Motivational Aspects of Test-Taking

Measuring Test-Taking Motivation in Swedish National Test Contexts

Eva Knekta
To all students participating in my studies
# Table of Contents

1. Introduction  
   1.1 Aim 3  
   1.2 Disposition of the thesis 4  
   1.3 Terminology 4  

2. Achievement motivation  
   2.1 Motivation theories 6  

3. The expectancy-value theory of achievement motivation  
   3.1 Why the expectancy-value theory of achievement motivation? 9  

4. Test-taking motivation  
   4.1 Assessment contexts where test-taking motivation may be relevant to consider 10  
   4.2 Previous research on test-taking motivation 12  
      4.2.1 Measures of test-taking motivation 12  
      4.2.2 Factors influencing test-taking motivation 13  
      4.2.3 Test-taking motivation and test performance 14  

5. The Swedish national assessment system  
   5.1 Research on students’ perception of the national tests 17  

6. Validity theory  
   6.1 Views of validity over time 18  
   6.2 Validity and validation in the thesis 20  

7. Method and methodological choices  
   7.1. Research design from a validity perspective 23  
   7.2 Data collections and participants 25  
   7.3 Instruments 25  
      7.3.1 The test-taking motivation questionnaires 25  
      7.3.2 Interview guide 28  
      7.3.3 The achievement tests 28  
   7.4 Statistical Analysis 29  
      7.4.1 Latent variable modelling 29  
      7.4.2 Structural equation modelling 29  
      7.4.3 Two level second order latent growth modelling 31  
   7.5 Thematic analysis 32  
   7.6 Ethical Considerations 33
8. The Studies, overview
  Study I 33
  Study II 35
  Study III 36
  Study IV 37

9. Discussion 38
  9.1 Main findings 39
  9.2. How can the test-taking motivation construct be theoretically understood, operationalized, measured and modelled? 39
    9.2.1 Towards a better theoretical understanding of test-taking motivation 40
    9.2.2 To operationalize and measure test-taking motivation 41
    9.2.3 To model the test-taking motivation construct 43
  9.3 Students’ self-reported test-taking motivation at tests with different stakes 44
    9.3.1 Implications for assessment practise 45
  9.4 A reflection on the research design from a validity perspective 47
  9.5 Limitations and future research 49
  9.6 Concluding Remarks 50

10. Författarens tack 51

11. References 53

Appendix 1.
Abstract

The overall aim of the work underlying this thesis was to improve the understanding of students’ test-taking motivation in connection to achievement tests for young adolescents. The thesis includes four studies and a summary. All four studies explore test-taking motivation and are all connected to validity in one way or another. The expectancy-value theory of achievement motivation was used as a theoretical framework in the operationalization and measurement of test-taking motivation and the achievement tests in focus were the Swedish national tests in science for grade nine students, a Swedish national test in mathematics for upper secondary students and field trials for both these tests.

In the first study psychometric properties of an expectancy-value based questionnaire measuring five aspects of test-taking motivation were evaluated. Findings provided support for construct validity, as internal structure of the data corresponded to the theoretical model, partial scalar invariance was obtained and correlations between test-taking motivation and test performance were found. In the second study the students’ test-taking motivation at the field trial and the national test were analysed in more detail. The analyses showed a significant increase in all test-taking motivation aspects between the field trial and the regular national test. Test-taking motivation at the field trial did also differ between school classes and test anxiety and expectancies differed between females and males. Further, effort, expectancies, importance and interest were significant predictors of test score after preliminary grades had been accounted for. The third study was an interview study aiming to examine students’ test-taking motivation at national test as well as the field trial in more depth. The interviews provided a rich understanding of the different aspects of test-taking motivation as described by the expectancy-value theory, for example why students find the national test important and what they mean by giving effort to a test. In the fourth study psychometric properties of an extended and revised test-taking motivation questionnaire were evaluated in a new sample of students. The analysis showed that the questionnaire had sound psychometric properties and supported the assumption that test-taking motivation consists of several distinctly different aspects, but also showed that the different subscales are highly related.

In conclusion, the results provide support for several aspects of construct validity of the test-taking motivation questionnaire. Further, all studies showed that test-taking motivation differ between test with different stakes.
as well as between students and classes taking similar tests. Test-taking motivation also seem to affect test results. Thus, all students do not seem to be equally motivated to do their best on all tests and consideration should be taken to how students approach and experience each specific test when planning and interpreting achievement tests.
Populärvetenskaplig sammanfattning


En viktig del av arbetet som ligger bakom denna avhandling var att utveckla och utvärdera en enkät för att mäta provmotivation hos ungdomar i samband med kunskapsprov. Baserat på litteraturstudier och befintliga enkäter utvecklades en enkät som inkluderade elevernas uppfattning av deras motivation att göra sitt bästa på proven, hur de tror att de gick för dem på provet, samt det upplevda värdet av att göra provet. Enkäten delades sedan ut till elever som genomförde svenska nationella prov i de naturorienterande ämnena i årskurs nio, svenska nationella prov i matematik för gymnasieelever samt i samband med utprövningar av provuppgifter för dessa prov. I samband med de nationella proven i de naturorienterande ämnena gjordes även intervjuer med eleverna om deras uppfattning av de nationella proven och utprövningarna inför dem.

Sammantaget visar alla studier att enkäten som utvecklades, administrerades och utvärderades hade lovande egenskaper. Enkäten är kort och lätt att administrera men innefattar ändå flera aspekter av provmotivation. Överlag stämde de insamlade data överens med den teoretiska modellen. Elevernas vilja att göra sitt bästa på provet kunde till relativt stor grad predikeras av hur bra de trodde det skulle gå för dem på provet och hur värdefullt de ansåg att det var att göra provet.
Resultaten från enkäter och intervjuer visade att eleverna överlag har en positiv upplevelse av och syn på de nationella proven. Eleverna anstränger sig på de nationella proven, de tycker de är viktiga och bra. En del elever rapporterar dock att de var så nervösa under provet att de glömde bort saker de egentligen kan. Samtliga empiriska studier visade att elevernas provmotivation skiljer sig mellan utprövningarna och de nationella proven samt mellan studenter och klasser som gör samma prov. Resultaten visade också att det finns ett samband mellan provmotivation och resultat på proven. Således verkar antagandet att alla elever gjort sitt bästa för att visa vad de kan inom det aktuella området inte alltid vara rimligt. Vid genomförande och tolkning av kunskapsprov bör man därför alltid reflektera över studenternas provmotivation vid det unika provtillfället.
Studies

This thesis is based on the following studies, which are referred to in the text by the enumeration used below¹


III. Knekta, E. & Sundström A. (2017). “It was perhaps the most important one” Students’ perception of national tests in terms of test-taking motivation. Submitted after 1st round of revision.


¹ Two of the studies (I and III) are co-authored and below, the authors’ respective contributions are specified.

Study I: I came up with the initial idea and design, performed the data collection and data analysis and was in charge of drafting the manuscript. The questionnaire was partly built on a previous questionnaire developed by HE. HE helped out in the planning of the study and analysis and interpretation of data. She revised the manuscript critically for important intellectual content and improved readability.

Study III: I came up with the initial idea and design. Further I planned and performed the interviews. Analysis of the interviews and writing were made in collaboration with AS.
1. Introduction

Imagine that you are asked to complete an achievement test. Your perception of the test will probably affect your approach to it. You might ask yourself why you should do the test and if you are able to complete it. If you can’t find any reason to complete the test or don’t think you will be able to complete it, you might feel unwilling to put effort into it. This illustrates the concept of test-taking motivation, which can be defined as a test-taker’s willingness to do her best in a specific test situation. The focus of this thesis is on test-taking motivation in the context of achievement tests.

If students are not motivated to do their best on a test, the test’s results may not only reflect the students’ knowledge of the subject in question but also their level of motivation, causing their knowledge and capabilities to be underestimated. This in turn can lead to inaccurate conclusions about issues such as which areas teachers should focus on in future, schools’ effectiveness, or item difficulty. Thus, test-taking motivation is an issue of validity of interpretation and use of test scores. Low motivation can affect test scores from any assessment but is likely to be especially concerning when the test is of little or no consequence to the students but is important to other stakeholders (low-stakes test). Such cases include tests that are primarily used for accountability purposes, to monitor student achievement on a system level, to try out items for high-stakes tests, or to compare educational systems. These tests can seem quite unimportant to students because their results have no consequences for the individual test-taker: they do not count towards grades, are not used for formative purposes and sometimes students don’t even get any feedback on their performance. Conversely, at tests that have high personal consequences (high-stakes tests) test-takers are assumed to be motivated and test-taking motivation is therefore not considered to be a great concern from a validity perspective. However, some students may be unable to do their best in high-stakes tests because of factors such as test anxiety or low expectancies, which could also lead to invalid interpretations of their knowledge. In addition to providing important insights into test validity, in terms of how motivation affects test performance, an improved understanding of test-taking motivation would contribute to a better understanding of test-taking psychology, in terms of the ways in which tests affect test-takers. Students’ perceptions of tests have shown to be related to their attitudes, learning and future achievements (Marton & Säljö, 1997; McMillan, 2016; Struyven, Dochy, & Janssens, 2005). It is therefore essential to develop and administrate achievement tests in a
way that ensures that the students’ test results can be validly interpreted while also promoting a positive attitude towards tests and learning. A good understanding of students’ test-taking motivation is essential for achieving both aims.

Although there is an awareness about the potential effects of test-taking motivation on achievement test results, and those of achievement tests on test-takers, students’ perceptions of tests are often neglected and most achievement test results are interpreted as if all students were motivated to do their best on the test (Elwood, 2012; Wise & Smith, 2016). There is a growing body of research in the test-taking motivation area, and a number of studies suggesting for example that there is a relationship between test stakes, test-taking motivation and test performance (Liu, Bridgeman, & Adler, 2012; Thelk, Sundre, Horst, & Finney, 2009; Wise & DeMars, 2005). Still, better understanding of test-taking motivation in different test contexts and student populations is needed. In the Swedish context, we know very little about how students perceive different tests in general, and even less about how they perceive the Swedish national tests in terms of test-taking motivation. Furthermore, the theoretical understanding of test-taking motivation is unsatisfying. Many studies on test-taking motivation are based on measures focusing mainly on the outcome, the unmotivated behaviour, or motivated behaviour (i.e. objective measures of response time or self-reported measures of effort; Wise & Kong, 2005; Sundre, 2007). Fewer studies have investigated the underlying beliefs and attitudes that cause students to be motivated or unmotivated when taking achievement tests. However, such knowledge is essential if one wishes to increase students’ willingness to do their best in low-stakes tests or enable them to perform as well as possible in high-stakes tests.
1.1 Aim

The overall aim of the work presented in this thesis was to improve the understanding of students’ test-taking motivation in the context of achievement tests for young adolescents and to thereby contribute to the research regarding theoretical, methodological, and practical issues associated with test-taking motivation.

More specific aims were:
1) to explore how the test-taking motivation construct can be theoretically understood, operationalized, measured and modelled in a valid way, and
2) to investigate self-reported test-taking motivation in different student populations and for tests with different stakes.

Using Stokes’ (1997) terminology, the research in this thesis can be characterized as *use-inspired basic research* because its main focus is on scientific understanding of the test-taking motivation construct but it also has a practical goal of generating knowledge that will be useful when developing and administrating achievement tests. The expectancy-value theory of achievement motivation was used as a theoretical framework in the operationalization and measurement of test-taking motivation. Validity theory was guiding the research design choices and considered in the interpretation and contextualisation of findings. Data for analysis were drawn from the Swedish national test in science for grade nine students, a Swedish national test in mathematics for upper secondary students, and field trials for both these tests. Four empirical studies form the basis of the thesis. The first study examined the psychometric properties of an expectancy-value based questionnaire measuring five aspects of test-taking motivation. The second focused on modelling and analysing changes in reported test-taking motivation from a low-stakes to a high-stakes test, as well as differences in test-taking motivation between classes. The third study was an interview study aiming to examine students’ test-taking motivation for the national test in science and the field trial in more detail. Finally, in the fourth study psychometric properties of an extended and revised test-taking motivation questionnaire was evaluated in a new sample of students using exploratory structural equation modelling.
1.2 Disposition of the thesis
This thesis consists of a summary and four empirical studies, all focusing on test-taking motivation. After this introductory chapter, chapter 2 introduces the concept of motivation and common theories used to understand achievement motivation. Chapter 3 describes the expectancy-value theory of achievement motivation, which is the theoretical framework used to conceptualize test-taking motivation in this thesis. Chapter 4 describes the test-taking motivation construct and research within this field in more detail. Chapter 5 briefly describes the Swedish national assessment system. Chapter 6 deals with validity, starting with a brief historic overview of the concept and then reflecting on validity and validation as applied in this thesis. Chapter 7 describes the method and methodological choices, and chapter 8 presents summaries of the four studies. In chapter 9, the main findings are first summarized, the results are then discussed in relation to my research aims as well as in terms of validity. Finally, the limitations of this work are discussed and some suggestions for future research are presented together with the concluding remarks.

1.3 Terminology
Before discussing the contextual and theoretical frameworks used in this thesis, I would like to define how I have used some important terms. This thesis concerns measurement of the construct test-taking motivation. By “construct”, I mean a concept that is not directly observable but is inferred from interrelated sets of observations. According to Crocker and Algina (2008; p. 5), “measurement of the psychological attribute occurs when a quantitative value is assigned to the behavioural sample collected by using a test”. In this thesis I use the term “measurement” in a broader sense that also includes qualitative data, i.e. data consisting of words (collected during interviews or by other means) rather than numbers. Thus, measurement of a psychological attribute occurs when an evaluative device or procedure is used to systematically collect information about an individual’s behaviour in a specific domain. The device and procedure could be an achievement test, a questionnaire, an interview or an observation, among other things.

Throughout this thesis, I discuss low-stakes and high-stakes tests. I define the terms “low-stakes test” and “high-stakes test” in relation to the test-taker’s perspective: a test that has no consequences for the test-taker is treated as a low-stakes test even if it is important to other stakeholders. Similarly, a high-stakes test is one with direct consequences for the test-taker. For the sake of simplicity, the terms are used dichotomously. However, in reality stakes probably vary in a more continuous fashion, from
very low to very high. Moreover, a test that is low-stakes to one test-taker will not necessarily be so for a different test-taker.

The achievement tests considered in this thesis were the Swedish national test in science for grade nine students, a Swedish national test in mathematics for upper secondary students, and field trials for these tests. Henceforth, the terms “the national test”, “the high-stakes test” or “the regular national test” refer exclusively to the national tests; the terms “the field trial” and “the low-stakes test” refer to the field trials.

Finally, when I use the term test-taking motivation, I refer to the whole construct, including both effort and underlying motivational beliefs. The definition of test-taking motivation is discussed further in chapter four.

2. Achievement motivation

Motivation is important for all school-related activities and is reciprocally related to learning and performance (Pintrich, 2003): motivation can affect our behaviour at school, and how we behave at school affects what we learn. At the same time, how, when, why and what we learn can affect our motivation. Motivation differs between individuals and within individuals across different subjects and activities, and over time depending on both personal and contextual factors. In this chapter I briefly define and describe achievement motivation in general and different theories that can be used to understand it.

The term motivation derives from the latin verb *movere* (to move), which implies that motivation makes us move, keeps us moving (Schunk, Pintrich, & Meece, 2010). Schunk et al. (2010, pp. 4) define motivation as ‘the process whereby goal-directed activity is instigated and sustained’. Motivation in school is described in more detail by Wentzel and Wigfield (2009, p. 1) as:

> the energy [students] bring to this tasks, beliefs, values and goal that determine which task they pursue and their persistence in achieving them, and the standards they set to determine when a task has been accomplished.

The focus of this thesis is test-taking motivation. Simply speaking it is about the students’ willingness to do their best on a specific test. It is a situation specific motivation supposed to differ between individuals and between different situations. In chapter 4 test-taking motivation will be described in more detail.
2.1 Motivation theories
As demonstrated by the above discussion, motivation is a complex construct and several theories have been developed to explain different aspects of it. Early views on motivation were introspective, focusing on inner forces such as desire and innate properties (Schunk et al., 2010). In contrast, many theories arising in the early 20th century were behavioural, with motivation being seen as an observable behaviour that is a function of environmental events or stimuli, not necessarily including thoughts or feelings (Schunk et al., 2010).

The dominant theories today are cognitive theories, which examine the underlying mental processes involved in motivation and their dependence on personal and environmental factors (Schunk et al., 2010; Wentzel & Wigfield, 2009). These theories assume that motivation involves cognition and that cognition directs actions. Early cognitive theories focused on the individual’s characteristics (e.g. their beliefs about their abilities, self-efficacy and expectancies for success, attributions, reasons for pursuing specific outcomes, and focus on particular goals). More recent theories also account for developmental, ecological and socialization factors that can influence motivational beliefs and intentions (Schunk et al., 2010; Wentzel & Wigfield, 2009). Some of the main contemporary motivational theories are expectancy-value theories (Wigfield & Eccles, 2000), attribution theories (Weiner, 1985), social cognitive theories (Pajares, 1996; Zimmerman, 2002), goal-oriented theories (Ames, 1992; Pintrich, 2000), interest and affect theories (Covington, 2000; Hidi & Harackiewicz, 2000; Pekrun, 1992), and intrinsic motivation theories (Ryan & Deci, 2000). These theories are not essentially contradictory to each other but rather focus on different aspects of the complex motivation construct. Several aspects of motivation, such as beliefs, values, goals, needs and emotions, recur in one way or another in many of the theories. In an overview of motivational research based on these theories, Pintrich (2003) highlights five social cognitive constructs that can explain what motivates students in classrooms: (1) adaptive self-efficacy and competence perceptions - students who expect to do well tend to try hard, persist, and perform better; (2) adaptive attributions and control beliefs – put simply, students who believe they have more personal control over their own learning and behaviour are more likely to do well; (3) levels of interest and intrinsic motivation; (4) perceived task value; and (5) goals - the students’ reasons for performing a task affect their motivation and behaviour.

One way to analyse motivation that is relevant to this thesis involves distinguishing between situation-specific and domain-specific motivation. The latter relates to students’ motivation within a specific domain (for
example, a particular academic subject), while the former concerns students’
motivation in a specific situation. Both domain-specific and situation-
specific motivation beliefs give meaning to learning and performance
activities (Boekaerts, 2002). Therefore, a student’s motivation to learn
chemistry cannot be regarded simply as a stable domain-specific trait; it will
also depend on the characteristics of the student’s specific learning situation.
For example, a student who likes chemistry might not always be motivated to
invest a lot of effort into a chemistry task if she finds the task uninteresting
or unimportant, and she is more likely to find the task interesting if she finds
chemistry interesting and or important. This thesis looks at students’
willfulness to do their best on a specific test, which is a form of situation-
specific motivation.

3. The expectancy-value theory of
achievement motivation

The work described in this thesis is largely based on the expectancy-value
theory of achievement motivation proposed by Eccles et al. (1983) and
further developed by authors including Wigfield and Eccles (1992, 2000,
2002). The expectancy-value theory is a social cognitive motivation theory
and thus assumes that motivation involves cognition that directs action, and
depends on individual as well as social and contextual variables. It is widely
used as a framework for understanding domain-specific motivation but has
also been used to understand situation-specific motivation. In the modern
expectancy-value theory proposed by Eccles, students’ performance, effort
persistence and task choice depend on their expectations of success and the
value placed on the task (Wigfield & Eccles, 2000; Figure 1). Expectancies
and values are directly influenced by the students’ goals, general self-
schemata, affective reactions and memories. These social cognitive variables,
in turn, are influenced by aspects such as the students’ perception of
socializers’ beliefs, gender roles and their interpretations of experience. The
studies in this thesis have focused primarily on expectations of success and
the value placed on the task. According to Wigfield and Eccles (2000), the
expectancy component of achievement motivation relates to the individual’s
beliefs about how well they will do on upcoming tasks (Can I do this task?).
The construct is similar but not identical to Bandura’s efficacy expectation
(Wigfield & Eccles, 2000). The component is usually divided into ability
beliefs and expectancies for success, where ability beliefs are conceived as
broad beliefs about competences in a given domain while expectancies for
success relate to the student’s perception of their likelihood of succeeding in a specific upcoming task (Wigfield & Eccles, 2002).

Figure 1. Simplified schematic picture of the expectancy-value theory of achievement motivation.

The value component of the theory is about the extent to which an individual values a task (Why should I do this task?). Eccles et al. (1983) and Wigfield and Eccles (1992, 2002) defined the perceived value of a task in terms of four major aspects: (1) the task’s attainment value or importance, (2) the interest value of the task, (3) the task’s future utility value, and (4) cost. The definition of importance is the personal importance of doing well on a task (Eccles et al., 1983). It has many dimensions, including the task's capacity to confirm or disconfirm salient aspects of the individual’s self-schema, to provide a challenge, and to allow the individual to fulfil achievement, power and social needs (Eccles et al. 1983). Interest value is “the inherent, immediate enjoyment one gets from engaging in an activity” (Eccles et al., 1983, pp 89). The interest value is similar to the construct of intrinsic motivation in other theories (Wigfield & Eccles, 2002). Utility is about the importance of the task for some future goal, such as admission to a desired course of study or some other career objective (Eccles et al., 1983). It is related to the concept of extrinsic motivation in other theories (Eccles et al., 1983). The last component of value is cost, which is the negative aspect of engaging in a task (Eccles et al., 1983). The cost aspects can be divided into three main parts; loss of time that could be used to engage in other valued activities, the psychological cost of failure, and the effort needed to succeed (Eccles et al., 1983).
It may be worth emphasizing the difference between the effort needed to succeed and the effort invested in the test. The amount of effort that the student thinks will be necessary to get a good result is the effort needed for success, and contributes to the cost. Conversely, the effort the student actually puts into the test is what I call the effort invested (or simply effort), and is regarded as an outcome of the student’s expectation of success and the value they place on the task. Test anxiety has been identified as one of the most important affective reactions in the school context (Pintrich & Degroot, 1990); in recognition of this, Pintrich and Degroot (1990) argued that a “test anxiety” component should be included in expectancy value theory. In Eccles expectancy-value theory, anxiety can be seen as part of the psychological cost of failure (Eccles et al, 1983).

Both expectancies and values must be understood to predict students’ willingness to engage in a task. Even if a student finds a task interesting and valuable, she may not engage with it if she doesn’t expect to do well at it. Similarly, a student may be confident of doing well at a task but nevertheless choose not to engage with it if she doesn’t value it. While both factors are considered important for predicting students’ performance, persistence and choice, research has shown that task value is more closely related to activity selection whereas expectancy is a better predictor of performance (Wigfield & Eccles, 2002, Eccles et al. 1983).

3.1 Why the expectancy-value theory of achievement motivation?
As noted above, the expectancy-value theory of achievement motivation is just one of many established theories of motivation. There are three main reasons for its selection as the theoretical framework for this study. First, it is one of the more comprehensive models, including many individual characteristics and various contextual factors that affect motivation, enabling a thorough understanding of how motivational aspects interact towards a motivated behaviour. Second, the expectancy-value theory is the motivation theory that most explicitly operationalizes task value (Pintrich, 2003). To understand test-taking motivation, especially in low-stakes tests (i.e. tests with no consequences and presumably no value to the test-takers), it is essential to understand students’ perceptions of the test’s value. Third, most earlier studies on test-taking motivation also refer to the expectancy-value theory (see for example Barry, Horst, Finney, Brown, & Kopp, 2010; Cole, Bergin, & Whittaker, 2008; Finn, 2015; Hawthorne, Bol, Pribesh, & Suh, 2015; Penk, Pöhlman, & Roppelt, 2014; Sundre &
Kitsantas, 2004). Hopefully my work will complement the existing body of knowledge in this field.

4. Test-taking motivation

In this thesis, test-taking motivation is defined as “the willingness to work on test items and to invest effort and persistence in this undertaking” (Baumert & Demmrich, 2001, p. 441). Based on the expectancy-value theory (Wigfield & Eccles, 2000), the willingness to invest effort in a specific test situation is hypothesized to be an overall measure of test-taking motivation that depends on expectancies for success on the test (expectancies), the personal importance of doing well on the test (importance), the student’s inherent immediate enjoyment of taking the test (interest), the importance of the test for some future goals (utility) and the negative aspects of engaging in the test (cost). Test-taking motivation is assumed to be distinct from more domain-specific traits like motivation, and to be related to performance in the test situation (see Penk & Schipolowski, 2015). This chapter begins with a description of some assessment contexts where neglecting the impact of test-taking motivation could make it difficult to form valid interpretations of test scores, and discuss the potential consequences of such neglect. I then will describe previous research about test-taking motivation from different perspectives.

4.1 Assessment contexts where test-taking motivation may be relevant to consider

The Standards for Educational and Psychological Testing published by the American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) in 2014 (henceforth referred to as “the Standards”) state that test-takers’ levels of motivation should be considered when interpreting test results, especially when individual scores are not reported to test takers or the scores have no consequences for the test-takers (low-stakes tests).

International comparative tests such as the Trends in International Mathematics and Science Study (TIMSS) and Program for International Student Assessment (PISA) tests are two examples of low-stakes tests. Individual results for these tests are not reported to the students or their teachers. However, the results of these tests attract lots of national and international attention, and have prompted several educational reforms.

Globally, there are also national assessment programs at the K-12 level and in higher education that use regular low-stakes tests to hold schools
accountable for the quality of the education (cf. Liu, Rios, & Borden, 2015; Hawthorne et al., 2015; Sessoms & Finney, 2015). Here again, these test results have no direct consequences for the students but could affect the school’s reputation or resource allocation. If students do not put effort into these low-stakes tests, analysis of their results may produce biased conclusions about their knowledge. For example, if students in different educational systems/schools put different amounts of effort into the tests, comparisons of test performance between different educational systems/schools might be distorted and so any conclusions drawn about the systems’ relative effectiveness might also be biased (Eklöf, Japelj, & Grønmo, 2014). Alternatively, if the test results are used to identify students’ instructional needs, a failure to account for differences in motivation could cause students to be re-taught material they already know.

The assessments considered in this thesis were Swedish national tests and field trials for these tests. Field trials, where new test items are tried out in order to evaluate item quality, are usually low-stakes tests for the test-takers. Consequently, some or all of the students may lack motivation to do their best, which could lead to inaccurate conclusions about issues such as item difficulty and differential item functioning. The national tests are high-stakes tests, but even so they may be conducted under conditions that prevent some students from doing their best. For example, high test-anxiety or low expectancies could harm the performance of some students and thereby affect the validity of any interpretation of the test results (Wigfield & Eccles, 1992; Segool, Carlson, Goforth, von der Embse, & Barterian, 2013).

In sum, there are a number of different tests that are used for important decision where the validity of the interpretations of the test results might be affected by low test-taking motivation. To determine when this is likely to be the case and identify effective ways of handling such situations, it is necessary to have a good understanding of test-taking motivation, e.g. how it can be theoretically understood and measured, how it varies between different test situations and between individuals, and how it affects test performance.

---

2Differential item functioning is a statistical characteristic of an item that shows the extent to which the item might be measuring different abilities for members of separate subgroups.
4.2 Previous research on test-taking motivation

4.2.1 Measures of test-taking motivation

There are several different methods for measuring test-taking motivation, for example observations, self-reports measures, interviews, and objective measures based on students' behaviour, such as item response time (Wise & Smith, 2016). In the present thesis, questionnaires and interviews were selected as the preferred tools for data gathering. There are many questionnaires for measuring student motivation, characteristics, self-efficacy etc. on a general or domain specific level but only a few questionnaires have been developed specifically to measure test-taking motivation. In the school context two widely used questionnaires for measuring test-taking motivation appears to be the student opinion survey (SOS; Sundre, 2007; Thelk et al., 2009) and the on-line motivation questionnaire (OMQ) developed by Boekaerts (2002). However other test-taking motivation questionnaires can also be found in the literature (e.g. Baumert & Demmrich, 2001; Butler & Adams, 2007; Cole et al, 2008; Eklöf & Nyroos, 2013; Liu et al., 2012; O’Neil, Abedi, Miyoshi, & Mastergeorge, 2005; Penk & Schipolowski, 2015). All these questionnaires include one or more items measuring invested effort. Some of them also include items relating to expectancies (Boekaerts, 2002; O’Neil et al., 2005; Penk & Schipolowski, 2015). The value aspect is usually represented by one or more items relating to importance (e.g. Boekaerts, 2002; Eklöf & Nyroos, 2013; Thelk et al, 2009). Aspects of interest, utility and test anxiety are represented by one or a few items in some of the questionnaires (e.g. Boekaerts, 2002; Cole et al., 2008; O’Neil et al., 2005). Although several of the previous instruments claim to be based in expectancy-value theory, none of the described questionnaires include all the expectancy and value aspects and few analysed the relationship as proposed by the theory. The cost aspect is particularly poorly researched: to the best of my knowledge there are no published studies exploring aspects of cost other than negative emotions in the context of test-taking motivation. Further, few studies have compared test-taking motivation between test with different stakes after controlling for measurement invariance of the questionnaire. The few studies that have sought to evaluate whether a larger expectancy-value theory seems to be a valid theoretical framework for the test-taking motivation construct examined slightly different combinations of the theory’s aspects, and do not give a uniform picture of the relationships between effort, expectancies, task values, and performance (Cole, et al., 2008; Penk, et al., 2014, Penk & Schipolowski, 2015).
In sum, a number of previous studies have investigated test-taking motivation through self-report measures in the form of questionnaires. However most previous studies have used rather narrow measures of test-taking motivation and thereby leaving out many of the underlying processes causing a motivated or unmotivated state, and empirical explorations of how the broader test-taking motivation construct can be conceptualized and measured is largely missing.

4.2.2 Factors influencing test-taking motivation

Good knowledge about the dynamics of test-taking motivation is essential if one wants to predict when and why test-taking motivation might be a threat to the validity of the achievement test score and also to know how potential low test-taking motivation can be dealt with.

Previous studies have consistently shown that students tend to report higher effort, greater perceived importance, and higher levels of test anxiety for tests that affect their grades than for those that do not (Eklöf & Knekta, in press; Segool, et al., 2013; Sundre & Kistansas, 2004; Thelk et al., 2009; Wolf & Smith, 1995). In some cases but not consistently, raising the perceived stakes of a test by motivational framing (e.g. informing the students about the test’s potential consequences for their school or institution) or offering performance-based compensation (such as free food or financial rewards for each item answered correctly) have shown to increase the effort invested by students (Baumert & Demmrich, 2001; Finn, 2015; Liu et al., 2012). Other factors that could increase perceived relevance of the test and thereby affect test-taking motivation that have been discussed by researchers are to clearly present the usage and importance of the test and to give feedback about the test results to the students (Finn, 2015; Sessoms & Finney, 2015). Although it seems reasonable to assume that feedback would increase perceived utility and also effort there is little empirical research that have been able to prove that it actually does (Finn, 2015; Wise & DeMars, 2005). Attitudes expressed by the test proctors, teachers and other students can also affect students’ willingness or ability to give good effort (Lau, Jones, Andersson, & Markle, 2009; Putwain, Connors, Woods, & Nicholson, 2012; Wise & Smith, 2016).

Characteristics of the test and its items have also shown to affect test-taking motivation. Constructed-response items and tests in essay format have been associated with lower levels of motivation and effort than multiple-choice items (DeMars, 2000). Researchers have also found that the more difficult, time consuming, and/or cognitively demanding an item is, the less motivation test-takers have to invest effort into it (Barry, et al., 2010;

Finally, previous research has indicated that the effort invested into a test seems to depend on the characteristics of the test-taker. For example, students with high reported agreeableness, conscientiousness, and openness made a greater overall effort (Barry et al., 2010) than students with lower scores on these variables. In a recent review, DeMars, Bashkov, and Socha (2013) showed that the existing literature indicates that there are small but consistent gender differences in self-reported test-taking motivation for low-stakes tests, with females generally reporting higher effort. Boekaerts (2002) identified several gender differences in test-taking motivation. For example, boys scored more highly on beliefs about how well they had done and on measures of emotional state (e.g. ease and being more enthusiastic).

In sum, a number of different factors seem to influence test-taking motivation and several methods for increasing test-taking motivation have been suggested in the recent years. Still, better understanding of test-taking motivation in different test contexts and student populations is needed. For example in the Swedish context, we know very little about test-taking motivation and factors influencing it.

4.2.3 Test-taking motivation and test performance

Even if test-taking motivation would differ between tests and groups of students, serious effects on the validity of the interpretation and use of test results are only likely if test-taking motivation affects performance. A review of a diverse set of studies revealed differences in test scores between equal tests with different stakes, and a correlation between motivation and performance (Wise & DeMars, 2005); the average effect size for differences in test scores between motivated and unmotivated groups across 12 studies was 0.59. The correlation between reported motivation (usually quantified in terms of effort) and test results ranged between 0.23 and 0.38. More recent studies have arrived at similar findings (Abdelfattah, 2010; Asseburg & Frey, 2013; Cole & Osterlind, 2008; Eklöf, et al., 2014; Eklöf & Knekta, in press; Liu et al, 2012). Some studies found that effort/persistence predicted test performance as strongly as (or even more strongly than) credits (Cole et al., 2008) and content knowledge (Boe, May, & Boruch, 2002). Eklöf and Knekta (2017) showed that in the PISA 2012 context, effort could explain variance in achievement test results, although it was modelled together with domain specific variables such as math interest, math efficacy, and math anxiety. The relationship between effort and test performance is not
necessarily simple and linear. Sundre and Kitsantas (2004) showed that test-taking motivation did predict student performance in low-stakes tests but not in high-stakes tests. Attila (2016) found that effort had an effect on performance, but this effect disappeared if students who invested very little effort (20% of the students) were excluded, suggesting a logarithmic relationship. There are also studies concluding that test-taking motivation has no or negligible effect on achievement test results (Eklöf, 2006; O’Neil et al. 2005). Thus, test-taking motivation seems to affect students’ test performance, at least in some contexts. When and how much it affects the test results still remains somewhat unclear. As test-taking motivation mostly is measured as effort only, there is little knowledge about how the other aspects affect performance.

To summarize the review above, several empirical studies have largely supported the core assumptions about test-taking motivation and its effect on test results. However, the current theoretical understanding of test-taking motivation remains unsatisfying. Moreover, as motivation and attitudes to schoolwork can be assumed to vary over ages, subjects, and cultures (Dotterer, McHale, & Crouter, 2009; Wigfield, Tonks, & Lutz Klauda, 2009), further empirical investigations into test-taking motivation in ecologically valid and varied achievement test situations are needed. My research has focused on the Swedish national tests, an assessment context in which we so far know little about student test-taking motivation and how it may affect the validity of interpretation and use of test scores.

5. The Swedish national assessment system

Children resident in Sweden are required to complete nine years of compulsory education, starting when they are seven years old. All youths who complete these compulsory studies are entitled to three years of upper secondary school education (grades 10 to 12; students are typically 16–18 years old), during which they can choose to follow one of 12 different vocational programmes or 6 different higher education preparatory programmes. Today, most youths in Sweden attend upper secondary school. The Swedish school system has a goal (criterion)-referenced system for grading, so grades are awarded based on predetermined achievement goals rather than relative to other students’ results. Students’ work starts being graded at the end of the sixth year of compulsory schooling. Grades from the final year of compulsory education are used as selection instruments for
upper secondary schools, and grades from upper secondary schools are used to determine eligibility for higher education. Grades and Swedish scholastic aptitude test results are the two most important selection instruments for admission to higher education.

One part of the Swedish assessment system is the national tests. These tests are administrated in grade 3 (Mathematics, Swedish), grade 6 (Mathematics, Swedish, English) and grade 9 (Mathematics, Swedish, English, Science, and Social Science) during compulsory schooling. There are also national tests in English, Swedish, and Mathematics for all upper secondary school programmes. The national tests are mandatory for all students participating in regular schooling; they are intended to help teachers assess students’ knowledge and to provide information on the extent to which curricular goals have been reached. National test scores are supportive of students’ grades rather than decisive, so one could argue that they are not high-stakes tests in the strictest sense. However, they are the highest-stakes tests in Swedish compulsory and upper secondary education: they are the only external tests used on a regular basis, and there is a high level of agreement between national test scores and final course grades (schools that do not achieve such agreement are criticized; Swedish National Agency for Education, 2014a). Teachers correct the national tests themselves, and cut-off scores for the different grades are established before the tests are released. The test’s difficulty is thus estimated before it is released. The national tests are taken on a single day in an auditorium or in several classrooms. They are usually proctored by teachers, but not necessarily by one of the students’ regular teachers. During the last decade educational reforms have led to national tests in more subjects and grade levels than ever before and there has been a lively debate in Sweden about the national tests, focusing on questions such as what subjects should be tested and how many national tests there should be (Eklöf, 2017).

To ensure high quality of the national tests, item quality (in terms of, item scoring rubrics, group differences, item difficulty, discrimination, and students’ understanding of the test items) is evaluated by performing try-outs with the target population. These item try-outs are administered and implemented in a separate process through so-called field trials. Randomly selected schools are invited to participate in field trials of sets of test items, and participation is voluntary. The field trials are usually administered by the students’ regular teachers during lecture time in a classroom. The teachers are asked to administer the field trial under test-like conditions and to urge the students to give their best effort to the test. However, the test developers’ experiences suggest that field trials often are performed under
conditions that are considered to be lower stakes than the regular national test (personal communication). Although the field trials are low-stakes tests from the test-takers’ perspective, it is important for the education system as a whole that the data from the field trials are reliable.

5.1 Research on students’ perception of the national tests

Although the national tests are important in the educational system and students must complete at least 10 national tests during their schooling, there has been relatively little research on how students perceive and react to these tests. There is some research on the students’ perspectives on national tests in grade 3 and upper secondary school available (Bagger 2015; Korp, 2006; Nyroos & Wiklund-Hörnqvist, 2011; Silfver, Sjöberg, & Bagger, 2015). The work underlying this thesis builds on two pilot studies conducted in 2009, when the year nine national tests in science were first introduced\(^3\), and in 2010\(^4\) (Åström & Eklöf, 2009; Åström & Eklöf 2011, Eklöf & Nyroos; 2013). The Swedish National Agency for Education (2014b) has performed some survey studies primarily focusing on teachers’ perception of the national tests but also looking at students’ perceptions. A deeper understanding of grade nine students’ perception of the national tests in general and in terms of test-taking motivation is however missing. Students’ test-taking motivation at field trials for national tests or at regular national tests in the upper secondary school have never been examined.

According to the Swedish National Agency for Education (2014b), teachers and students have positive overall opinions about the national tests. In studies on students’ perceptions of the Swedish national test in science, many students reported that they invest effort into the national test and consider it to be important (Eklöf & Nyroos, 2013; Åström & Eklöf 2011). Despite finding the test important, several students also reported that they considered it unnecessary (Åström & Eklöf 2011). A recent governmental investigation found that the role of the national tests in grading is unclear to both teachers and students (SOU 2016:25). Studies on national tests in grade three and upper secondary school have arrived at similar conclusions (Bagger, 2015; Korp, 2005). Most studies in the Swedish national test context have indicated that some but not all students experience test anxiety during the national tests (Eklöf & Nyroos, 2013; Nyroos & Wiklund-Hörnqvist, 2011; Sjöberg, 2006; Swedish National Agency for Education, 2014b).

---

\(^3\) on this occasion, test participation was voluntary
\(^4\) the first year that test participation was compulsory
6. Validity theory

The quality of measurements is important in all sciences and is an issue at the very core of this thesis. Although different terms are used in different disciplines, the principles and problems are the same in all cases. In physics, the terms “accuracy” and “precision” are commonly used, whereas the terms “trustworthiness” and “transparency” are used to discuss the validity of interview studies. In educational and psychological testing, “reliability” and “validity” are key concepts in the quality discourse. Briefly, reliability refers to the consistency of a measurement. For example, will the measured difficulty of a test item be the same across different measurement occasions and regardless of who made the interpretations? Simplified, validity relates to whether a test measures what it is supposed to measure. For example, do achievement test scores accurately reflect the knowledge of interest without being systematically affected by other factors? This chapter begins with a brief historic overview of the concept of validity, which is followed by some reflections on validity and validation in the context of this thesis.

6.1 Views of validity over time

The debate about how to define the concept of validity and what it should encompass has been going on for a long time. Originally, validity was primarily a narrow empirical concept based on the correlation between a measurement (e.g. a job selection test or achievement test score), and a criterion of interest (e.g. job performance or success in a later course). Guilford (1946) stated that:

In a very general sense, a test is valid for anything with which it correlates. (p. 429)

Validity was seen as a binary property of the test, and what we now call criterion-related validity was the essential measure of quality. However, it can be quite hard to find a plausible criterion that is better than the assessment itself and at some point the criterion itself must be validated (Kane, 2006). Furthermore, focusing solely on correlations with external criteria can cause the test's content (i.e. the thing the test actually measures) to be ignored. Consequently, the concept of content validity became increasingly prominent in the mid-twentieth century (Kane, 2006). Content validity concerns the extent to which test items are a representative of the whole domain of interest, addressing questions such as “does this mathematics test reflect the relevant parts of the mathematics curriculum?”
However, many tests (e.g. personality tests and attitude inventories) lack both criteria for prediction and content domains to sample (Kane, 2006). To address this deficiency, the term “construct validity” was introduced in the nineteen fifties (Cronbach & Mehl, 1955). Construct validity referred to the alignment between the measurement model and the theoretical construct, and was to be used whenever a test was to be interpreted as a measure of some attribute or quality that is not “operationally defined” (Cronbach & Mehl, 1955). This traditional view of validity referring to the innate properties of the test and that encompasses criterion-related, content and construct validity was dominating for a long time.

In 1989, Samuel Messick presented a new and very influential broader view on validity (Messick, 1989). He argued that earlier views were too narrow and that validity should include not only the instrument but also the use of the scores from the instrument and its consequences. According to Messick (1989) validity is a unified concept in contrast to the previous view with three separate validities. The unifying force is construct validity, which subsumes both criterion-related and content validity, as well as reliability. The main sources of invalidity mentioned by Messick (1995) are construct underrepresentation (which occurs when the assessment is too narrow and fails to include important dimensions) and construct-irrelevant variance (which occurs if the test score is influenced by irrelevant factors). These two sources of invalidity can cause unintended social consequences when a test is used, and Messick argues that these consequences must be considered when evaluating construct validity. Another influential view on validity was presented by Kane (1992), whose approach closely resembles Messick’s but provides a more detailed description of the validation procedure and is more inclusive with respect to social consequences. Since Messick and Kane presented their views on validity, some researchers have argued that the current view of validity is too complex (e.g. Borsboom, Mellenbergh, & van Heerden, 2004; Lissitz & Samuelsen, 2007; Mehrens, 1997). Nevertheless there is today a broad consensus that validity is a unitary concept and that validation is ongoing evaluation of whether the proposed interpretation and use of the test is supported (Cizek, 2016; Kane, 2016). The definition of validity given in the latest Standards (AERA, APA, & NCME, 2014) is:

Validity refers to the degree to which evidence and theory support the interpretations of the test scores for the proposed use of the tests. (p. 11)

Thus validity is not a property of the test, but rather refers to its proposed interpretation and use. Validation refers to the process that “involves
accumulating evidence to provide a sound scientific basis for the proposed score interpretation” (AERA, APA, & NCME, 2014, p. 9). The validity evidence needed in any given case depends on the proposed interpretation and use of the test’s scores. For example, the validation of a test as a reasonable measure of sixth grade mathematical achievement does not automatically validate the test’s use for placement and promotion decisions (Kane, 2016). According to the latest edition of the Standards (AERA, APA, & NCME, 2014), sources of validity evidence include 1) Evidence based on test content, i.e. analyses of the relationship between a test’s content and the construct it is intended to measure; (2) Evidence based on response processes, i.e. information on how test-takers answer the test’s questions; (3) Evidence based on internal structure, i.e. analyses of internal relationships between test items and test components, and how they conform to the intended construct; (4) Evidence based on relations to other variables, i.e. analyses of the relationships of test scores to variables external to the test, and to other tests that measure the same construct or related constructs; and (5) Evidence based on consequences of testing, evidence about consequences that can be traced to sources of invalidity such as construct underrepresentation and construct-irrelevant components are relevant as well as evidence that support proposed use.

6.2 Validity and validation in the thesis
Validity is a central concept in this thesis, and one that is important at multiple levels of my work. As discussed in chapter four, one of the rationales behind the work presented in this thesis is the possible effect of test-taking motivation on the validity of the interpretations of test scores and what the consequences of invalid interpretations may be. Additionally, all the included studies concern the latent construct of test-taking motivation and finding valid ways to theoretically understand it, operationalize it, measure it, and model it, and different validity evidences were collected in the different studies.

The definition of validity used throughout the work presented in this thesis is that of the Standards from 2014 (AERA, APA, & NCME, 2014). To me, the use of a unified validity concept clarifies the point that validity is an overall quality indicator and that one must consider several aspects during the validation process. Moreover, I like the Standards’ emphasis on validating interpretations based on proposed use: I find it hard to detach a test’s validity from its proposed uses. A measurement always has a purpose, and is always intended to be used in some way. However, this use may be narrow (e.g. measuring motivation for research purposes such as
understanding students’ test-taking motivation in different test situations) or broad (e.g. using motivation scores to correct students’ scores on international comparative studies). These two goals do not necessarily require different amounts or qualities of validity evidence, but they clearly do require different kinds of validity evidence. Many potential sources of validity evidence are described in the Standards, and several studies utilizing different approaches, different samples, and different populations may be required to support the validity of a given interpretation of the scores from an instrument. The Standards also state that considering the vast number of tests and circumstances that exist in the world, some types of evidence will almost certainly be quite important in any given case while others will be less useful (AERA, APA, & NCME, 2014).

In this thesis I used the demonstration of a strong program for validation provided by Benson (1998) as a guide when validating the measures of test-taking motivation, and the test-taking motivation questionnaire in particular. The strong program for validation is based on Messick’s view on validity and corresponds largely to the latest Standards. The program places great emphasis on the affiliation between theory and the empirical model and is, as I see it, well adapted for measuring personal traits. In her article, Benson (1998) describes three stages of the validation process; the substantive stage, the structural stage and the external stage. The substantive stage concerns the definition of the theoretical and empirical domains of the construct and includes processes such as gathering content-related evidence and generating theoretical and empirical definitions. The theoretical domain represents our best understanding of the construct and evolves from theory, previous research and the researcher’s own values and observations. The empirical domain relates to how the construct is operationalized, and contains all possible observed variables and ways in which these variables can be measured. The empirical domain should reflect the theoretical domain, and the theoretical domain should be continuously revised as new knowledge is obtained through empirical studies. The substantive stage corresponds closely to the analysis of evidence based on content and responses as described in the current Standards (AERA, APA, & NCME, 2014).

The structural stage concerns internal relations among observed variables and includes factor analysis and generalizability theory among other things. The aim at this stage is to determine the extent to which the observed variables covary among themselves and how they covary with the intended structure of the theoretical domain. Studies of internal structure can provide information on issues such as the construct’s dimensionality but do not
provide evidence on what exactly is being measured. The structural stage corresponds closely to the analysis of evidence based on internal structure as defined in the current Standards (AERA, APA, & NCME, 2014).

The external stage of construct validation concerns the relations of the construct of interest to other constructs. External validity evidence can be obtained through analysis of assumed group differences, assumed effects of external manipulations, and correlations to other measures of the construct. The external stage corresponds closely to the analysis of evidence based on relations to other variables in the current Standards (AERA, APA, & NCME, 2014). Benson considers evidence-based consequences of testing in terms of consequences that can be traced to sources of invalidity such as construct underrepresentation and construct irrelevant components. In the studies presented in this thesis, I followed this approach but also considered evidence about consequences that support proposed uses, of in my case the Swedish national tests.

7. Method and methodological choices

The research presented in this thesis is built upon a pragmatic worldview (Creswell, 2014). Unlike positivism and constructivism, pragmatism is not committed to any one system of philosophy and reality. Instead of trying to find an objective reality (as in positivism) or meaning (as in constructivism), pragmatism seeks practical solutions to practical problems. Pragmatism does not focus specifically on any one method; instead, the focus is on the research problem, which is to be understood using pluralistic methods. The focus of the research in this thesis is on test-taking motivation and related validity issues. Questionnaires and interviews were used to collect quantitative and qualitative data to obtain a better understanding of test-taking motivation. According to the definitions given by Johnson, Onwuegbuzie, & Turner (2007), the research approach used can be labelled as quantitative dominant mixed methods research because it includes qualitative data, but the emphasis is on the quantitative data. Mixed methods approaches that combine elements of qualitative and quantitative research are increasingly being used in educational research because they combine the strengths of both approaches, providing a broad and deep understanding of the studied phenomenon (Johnson & Onwuegbuzie, 2004; Johnson et al., 2007). Despite their many benefits, mixed methods approaches have not been widely used in test-taking motivation research. In this chapter I will first
describe the overall research design for the four studies included in the thesis from a validity perspective. Then the data collection, participants, instruments and analysis will be described in more detail.

7.1. **Research design from a validity perspective**

The overall research design for the work underlying this thesis is illustrated in Figure 2. One of the work’s main aims was to explore how the test-taking motivation construct can be theoretically understood, operationalized, measured and modelled, and an important step towards achieving this aim was to develop, use and evaluate a questionnaire. Guided by Benson’s demonstration of a strong program for validation (presented in chapter 6), I first reviewed the literature on motivation theories, with particular emphasis on the expectancy-value theory, and then performed a separate review of previous studies on test-taking motivation and instruments used to measure test-taking motivation. Based on the knowledge obtained about the theoretical and empirical domains of test-taking motivation, a questionnaire was developed with the aim of achieving a good affiliation between theory and the empirical measurement. The questionnaire was administered to a sample of students taking a low-stakes and a high-stakes test. Structural equation modelling was applied to the collected data to collect validity evidence on the internal relationships between test items and components. Additionally, to determine whether the relationships among the variables were the same for high- and low-stakes tests, an invariance analysis was performed. In parallel with the administration and analysis of the questionnaire, cognitive interviews and semi-structured interviews were designed based on the literature review and performed with a smaller sample of students. Results from the interviews were used to obtain a deeper understanding of the test-takers’ test-taking motivation and to collect validity evidence based on response processes for the questionnaire. Based on the analysis of the resulting data, a revised questionnaire was developed and distributed to a new sample of students taking low- and high-stakes tests. To evaluate the relationship between test-taking motivation and other constructs, I related test-taking motivation to test performance, compared test-taking motivation between different tests and different classes, and compared the results from the test-taking motivation questionnaire to transcripts of interviews with the students.
Figure 2. Illustration of the overall design of the work underlying the thesis.
7.2 Data collections and participants
The work presented in this thesis is based on two main data collections. The first data collection was made in spring 2012 in connection to a national test in science for grade nine students (15-16 years old) and a field trial for this test. In total, 1422 students at 47 schools completed the questionnaire during the field trial, and 536 students at 17 schools completed the questionnaire after taking the regular national test in science. Twelve students from two of the schools were recruited for in person interviews. The interviews were conducted with one student at a time, one and six days after taking the national test, and lasted between 20 and 45 minutes.

The second data collection was made in spring 2014 in connection to a national test in mathematics for upper secondary students and a field trial for that test. In total, 222 students from 8 schools completed the questionnaire after the field trial, 62 students from 4 schools completed the questionnaire again after the regular national test. All the students were in grade 11, between 17 and 18 years of age, and studying the upper secondary economics, aesthetics, humanities, or social science programs.

During both data collections, the questionnaire was first administered to all the schools that agreed to participate in the relevant field trial, i.e. the total population of students participating in each field trial was sampled. The same schools were then asked to administer the test-taking motivation questionnaire again in connection to the regular national test, one to three weeks later. The largest dropouts during both data collections were due to schools and teachers that were asked to participate in the field trials but chose not to, and teachers choosing to participate during the field trial but not the regular national test. In contrast, non-response rates from students or for individual items were relatively low.

7.3 Instruments
The data underlying the work presented in this thesis were collected using three main instruments, the test-taking motivation questionnaire, the interview guide and national tests. These instruments are briefly described below.

7.3.1 The test-taking motivation questionnaires
The aim of the questionnaires was to measure students’ test-taking motivation in different test situations. The results were primarily intended to be useful for basic research and understanding the test-taking motivation construct, but they could also provide useful information to guide the process of developing and administering achievement tests. The target group for the questionnaire
was Swedish students in secondary school and upper secondary school aged 13 to 18. The questionnaire was based on the expectancy-value theory of achievement motivation and was designed to cover effort\textsuperscript{5}, expectancies, and different aspects of task value. Despite being comprehensive, the questionnaire still had to be short enough for students to be able to complete it immediately after taking a test, and to be easily administered in different test situations. The questionnaire was developed, evaluated and refined through an iterative process involving literature reviews, psychometric evaluation, and interviews among other things (Figure 2).

The first version of the questionnaire included 19 items in total and aimed to cover the latent constructs effort, expectancies, importance, interest, and cost (Table 1; Study I). The cost aspect was conceptualized as test anxiety. All items were rated on a 4-point Likert scale ranging from 1 (\textit{strongly disagree}) to 4 (\textit{strongly agree}). The quality of a measurement can always be improved and before the second data collection the first questionnaire was revised based on the interpretations of the findings from psychometric evaluation, discussions in an expert group and cognitive interviews\textsuperscript{6}. The revised questionnaire included 22 items and consisted of seven brief subscales representing six aspects of the expectancy-value theory, effort, expectancies, importance, interest, utility, and cost. (Table 1; Study IV). The cost aspect was conceptualized as both test anxiety and a more general cost aspect. The test anxiety items were modified to only reflect one dimension of test anxiety i.e. cognitive interference (“the degree to which test anxiety interferes with organizing one’s thought, attending, concentration and remembering before and during an examination”; Lowe, 2014, p. 407).

\textsuperscript{5} It is worth pointing out that, the subscale concerning the outcome of the expectancy and value aspect (which I called “effort”, in keeping with many other studies) relates to both effort invested and to feeling motivated. It is possible for a student to feel motivated but nevertheless not need to or be unable to put effort into a test for some reason. It might therefore be more accurate to call this scale “Motivation.”

\textsuperscript{6} The results from the cognitive interviews were presented at the AEA conference (Knekta, 2012), but are not included in any of the studies presented in this thesis. In brief, during the cognitive interviews the students pointed out some sources of error that could introduce construct-irrelevant variation to the items’ scores. These errors could be grouped into six main themes; (1) Uncertainty about whether the items’ reference period should include the time before the test or be limited to the test’s duration, (2) Uncertainty about what they should compare their experiences to – all tests they had previously taken in school or just tests in the subject in question. (3) Having insufficient knowledge answer some items – for example, the students didn’t know whether it was important to their teacher for them to get good results. (4) Problems with interpreting the item or terms used in the items – for example, “what does the term ‘relaxed’ mean?” (5) Some items were considered to address the same topic (i.e. appeared redundant to the students), such as items Im1 and Im2, and (6) Incorrect assumptions such as assuming that the students knew about the test before taking it.
Table 1. Items included in the first and second versions of the test-taking motivation questionnaire. Deleted and new items are indicated by empty rows in one column or the other

<table>
<thead>
<tr>
<th>Version 1</th>
<th>Version 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1 I did my best on this test</td>
<td>E1 I did my best on this test</td>
</tr>
<tr>
<td>E2 I worked with all items in the test without giving up, even when an item was difficult</td>
<td>E2 I worked with all items in the test without giving up, even when an item was difficult</td>
</tr>
<tr>
<td>E3 I felt motivated to do my best on this test</td>
<td>E4 I felt motivated to do my best on this test</td>
</tr>
<tr>
<td>E4 I spent more effort on this test than I do on other tests we have in school a</td>
<td></td>
</tr>
<tr>
<td>Ex1 I did well on this testb</td>
<td>Ex1 I think I did well on this test</td>
</tr>
<tr>
<td>Ex2 This was a difficult test a</td>
<td>Ex2 This was about things that I know</td>
</tr>
<tr>
<td>Ex3 Compared to other students I think I did well on this testb</td>
<td>Ex3 Compared with other students in my class, I think I did well on this test</td>
</tr>
<tr>
<td>Im1 This was an important test to me</td>
<td>Im1 This was an important test to me</td>
</tr>
<tr>
<td>Im2 It was important to me to get a good result on this test</td>
<td>Im2 It was important to me to get a good result on this test</td>
</tr>
<tr>
<td>Im3 I am very curious about the result I received on this test a</td>
<td>Im3 This test was more important than other tests we have had in mathematics</td>
</tr>
<tr>
<td>Im4 It is important for my teacher that I get good results on this test</td>
<td></td>
</tr>
<tr>
<td>In1 I looked forward to doing this test</td>
<td>In1 I liked taking this test</td>
</tr>
<tr>
<td>In2 It was fun to take this test</td>
<td>In2 It was fun to take this test</td>
</tr>
<tr>
<td>In3 I learned something new by doing this testc</td>
<td>In3 The questions in this test were interesting</td>
</tr>
<tr>
<td>T1 Before taking this test, I worried about how difficult it would be</td>
<td>T1 While taking this test, I thought about how badly I was doing a</td>
</tr>
<tr>
<td>T2 I was scared of failing on this testb</td>
<td>T2 I was so scared of failing this test that I had a hard time focusing my thoughts</td>
</tr>
<tr>
<td>T3 I felt relaxed while taking this test a</td>
<td>T3 I was so nervous when taking this test that I forgot things I actually know</td>
</tr>
<tr>
<td>T4 I was so nervous when I took this test that I forgot things I usually know</td>
<td></td>
</tr>
</tbody>
</table>

Note: E = Effort, Ex = Expectancies, Im = Importance, In = Interest, U = Utility, C = Cost, and T = Test Anxiety. a. Items excluded after psychometric analysis. b Items that have been modified between version 1 and 2. c. Item was reclassified from interest to utility.
7.3.2 Interview guide
The interviews consisted of a semi-structured interview part and a cognitive interview part (Kvale & Brinkmann, 2014; Willis, 2005). The aim of the semi-structured part was to examine grade nine students’ perceptions of the Swedish national science test in terms of test-taking motivation, and to explore students’ beliefs about different aspects of test-taking motivation, which is usually measured in large-scale studies, in more detail. The interview guide for the semi-structured part was designed to cover different aspects of the expectancy-value theory as conceptualized in the test-taking motivation questionnaire (Appendix 1; Study III). The main focus of the interview was on the regular national test, but the students were continuously asked to also reflect on the field trial. The aim of the cognitive interview was primarily to get a better understanding of how the students thought when answering the test-taking motivation questionnaire, i.e. to gather validity evidence relating to their response processes. For each item, the interviewer read the questionnaire item aloud and the students were asked to explain their thoughts when reading the statement (Appendix 1).

7.3.3 The achievement tests

The national test in science for grade nine students
The first data collection related to the Swedish national test in science for students in grade nine that was administered in 2012. Each school was assigned a test in one of the subjects Biology, Physics, or Chemistry. The tests in each subject were based on the same test model and were designed to assess the same general skills and to have the same degree of difficulty. The national tests in 2012 included 15-17 items, awarding a maximum of 39 points. About a fifth of the subitems were multiple choice or matching items, another fifth were short answer items, and three fifths were longer constructed response items. The coefficient alpha values for the tests were .87 for Biology, .90 for Physics and .89 for Chemistry. The students had 210 minutes to complete each test. The field trial consisted of ten different sets of items with between four and seven items each. The diversity of the field trial tests means that the analysis of the correlations between test-taking motivation and effort should be interpreted with care.

The national test in mathematics for upper secondary school students
The second data collection related to the national test in mathematics for upper secondary school students in the economics, aesthetics, humanities, or the social science programs that was administered in spring 2014. The test
included 23 items, awarding a maximum of 57 points (17 points for short answers and 40 points for long answers). The coefficient alpha for the whole test was .90. The students had 240 minutes to complete the test. The field trials consisted of four different sets of items with between 5 and 7 items.

7.4 Statistical Analysis
During both the structural and external stages of validation, multiple statistical analyses were applied to the collected empirical data. During the structural stage, structural equation modelling was used to determine whether the empirical data covaried with the intendent structure of the empirical domain and whether the questionnaire could be validly used to compare test-taking motivation between low- and high-stakes tests (Studies I & IV). During the external stage, second order latent growth modelling was used to analyse possible differences in test-taking motivation between test stakes, gender, and subject, and to identify potential relationships between test-taking motivation and achievement test results (Study II).

This section provides a brief overview of the principles of latent variable modelling, which formed the basis of much of my work. This is followed by description of the main statistical methods used.

7.4.1 Latent variable modelling
Like many psychological traits, test-taking motivation is a so-called latent construct that cannot be directly observed (unlike, say, height) and must therefore be inferred from other observable variables (Schumacker & Lomax, 2010). The idea behind latent variable modelling is that the latent value (e.g. students’ true perception of the importance of a test) causes the students to respond as they did on the questionnaire items. Mathematically, the value of the latent variable is estimated by analysing the common variance among the questionnaire items; the shared variance is assumed to represent the latent variable, while the non-shared variance is treated as a random error. The information from the observed variables (e.g. the three importance-related items) is thus reduced to a single common factor (e.g. “importance”) that is free of measurement errors.

7.4.2 Structural equation modelling
Structural equation modelling (SEM) and exploratory structural equation modelling (ESEM) were used to determine the extent to which the collected data support the theoretical assumption of the test-taking motivation construct described by the expectancy-value model (Studies I & IV).
In brief, when applying a SEM one must first specify the number of factors (in my case, effort, expectancies, importance etc.), their relationships to one- another, and which items are supposed to measure which factor. In study I, a five factor SEM with expectancies, importance, interest, and test anxiety as predictors of effort was specified. This theoretical model was then applied to the collected data and the maximum shared variance among items (e.g. the four effort items) that were supposed to measure the same underlying factor (e.g. effort) was extracted. 7 Simultaneously, the shared variance was analysed to describe the relationships between the latent factors (e.g. effort and importance) and between latent factors and observed variables. Structure and pattern coefficients, communalities, residuals and modification indices were then examined and multiple fit indices were consulted to evaluate to what extent the theoretical model was supported by the sampled data (Schermelleh-Engel, Moosbrugger, & Müller, 2003).

SEM models are frequently used in a confirmatory fashion as described above, i.e. the analysis is based on the theoretical model and the aim is to investigate the degree to which the empirical data matches the specified model. Confirmatory models are suitable when the theoretical domain of the construct is relatively well known and the relations between items and latent factors can thus be specified by the researchers. However, when the theoretical domain of the construct of interest is not well understood or during the early stages of instrument development, an exploratory approach might be more suitable. ESEM is a relatively new and more flexible alternative approach for evaluating factor structures that combines exploratory analysis with common structural equation modelling (Asparouhov & Muthen, 2009; Marsh, Morin, Parker, & Kaur, 2014). It offers a more direct way of discovering mis-specified loadings than confirmatory SEM. Moreover, a common basic assumption in SEM analyses is that that each item loads solely on one factor. For example, the importance items are assumed to load only on the latent variable importance. However, measures of multidimensional constructs often have many small cross- loadings that are well motivated by either substantive theory or the formulation of the measurements (Asparouhov & Muthen, 2009). In ESEM it is possible to allow items to load on multiple subscale, which can yield a better model fit and unbiased factor correlations (Asparouhov & Muthen, 2009; Marsh, et al., 2014). Because test-taking motivation is a multidimensional construct and my revised questionnaire included new

---

7 This is done simultaneously for the whole model with the aim of minimizing the remaining total error variance.
untested items and subscales, I chose to apply exploratory structural
equation modelling (ESEM) when evaluating my revised questionnaire
(Study IV).

Measurement invariance
To be able to validly compare students’ test-taking motivation between
low- and high-stakes situations, the scores from the questionnaire should
represent the same thing in both test situations, i.e. the questionnaire
should be invariant (Bashkov & Finney, 2013; Vandenberg & Lance, 2000).
For example, each item should be equally representative of the constructs
(e.g. importance) in both test situations, and the relationships between
effort and the other latent variables should be identical across test
situations. To determine whether the test-taking motivation construct (as
conceptualized in my questionnaire) really is the same in the low- and
high-stakes settings, measurement invariance was analysed using data
from the first data collection (Study I). This was done by specifying a
multigroup SEM where each test situation represents one group. Then,
different parameters of the model were constrained to be equal across the
groups in a stepwise fashion, and differences in model fit between the more
and less constrained models were tested.

7.4.3 Two level second order latent growth modelling
In study II, a two-level second order latent growth modelling approach was
used to analyse differences in test-taking motivation between the field trial
and the national test and variance in test-taking motivation between school-
classes. Also differences in test-taking motivation between students taking
tests in different subjects (Physics, Biology and Chemistry) as well as
between genders were analysed. I chose to use a multilevel approach because
the data consisted of observations nested within individuals and individuals
nested within classes. I could not assume homogenous variance over the two
time points neither over the different classes. Students in the same class are
instead expected to influence each other and thus to be more highly
correlated than students in other classes, and the variance was expected to
be higher in the field trial than the regular test. Therefore, the basic
assumptions required for paired sample t-tests and repeated ANOVA are
violated, and using any of these methods to estimate differences between the
field trial and regular test or between schools would give downward-biased
standard errors (Hox, 2010). Multilevel modelling also makes it possible to
include variables at several levels, (e.g. gender at the individual level and
different subjects at different time points and for different classes).
Consequently, the data do not need to be aggregated or disaggregated, which could lead to loss of power, high operational alpha levels or interpretation of effects at the wrong level (Hox, 2010).

Second order latent growth modelling was used to estimate changes over time (Duncan & Duncan, 2009; Newsom, 2002; Voekle, 2007). Latent growth models were chosen because they can handle heterogeneous error structures and the growth curve can be incorporated into more complex models (Newsom, 2002). For example, latent constructs can be defined by multiple items so as to account for measurement errors (Newsom, 2002). Latent growth models based on only two time points (dyadic data/two-wave data) are not optimal because in such cases the approach is limited to analysis of straight lines and includes few degrees of freedoms (Duncan & Duncan, 2009; Voekle, 2007). However, with the addition of some restrictions it can be used to analyse pre- and post-test differences (Newsom, 2002; Voekle, 2007). In fact, the paired sample t-test can be seen as a special case of latent growth modelling (Voekle, 2007).

The second level sample size was small ($n = 26$), and small second level units can lead to underestimation of group level variance and their standard errors (Maas & Hox, 2004). It follows that the differences between school classes might be larger than estimated by the model, and at the same time we run the risk of erroneously concluding that observed differences are significant. However, the standard errors would also have been underestimated if we ignored the multilevel structure of the data.

### 7.5 Thematic analysis

In study III, the interview transcripts were subjected to a thematic analysis to explore the grade nine students’ test-taking motivation during the field trial and the national test in more depth. Thematic analysis is a method for identifying and interpreting underlying patterns in data (Braun & Clarke, 2006). Because the students’ responses in the cognitive interview also turned out to provide valuable information about students’ test-taking motivation, transcripts from both the semi-structured interviews and the cognitive interviews were included in the thematic analysis.

Verbatim transcripts from the interviews were read carefully to search for themes in the data. The search for themes was primarily conducted using an inductive approach in which codes and themes were developed empirically from the data. Coding focused on the semantic content of the interview excerpts, i.e. the explicit meanings of the data. The reliability of the coding was assessed by comparing the codings produced by two researchers for a sample of four transcripts. The codes were then organized
into groups of similar codes, which in turn were organized into sub-themes and themes. This was first done independently by two researchers, resulting in two thematic maps that were then discussed and compared to produce a common thematisation. Thereafter, codes and interview excerpts were reread to validate the themes. This resulted in further refinement of the themes.

7.6 Ethical Considerations
The ethical research principles described by the Swedish research council were upheld at all stages during this work (Forskningsrådet, 1996). Special attention was paid to; i) information about the research project and use of data, ii) confidentiality and iii) consent. Informed consent was obtained from the students completing the questionnaire and the students participating in the interviews. The students were informed that participating was voluntary, that the data collected were to be treated confidentially, and data would be presented in such a way that individual respondents could not be identified. The aim of the research project and the use of the data were explained. The students completing the questionnaires were given written information in a cover letter and the interviewees received written information before the interview and oral information during the interview. During the interviews, special consideration was paid to the fact that the interviewees were young people as specified by Leeuw, Borgers, and Smits (2004). Interviews were performed in a separate room, their aim and structure were clearly explained, and the point that there were no right or wrong answers was emphasised. The interviewees gave their consent for the interviews to be recorded.

8. The Studies, overview
This chapter summarizes the four studies that form the basis of this thesis. Studies I, II and III are all based on the same data collection, which relates to the test-taking motivation of grade nine students for the national test in science and a field trial for this test. Study IV is based on the data collected in connection to a national test in mathematics for upper secondary students and a field trial for that test.

Study I
Several previous studies have investigated test-taking motivation using self-report measures (e.g. Cole et al., 2008; Liu et al, 2012; Thelk et al, 2009). However most previous studies have used rather narrow measures of test-
taking motivation, and thus neglect many of the processes that can give rise to a motivated or unmotivated state. Further, few studies have compared test-taking motivation between test with different stakes after controlling for measurement invariance.

The aim of the first study was to evaluate the psychometric properties of an expectancy-value based instrument for measuring test-taking motivation. The instrument's measurement invariance was investigated as well as possible differences in reported test-taking motivation for a low-stakes test (national test field trial) and a high-stakes test (regular national test). A test-taking motivation questionnaire including 19 items was assembled by sampling or adapting items from existing test-taking motivation scales, creating new items inspired by items from domain-specific motivation scales, and creating a few new items. The questionnaire represented five aspects of the expectancy-value theory: Effort, Expectancies, Importance, Interest and Cost (conceptualized as Test Anxiety). The questionnaire was distributed to grade nine students taking a national test in science and a field trial for this test. Model fit of the data to the theoretical model was tested by means of structural equation modelling for the low-stakes (field trial; \( n = 1,047 \)) and high-stakes (regular national test; \( n = 536 \)) data, respectively. In addition, the measurement invariance across the test situations was investigated.

After excluding four items, the model achieved adequate model fit for both test situations. Effort was strongly predicted (\( R^2 = .69 \) and \( .73 \)) by the four other subscales. But only expectancies and importance were significant predictors. The other subscales showed significant correlations to expectancies and importance. Test performance exhibited significant bivariate correlations (Pearson’s \( r \) between .20 and .48) with effort, expectancies, importance and interest in both test situations. Partial scalar invariance was obtained, meaning that the instrument can be used to measure and compare test-taking motivation between low- and high-stakes assessment situations. Subsequent latent mean comparisons showed significantly higher effort, expectancies, importance, interest, and test anxiety in the high-stakes context. In sum, the questionnaire had promising properties, is brief and easily administered, and covers several relevant aspects of situation-specific motivation. It could thus contribute to deeper understanding of students’ test-taking motivation for tests with different stakes and be a useful tool for identifying ways to increase effort in low-stakes tests.
Study II
Few studies have compared reported levels of test-taking motivation for the same samples of students across authentic low- and high-stakes settings. Even fewer have investigated differences in test-taking motivation between different school classes.

The main aims of study II were to (1) investigate changes in reported test-taking motivation for one group of students between the field trial (low-stakes) and the regular national test (high-stakes), and (2) determine whether there are differences in reported test-taking motivation between school classes. The role of gender and subject in relation to test-taking motivation, and the relationship between test-taking motivation and test performance were also analysed. Two-level second order latent growth models were used to analyse the data collected in connection to the national test in science in spring 2012. The analysis was based exclusively on data for students who completed the test-taking motivation questionnaire after both the field trial and the regular national test in science \((n = 375)\).

The analyses revealed significant increases in the scores for all test-taking motivation aspects between the field trial and the regular national test. The aspects showing the greatest increases were effort, importance and test anxiety. That is, the students reported that they invested more effort into the national test, found the national test more important, and experienced more test anxiety relating to the national test. Smaller differences were found for the interest and expectancies aspects, both of which should in theory be less sensitive to external situational factors. The variability in the test-taking motivation construct scores for the field trial was higher than for the national test, and between 3 and 11% of the variance in the test-taking motivation construct scores existed at the class level. Females reported higher test anxiety and lower expectancies than males. Subject did not have a substantial effect on test-taking motivation. Effort, expectancies, importance and interest were significant predictors of test scores after preliminary grades had been accounted for.

In conclusion, the study showed that test-taking motivation varies between and within tests with different stakes, and that test-taking motivation seems to affect test performance. These factors could affect the validity of the interpretations and uses of the test results from low-stakes tests. Moreover, if the low-stakes test is a field trial for a high-stakes test, misinterpretation and misuse of its results could threaten the validity of the interpretation and use of the high-stakes results.
Study III

Although there are some qualitative studies of students’ perceptions of specific tests (e.g. Brookhart & Bronowicz, 2003; Hopfenbeck & Kjærnsli, 2016; Triplett & Barksdale 2005), there is a lack of studies that provide a deep and rich understanding of students’ perceptions as described by the expectancy-value theory. Study III was the first such study undertaken from a Swedish perspective, and provides a rich understanding of students’ test-taking motivation in the national test and field trials.

The aim of study III was to explore in more depth grade nine students’ perceptions of the Swedish national test in science, in terms of test-taking motivation, using the expectancy-value theory as theoretical framework. The study primarily focused on perceptions of the regular national test but also included perceptions of the preceding field trial. The study is based on semi-structured interviews with twelve students, and the interview data were analysed using thematic analysis.

This resulted in the identification of five main themes; (1) opinions about the test, (2) motivation to do their best, (3) how did I do on the test, (4) the value of the test and (5) feelings about the test. Overall, the students had positive perceptions of the national test, they were motivated to do their best, believed they did well, and found the test important and good. They considered the test important for a variety of reasons including grades, knowledge, and its important appearance. Although the students found the national tests important, the interviews revealed that they were uncertain about the purpose and importance of both the field trial and the national test. Several students were nervous before taking the test and some during the test, whereas other felt OK or even enjoyed taking the national test. The students talked about less effort and test anxiety during the field trial, and perceived the test to be less important than the national test. Although they claimed to have invested less effort into the field trial, they still said that they tried to answer all its questions and none said that they didn’t care about the trial at all.

The interviews provided a rich understanding of the different aspects of test-taking motivation as described by the expectancy-value theory, for example by clarifying why students find the national test important and what they mean by putting effort into a test. They also suggested some ways of modifying test administration to increase test validity. For example, it may be beneficial to clearly inform students about the aims and purposes of low-stakes field trials, and to strive to create positive and permissive test environments.
**Study IV**

The rationale behind the fourth study was that few studies have tried to model a more comprehensive test-taking motivation construct, and those that have done so used rather restrictive assumptions about the relationships between the questionnaire items and the latent factors.

The aim of study IV was to explore the psychometric properties of an extended and revised version of the test-taking motivation questionnaire evaluated in study I and to test the theoretical assumptions underlying the instrument’s design. Some of the items of the first questionnaire (see study I) were revised and modified based on psychometric evaluation and cognitive interviews, and two new subscales were added: a utility subscale and an overall cost subscale. In total, the revised questionnaire included 22 items and consisted of seven brief subscales representing six aspects of the expectancy-value theory: effort, expectancies, importance, interest, utility, and cost (overall cost and test anxiety). The questionnaire was administered to a sample of upper secondary students taking a field trial for a national test in mathematics \( (n = 222) \). Sixty-two students also answered the questionnaire three weeks later after completing the regular national test in mathematics. The field trial data were analysed by using common structural equation modelling (SEM) and exploratory structural equation modelling (ESEM). Additionally, differences in test-taking motivation between the low-stakes field trial and the high-stakes national test were tested for the group of students that completed the questionnaire on both occasions.

The ESEM analysis suggested that the original hypothesized seven factor model did not fit the data well, but a six-factor model excluding the overall cost items and two other items achieved a good fit. The analysis supported the theory that test-taking motivation consists of several distinctly different aspects, but also showed that the different subscales are highly related. For example, some items were theoretically aligned to a certain subscale but empirically loaded on another subscale, and several items loaded on multiple subscales. The ESEM model provided a better fit to the data than a comparable SEM model, and also provided a more direct method for discovering mis-specified loadings. The analysis showed that effort was quite strongly \( (R^2 = .44) \) predicted by the five other subscales, with importance and test anxiety being significant predictors. The students reported lower effort, expectancies, importance and test anxiety for the field trial than for the national test. In conclusion, the revised instrument appears to be a promising tool for measuring test-taking motivation and ESEM seem to be a suitable approach for evaluating factor structures for the test-taking motivation construct.
9. Discussion

The aim of the work underlying this thesis was to improve the understanding of student test-taking motivation in connection to achievement tests. Special focus was to explore how the test-taking motivation construct can be theoretically understood, operationalized, measured and modelled and to investigate self-reported test-taking motivation in different student populations and for tests with different stakes. In a wider sense, the research in this thesis concerns the potential effect of test-taking motivation on the validity of the interpretation and use of achievement test results. The thesis is based on four empirical studies, all exploring test-taking motivation and all connected to validity in one way or the other.

In studies I and IV, the test-taking motivation construct was modelled and the psychometric properties of an expectancy-value based test-taking motivation questionnaire were investigated. In all studies, students’ reported test-taking motivation for low- and high-stakes tests was investigated using questionnaires (Studies I, II, IV) and interviews (Study III). The assessment context was Swedish national tests in science and mathematics as well as field trials for these tests.

Increased knowledge of test-taking motivation may be important in terms of test validity (because of the way motivation may affect test performance) and test-taking psychology (because of the way the test may affect the test-taker). Despite this, students’ test-taking motivation is rarely considered when planning, performing and interpreting achievement tests (Wise & Smith, 2016). Further, most studies on test-taking motivation claim to be based on the expectancy-value theory but none has tried to measure and model expectancies and all task values. Consequently, the theoretical understanding of the constructs is currently rather underdeveloped, and little is known about why students are motivated (or not motivated) to do their best on a test. Also, in a Swedish perspective few previous studies have considered students’ perceptions of the national tests and none have investigated how the students approach national test field trials.

In this concluding chapter I first briefly summarize the main findings from the studies. I then discuss the results in relation to the two research aims and issues of assessment validity. Finally, I reflect on the research design from a validity perspective, discuss the limitations of this work, present some ideas for future research, and offer some concluding remarks.
9.1 Main findings
The test-taking motivation questionnaire developed, administered, and evaluated in this work appears to have promising properties. It is short and easily administered but still covers several relevant aspects of situation-specific test-taking motivation. The first version of the questionnaire had acceptable properties, while the second version displayed better model fit to the data and included additional value aspects. Overall, findings further suggest that the expectancy-value theory seems to be a viable framework for interpreting the test-taking motivation construct. A theoretical model in which effort is treated as an outcome of expectancies and task value achieved a good fit to the data across student samples and tests, all of the proposed subscales except for overall cost seem to constitute separate factors, and effort was quite strongly predicted by the other latent variables (Studies I & IV). However, the findings also show that test-taking motivation is a complex construct that is challenging to operationalize, measure and analyse.

Findings related to the second research aim clearly suggest that test-taking motivation varies between and within tests with different stakes. Students in general, in both samples, reported positive attitudes towards the regular national tests, they put effort into the tests, and they found the tests important and good. A fair amount of students did however also report test anxiety at the national tests (Studies I & IV). In contrast, students reported putting less effort into the field trials. Students also seemed to find the field trials less important than the national tests and experienced less test anxiety during the former (Studies I-IV). Additionally, test-taking motivation seems to differ between school classes and between genders (Study II). Finally, the results obtained suggest that test-taking motivation, or at least some aspects of the test-taking motivation construct, is related to test performance. Positive relationships were identified between achievement test results and effort, expectancies, importance, and interest (Studies I & II). In all, main findings largely support previous research in other assessment contexts and suggest that test-taking motivation is an aspect worth considering in assessment situations. Below, these main findings are discussed in somewhat more detail in relation to the research questions and to assessment validity issues.

9.2. How can the test-taking motivation construct be theoretically understood, operationalized, measured and modelled?
Validation of a measurement involves a continuous accumulation of relevant evidences supporting the proposed interpretation and use of an instrument
or assessment system. As noted in chapter 6, the validation procedure in the current thesis was guided by the Standards (AERA, APA, & NCME, 2014) and their conception of validity and validation, and by Benson’s demonstration of a strong program for validation (Benson, 1998). Although the validity discussion below mainly relates to the measurement of test-taking motivation, the general principles and conclusions that are drawn are applicable also when validating tools for measuring other latent constructs.

9.2.1 Towards a better theoretical understanding of test-taking motivation

A well-articulated theoretical domain is an important part of construct validity (Benson, 1998). Without this, it becomes difficult to make a proper evaluation of an instrument or assessment system. Still, in much empirical research, the theoretical domain gets little attention. The theoretical domain is defined during the substantive stage described by Benson (1998), and it is continuously revised based on knowledge obtained during empirical studies. This thesis uses expectancy-value theory as its theoretical framework for studying test-taking motivation.

The structural equation modelling showed that effort was rather well predicted by the other subscales considered in this work, in particular in study I (Studies I & IV). Thus the results obtained in this work support the assumption that effort is an outcome of the expectancy and value aspects and can be used as an approximation of test-taking motivation. Previous research has arrived at similar findings (Cole et al., 2008; Penk & Schipolowski, 2015; Penk et al., 2014). Thus, if the main goal of a test-taking motivation measure is to determine whether the students were motivated to do their best on a test, a measure of effort may be sufficient (e.g., whether one wants to know if scores from tests like PISA or the national test field trial suffer from construct-irrelevant variance due to low motivation). However, if one wants to understand why students are motivated or not and how effort could be modified, it may be valuable to also consider expectancies and task values.

Results from studies I and IV in line with previous studies consistently show that effort invested in a test is positively correlated to how important the students find the test (Cole et al., 2008; Penk & Schipolowski, 2015, Thelk et al., 2009). How expectancies and the other value aspects relates to effort are less studied and existing research, including the studies in this thesis, show inconsistent results (see Studies I, III, & IV; Cole et al., 2008; Penk & Schipolowski, 2015; Penk et al., 2014). For example, both in study I and in Penk and Schipolowski’s (2015) study expectancies had a relatively high correlation to effort while in study IV no correlation between
expectancies and effort was found. Expectancies and task values are commonly treated as two separate components of the expectancy-value theory (Wigfield, et al., 2009). However, my results suggest that the theoretical division of expectancies and task values is not completely self-evident, as evidenced by the fact that some value aspects were more closely related to expectancies than to other value aspects (Studies I & IV). As also Eklöf (2006) concludes it is possible that the expectancy-value theory should be interpreted slightly differently when measured on a task specific level. The good fit between the tested theoretical model and the empirical data does not exclude the possibility of plausible alternative models based on different conceptualizations of the relationships between the motivational aspects.

Even if many questions still remain to be resolved, the current thesis provides one step towards a more thorough understanding of the theoretical domain of test-taking motivation. In the future, I would like to see more theoretical discussions about how the different aspects of the expectancy-value theory should be interpreted in a test-taking motivation context together with studies looking more closely at the internal relationships between expectancies, values and effort. For example discussions about how the factors identified as influencing test-taking motivation in previous empirical studies (see section 4.2.2) could be understood within the expectancy-value framework would be interesting.

Choosing the expectancy-theory as the theoretical framework for this thesis meant emphasizing some aspects of test-taking motivation while paying less attention to others. Consequently, it might be possible to obtain complementary insights by studying test-taking motivation from another theoretical perspective, such as the self-determination theory (Ryan & Deci, 2000). Studies using the concepts of autonomy and relatedness within the framework of self-determination theory could clarify how a student’s willingness to do their best is associated with their desire to feel connected to significant others and to feel a sense of control. Such an analysis could potentially reveal ways to increase motivation for low-stakes tests by giving students an increased sense of control and stronger feelings of relatedness without raising the perceived stakes.

9.2.2 To operationalize and measure test-taking motivation
A valid measure of a latent construct requires a carefully considered operationalization of the construct, i.e. a definition of the empirical domain (Benson, 1998). Too often, researchers state that they have measured a certain construct without really capturing it in the way it was theoretically defined (Wigfield et al., 2009). Problematisations of how the different
aspects of test-taking motivation should be operationalized is seldom seen. Overall, the empirical results in this thesis indicate a good agreement between the theoretical and empirical domains. The interviews and empirical data however also prompted some new thoughts about how to operationalize some aspects of test-taking motivation. The empirical analysis of the questionnaire data showed that it is challenging to construct items that capture a single aspect of a multidimensional construct such as test-taking motivation without impinging upon some other aspect\(^8\). A careful analysis of factor structure is therefore essential when interpreting measurements of latent constructs. The specificity of the items could also be problematized. For example, the operationalization of the importance construct in several test-taking motivation studies, including the ones in this thesis, reflects importance on a very general level (e.g. “this test was important to me” as used by authors such as Baummert & Demmrich, 2001; Eklöf & Nyroos, 2013; Sundre, 2007). Conversely, the importance aspect in the expectancy-value theory relates to the personal importance of doing well on a task (e.g. because the test provides an opportunity to demonstrate aspects of one’s ideal self-schema, because it provides a challenge, or because it provides a means of fulfilling achievement, power and social needs; Eccles et al., 1983). The interviews presented in study III showed that students had several possible reasons for agreeing with the statement “this was an important test to me”, including a desire to get good grades to secure admission to a particular course (utility), to show off what they knew (importance according to the expectancy value theory), or because the test sounded important (subjective norms). Students also noted that because they perceived the test as important, preparing for it was not a waste of time (cost). From this perspective, it is quite logical that importance as conceptualized in my questionnaire has a direct effect on effort and that some of the other value-related variables affect importance. In future studies, it would be interesting to include some items designed to more explicitly conceptualize the importance aspect of task value as described by Eccles et al. (1983). The specificity of the formulations of the utility items could also be discussed. Utility can mean very different things in the contexts of high- and low-stakes tests\(^9\). It is therefore not necessarily true that utility can be defined and measured in a sufficiently general way to be useful for

\(^8\) For example, “a good result on this test might contribute to me getting better grades” was assigned to utility but loaded on importance, and “while taking this test, I thought about how badly I was doing” was assigned to test anxiety but loaded on expectancies.

\(^9\) The utility of completing a low-stakes test could derive from getting feedback about one’s strengths and weaknesses or practicing for an upcoming high-stakes test. Conversely, the utility of a high-stakes test could be due to getting the grades needed for future study.
both high- and low-stakes tests, and it may be necessary to develop specially adapted items for each test type. During the interviews, several students mentioned aspects connected to utility when talking about the field trial and the national test. For example, they talked about repetition, learning from mistakes, learning about strengths and weaknesses, or learning more about the national test format. However, they did not associate that with “learning something new”. It may be beneficial to add some more specific items to capture these aspects.

9.2.3 To model the test-taking motivation construct
Valid interpretation of empirical data requires the use of statistical methods whose properties complement the characteristics of the data (Benson, 1998). Often, researchers apply default methods without considering the characteristics of their data or the proposed uses of the findings. When modelling the test-taking motivation construct, a range of methodological choices and considerations had to be made (see chapter 7). These considerations are similar regardless of which latent construct that is measured. Because all the items in the test-taking motivation questionnaire were rated on a 4-point Likert scale, the gathered data are on an ordinal level, so one could argue that non-parametric methods should be used (Henriksson, 1992). However, analysis of the data showed that the item distribution could be assumed to be normal and could thus be analysed using parametric methods. In many of the analyses, non-parametric and parametric methods were applied in parallel; in all such cases, both methods yielded very similar results. Study II showed that a significant amount of the variation in test-taking motivation exists at the class level, which implies that clustering should be taken into account when analysing this construct in samples clustered into classes. More generally, many measurements in educational research are performed on students clustered into classes, so clustering should always be at least considered when modelling educational data. If one want to analyse variables that are measured on the class level or the distribution of the variance between the individual and class level, multilevel analysis might be suitable. Otherwise, statistical methods taking clustering into account could be sufficient. My studies showed that several items partly loaded on multiple test-taking motivation scales, so I would recommend comparing traditional analytical tools such as SEM to methods such as ESEM (see study IV) when modelling multidimensional constructs such as test-taking motivation. The work in the current thesis shows that valid modelling of complex phenomena measured in real-life contexts requires advanced statistical methods in order to match the “true”
characteristics of the data. However, results from such analysis might be difficult to interpret in terms of real-world meaning, why researchers should strive to find a balance between statistical complexity and staying close to the data collected.

9.3 Students’ self-reported test-taking motivation at tests with different stakes

Measures of test-taking motivation at tests with different stakes is useful both for the basic understanding of test-taking motivation and for the process of developing, administering, interpreting and using achievement tests.

In line with previous pilot studies in the national test context, the results from the questionnaires showed that a majority of the students found the regular national tests important and that they spent effort on the regular national test (Åström & Eklöf, 2011). A fair amount of students did however also report feelings of test anxiety in connection with the regular national test. This supports previous research showing that some students experience test anxiety when taking the national tests (Bagger, 2015; Eklöf & Nyroos, 2013). Test anxiety is generally considered to negatively affect students’ performance in the short term and also their motivation for learning in a longer perspective (Eum & Rice, 2011; Putwain et al, 2012). There is however research indicating that the relationship between test anxiety and performance is curvilinear, i.e. a certain amount of test anxiety might be positively related to performance (Kofman, Meiran, Greenberg, Balas, & Cohen, 2006; Pintrich & Degroot, 1990; Sung, Chao, & Tseng, 2016). Thus, even though no or weak linear correlation between test anxiety and performance were found in studies I and II it cannot be excluded students with high test anxiety is negatively affected.

The current thesis provides rather clear evidence that invested effort, perceived importance, and experienced test anxiety were lower for the low-stakes field trial than the high-stakes national tests (Studies II & IV). However, as also shown in other studies, most of the students still reported that they invested effort into the low-stakes test (Hopfenbeck & Kjaermsli, 2016; Wise & Smith, 2016). This is a reassuring finding suggesting that students do not simply dismiss low-stakes tests in terms of their willingness to invest effort. Nevertheless, as differences in test-taking motivation were found, it is still important to determine how the lower test-taking motivation might affect the validity of the interpretation of achievement test results. In line with previous research (Abdelfattah, 2010; Asseburg & Frey, 2013; Cole & Osterlind, 2008; Eklöf et al., 2014; Liu et al, 2012; Wise & DeMars, 2005), the current thesis indicated a relationship between achievement test scores
and effort, expectancies, importance and interest for both low- and high-stakes tests. Meaning that lower test-taking motivation yields lower performance. According to calculations presented in study II the difference in effort between the field trial and the national test on average caused a relatively small decrease in performance. Considering that the calculation represents an effect averaged over individuals and different item types and that test-taking motivation has been shown to vary depending on personality and to have different effects on different types of item (Barry, et al., 2010; Demars, 2000) it is likely that the effect of test-taking motivation is small for some item/students and more substantial for other items/students.

As concerns possible consequences for the national tests, if the proportion of students who correctly answer some item in a field trial is lower than would be the case in a national test due to low motivation, the field trial results may not accurately reflect the item’s difficulty. Because the grade boundaries for the regular national test are set prior to test release, incorrect conclusions about item difficulty based on the field trial could also affect the validity of the interpretation of regular national test results. Lower effort could also potentially affect the quality of the students’ answers on constructed response items. The usefulness of the field trial as a tool for characterizing students’ understanding of the test items and a source of illustrative answers for scoring rubrics would be compromised if the students’ answers to the trial items were of lower quality than those they would produce for a genuine national test. Additionally, this work has shown that effort varies between school classes (Study II). If effort varies between school classes and affects test results, the conclusions drawn from a field trial may depend on which class participated in the trial, making the reliability and the validity of those conclusions questionable.

In sum, the results showed that lower test stakes resulted in lower test-taking motivation and that test-taking motivation correlates with test performance. Low test-taking motivation may thus affect the validity of any interpretation of field trial results and, in the longer term, the national tests. Further research is however needed to clarify the specific effect of lower test-taking motivation on the validity of the interpretation of test results.

9.3.1 Implications for assessment practise

Findings from the current research suggest that test-taking motivation may be important to consider in different assessment contexts. This fact has been pointed out also by others (AERA, APA & NCME, 2014, Wise & Smith, 2016), but still, most achievement test results are interpreted under the assumption that students are always motivated to do their best on any test (Wise &
Smith, 2016). My opinion is that measures to increase test-taking motivation should always be taken because of the potential effects of motivation on the validity of achievement test scores (as noted in the Standards) and for the students’ sakes. Students participate in vast numbers of tests in school and their experiences of the tests has shown to have significant association with both their approaches to learning (Marton & Säljö, 1997; Struyven et al., 2005), as well as their view of themselves as learners (Putwain et al, 2012). It could therefore be considered disrespectful to demand that students invest time and energy into tests they don’t find meaningful and whose results might not be very useful. An important question here is how test-taking motivation should be considered, i.e., what can we do to raise, adjust, or control for the effect of test-taking motivation?

As mentioned in chapter 4, several methods for increasing test-taking motivation (e.g. Finn, 2015; Lau et al., 2009, Sessoms & Finney, 2015) or statistically accounting or adjusting for “non-effortfull behaviour” (Swerdzewski, Harmes, & Finney. 2011; Wise & Kong, 2005) have been explored in recent years. The section below discusses five methods of this sort based on the national test context and the results presented in this thesis. First, one obvious way to increase test-taking motivation is to increase the stakes of the test by making the field trial result count towards the students’ grades (Finn, 2015; Wise & Demars, 2005; Wolf & Smith, 1995). However, this isn’t always possible because of administrative constraints such as high costs, or security and exposure concerns. Increasing the stakes by making the results count towards grades could also have other undesirable effects such as increased test anxiety (Eum & Rice, 2011; Segool, et al., 2013). Second, an alternative way to increase students’ willingness and ability to invest effort into field trials and national tests is to establish a positive and permissive test environment by, for example, actively encouraging students to produce their best effort and avoid emphasis on competition (Lau, et al., 2009: Putwain et al., 2012). The results from study III (see also Silvfer et al., 2015) indicated that students feel some kind of pressure to finish tests as quickly as possible because they believe that being one of the last to finish is perceived negatively by their classmates. To overcome this, the test developers could provide the teacher with a recommended time-frame for completion of the specific field trial and tell the teacher that the students should work on the field trial for the whole recommended time. Third, my studies showed that not all students had a clear understanding of the purpose of the field trial and national tests or how their results were used, and the understanding of these factors seemed to differ between classes (Studies II, III). This is consistent with earlier studies
on the Swedish national tests for grade three students (Silvfer et al., 2015). The purpose of low-stakes tests and the intended uses of their results should always be clearly explained (Sessoms & Finney, 2015) because control over the situation is assumed to increase students’ motivation (Ryan & Deci, 2000) and increasing students’ control over their situation could also improve their attitudes towards tests (Urdan & Schoenfelder, 2006). Fourth, although no correlation between utility and effort was identified in the studies included in this thesis and there is little empirical evidence of such a correlation (Wise & Smith, 2016; see however Cole et al., 2008), I think it would be preferable for the students to derive some clearly perceived utility from participating in a field trial – for example, they could get feedback about their strengths and weaknesses and what areas they need to work on, or it could be stressed that participation represents a good opportunity to practice in advance of the national test. Finally, a recent study showed that is primarily students who put forth very little effort that affect achievement test results (Attila, 2016). Therefore, it may be useful to apply a single simple measure of effort during the field trials and exclude the results of low-effort students from subsequent analyses, as suggested by Swederdzewski, et al. (2009). However, this requires a valid way of determining what constitutes very low effort.

9.4 A reflection on the research design from a validity perspective

In contrast to previous studies in this area, I used a mixed methods approach and included a wider range of value aspects as well as expectancies when measuring test-taking motivation, trying to overcome the construct underrepresentation that has likely been present in most previous studies. In such this work has contributed with important new understanding of the test-taking motivation construct. The large-scale surveys gave a general picture of the students’ test-taking motivation at both high- and low-stakes tests (Studies I, II, & IV). The interviews on the other hand gave deeper understanding of the students’ thoughts concerning test-taking motivation, e.g. why they find a test important or not and how nervous they were (Study III). The interviews also gave a better understanding of how to interpret the large-scale data and how to improve the questionnaire, both resulting in more valid interpretations of the questionnaire data. Further the interviews served as external validity evidence for the questionnaire. The results from the questionnaires and the interviews with respect to the students’ test-taking motivation at the low-stakes and the high-stakes test showed high concordance in general.
During the work a number of other validity evidence have been collected to support the validity of the interpretation and use of the results from test-taking motivation questionnaire. Cognitive interviews and SEM modelling provided evidence based on response processes and the internal relationships between test items and components, supporting the validity of the interpretations and use of the questionnaire. Measuring test-taking motivation in different contexts provided important external validity evidence for the test-taking motivation questionnaire. The questionnaire revealed variation in test-taking motivation that is consistent with the assumptions of the expectancy-value theory and the students’ test-taking motivation correlated with their test performance as expected. Although the first version of the questionnaire had acceptable properties I decided to modify the test-taking motivation before the second administration in order to better reflect the expectancy-value theory and to include more value aspects. The second version displayed better model fit to the data and included additional subscale. However the findings also show that test-taking motivation is challenging to measure. Factors that are theoretically distinct were sometimes difficult to separate empirically, the empirical relationships between the latent aspects were not always consistent with those assumed by the theory, and even though the instrument was developed with great care, there were items that seemed to be mis-specified or otherwise did not work well (Studies I & IV). Still, if complementing the revised questionnaire with a few new items I would recommend the revised questionnaire as a basis for future studies on test-taking motivation.

Through the research questions, the choice of design and research context I have also put the limelight on possible consequences of construct irrelevant variance caused by non-cognitive variables like motivation on tests like the field trials and national tests. Do these tests only measure knowledge? The national tests are an essential part of the Swedish national assessment system and more research on their potential strengths and weaknesses is important. This has not been the prime focus of this thesis but I believe that the research has raised some questions and presented some findings that might be important to consider in the future, in the national test context but also in other large-scale, low-stakes assessment contexts.
9.5 Limitations and future research

All empirical studies have limitations, and those included in this thesis are no exception. Some of these limitations are discussed in earlier sections of the thesis summary and the empirical studies. Here, I will discuss some of the main limitations of this work and suggest some directions for future research.

First, I limited my studies to the Swedish national test context. My choice of test context influenced which and how many students I got access to, and the test characteristics. By choosing the national test context I got access to groups of students who took both low- and high-stakes tests in real world settings. However, I also had to adjust my methods to fit in with the national test administration procedures. The schools that were invited to participate in the field trial were randomly selected, but the fact that participation was completely voluntary increases the risk that only the most “test-positive” teachers participated; such teachers may have unusually “test-positive” students. However, the sample sizes (especially in study I) were rather large and the students’ preliminary grades were consistent with the final course grades in the Swedish population of Grade 9 students. Nevertheless, further work on test-taking motivation in Swedish national test contexts in different age groups and in other subjects would be valuable to better understand the theoretical underpinnings of test-taking motivation and the dynamics of test-taking motivation in the national test context. Also studies of a more comprehensive test-taking motivation construct in other types of test such as the international comparative tests (e.g. PISA or TIMSS context) would be valuable.

Another draw-back of working in the national test context was that the field trials (the low-stakes test) consisted of several different test versions and these were not identical to the regular national test (the high-stakes tests). Consequently, I cannot not say whether the variation in test-taking motivation between school classes detected in study II depended on that the students took different field trials or that they took the field trials under different conditions. Also, the change in test-taking motivation from the field trial to the national test reported in studies I, II and IV might be partly due to the fact that the field trial tests contained fewer items and different kinds of items than the national test. In study II, the existence of different field trial versions was partly accounted for by doing a multilevel analysis controlling for subject, which did not eliminate the differences between the high- and low-stakes tests, and thus supports the validity of my interpretation of the findings. Also, it should be noted that from a validity perspective, the possible effects of low test-taking motivation on test results
are problematic no matter what the reason for its variation within field trials or between field trials and regular national tests.

Most my results derive from quantitative data that were collected using questionnaires and analysed by means of rather complex statistical methods. The results of these analyses and their implications were not always easily interpreted in terms of real-word meaning. Such interpretation was greatly facilitated by the knowledge gained from the interviews. However, both the questionnaires and interviews are (post-test) self-report measures and thus suffer from all the inherent limitations of self-reporting (Wise & Cotton, 2009). In future studies, it would be interesting to combine questionnaires and interviews with an objective measure such as response time measurements. Besides giving the possibility to compare students’ own perceptions of effort to an objective measure, this would also make it possible to determine how effort varies over the course of a test and over different item types (Wise & Smith, 2016). As more tests are becoming computerized, the scope for analysing response times and other test-taking behaviours will probably increase in the future.

Finally, it should be emphasised that all models in my studies are so-called correlation models. Therefore, the conclusions I can draw from my results relate to associations between variables and not causality. Thus, although I assume throughout this work that effort is predicted by expectancies and task value, experimental studies would be needed to draw firm conclusions about the causality in these relationships.

9.6 Concluding Remarks
The overall aim of the work presented in this thesis was to improve the understanding of test-taking motivation in the context of achievement test for young adolescents. I believe my work has contributed with some important bricks in the building of the knowledge of how to measure and understand test-taking motivation as well as with new knowledge of how Swedish students perceive national test field trials and regular national tests in terms of test-taking motivation. I hope that the work in this thesis may inspire researchers to discuss the theoretical concept of test-taking motivation and to carry out more empirical studies concerning the construct. I also hope that it may encourage more test developers and test users to consider test-taking motivation when designing, administrating and interpreting achievement tests. In the long run that could contribute to a more fair, stimulating and non-threatening assessment environment for the central actors in educational assessment, the students.
10. Författarens tack

Nog finns det mål och mening i vår färd - men det är vägen, som är mödan värld.10

Att skriva denna avhandling har varit en lång, lärorik, stundtals krävande men oftast mycket angenäm resa. Jag har under hela min doktorandtid varit omgiven av helt fantastiska människor som alla på olika sätt bidragit till att denna resa nu nått sitt mål.

Alla elever som ställt upp och delat med er av era tankar och åsikter, ni har utgjort själva landskapet för min resa och jag är oändligt tacksam för det ni gett mig. Utan er fanns inte min avhandling.


Min familj, släkt och vänner, ni är mitt drivmedel, under hela resans gång har ni tankat mig full med trygghet, glädje och energi. Utan alla kramar, kloka ord, klättring, resor, skidturer, löppass, middagar, brödbak,

Bryt upp, bryt upp! Den nya dagen gryr. Oändligt är vårt stora äventyr.10

Eva Knekta, februari, 2017
11. References


Appendix 1.

Short description of the interview guide

**Introductory part**
The aim and structure of the interviews were clearly explained and special emphasis was put on that there is no right or wrong answers. Informed consent was obtained from the students participating in the interviews.

**Background questions**
Questions about name, age, mother language and which national test and field trial they had done.

**Part 1. Perceptions of the national test and the field trial**
First part consisted of seven main questions about their perceptions of the national test.

1. Could you tell me about how you did prepared for the test?
2. What did you think about the test?
3. How important do you feel that the national test in biology/physics is?
4. How did you feel before taking the test?
5. Do you feel like you learned something by doing the test?
6. How do you think you did at this test?
7. How much effort did you put into doing this test?

Possible supplementary questions were listed for each main questions such as; ‘what is the difference between a good and a bad test?’ and ‘how did you feel when taking the test’. Throughout the interview the students were continuously asked to also reflect on the field trial.

**Part 2. Cognitive interviews**
For each item in the questionnaire the students were asked to explain what they thought when answering the item. For each item special probes were prepared such as; ‘what is a good results for you?’ or ‘tell me more about how you were thinking’.