analyzed, and then which method that should be used to extract factors and how many factors that should be extracted. The researcher also has to decide how these factors are to be rotated, if they are to be rotated at all. Last, but not least, the researcher has the delicate task of interpreting the results in terms of their meaning and reasonableness.

The most common matrix of association, and the matrix used in all EFAs in this thesis, is the Pearson product-moment bivariate correlation matrix. There are, however, other matrices that could be used (see Thompson, 2004 for a more detailed presentation of EFA as well as CFA).

As concerns factor extraction method, there are also many possible choices. The most frequently used extraction method is principal components analysis (PCA). PCA assumes that the scores on measured variables have perfect reliability (1’s on the diagonal of the correlation matrix). A second common factor extraction method is principal axis factoring (PAF), where perfect reliability is not assumed, but communality coefficients (which are lower-bound estimates of score reliability) are put on the diagonal of the correlation matrix. In the empirical papers attached to this thesis, PAF was used in study I and PCA was used as the main extraction method in studies III and IV.

Deciding on the number of factors to be retained is, obviously, a critical decision in EFA. Two common factor retention rules are the eigenvalue-greater-than-one rule and the scree test. These two methods are easily available in statistical programmes like SPSS and are therefore widely used. However, there are other, more promising, factor retention rules. Parallel analysis (Thompson and Daniel, 1996) involves computing eigenvalues for randomly generated data sets of the same size as the sample data set. The randomly generated eigenvalues are then compared to the sample eigenvalues and only those eigenvalues in the sample data set that exceed the corresponding values from the random data set are retained. The Velicer’s MAP (minimum average partial) test performs a principal components analysis and then examines a series of matrices of partial correlations. The MAP test is concerned with the ratio of systematic variance to unsystematic variance (O’Connor, 2000). Components are retained as long as the variance in the correlation matrix represents systematic variance.

The use of multiple factor retention rules is recommended (Thompson, 2004) and in my EFAs, I have chosen to use these two latter methods together with the eigenvalue-greater-than-one rule to inform me about the number of factors to retain.
Rotation of extracted factors is usually essential to interpretation (Thompson, 2004). There are two major classes of rotational techniques, those who assume that factors are perfectly uncorrelated (orthogonal rotational techniques) and those who allow factors to correlate (oblique rotational techniques). The most common orthogonal rotation method is \textit{varimax} (see Pedhazur & Schmelkin, 1991; Thompson, 2004 for a review of different rotational techniques) and this was the method I used in the orthogonal rotations.

Orthogonal rotation is probably the most common way of rotating factors. However, the assumption of perfect orthogonality of factors is not always valid when the constructs of interest are latent psychological variables that can be assumed to share some variance. Therefore, oblique rotation is advisable, at least as a first step to enable the researcher to evaluate if it is suitable to report an orthogonal solution. In my oblique rotation of factors, I used the direct oblimin rotation method.

As part of the construct validation effort, and to investigate the relation between motivation and test performance, the scales obtained through factor analysis were related to student achievement in TIMSS 2003 through linear regression performed in SPSS (studies II, III, and IV).

In studies III and IV, confirmatory factor models were specified and applied to data. A major difference between EFA and CFA is that in EFA, meaningful relationships between the measured variables are sought with the purpose of reducing the number of variables or “discovering” one or more latent variables. In CFA, on the other hand, relations between measured variables (indicators) and latent variables (factors) are specified by the researcher and the purpose of the analysis is to investigate the degree to which empirical data actually matches the specified model. In my empirical studies using confirmatory factor analysis, only measurement models were considered and no path models were specified. In the CFAs, the covariance matrix was used as the measure of association, and the estimation method was maximum likelihood. Each measurement model was an independent clusters model which means that no indicator was allowed to associate with more than one common factor. All factor variance parameters in all models were fixed to 1 which means that all factor pattern coefficients were free to be estimated.
7. Summary of the Papers

In this chapter the five papers attached to this thesis will be summarized. All five papers are related to the measurement of motivational beliefs (domain-specific and situation-specific), and they are all concerned with construct validation in one form or another. Papers I-IV are empirical investigations of motivational constructs in the TIMSS 2003 context, their structure and association with achievement in mathematics and science. Paper V has a more descriptive character, and issues related to perceptions of test stakes and test-taking motivation are discussed, using interviews and an open-ended questionnaire item as illustrative examples.

The first two papers represent an exploration into the concept of test-taking motivation, using a subsample of the Swedish TIMSS 2003-participants \((n = 350)\). The impetus for these two papers was a discussion about student motivation in low-stakes testing situations (see Wainer, 1993; Wise & DeMars, 2003). The first paper reports a rather traditional validation study, where the development and validation of scores from a questionnaire intended to measure student test-taking motivation is described. The next obvious step was to relate ratings of test-taking motivation to test performance in TIMSS and this was done in paper II, which focused on mathematics. In this study, two motivational scales derived in TIMSS 2003 (self-concept and valuing of mathematics) were included in the regression analysis. This created an interest in these scales, their theoretical base, internal structure and relation to performance, as well as an interest in how these scales are scored and used in TIMSS 2003. This led to the third study, where all Swedish students participating in TIMSS 2003 \((n = 4,256)\) were used as the study sample. The internal structure and relation to performance of the mathematics motivational items in the TIMSS 2003 Student Background Questionnaire was investigated and compared to the scaling and scoring done in TIMSS internationally. The purpose of the fourth study was to investigate the internal structure and relation to performance of the science motivational items in the TIMSS 2003 Student Background Questionnaire. The fifth study used interview data and answers to an open-ended questionnaire item to describe students’ perceptions of test stakes and their perception of the TIMSS study in terms of motivation.

The empirical studies are not isolated from each other but are interrelated and build upon each other in the sense that the first study inspired the second, the second study raised questions that were investigated in the third, and so on.
Also, as all five studies take advantage of TIMSS 2003 data, there are cases of overlap between the studies. As I see it, these overlaps are not redundant, but rather an important part of the construct validation effort.

PAPER I. Development and Validation of Scores From an Instrument Measuring Student Test-Taking Motivation

It has often been argued that students do not try their bests on tests like TIMSS, in which the results do not have any consequences for the students. This argument was the impetus for the first study, which was concerned with construct validation in a rather traditional sense. As the literature on test-taking motivation is scarce, and as no established instruments measuring test-taking motivation were available, a self-report inventory asking for motivational beliefs in the TIMSS 2003 context (the Test-Taking Motivation Questionnaire, TTMQ) was constructed with the expectancy-value theory of achievement motivation as a basis (see Figure 2 for study design). The purpose of paper I was to describe the development and application of this instrument, as well as to report indices of internal consistency and discriminant validity. The instrument was completed by a subsample of the Swedish eight-grade students participating in TIMSS 2003 (n = 350). Internal structure and discriminant validity was investigated through exploratory factor analysis (EFA). Through the EFAs an investigation was undertaken on whether the instrument seemed to measure different aspects of test-taking motivation, and if this test-taking motivation construct seemed to be distinct from self-concept and valuing of the school subjects. It was concluded that it was suitable to collapse a number of the items in the test-taking motivation questionnaire into a scale labeled test-taking motivation, that this scale was distinct from the domain-specific motivational beliefs measured in TIMSS 2003, and that the theoretical assumptions were mainly, but not entirely, confirmed.

PAPER II. Test-Taking Motivation and Mathematics Performance in TIMSS 2003

Are students not motivated to do their best in TIMSS 2003? How is test-taking motivation related to test performance? These issues were explored in paper II using data sampled through the test-taking motivation scale described in the first study. The focus of this study was on mathematics.
In this second paper, Swedish students’ \((n = 343)\) ratings of test-taking motivation (pretest and posttest), planned effort and perceived importance of a good test performance were presented. A majority of the students reported that they were well motivated to do their best on the TIMSS test, before as well as after they completed the test, that they planned on investing a fair amount of effort when completing the test and that they valued a good performance in TIMSS. The test-taking motivation scale derived in the first study (paper I) was regressed on mathematics score (national Rasch score). Reported level of test-taking motivation was positively but rather weakly related to achievement. The multiple linear regression model included the mathematics self-concept and task value scales used in TIMSS internationally, and when these variables were held constant, the association between test-taking motivation and achievement was not statistically significant. There was a stronger association between test-taking motivation and achievement for boys than for girls.

It was concluded that the Swedish students in general seemed to have been well motivated to do their best on the TIMSS mathematics test, even though the test is regarded to be low-stakes for the students, and that the Swedish results in TIMSS are unlikely to be affected by a lack of student motivation.


In paper II, two scales derived in TIMSS internationally - mathematics self-concept and valuing of mathematics, were included in the regression analysis to investigate whether test-taking motivation could explain any variance in the TIMSS achievement score not accounted for by the domain-specific motivational beliefs measured in TIMSS 2003. This created an interest in these scales. What was the theoretical rationale behind the inclusion of these scales in TIMSS? How were the items scaled and scored? How were they related to achievement? Were they suitable for application in a Swedish context? The purpose of the third study was therefore to investigate the scale structure and relation to performance of the items in the TIMSS 2003 student background questionnaire assumed to measure mathematics self-concept and valuing of mathematics. Unlike the two previous studies, data in this study was collected from the entire sample of Swedish TIMSS 2003-participants \((n = 4,256)\). No theoretical rationale for the inclusion of these specific items is explicitly given in
TIMSS, but looking at the content of the items, the expectancy-value theory of achievement motivation seemed as a suitable theoretical framework. Exploratory (principal components analysis) and confirmatory factor analyses were performed on data and different scoring techniques were compared. The theoretical assumptions of the expectancy-value theory were basically confirmed. The results indicated that the self-concept scale was internally consistent and positively related to achievement (plausible values), while the value scale was rather heterogeneous and unrelated to achievement. The analysis largely supported the scaling performed in TIMSS internationally but the scoring of the scales for presentation in the international reports was questioned as being useful for secondary analyses.

PAPER IV. Science Motivational Beliefs in a Swedish TIMSS 2003 Setting: Scale Structure and Relation to Performance.

The focus of the first three studies (papers I, II, III) was mainly on mathematics. In TIMSS, science proficiency and student motivation to learn the science subjects is also measured and therefore the scale structure and relation to performance of the science motivational beliefs were examined in this fourth paper. The general purpose was similar to that presented in paper III, but in paper IV a more explicit attempt to test a proposed factor structure was made, based on theoretical assumptions and the results from paper III. The internal structure and relation to performance of the self-concept and task value scales for each science subject (physics, chemistry, biology, and earth science) was investigated through confirmatory factor analysis, exploratory factor analysis (principal components analysis), and regression analysis. It was shown that the factor solution for the science subjects was not as simple as for mathematics, and that the internal structure of the two motivational scales was not identical to the findings for the corresponding mathematics motivational scales in study III. The association between ratings of self-concept and achievement was weaker than expected, while ratings of task value were basically unrelated to achievement. Thus, the theoretical assumptions could not be entirely confirmed, and the proposed two-factor model could not be entirely supported. The fact that Swedish eighth-grade students have only had a few years’ experience of science taught as separate subjects was put forward as a possible explanation for the somewhat fuzzy factor solution. The need for further research on motivational beliefs in the science subjects was stressed.
PAPER V. Student Motivation on Low-Stakes Tests: An Example from TIMSS 2003

Papers I–IV all had a correlational design and investigated the internal structure and relation to performance of domain-specific (papers III and IV) and task-specific (papers I and II) motivational beliefs. However, none of these studies reveals much about what eighth-grade students actually perceive and think about a study like TIMSS. Indeed, the examinee perspective and their perceptions of tests and test situations is often ignored and this fact inspired the fifth study. Furthermore, a proper construct validation is dependent on multiple sources of evidence. The fifth paper elaborated on the findings from papers I and II and was concerned with student test-taking motivation on low-stakes tests. The approach was more qualitative and data used in this study consisted of interviews with thirty students and answers to an open-ended questionnaire item \( (n = 329) \). The purpose was to explore the students’ perceptions of TIMSS in terms of test stakes, their motivation to do their best and the reasons for this motivation or lack of motivation.

The students’ answers to the open-ended item largely corroborated the results from the questionnaire items with a closed item format. A majority of the students reported that they were rather well motivated to do their best in TIMSS. The answers further indicated that many of the students did not perceive TIMSS as a low-stakes test in the sense that it was unimportant to them. Many students put forward a comparative/competitive perspective or a social responsibility perspective as motivating. Some students reported the low stakes of the test as detrimental to their motivation but these students constituted a minority of the total sample.

The interviews, which were performed a few weeks after TIMSS, mainly-supported the validity of the inferences made from study II and the results from the open-ended item. In the interviews, the majority of the students said they had been motivated to do their best in TIMSS and that they perceived TIMSS as a rather important test. It was concluded that most students do seem to care about how they perform on studies like TIMSS, but that they nevertheless are aware of the fact that the test result does not count for them personally. The need for similar studies in other countries, cultures and contexts was stressed as the results are not generalizable due to their individualized character and the small study sample.
8. Discussion

The main purpose of the present thesis was to study aspects related to student achievement motivation in the TIMSS 2003 context. In papers I and II, the structure of the test-taking motivation construct and the association between test-taking motivation and performance were explored. In papers III and IV, the structure of domain-specific motivational beliefs, in terms of self-concept and task value, and their relation to performance were investigated. In paper V, students’ perceptions of TIMSS 2003 and their motivation to perform well in the study were investigated through interviews and an open-ended questionnaire item. Swedish eighth-grade students participating in TIMSS 2003 formed the sample in all studies. In the introductory part of the thesis, these empirical studies have been summarized and contextualized. The rationale behind the thesis has been presented from the perspective of a larger validity theory framework and an achievement motivation framework as well as from the perspective of an international comparative study. In this concluding chapter, the main findings of the thesis will be presented and discussed from a theoretical and an applied perspective and some suggestions for future studies will be made.

Main Findings From the Empirical Studies

First, contrary to commonly held preconceptions, the questionnaire study (papers I and II) as well as the interview study (paper V) indicated that the Swedish students participating in TIMSS 2003 in general were well motivated to do their best on the TIMSS test. Thus, the argument that students lack motivation to do their best on low-stakes tests like TIMSS could not be supported from the results presented in this thesis. Also, the test-taking motivation construct in terms of task value perceptions and perceived motivation to do well seems to be a construct distinct from domain-specific motivational beliefs in terms of self-concept and valuing of school subjects.

Second, the Swedish mathematics result in TIMSS 2003 does not seem to be affected by a lack of motivation among the participating students, as the students reported that they were well motivated to do their best and as ratings of test-taking motivation were positively but rather weakly related to performance. Further, boys’ ratings of test-taking motivation were more strongly related to how they actually achieved on the TIMSS test than girls’ ratings were.
Third, although the relationship between test-taking motivation and test performance was rather weak, the test-taking motivation variable still explained some of the variance in the result that could not be explained by the domain-specific motivational variables actually measured in TIMSS.

Fourth, the scales derived for mathematics self-concept and value of mathematics in TIMSS internationally in general seemed appropriate also for the Swedish sample, and the associations between ratings of self-concept, task value and achievement were in accordance with theoretical assumptions. The internationally derived scales assumed to measure self-concept in and valuing of the science subjects proved to be less consistent with theoretical assumptions and not as easy to interpret as the corresponding scales for mathematics.

Beyond the Main Empirical Findings

This thesis has presented a way of conceptualizing and measuring student test-taking motivation. The thesis also exemplifies how an instrument can be developed, how scores from this instrument can be validated, and why it is important to consider and reflect upon different aspects in the measurement process. The thesis has also proposed a way of interpreting the motivational scales measured in TIMSS 2003. Together, the empirical papers and the introductory section have shown that measurement in the behavioral sciences is a process that includes several considerations, none of which can be ignored if the interpretation and use of obtained scores are to be valid.

The implications of this thesis can be conceptualized on three different levels. The first (micro-) level is mainly concerned with aspects important to consider in the measurement of latent constructs like motivation. The second (applied) level is concerned with what the findings (on the micro level) mean to the interpretation and use of scores from studies like TIMSS. The third (macro) level is concerned with quality issues in large-scale comparative studies in a wider sense. Below, the micro level is discussed mainly in terms of theoretical assumptions and explanations. The second level is discussed in terms of applied implications of the thesis. The macro level is briefly discussed in terms of validity and validation in the larger TIMSS context.
Theoretical Implications of Obtained Results

The introductory part of the thesis stressed the need for theory in the measurement of latent constructs and in the validation of this measurement. All five empirical papers attached to this thesis used the expectancy-value theory of achievement motivation as the general theoretical framework. The theory predicts that performance expectancies and task value perceptions are important determinants of achievement behavior and achievement choices, and that these perceptions are influenced by a variety of contextual and psychological factors (see Figure 1, p. 20). This theory has been extensively validated, but a theory that is valid in some contexts, for some purposes, and for some time period, is not necessarily valid in others (Benson, 1998; Messick, 1989). In particular, test-taking motivation can be seen as a special case of achievement motivation as it is situation-specific. The construct is not well researched, and it is not obvious that the expectancy-value theory is the best theory for explaining the test-taking motivation construct (see Eccles & Wigfield, 2002, for a review of motivational theories focusing on expectancies, values, or both).

One issue worth discussing is whether performance expectancies and task value perceptions are the best indicators of test-taking motivation. It may be difficult to empirically differentiate task-specific performance expectancies from domain specific ability beliefs, as shown in paper I (see also Wigfield & Eccles, 2002, 2000), and the focus in papers II and V was therefore mainly on task value perceptions. However, previous studies have shown that domain-specific task value perceptions are not very strong predictors of actual achievement, but rather predictive of future achievement choices (Eccles & Wigfield, 2002). Future achievement choices are not relevant in the test-taking motivation context, as this kind of motivation is task-specific. For the task value component to be meaningful, appropriate, and useful (i.e. valid) on a task-specific level as a predictor of achievement, task-specific value perceptions should be at least moderately related to achievement. The results in paper II indicated that there was a positive but rather weak correlation between ratings of test-taking motivation and test performance. Here, it should be remembered that the variance in the test-taking motivation variable was rather restricted which may have caused an underestimation of the association between motivation and achievement. Whether task value perceptions are predictive of achievement or not may also be dependent on which specific questions are asked and how task value perceptions are operationalized (Bong, 1996).
The work presented in the thesis is based on two rather comprehensive theoretical frameworks: Eccles and Wigfield’s expectancy-value theory and Messick’s validity theory. The empirical investigations explored only part of the Eccles & Wigfield’s expectancy-value model, and only part of the validity theory was applied. For comprehensive theories of achievement motivation as well as for the comprehensive validity theory, it is often not feasible to include all parts of the model in a single study. However, motivation to do well on a given test is most likely not only determined by the characteristics of the specific test and the test situation but also by students’ goal orientations, personal reasons for performing in one way or another, and achievement history. Many important aspects of the expectancy-value model and their relation to students’ motivational disposition in test situations thus remain to be studied. A thorough development of the theoretical as well as the empirical domain of the test-taking motivation construct should consider other relevant aspects of the expectancy-value model and perhaps other theoretical frameworks (see for example Covington, 2000; Deci, Koestner, & Ryan, 1999; Eccles & Wigfield, 2002)

For now, based on the measure developed, used and interpreted in this thesis, it can be concluded that test-taking motivation, in terms of task-specific perceptions of task value and motivation to do well on a test, seems to be a construct empirically distinct from domain-specific motivational beliefs. The test-taking motivation construct also seems to be positively albeit not very strongly related to achievement.

According to modern perspectives on validity, empirical findings should be related to theoretical assumptions in order to increase the meaningfulness and interpretability of results (Messick, 1989). The theoretical framework behind the motivational scales in the TIMSS 2003 Student Background Questionnaire is never made explicit and it is therefore difficult to support the validity of interpretations made. The results presented in this thesis indicate that the expectancy-value theory of achievement motivation can be a suitable theoretical framework within which the motivational scales used in TIMSS can be interpreted. This is true for the mathematics motivational scales as well as for the science motivational scales, although the science motivational scales did not behave entirely in accordance with the theoretical assumptions.

In summary, the results presented in this thesis indicate that the expectancy-value theory of achievement motivation is a suitable theoretical framework for describing the test-taking motivation construct as well as the motivational scales used in TIMSS 2003. Nevertheless, the results also indicate that all measures of
motivation used in the empirical studies need further development and refinement, and that rival hypotheses, theoretical and empirical, should be specified and tested as well.

Applied Implications of Obtained Results

What do the thesis’ findings mean for the interpretation and use of scores from studies like TIMSS? First, measurement in the behavioral sciences in general has the purpose of obtaining information about individuals or groups, information that can then be used for various purposes. The fact that the obtained information is often used for one purpose or another makes the question of measurement quality and validity a fundamental concern also in an applied sense (Kane, 2004) as illustrated in the introductory part of the thesis.

The present thesis explored an issue that has largely been ignored in the literature: student motivation to do their best in low-stakes contexts. Test-taking motivation can be relevant to the validity of interpretation and use of test results and ignoring the test-takers and their views is not compatible with modern conceptions of validity (Messick, 1989). If test-taking motivation would contaminate responses to tests, this is an example of construct-irrelevant variance that affects the validity of test score interpretation (Benson, 1999; Messick, 1989). It follows that students’ reactions to tests and their task-specific motivation to do well should be acknowledged in the interpretation and use of test scores, also in TIMSS.

The results obtained in the present thesis imply that even a short measure of test-taking motivation can provide important information, and it would be possible to include a measure of student effort, performance expectancies, perceived importance of a good performance, or level of motivation in the TIMSS test battery. Including such a measure could contribute to the understanding of score meaning, and to the validity of score based inferences.

As noted in chapter 2, the number of items in the TIMSS Student Background Questionnaire measuring aspects related to student motivation has decreased over time. Also, the theoretical rationale for the inclusion of particular items is not made clear. As student motivation is important to achievement behavior and future achievement choices (Pintrich & Schunk, 2002), inclusion of an additional number of relevant and theoretically motivated items measuring student achievement motivation in mathematics and science could add to the understanding of obtained results. This must not necessarily be done in the
internationally standardized instruments. Each participating country has the possibility of adding national options to the questionnaires and more nations, or possibly collaborating clusters of nations, should take advantage of this possibility.

The TIMSS administration encourages researchers using TIMSS data to analyze the internationally derived scales and to derive their own scales but few researchers seem to have appreciated this possibility. Instead, either the internationally derived scales or single items are used as indicators of latent constructs. The present thesis illustrates that it is important to empirically examine the appropriateness of internationally derived scales in national contexts.

In summary, the results presented in this thesis imply that contextual variables and the appropriateness of these variables need to be considered when interpreting and evaluating results from studies like TIMSS. Further, the thesis illustrates the possibility of adapting or adding national options to the TIMSS test battery. All nations participating in TIMSS have unique characteristics and adding national options to the anchor instruments administered by TIMSS could mirror these unique characteristics and enable large-scale investigation of questions of particular interest. I believe that this would be beneficial not only to a given country, but for the dynamics and vivacity of international studies as such.

Validity and Validation in the Larger TIMSS Context

In a larger TIMSS context, the studies attached to this thesis concern a micro-level. They are concerned with student motivation only, with one national context only, and they have examined only two out of many possible scales in the TIMSS background questionnaires. There are of course many other relevant validity issues in the TIMSS context, as was noted in chapter 5.

Samuel Messick (1987) has listed a number of central features of good large-scale policy research. The first obvious feature is a capacity to provide data that are comparable across time and student populations. A second central feature is a capacity to provide correlational evidence that support construct interpretations. A third central feature is the measurement of background factors that illuminate context effects. These features are all important to consider for researchers using TIMSS data and for those designing the instruments used.

As concerns validity and validation in the general TIMSS context, as well as in the context of any large-scale comparative study, Messick’s model of validity with two dimensions, one evidential and one consequential, appears to be a suitable model for validation (see Messick, 1988, 1989, 1995) as questions about value
implications and intended as well as unintended consequences are highly relevant for this kind of study. Studies like TIMSS can have an impact on what is regarded as desirable knowledge and how this knowledge is best measured and taught. Large-scale, standardized achievement testing is a normal feature in many but not all of the national school systems that participate in TIMSS. TIMSS may have consequences for how national school systems structure their system for testing their students, and ultimately how they teach their students. This is one of TIMSS’ purposes, but for the consequences to be of an intended sort, it is crucial that the inferences and interpretations made from TIMSS scores, and the use of these scores, are valid. Thus, regardless of whether one thinks that the social consequences of a measurement should count to a unified view of validity or not, the potential consequences of large-scale international studies like TIMSS should be recognized and evaluated to a larger extent than they have been so far. This does not mean that the ordinary analysis and reporting of TIMSS data is not relevant or useful. On the contrary, TIMSS and other IEA studies are in many respects examples of an ideal large-scale measurement practice. They are the result of many years’ experience of conducting comparative studies, and much care is taken to make the measurement as reliable and valid as possible for as many nations as possible. Much care is also taken to make all assessment frameworks, study rationales and data bases available to anyone that is interested in performing secondary or extended analyses. Still, with such rich data bases that are open to and shared by so many researchers in so many countries, more thorough investigations into the meaning of obtained scores would be possible and desirable. There is always room for improvement, and the present thesis can be viewed as a small-scale example of how the interpretation and use of scores obtained from large-scale comparative studies could be improved.

Limitations and Generalizations

The empirical studies presented in this thesis are not without limitations. First, the studies of student test-taking motivation were explorative in nature. There was no given theoretical domain for this construct; there were few previous studies performed and no established instrument available. It follows that the instrument is somewhat tentative and needs continued development and continued validation. Further, a theoretical limitation is that papers II and V mainly focus on task value perceptions and perceived motivation and not
performance expectancies or other parts of the expectancy-value model. Although it is acknowledged that performance expectancies and task value perceptions are interrelated, it could be argued that the task value component is a more critical indicator of test-taking motivation than the expectancy component is. It can also be noted that if one wants to increase student motivation for a given test, task value perceptions are possibly more amenable to change in the short perspective. It should be remembered that the focus of the studies on test-taking motivation was on a specific context, TIMSS 2003, and how this low-stakes context may influence student motivation. A further development of the instrument could include a stronger focus on testing parts of the expectancy-value model or other theoretical models, and possibly include items that use more relative ratings, where students are asked to compare their level of motivation for a given test compared to other tests or other activities (see Nevo, 1993).

As the test-taking motivation scale used was short and tentative, construct underrepresentation (Messick, 1989, 1995) might be present. This is difficult to know, however, as the theoretical and the empirical domain of the construct is largely unknown, and as the construct "motivation" is difficult to define and demarcate to begin with (see Bong, 1996). The scales used in TIMSS internationally, particularly the task-value scale, may also suffer from threats to validity in the form of construct underrepresentation as suggested by confirmatory analyses. Still, all findings in all empirical studies point in the same direction, which speaks for the credibility of results.

From a TIMSS perspective, an obvious limitation of the research presented in this thesis is that it only includes Swedish TIMSS participants. The results in studies I, II, and to some extent paper V are generalizable to all Swedish students participating in TIMSS 2003 as the sample used was representative of all Swedish TIMSS-participants. They are not generalizable to other student populations or other settings, however. Studies III and IV used the total sample of Swedish students participating in TIMSS. As TIMSS sample weights were added to data, the results are generalizable to the entire population of Swedish eighth-grade students. Thus, the structure and relation to performance of the domain-specific self-concept and task value scales can be assumed to be similar in any Swedish sample of eighth-graders, and for any mathematics or science achievement measure.
Suggestions for Future Research

Some suggestions for future research have already been made in this chapter and below a number of these suggestions are emphasized and summarized. As noted, the need for further research on the test-taking motivation construct is obvious. Future studies should have a theoretical as well as an applied focus as it is important to develop a sound theoretical model for the construct as well as to further study the relationship between test-taking motivation and test achievement as well as the relationship between test-taking motivation and other variables related to achievement motivation. Also, no study has attempted to conduct a cross-national study of student test-taking motivation and this is an important task to pursue, not least in light of the increasing number of and impact of international comparative studies. As indicated by study II, group differences in test-taking motivation such as gender differences or differences between ethnical and social groups in national and international contexts are also worthy of systematic investigation.

In general, there is a need for more research on students’ reactions towards tests and testing, not least in a Swedish context where these issues have been largely ignored. Such research would be important from a theoretical, psychological perspective as well as from an applied, didactical perspective.

In the TIMSS context, there is a need for more detailed analyses of the instruments used in TIMSS, and for more detailed analyses into the meaning of test scores. Such analyses could be done in national as well as in international contexts. In Sweden, very few studies have used TIMSS data for secondary analyses, but the instruments used and the relationships between background variables and achievement variables ought to be investigated in order to understand student performance over time and in an international perspective. Given the wealth of information that is present in the TIMSS data bases, well performed studies using this information and any additional information gained from using the possibility of adding national options could be very useful.
References


