UMEÅ UNIVERSITY

Dimensions of validity

Studies of the Swedish national tests in mathematics

Anna Lind Pantzare

Department of Applied Educational Science
Educational Measurement
Umeå 2018

Dissertation for PhD
ISBN: 978-91-7601-936-8
ISSN: 1652-9650
Cover art and design by: Björn Sigurdsson
Electronic version available at: http://umu.diva-portal.org/
Printed by: UmU Print service, Umeå University
Umeå, Sweden 2018

Finally...

Table of Contents

Abstract

The main purpose for the Swedish national tests was from the beginning to provide exemplary assessments in a subject and support teachers when interpreting the syllabus. Today, their main purpose is to provide an important basis for teachers when grading their students. Although the results from tests do not entirely decides a student's grade, they are to be taken into special account in the grading process. Given the increasing importance and raise of the stakes, quality issues in terms of validity and reliability is attracting greater attention. The main purpose of this thesis is to examine evidence demonstrating the validity for the Swedish national tests in upper secondary school mathematics and thereby identify potential threats to validity that may affect the interpretations of the test results and lead to invalid conclusions. The validation is made in relation to the purpose that the national tests should support fair and equal assessment and grading. More specifically, the focus was to investigate how differences connected to digital tools, different scorers and the standard setting process affect the results, and also investigate if subscores can be used when interpreting the results. A model visualized as a chain containing links associated with various aspects of validity, ranging from administration and scoring to interpretation and decision-making, is used as a framework for the validation.

The thesis consists of four empirical studies presented in the form of papers and an introduction with summaries of the papers. Different parts of the validation chain are examined in the studies. The focus of the first study is the administration and impact of using advanced calculators when answering test items. These calculators are able to solve equations algebraically and therefore reduce the risk of a student making mistakes. Since the use of such calculators is allowed but not required and since they are quite expensive, there is an obvious threat to validity since the national tests are supposed to be fair and equal for all test takers. The results show that the advanced calculators were not used to a great extent and it was mainly those students who were high-achieving in mathematics that benefited the most. Therefore the conclusion was that the calculators did not affect the results.

The second study was an inter-rater reliability study. In Sweden, teachers are responsible for scoring their own students' national tests, without any training, monitoring or moderation. Therefore it was interesting to investigate the reliability of the scoring since there is a potential risk of bias against one's own students. The analyses showed that the agreement between different raters, analyzed with percent-agreement and kappa, is rather high but some items have lower agreement. In general, items with several correct answers or items where different solution strategies are available are more difficult to score reliably.

The cut scores set by a judgmental Angoff standard setting, the method used to define the cut scores for the national tests in mathematics, was in study three

compared with a statistical linking procedure using an anchor test design in order to investigate if the cut scores for two test forms were equally demanding. The results indicate that there were no large differences between the test forms. However, one of the test taker groups was rather small which restricts the power of the analysis. The national tests do not include any anchor items and the study highlights the challenges of introducing equating, that is comparing the difficulty of different test forms, on a regular basis.

In study four, the focus was on subscores and whether there was any value in reporting them in addition to the total score. The syllabus in mathematics has been competence-based since 2011 and the items in the national tests are categorized in relation to these competencies. The test grades are only connected to the total score via the cut scores but the result for each student is consolidated in a result profile based on those competencies. The subscore analysis shows that none of the subscores have added value and the tests would have to be up to four times longer in order to achieve any significant value.

In conclusion, the studies indicate that several of the potential threats do not appear to be significant and the evidence suggests that the interpretations made and decisions taken have the potential to be valid. However, there is a need for further studies. In particular, there is a need to develop a procedure for equating that can be implemented on a regular basis.

Keywords: national tests; validity; interrater reliability; standard setting; linking; subscores; test development

Abbreviations and variables

CAS	Computer Algebra System
EA	Extended answer
LMEAN	Mean Levine observed score method
LLIN	Linear Levine observed score method
MC	Multiple-choice
NAE	The Swedish National Agency of Education
NAEP	National Assessment of Educational Progress
NEAT	Non-equivalent group anchor test
NCTM	National Council of Teachers in Mathematics
PISA	Programme for International Student Assessment
PRMSE	Proportional reduction in mean square error
SA	Short answer
SAT	Scholastic Assessment Test
SD	Standard deviation
SweSAT	Swedish Scholastic Assessment Test
SSI	The Swedish School Inspectorate
TIMSS	Trends in Mathematics and Science Study
TLIN	Linear Tucker method
TMEAN	Mean Tucker method
VAR	Value added ratio
UmU	Umeå University

Variables used in the papers. Number of the paper in brackets

c1, c2	Class 1, class 2 (I)
f, m	Female, male (I)
K	Cohen's kappa (II)
A	Coefficient alpha (II)
$P(a)$	Percent agreement among judges (II)
$P(e)$	Percent agreement by chance (II)
σ^2	Variance (II)
K	Number of judges (II)
U	Utility (IV)
$\tilde{U}$	Relative utility (IV)

Populärvetenskaplig sammanfattning

Att prov och bedömningar har en stor inverkan på skolans verksamhet kan nog de flesta hålla med om. Via proven ska eleverna på ett likvärdigt och rättvist sätt kunna få visa vad de kan och lärarna ska kunna använda resultaten som ett stöd i betygssättningen eller i planeringen av den fortsatta undervisningen. Oavsett vilket syfte proven har ska de vara av hög kvalitet och generera reliabel information, annars kan de slutsatser som dras utifrån resultaten vara missvisande. Denna typ av kvalitetsfrågor blir ännu viktigare för externa prov, så som de nationella proven, som ska användas som ett stöd för betygssättning. I och med att denna typ av prov konstrueras av någon annan än den undervisande läraren blir det än viktigare att frågorna är begripliga och inte kan missförstås, att bedömningsanvisningarna ger förutsättningar för att göra en likvärdig bedömning, att gränserna för de olika provbetygen på ett bra sätt speglar kravnivåerna i kunskapskraven så att materialet inte inbjuder till tolkningar som egentligen inte är möjliga att göra.

När man diskuterar kvalitet hos prov är det två begrepp som ofta nämns nämligen validitet och reliabilitet. Enkelt kan man säga att validitet handlar om att undersöka att provet mäter det som ska mätas och reliabilitet handlar om tillförlitligheten i resultaten, det vill säga att de skulle bli desamma om provet upprepades eller om någon annan bedömde svaren. Utifrån den syn på validitet som råder idag är reliabiliteten en förutsättning för validiteten men det är inte ett tillräckligt krav. Bara för att reliabiliteten är hög behöver det inte innebära att de slutsatser som dras utifrån resultaten är valida. Det är också nödvändigt att validiteten undersöks utifrån provets syfte och om det finns flera syften måste validiteten undersökas utifrån vart och ett av dem.

Syftet med denna avhandling är att validera de svenska nationella proven i matematik på gymnasiet utifrån att de ska vara ett stöd för likvärdig och rättvis bedömning och betygssättning. Mer specifikt har arbetet inriktat sig mot att undersöka eventuella hot mot validiteten, hot som skulle kunna påverka de tolkningar som görs utifrån resultaten. Valideringen har fokuserats på några olika delar av provprocessen, från det att proven genomförs, via bedömningen till sammanställning och hantering av resultaten. Som stöd för detta arbete har ett ramverk utgående från åtta olika kritiska punkter använts som stöd. Här nedan presenteras de studier som ligger till grund för avhandlingen.

De nationella proven i matematik innehåller sedan år 2000 provdelar som eleverna ska genomföra utan hjälpmedel och provdelar där det är tillåtet att använda digitala verktyg. Sedan 2007 är det tillåtet att använda alla typer av digitala verktyg men det har inte krävts att eleverna haft annat än funktionsräknare eller grafritande verktyg. Det finns dock betydligt mer avancerade verktyg som till exempel symbolhanterande verktyg som klarar av att lösa ekvationer algebraiskt, någonting som skulle kunna minska risken för att

göra fel. När dessa verktyg tilläts var att de var relativt dyra att köpa och då de inte krävdes vid genomförandet av nationella proven valde många skolor att inte köpa in den typen av verktyg till eleverna. Det blev därmed en fråga om resurser då vissa elever valde att, på eget bevåg, köpa ett sådant verktyg medan andra fick hålla tillgodo med de verktyg som skolan tillhandahöll. Utifrån ett provperspektiv blev det extra problematiskt då det även fanns ett uppdrag att elever med symbolhanterande verktyg inte skulle ha någon fördel av sitt verktyg. Syftet med studie I var därför att ta reda på vad elever som har tillgång till symbolhanterande verktyg och som dessutom är vana att använda dem även i undervisningen gör med verktyget. För att kunna veta hur eleverna resonerar och för att i detalj kunna studera vad de gör genomfördes en tänka högt studie där åtta elever fick lösa fem uppgifter som innehöll olika typer av ekvationer. Resultatet visade att alla eleverna försökte, till att börja med, lösa ekvationerna för hand och det var i stort sett bara de högpresterande eleverna som hade någon nytta av sitt verktyg, de lyckades lösa ytterligare några uppgifter. Slutsatsen i studien var att de symbolhanterande verktygen inte påverkade resultaten i någon större utsträckning. Det skulle dock vara intressant att upprepa studien idag då verktygen finns mer lättillgängliga och är dessutom mer vanligt förekommande i klassrummen.

Nästa steg i valideringen handlade om att undersöka bedömningen av de nationella proven. I det svenska systemet sköts bedömningen av de nationella proven ute på skolorna och det är rätt så vanligt att den undervisande läraren bedömer sina egna elever. I många system internationellt innehåller bedömningsprocessen av ett batteri av kontroller. De som ska bedöma proven får en utbildning och sedan övervakas arbetet så att bedömningarna blir likvärdiga. I Sverige får lärarna en bedömningsanvisning som ska vara självinstruerande sedan får de klara sig själva, möjligen med stöd av andra lärare på skolan. Det skulle därmed kunna vara en risk för att bedömningen inte sker på samma sätt överallt. Studie II visar dock att på det stora hela är svenska matematiklärare duktiga på att bedöma de nationella proven och resultaten är i linje med de värden man har internationellt efter utbildningar och kontroller. Det finns en del uppgifter som är problematiska att bedöma, främst sådana där det finns olika lösningsmetoder som alla kan vara lika bra men där bedömningsanvisningen bara kan ge ett stöd för den vanligaste metoden. Även uppgifter där det är möjligt att få ett korrekt slutgiltigt svar trots att det finns felaktigheter i lösningen är problematiska. Slutsatsen var dock att på det stora hela har de nationella proven en hög bedömaröverensstämmelse.

Studie III handlar om kravgränssättningen av proven. En viktig del i likvärdigheten är att det ska vara lika lätt eller lika svårt att nå ett visst provbetyg oavsett vilken termin som kursprovet ges. Idag görs kravgränssättningen med en metod där bedömare som är väl insatta i styrdokumenten och som också har en god kännedom om vad som krävs för att uppfylla kunskapskravet på respektive betygsnivå får bedöma vad som ska krävas för det enskilda provet. Denna typ av

metod bygger på subjektiva bedömningar av uppgifterna i relation till kunskapskraven och det finns alltid en risk för att det sker glidningar i tolkningarna av vad som ska krävas. Ett annat sätt att fastställa kravgränserna är att använda statistiska metoder där information om det nya provet jämförs med det gamla provet så att gränserna blir lika krävande. Detta kallas för att man ekvivalerar proven. För att kunna göra en ekvivalering krävs det antingen att elevgrupperna som genomför proven är statistiskt ekvivalenta eller att det finns gemensamma uppgifter, så kallade ankaruppgifter, som båda grupperna av elever gör. I denna studie används den senare metoden. Resultaten indikerar att kravgränserna i de båda proven är i det närmaste likvärdiga och att det endast skiljer något enstaka poäng. Det som dock kanske är det tydligaste resultatet är att det behöver göras fler studier för att finna en metod för ekvivalering som är långsiktigt hållbar i relation till de förutsättningar som de nationella proven ger. I och med att kravgränserna måste vara satta innan provet genomförs krävs det att de gemensamma uppgifterna genomförs i anslutning till närmast föregående prov eller i någon annan form av utprövning. Detta ställer höga krav på att de deltagande elevgrupperna är tillräckligt stora och att ankaruppgifterna inte sprids eftersom de då inte fungerar som de ska längre.

Den sista studien handlar om vilka resultat som är reliabla att generalisera utifrån och som är möjliga att rapportera och inte. I och med att de svenska ämnesplanerna i matematik är uppbyggda utifrån förmågor har även de nationella provens uppgifter kategoriserats till dessa förmågor, men provbetygen har baserats på totalpoängen. I återrapporteringen av resultat till provinstitutionen har läraren för varje redovisad elev fått en sammanställning i form av ett stapeldiagram där elevens resultat per förmåga eller par av förmågor har jämförts med det totala antalet möjliga poäng i respektive grupp. Syftet med studie IV var att undersöka om denna uppdelning av förmågepoäng ger någon ytterligare information om elevens kunnande än vad totalpoängen gör. Resultaten visar att ingen av förmågegrupperna bidrar med någon ytterligare information och därför borde de inte redovisas. Däremot är det utifrån ett provkonstruktionsperspektiv viktigt att se till att provet har en rimlig täckning av förmågorna för att totalpoängen ska kunna användas för att stödja en likvärdig betygssättning.

Sammanfattningsvis visar studierna att det finns validitetsbevis till stöd för den utformning som de nationella proven har men det finns också en del hot som behöver undersökas vidare.

Studies

This thesis is based on the following studies, which are referred to in the text using the following enumeration[1]

I. Lind Pantzare, A. (2012). Students' use of CAS calculators: effects on the trustworthiness and fairness of mathematics assessments. International journal of mathematical education in science and technology, 43(7), 843-861.

II. Lind Pantzare, A. (2015). Interrater reliability in large-scale assessments – Can teachers score national tests reliably without external controls? Practical Assessment, Research & Evaluation, 20(9).

III. Lind Pantzare, A. (2017). Validating Standard Setting: Comparing Judgmental and Statistical Linking. In S. Blömeke & J.-E. Gustafsson (Eds.), Standard Setting in Education: The Nordic Countries in an International Perspective (pp. 143-160) New York, NY: Springer.

IV. Lind Pantzare, A., & Wikström, C. (2018). Using summative tests for formative purposes. An analysis of the added value of subscores. Manuscript submitted for publication.

[1] Study IV is co-authored. I was in charge of drafting most of the manuscript, performing the data analyses and interpreting the results. CW added parts in the introduction and the discussion. She also provided valuable revisions of the manuscript.

1 Introduction

Most people probably agree that tests play an important role in schools. For the students to have an opportunity to show what they have learned; for the teachers as a support for grading or for planning forthcoming teaching; and also for those who have the overall responsibility for the school, such as principals and politicians who all want to make different types of evaluations of school performance. Regardless of the reasons for using a test, it is important that the test is of high quality in relation to the purpose, fair to the test taker and yields reliable information. Otherwise, a lack in quality might influence the conclusions drawn and the decisions taken on the basis of the result and that might cause negative consequences.

These quality issues become even more important in relation to external tests – that is, tests developed by someone else than those who use them, like the national tests – where the same test or parallel test forms are administered to many schools and students. Irrespective of whether the national tests in a subject are to be used to support grading or to evaluate the school system, the choice of test items have to be representative of the syllabus. Moreover, it must be possible to score the answers consistently, interpret the results in the same way and compare results between test forms.

Sweden has criterion-referenced national tests in some of the core subjects. In upper secondary school there are such tests in Swedish, English and mathematics. The Swedish national test system is highly decentralized with a high trust in teachers and other school personnel to administer, score and report the results from the tests without any moderation or control (Dufaux, 2012). The Swedish National Agency of Education (NAE) has commissioned a number of Swedish universities to develop the tests (Erickson, 2017). These universities have had the test development commission for twenty years or more and there are elaborated processes for developing and maintaining tests of high quality.

Today, the main purpose of the national tests is to support fair and equal grading but it has not always been this way. From the beginning of the era of national tests, the use of the tests was optional and the tests focused on diagnostic purposes such as identifying the students' strengths and weaknesses and supporting an interpretation of the syllabuses. Giving the task of developing the tests to the universities was seen as an appropriate way to achieve the quality required for a centralized test. In addition, there was a rather naive belief that teachers would understand, simply by using the tests, how to develop good tests themselves. Over the years much has happened, the tests have become mandatory, the stakes for students and schools have increased, and the diagnostic purposes have been removed. With the shift in purposes and importance of the tests, quality issues have over time become much more important since the

results can be, and also are, used for comparisons, both locally and nationally (C. Lundahl, 2016).

Two key concepts are involved when considering the quality of a test: validity and reliability. Validity, in simple terms, is concerned with the degree to which a test measures what it claims to measure. Reliability has to do with trust and an assurance that the test results will be the same if the test is repeated or if someone else is scoring the answers (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). High reliability is a necessity for validity but it is not enough, and there are sometimes misconceptions regarding this. A test can be very reliable but still not measure what it aims to measure.

It is generally accepted that the validation of a test, that is the process of investigating validity, should be done in relation to the purpose of the test (American Educational Research Association et al., 2014). A large part of the validation takes place during the development process of the tests, and all the steps in the validation process are important. From defining the construct, which is what the tests purport to measure; via a realization of the construct leading to a blueprint that is a general definition of how a test should be composed; to the concretization in the form of items in a test. However, the validation does not end when the test form is settled. It is also important to investigate the validity of the whole chain from administration via scoring and interpretation of the results to decision-making.

Considerable evidence for the increased interest in quality issues, validity and reliability, connected to the Swedish national tests has emerged in the last couple of years. A governmental investigation into the national tests was reported in 2016 (SOU 2016:25) highlighting the need for systematic quality control. As a result of the investigation NAE has drawn up a general framework for the national test program (Skolverket, 2017b), something that did not exist before, even though the tests have existed since the middle of the 1990s, which is rather remarkable. Recently NAE also revised the teacher information and the instructions for administering the tests in order to gain greater consistency between schools.

Even though the national tests have existed since the middle of the 1990s, very few studies focus on quality issues connected to the tests. There are annual reports containing analyses of the results, descriptive statistics on teacher opinions, and the relation between the results and the grades (See e.g. Skolverket, 2017a). However, the ambitions in the general framework for the national tests show that it is necessary to investigate the validity with specially designed studies, studies that are not conducted on a regular basis as is the case at the moment.

1.1 Aim

The overarching aim of this thesis is to examine the validity of the Swedish national tests in upper secondary school mathematics. The validation is made in relation to the main purpose of the national tests, that they should support fair and equal assessment and grading. Even though the development process is thorough, there are validity issues that have to be investigated further. The papers in this thesis focus on the potential threats to validity that may affect the interpretations of the results, giving rise to invalid conclusions.

More specifically, the thesis considers the following research questions:

1) How do different prerequisites connected to digital tools, scorers and standard setting affect the result?
2) To what degree could the national tests be used for formative purposes, in addition to their usual summative use, by investigating whether subscores can be used when interpreting the results.

A framework supporting a structural validation process has been used to identify the threats investigated empirically.

1.2 Disposition of the thesis

This thesis consists of four papers and an introductory text, which aims to put the papers in context. Chapter 2 contains an overview of how national tests are used in Sweden and internationally. In chapter 3 there is a description of the Swedish school system and the system of national tests in general and national tests in mathematics in particular. In addition, chapter 3 describes the process of developing the national tests in mathematics. Chapter 4 addresses validity, the theoretical foundation of the thesis. The chapter begins with a summary of how the concept of validity has evolved over time. It continues with a description of the validity framework that has been used as the basis for the validity analysis in this thesis and how the studies are connected to the framework. Chapter 5 contains descriptions of the data and the participants included in the studies. Chapter 5 also include some elaborations on the chosen methods. In chapter 6 summaries of the papers are presented and in chapter 7 the main findings are reported and discussed. Chapter 7 also includes suggestions for future research and concluding remarks.

1.3 Terminology

As a non-native English speaker it becomes obvious when summing up the work that different words for the same concept have been used. In most cases this is connected to differences in British and American English. For example, ‘mark’,

'point' or 'score' all refer to the value of a correct answer for a test item or the total result from a test. 'Marking', 'rating' and 'scoring' are all used to denote the process of evaluating an answer to a test item. Hence, 'mark scheme' and 'scoring guide' are concepts that stand for instructions describing how to evaluate an answer and how many marks/scores an answer can be awarded.

In the thesis, 'test' and 'assessment' are used as parallel concepts, even though they are seen as very different in many systems. Also, Popham (1999, 2003) argues that the concepts are to be seen as parallel and claims that the introduction of the word 'assessment' was a way to distinguishing tests that can be scored automatically (such as multiple choice) from standardized tests that have to be scored by humans. In the UK, 'assessment' is defined as a broader concept including different methods for evaluating performance and attainment (Gipps, 1994; Gipps & Stobart, 2009). With this definition, the concept of assessment encompasses tests and examinations. However, in the Swedish school system there has never been, at least not until recently, a discussion about the difference in the meaning of these concepts. The national tests investigated in this thesis – following Gipps (2002) and Black and Wiliam (1998), for example – would probably better be described as national assessments since the national tests include performance-based items like essay writing, oral communication, laboratory work and problem solving, items that cannot be automatically scored.

The person who takes a test is often denoted 'test taker' but in all of the studies in the thesis, the test takers are also students in upper secondary school and therefore 'student' is sometimes used to denote the person taking the test.

The terms 'grading criteria' and 'knowledge requirements' have both been used in the Swedish criterion-referenced syllabuses to denote the description of what it takes to reach a certain grade level. The term 'grading criteria' was used from 1994 to 2011 and the term 'knowledge requirements' has been used from the revision in 2011. On the one hand, one could argue that these two terms denote the same part of the syllabuses; on the other hand, there is a difference between them. The syllabuses from 1994 to 2011 included defined goals for students to attain, which also were the same as the requirements for the pass grade. The grading criteria only described the demands for the higher grades. Syllabuses from 2011 does not specify goals to attain; instead there are general descriptions of the competences that the students are to develop in the course. The knowledge requirements describe the requirements that must be met for each grade level. The investigation that preceded the revision of the syllabuses identified a need to clarify the requirements for each grade (SOU 2007:28). Since they were to describe the requirements for knowledge in a subject, it became natural to call them knowledge requirements.

2 National testing

2.1 The Swedish assessment tradition

Sweden has a long tradition of giving teachers considerable autonomy, trusting them to teach, assess and grade their own students (Dufaux, 2012; Erickson, 2017). The grades are important since they are used for admission to higher education, determining both eligibility and selection. In Sweden, like many other countries in Europe, a combination of exit examinations along with teacher assessment and grading has been used for certifying and governing learning and instruction. However, the use of final examinations was terminated nearly 50 years ago and replaced with central tests. The idea of only using tests for standardizing the grading, instead of examinations by external inspectors, was introduced by Fritz Wigforss, a former minister of education, in a governmental investigation (SOU 1942:11). This report has become very influential in the development of Swedish grading and assessment practices. In the investigation the high trust in the teachers' capability to assess and grade their own students was very apparent. Wigforss (1942) concluded that teachers are very good at rank ordering the students in their own classroom, but that they have difficulties to relate the abilities of their own students to students in other classes, therefore indicating the need for standardized tests. However, it was not until the end of the 1960s that the use of exit examinations terminated. Then, a system with norm-referenced central tests was introduced in Swedish, mathematics, English and other foreign languages, and natural science subjects, Physics and Chemistry, in order to regulate the grading at a national level. When the current criterion-referenced system was put in place the central tests were replaced with the so called national tests that were only intended to be supportive and informative. The Swedish system of not using the results from examinations or other standardized tests as the basis for admission to higher education is rather uncommon.

2.2 National tests and examinations in the Nordic countries

The concept 'National tests' is not always optimal, as there are often misconceptions regarding what types of tests these are and what they should do. This has to do with the fact that the national tests in many countries have other purposes and content. In order to understand the features of the Swedish assessment tradition and the current national tests it is useful to start with an international comparison, beginning with Sweden's neighbors.

In Norway, which probably is the country that is most similar to Sweden, the national testing system differs in a number of ways. There are national tests in compulsory school, in year 4 and 7 (Tveit, 2018), and the purpose of these tests is to monitor whether schools are teaching the specified general competencies

defined in the curriculum. It is only the core competencies – reading, calculating and English – that are monitored via the national tests. These tests are important for the schools but low-stakes for the students since it is only the aggregated results in each class and at the school that are reported. These tests are computer-based and automatically scored (Tveit, 2014). Also, unlike Sweden, where end-of school examinations were abolished many years ago, Norway uses exit examinations at the end of upper secondary school. These are externally scored and the test result, the test grade, is used as one part of the final grade and the teacher grading is used as the other part (Eurydice, 2018a; Tveit, 2014). These two grades are combined and used for admission to higher education.

In Finland there are no national tests at all during compulsory school and it is only in the final year of upper secondary school that students meet external tests. In order to receive a degree from the general upper secondary education the students have to sit at least four so-called matriculation exams. It is mandatory to sit the exams in the native language, one foreign language, mathematics and one of the exams in natural science or social science subjects. The grades from the matriculation exams are used for admission to higher education (Finnish parliament, 1998).

There are also national tests in the Danish compulsory school, folkeskolen. The overall purpose of these tests is to support an evaluative culture at schools but the test scores are also used as information regarding the progress of each student in the specific subject while aggregated results are used for evaluating schools (Rambøll, 2013). These national tests are digital, automatically scored and adaptive, which means that the difficulty of the upcoming questions is adapted to the knowledge level of each student. In upper secondary school, the students have to sit ten examinations to pass the complete examination. The students have to take examinations in those subjects that are necessary for admission to higher education. The examinations are of different kinds both written and oral (Danish Ministry of Education, 2017).

Iceland, the fifth Nordic country, having the smallest population, has no national tests. Instead, teachers, under the supervision of head teachers, are responsible for the examinations administered at the end of each semester. These examinations, developed by teachers, are mandatory for the students. There are reference examinations in Icelandic, English and Mathematics to support teachers when preparing the examinations (Eurydice, 2018b).

This short review of the assessment systems in the neighboring countries shows that, although the national tests differ in content, format and purposes, there are also commonalities, such as generally high trust in the ability of teachers to assess their own students in compulsory school and to some extent also in upper secondary school. However, in upper secondary school, only Sweden has no national mandatory exit examinations.

2.3 National tests and examinations internationally

Outside the Nordic countries, it is common to use exit examinations at the end of upper secondary school. A recent publication provides a comprehensive overview of the assessments and examinations used in a number of countries around the world (Baird, Isaacs, Opposs, & Gray, 2018). Taking a few, but representative, European examples, England has their A-level exams, Germany the so-called Abitur examinations and France the baccalaureate, giving students certifications to be used for eligibility and selection to higher education. All of these examinations are curriculum-based.

When discussing national testing and international examination it is necessary to include the USA, where the use of standardized tests go back to the beginning of the twentieth century (Ferrara & DeMauro, 2006), when universities began admissions testing, the Scholastic Assessment Test (SAT) first being used in 1926. The increasing number of people going to college after World War II increased the interest in using standardized multiple-choice tests that were efficient and reliable. The use of admissions testing in the USA was a source of inspiration in Sweden, with the Swedish Scholastic Assessment Test (SweSAT) being introduced in 1977. The SweSAT is quite similar to the SAT and is used as an alternative to and parallel with the grade point average for selection and admission to higher education (Lyrén, 2009).

The American national tests are called the National Assessment of Educational Progress (NAEP). As the name indicates, NAEP assesses what students know and can do in different subjects. The test is administered in the school years 4, 8 and 12. NAEP was put in place 1969 and remains the only assessment to provide comparative information across the nation. NAEP is used for educational evaluation and to support educational improvement in the different states and in the nation but it does not provide any information at student or school level.

In the USA there are also other tests that give information on student achievement in relation to a curriculum. There are standardized tests developed in relation to the different state curriculums (Ferrara & DeMauro, 2006). These tests give the teachers within a state the opportunity to compare their students with other students following the same curriculum, but it is not possible to make comparisons between states or nationwide. However, this system of assessments is not the basis for admission to higher education. Instead different kinds of college admissions tests, such as the Advanced Placement test (AP) or the Scholastic Assessment Test (SAT), are used (Zwick, 2006).

3 The development of the Swedish school system and the national tests

In order to make it easier to understand the features of the Swedish national test system a brief overview of Swedish school system is required. It is however not manageable to, within the scope of this thesis, make an all-embracing historical expose since a detailed description of how the Swedish school system has evolved would require a thesis of its own. For a more thorough description C. Lundahl (2006) or Lundgren, Säljö and Liberg (2017) are recommended. This chapter will attempt to describe some of the main features of the school system that are relevant to the national tests and describe the development of the national test system in general and in mathematics in particular.

3.1 A brief historical description of the Swedish school system

Education has always been important in Swedish society, visible in a long history of public education and ambitions to provide equal opportunities for all. Since the end of the 1960s, Sweden has had a unitary nine-year compulsory school education, which starts when the children are seven years old. In addition, an optional pre-school class for six-year olds was introduced in 1998. From autumn 2018, compulsory school has been extended by one year when the pre-school class became compulsory (Prop 2017/2018:9). After compulsory school, it is possible to continue with upper secondary school, which is optional. However, most students attend upper secondary school (Skolverket, 2018).

At the beginning of the 1990s, a number of major reforms fundamentally changed the Swedish school system. The most prominent was a shift from centralization to decentralization, from national governance to responsibility being delegated to municipalities (Wikström, 2006). However, despite the decentralization, the national curricula maintained a centralized definition of goals in general, as well as of subject-specific goals and grading criteria (Erickson, 2017). Sweden went from being one of the most centralized school systems in the world to a very decentralized system in just a few years (Bunar, 2010; L. Lundahl, 2002). In addition, a system with private independent schools was introduced, and it became possible for parents and students to choose a particular school instead of being allocated a school, typically the one within the catchment area (Blanchenay, Burns, & Köster, 2014). Like municipal schools, independent schools are financed through tax revenue via a voucher system. One of the reasons for introducing the independent schools reform was to build a system where parents would have the opportunity to choose the schools with highest quality, which was supposed to increase efficiency, quality and learning outcomes (Yang Hansen & Gustafsson, 2016). Even though Sweden does not have a traditional accountability system, unlike many other

countries (Eklöf, Andersson, & Wikström, 2009), independent schools reform combined with parents choosing schools has introduced competition and resulted in grade inflation (Gustafsson, 2006; Gustafsson, Cliffordson, & Erickson, 2014; Wikström, 2005; Vlachos, 2018). During the first ten years of the 21st century the number of students attending free schools rose from about 5% to about 25% (Skolverket, 2018).

3.2 Changes in curriculum and syllabuses

The shift from a centralized to a decentralized school system is also reflected in the changes in the curriculum and the syllabuses. The upper secondary school curriculum that applied from 1970 to 1994 was seen as highly centralized and regulated (Skolöverstyrelsen, 1970a, 1970b). These documents specified general aims and goals of education and subject-specific goals, together with a highly detailed list of content, at least for content-oriented subjects like mathematics, social science and natural science subjects. The curriculum also included suggestions and recommendations of what teaching methods to use and how much time should be allocated to each part of the syllabus. Even though the centralized school system was implemented to ensure equal education in the country as a whole, voices were raised about it restricting local initiatives and freedom for teachers (L. Lundahl, 2002). Nevertheless, even if the curriculum was regarded as being normative and restrictive, a scrutiny of what is actually written gives a slightly more nuanced picture. The curriculum text about mathematics, for example, states that the teacher should not only rely on the textbook but also use other materials and a focus on problem solving is apparent (Skolöverstyrelsen, 1970a).

In parallel with the decisions to decentralize the school system, there were also discussions on how the grading system could be improved (DS 1990:60; SOU 1992:86). At that time, the grading system was norm-referenced. The main critique of this system was that while the grades made it possible to compare students they did not adequately reflect the knowledge of individual students. A major problem from a policy perspective was also the inability to evaluate shifts in the students' level of knowledge (Gustafsson et al., 2014). Another critique concerned how the grading system was understood and implemented, with some teachers and schools grading on the curve within each class instead of the national cohort, as intended (Wikström, 2006). A governmental investigation (SOU 1992:86) suggested changing to a criterion-referenced grading system but highlighted that norm-referenced grades are better if the grades are used for selection, which was and is the case in Sweden. The use of the grades for admission to higher education was probably one of the reasons why it lasted so long before any changes to the curriculum and grading system were made. The basis for the decision can be found in a governmental investigation that was reported in 1992 (SOU 1992:86). This inquiry proposed that upper secondary

school should become course-based where the subjects should be divided into a number of courses. There should be a number of compulsory courses in some core subjects but the students should also have the possibility to choose some of the courses out of interest. In the same investigation, the decentralization ambitions with regards to teaching and learning became apparent, stating that the local authorities and also the teachers should have much more freedom to decide how to teach and what to focus on in the classroom.

The grading system finally changed in 1994, in spite of the concerns regarding selection to higher education, and a goal- and criterion-referenced system was implemented (Utbildningsdepartementet, 1994b). For upper secondary school there were four grade levels: fail, pass, pass with distinction, and pass with special distinction. However, initially only goals for pass and grading criteria for pass with distinction were defined in the syllabuses. Teachers were expected to discuss and define the demands for pass with special distinction locally (Erickson, 2017; Tholin, 2006). Also, the syllabuses for each subject and course no longer included lists of specified content, emphasizing the shift from centralization to decentralization.

These reforms were introduced very quickly and teachers and schools did not receive adequate resources or support for the implementation and therefore the norm-referenced approach was still commonplace in the discussions about grading (L. Lundahl, 2002; Yang Hansen & Gustafsson, 2016). It also soon became obvious that the lack of defined grading criteria for all grades was problematic in terms of equality. It goes without saying, that it was very naïve to believe that teachers would make the same interpretations of documents that were incomplete, leaving teachers to make their own interpretations. Revised syllabuses were introduced in autumn 2000, in which grading criteria for pass with special distinction were defined for each subject and course (See e.g. Skolverket, 2000). Despite the definitions of the grading criteria for all grade levels, they were not so clear that they did not need any interpretation.

During the first years of the 21st century several reports showed that large differences still existed in the grading (See e.g. Gustafsson, 2006; Wikström, 2005). A new governmental investigation was introduced in the middle of the first decade and the report was published in 2007 (SOU 2007:28). The main suggestions in this report, that probably did not surprise anyone, were that subject-specific content needed to be better defined in the syllabuses, and with clearer goals and grading criteria or 'knowledge requirements' as they were recommended to be called. Although this investigation focused on compulsory school, the critique was the same for upper secondary school, and was highlighted in the governmental investigation presented the following year (SOU 2008:27).

It took until 2011 before a revised curriculum and revised syllabuses were introduced (See e.g. Skolverket, 2011). The main changes were aims written from the perspective that education should give students the opportunity to develop certain competencies. Lists of central content were reintroduced and

the grading scale was extended by the addition of two grades. The new grading scale runs from F to A, where F denotes fail and E to A are pass grades with A being the highest. There are now five pass grades but only three of them, E, C and A, have defined knowledge requirements (The Swedish National Agency for Education, 2012a, 2012b). The grades D and B are to be assigned if the student has met all the criteria at the grade below and more than half of the criteria of the grade above.

This grading system has also been criticized and for similar reasons as before: the knowledge requirements are too vague and difficult to interpret. In addition, since the knowledge requirements have to be fully met in order to reach a grade it is the students' weakest effort that will be decisive, something that has also been criticized (Skolverket, 2016). NAE has recently started a project to review the knowledge requirements with a view to revising them.

3.3 Assessment traditions in Sweden

As mentioned in Chapter 2, the assessment system in Sweden is highly decentralized. External examinations were abolished with the introduction of the nine-year compulsory school. Towards the end of the 1960s, so-called norm-referenced central tests were introduced with the main purpose of standardizing the grading (C. Lundahl, 2006). From 1968 until 1994 the central tests were taken by students during their second or third year of upper secondary school. A small number of test booklets were centrally scored in order to finalize the scoring guide, to evaluate the difficulty of the test form and to standardize the results to the norm-referenced curve. The teacher was then supposed to follow the mean of the central test results with a maximum deviation of 0.2 when assigning the grades, which made the tests quite high-stakes both for the students but also for the teachers and schools (Pettersson, 2004; Ramstedt, 1996). However, since the results were mainly normative at the group level the stakes for the students were lower than for the former exit student examinations. The ambition with the central tests was to have an objective national measure of the students' knowledge. However, these ambitions gradually led to a discussion about the alignment between the central tests on the one hand and the teaching and teacher assessments on the other (C. Lundahl & Tveit, 2014). This debate was probably a consequence and a part of the ongoing decentralization discussions, which also influenced the thoughts about and the design of the criterion-referenced national tests.

3.4 The introduction of the national tests

The decentralization of the school system had a significant influence on the development of the new national testing system. The purpose of the national testing system was considered by a government investigation prior to the

changeover from the norm-based to the criterion-based system (SOU 1992:86). The investigation concluded that a national test system was required in order to help teachers interpret the syllabuses, at least to begin with. It also highlighted that such a system should not restrict local freedom for schools and teachers when it came to planning education. In addition, in the investigation *Skola för bildning* (SOU 1992:94) the ambition to change focus from content to competences and skills is obvious. Skills like independence, ability to analyze and creativity but also ability to communicate and cooperate were seen as important in a modern society and should be an integral part of the curriculum. It was also emphasized that tests and other forms of assessment materials should have a broad approach and not only cover knowledge in a narrow sense.

The governmental proposal for a criterion-referenced curriculum stated that national tests would be needed as a support for teachers to interpret the curriculum and the syllabuses (Prop 1992/93:250). It also claimed that the need for centrally administered tests would decrease since the grades would be determined and standardized by the goals and grading criteria. This is another example of the rather naïve view that there would be no need for central regulations to ensure an equivalence in the grading. Following publication of the proposal, NAE received a commission to develop, with the help of test development groups at Swedish universities, a system of national tests in some of the courses in Swedish, English and mathematics (Utbildningsdepartementet, 1994a). The use of these national tests was to be recommended but not mandatory.

The general ambition of the national tests was that they would be different from the former central tests in both purpose and content. The tests in Swedish and English included an oral communication part; and, in mathematics, the focus was on problem solving. The general view was that it was more important to via the test items focus on exemplarity, even if that interfered with the possibility of having many items, indicating that reliability issues were not considered to be important (Erickson, 2017).

3.5 Changes of the purposes and stakes of the national tests

At the beginning of the new system, the purposes of the national tests were not so obvious. The various investigations, propositions and commissions from this time suggest the general view was that only by offering national tests would teachers receive the information and support they needed to assign equal grades. The directive to NAE concerning the development of a national testing system reiterated that centrally developed tests were necessary to support equal and fair grading but did not specified how this support should be designed (Utbildningsdepartementet, 1994a). In the teacher information that accompanied the first tests the main purpose of the national tests was to provide possible interpretations of the goals and to clarify the grading criteria (Skolverket,

1995). A further purpose of the tests was to stimulate discussions about subject-specific content and methods. These purposes contribute to the picture that the support provided by the national tests was not explicit, and that the tests would only be good examples of how the syllabuses could be interpreted and translated into assessments. In addition, the use of the tests was voluntary during the 1990s, they were to be taken by students sometime during a test period and the secrecy of the tests ceased in the end of the semester. However, despite the introduction of the voucher system and the independent schools reform, which introduced an element of competition for students both between independent and municipal schools and between different municipal schools (Wikström & Wikström, 2005), there were almost no comparisons or league tables of the results from the national tests at this time. Mainly the grades were simply reported and compared. As a result, the national tests were fairly low-stakes, for the students, the teachers and the schools.

Towards the end of the 1990s, the first reports on differences in the grading were published, strongly emphasizing the importance of having systematic quality assurance at a national level in order to achieve equality in education across the country (Prop 1997/98:169). One action deemed necessary was the development of national grading criteria for the highest grade in each course (i.e. pass with special distinction). In addition, a revised commission for NAE to develop new tests emphasized the importance of having the national tests as a support for grading (Utbildningsdepartementet, 1999). It was decided that the national tests in the first and final course in Swedish, English and mathematics in upper secondary school should be mandatory. It was still quite clear that the main purposes of the tests were to support the interpretation of the syllabuses and indicate the students' strengths and weaknesses. However, supporting fair and equal assessment and grading was now included as a specific purpose of the national tests. The stakes increased somewhat when the tests became mandatory, but the tests were still to be seen as rather low-stakes. However, with mandatory national tests it now became possible to report national statistics, since all schools participated, and parents and students had the possibility to review the national test results before they chose a school.

During the first years of the 21st century the Swedish government's ambitions to monitor and follow up results increased. In the case of education the national tests were already in place and generating data. In 2004, the government added a new purpose to the national tests: they were also to be used to evaluate how well the knowledge objectives were reached within each school and municipality, as well as on a national level (Utbildningsdepartementet, 2004). However, this purpose was last in a list that now included five purposes; the original purposes relating to the interpretative role of the tests were still seen as the most important. In the years that followed, NAE regularly published comparisons based on national test results and grades and the comparisons became a matter of public interest. In addition, the results from the PISA study in 2006 showed that

Swedish students performed slightly worse than before (Skolverket, 2007) and the PISA study in 2009 showed even lower results (Skolverket, 2010). These studies naturally generated considerable discussion about the quality of Swedish education, discussions that were similar to those taking places in other countries such as Germany (See e.g. Gruber, 2006).

One example of governmental monitoring was the commission received by the Swedish School Inspectorate (SSI) in 2009 to re-evaluate a number of student booklets from the national tests. In their report they concluded that the teacher ratings and the re-rating differed to an unacceptable extent (Skolinspektionen, 2011). SSI also concluded that there were now too many purposes and that they were contradictory. The SSI report had a significant impact even if there has been some criticism of the methods used by SSI (Gustafsson & Erickson, 2013). One effect of the study was to reduce the number of purposes when the national tests were re-introduced after syllabus revisions in 2011. Only two purposes remained: the tests should support fair and equal assessment and grading; and the tests should provide information about the extent to which knowledge requirements are reached at different levels of the school system. However, it was still mentioned that the tests could support an interpretation of the syllabuses (See e.g. Skolverket, 2012).

As mentioned earlier, a new curriculum and new syllabuses were introduced in 2011 and a discussion regarding the quality of the tests and how they could fulfil the purposes had commenced. As a result of the discussions, a governmental investigation focusing on the national tests was assigned in 2014 and the report was published in March 2016 (SOU 2016:25). Among the recommendations on the national tests, two targeted the purposes and stakes of the tests. In the investigation it was suggested that the national tests should only support fair and equal grading. This suggestion clarified the changes in the discourse that had been ongoing for a while, where the formative and diagnostic purposes were no longer expressed. All discussions about the national tests as a tool to implement the curriculum and syllabuses were removed and the tests were no longer required to support teachers develop better classroom assessments, at least not as an explicit purpose. Further, it was also recommended that the student booklets should be anonymized and that the teachers should no longer score their own students' tests. It was also proposed that teachers should pay particular attention to the results from the national tests in the grading, which probably raised the stakes even more. These suggestions were included in the governmental proposition (Prop 2017/18:14) which was accepted and its recommendations will be implemented from autumn 2018.

In recent years, an emerging threat to validity has appeared, especially in mathematics but also in other subjects, since several of the test forms have been disclosed the day before the test day. This type of unauthorized disclosure is probably another sign of the increased stakes.

As presented above, the school system and the national testing system have undergone several changes, especially since 1994. In Figure 1, the changes in the Swedish national testing program from the 1960s until today are illustrated. The specified years in the figure show when changes in the testing system were introduced, changes that, as I see it, have affected the stakes of the tests. The level of the stakes in each section of the figure is not to be seen as an absolute value but rather an attempt to show my interpretation of how the stakes during different periods relate to each other.

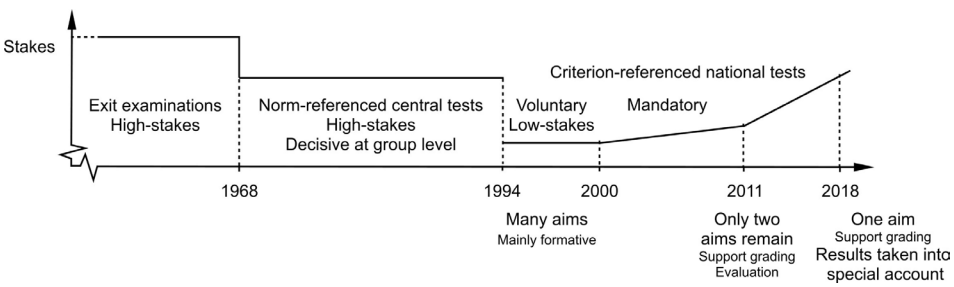

Figure 1. A schematic picture of how the stakes for the national tests has changed over time.

3.6 The national tests in mathematics

As presented in the previous chapters, many changes have been made to the Swedish national testing system over the last 50 years. These changes have also affected the subject-specific national tests and since this thesis is about validity connected to the tests in mathematics it is also necessary to also comment on the specific changes to these tests.

Mathematics is a subject that traditionally, in Sweden as well as internationally, has focused on content – such as geometry, arithmetic and algebra – but with an ambition to include problem solving activities (National Council of Teachers of Mathematics, 2000; Niss & Højgaard, 2011; Skolöverstyrelsen, 1970b)

Mathematics probably experienced the largest changes of any subject when the syllabus changed in 1994. This was because content was no longer specified in exhaustive lists. Instead, more generic areas could be found in the goals, while the grading criteria did not include any specified content at all.

Due to the specified purposes of the national tests at the time, problem-solving items dominated the Swedish national tests in the 1990s in order to give good examples of the interpretation of the syllabus. This meant that there were few but rather extensive items in the tests. It was also the intention to have open-ended items, which were possible to solve with different kinds of solution strategies and still reach a correct answer. As mentioned before, the tests were not mandatory and had very few regulations. It was possible to administer the

tests during a test period of two to four weeks depending on the course. The scoring guides were not particularly detailed and teachers were advised to score the national tests in the same way as they scored their own tests. For many items only the total score was defined and the teachers were to decide locally how far the student had to progress in order to receive a partial score. Cut scores for each grade were expressed as an interval, and teachers could decide which score to use from that interval.

When the tests became mandatory in autumn 2000, the administration of the tests became more standardized, and all tests were to be taken by students on a given test day. The number of items was increased slightly due to the inclusion of some short answer items and the scoring guide became more prescriptive by defining the minimum requirements for each score. Nevertheless, the system was still open to local interpretation since many of the items retained an openness regarding solution strategies.

In 2003 the first significant unauthorized disclosure of test materials occurred, which could suggest the importance of the tests to students and schools had increased considerably. Therefore, between 2004 and 2007, the mathematics tests were produced in two versions so that the consequences of unauthorized disclosure of test items would be reduced. Fortunately, during this time period none of the tests were disclosed before the test day. However, it was difficult to develop parallel versions of the tests each semester and it was expensive. There were also problems in the administration of the tests. All of these problems were reasons to return to a system with only one version of the test in each course.

From 2007, students were permitted to use advanced digital tools, for example tools with a Computer Algebra System (CAS) when solving certain items in the test. With CAS it is possible to solve equations algebraically instead of doing it by hand, which might reduce the risk of making errors. There were probably many reasons for this decision but one can assume that it was an ambition to increase the digital competence in mathematics and also show that digital technology is important in a modern society. There was, however, no discussion on how this decision could affect the national tests. At that time very few students had advanced calculators, most of them had function calculators, and students who were reading a lot of mathematics had graphic calculators. It had become more common that students had their own school computer but the computers were rarely used in the mathematics classroom, therefore the permission to use the advanced tools did not significantly influence the tests at that time. In the 2011 syllabuses the ambitions to increase the digital competence were more prominent and in mathematics it was highlighted that the students should work with and without digital tools (The Swedish National Agency for Education, 2012a). In the highest course it became necessary to have tools with a Computer Algebra System (CAS) in order to work with differential equations.

The changes relating to the use of digital tools and the changes to the syllabus in 2011, together with the change of purposes, where the national tests no longer

should support teachers to interpret the syllabuses had a larger impact on the national tests in mathematics than the changes in 2000. It became important to increase the reliability of the tests, which resulted in the inclusion of a larger number items that also were less extensive, in combination with longer testing times. The decision to allow the use of advanced digital tools together with the demand that students with advanced calculators should not have an advantage resulted in more items having to be solved without assistance from technology. This might be surprising since the purpose of allowing the use of more advanced digital tools was to increase digital competence. However, it became problematic in the test situation since all tools became allowed but the students were only required to have a scientific calculator for the lower courses and for course 4 a graphical calculator. Therefore, in the interests of fairness, it was seen as appropriate to include items that could be solved without tools or with a scientific calculator.

3.6.1 The development process of the national tests in mathematics

Ever since the mid-1990s and the introduction of the criterion-referenced grading system, the test development process for the mathematics tests for upper secondary school has largely remained the same. The process is guided by measurement theory, and adopts many of the recommendations in the latest version of the Standards for Educational and Psychological Testing (American Educational Research Association et al., 2014), hereafter called *Standards,* and other assessment literature (See e.g. Gronlund & Waugh, 2009; Haladyna & Downing, 2011). The challenges in the Swedish system are probably the same as in many other systems, such as being able to develop high quality test items, field testing the items in a setting that is similar to the setting when the items are administered in a national test, and defining cut scores that do not change the standards. Two of the benefits with the Swedish decentralized school system is that teachers are responsible for developing their own assessments and they are well aware of the knowledge requirements since they grade their students, which are important conditions in the test development work.

The development process of the national tests in mathematics includes a combination of qualitative and quantitative approaches but the foundation has always been to discuss and scrutinize the items and scoring guides in several rounds of review meetings with teachers and subject experts (Lindström, 2003). The compositions of the different test forms are based on a test specification for each course test. The test items are field tested several times but there are no formal equating procedures, procedures for ensuring that different test forms are of similar difficulty. The reasons for not equating the tests are many. From the beginning, equating was not discussed at all, probably because the tests would mainly be used to demonstrate possible interpretations of the syllabus and it was not so important that the difficulty of different test forms was comparable. At the beginning of the 21st century, when the grading

support purpose of the tests became more important, a discussion arose about the differences in the level of requirements for different test forms and there were suggestions for using anchor items, items with known characteristics, in the regular tests or in the field trials. The proposal to include anchor items in the regular tests was rejected because teachers do the scoring and therefore it would be difficult to keep the anchor items secret over time. Some attempts were made to include anchor items in the field trials but it proved difficult to include enough items to make it possible draw any useful conclusions. So far, since all data on the items in earlier test forms are saved in a database it has been possible and regarded as sufficient to compare and evaluate similar items and therefore develop test forms with roughly the same difficulty. The cut scores for each test grade in each test form are finally defined through a modified Angoff standard setting procedure (Angoff, 1971; Hambleton & Plake, 1995).

Even though the test development process for the Swedish national tests in mathematics is extensive, it is not sufficient to ensure quality in the test results. There are a few published studies focusing on quality issues connected to the national tests and one of them in mathematics is on reliability, more precisely classification accuracy (Nyström, 2004). In another study connected to quality, teacher made tests are compared to national tests in mathematics, focusing on what content and which competencies they measure (Boesen, 2006). This study indicates that the national tests include more problem solving and reasoning items, where the student has to use some creativity in order to solve them, than tests developed by teachers. In a recent publication, the national tests were analyzed in relation to what competencies they measure (Boesen, Lithner, & Palm, 2018). The analysis shows that all the competencies in the syllabus are fairly well represented in the tests. However, it also shows that very few, if any, of the items measure if the students are able to interpret or judge mathematical arguments, and argues that the purpose of the national tests concerning grading support would be better served by the inclusion of such items.

Other studies related to national tests consider which impact the tests has on students and teachers, mainly in compulsory school (See e.g. Bagger, 2015; Korp, 2006; Olovsson, 2015), or motivational aspects (Knekta, 2017). However, most common are descriptive reports with comparisons of the national test results in relation to the grades (See e.g. Skolinspektionen, 2012; Skolverket, 2009, 2017a). In these comparisons, a common picture that emerges is that the difference between test results and grades and the variation in difficulty between years is greatest in mathematics and least in English. There are probably many reasons for these results. One of the explanations is that there are fewer items in the mathematics tests than in the English tests. Therefore, the mathematics tests are, to a greater extent, affected by each item and by that a risk of variation in difficulty arises. The differences between test results and grades also indicate that the teachers are not using the tests as a support in the same way, which is not so surprising since how much and in what way the tests are to support the grading

is not regulated. However, the difference in the interpretations of the results from the national tests might be a threat to validity.

As described in this chapter, the importance, and therefore the stakes, of the national tests has increased, which implies a greater need to ensure that the test results are reliable and valid. It has also been said that the results are to be taken into special account in the grading, despite the fact that the test development process was put in place when the purposes of the tests were more formative and diagnostic and the stakes were lower. Therefore, there is a need to scrutinize whether the national tests in mathematics are able to support fair and equal assessment and grading and, with that scrutiny, highlight potential threats to validity.

4 Validity

Validity is a key concept and of central importance for all testing, assessment and evaluation practices. At the same time, validity is probably the most complex consideration in the process of developing tests. The understanding of validity has been the subject of considerable debate and the definition of validity has shifted over time. As the title suggests, validity is a central concept in my thesis. The main aim is to investigate and find evidence for the validity of the Swedish national tests in mathematics but also detect possible threats to the validity of those tests.

The definition of validity used in the thesis comes from the latest edition of the *Standards*: "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." (American Educational Research Association et al., 2014, p. 11). There are some important parts in the definition that need to be highlighted. Firstly, validity is connected to how the test results are *interpreted* and *used* and not to the test itself. In addition, if a test has several purposes the validity has to be investigated in relation to each purpose. Secondly, validity is a matter of degree: that is, the interpretations and uses can be more or less valid. In relation to the national tests, the definition can be understood as investigating whether the test scores reflect the knowledge of the syllabus and that the scores are not systematically affected by other factors. Validation is therefore the process of accumulating evidence that demonstrates the interpretations and uses are sound and trustworthy (and fails to find evidence to the contrary).

In order to understand the latest version of the *Standards* this chapter begins with a historical overview of how the concept of validity has developed over time. This is followed by a section that deals with how validation can be performed systematically. The chapter concludes with a justification for the choice of validation framework used in the thesis and a review of the validity evidence that already exists.

4.1 Historic development of the concept of validity

Even if the history of measurement goes back more than 1500 years to the examinations to become an official in Imperial China (Gregory, 2004), it was not until the beginning of the 20th century that validity became a concept in the context of measurement. Initially, the concept was used to describe the representativeness of chosen test items. In 1924, Giles Ruch wrote what has become the classic definition of validity:

> By validity is meant the degree to which a test or examination measures what it purports to measure (Ruch, 1924, p. 13).

A rather common approach to finding evidence of validity was to relate the test results with some external criterion that the test was supposed to measure and the general statistical approach was correlation. Over time, the belief in correlations as a proof for validity went so far that Guilford (1946) stated "In a very general sense, a test is valid for anything with which it correlates." (p.429). Thus high correlation between the observed score from a test and the criterion was taken as evidence of high validity. The early traditional view of validity is therefore often referred to as *criterion-referenced validity* (See e.g. Kane, 2001a; Shepard, 1993). However, Newton and Shaw (2014) argue that the discussion of validity in the early twentieth century was more elaborate than just criterion referencing and there is much written about quality in tests from that time; however, there was no consensus in the discussions and therefore no seminal papers emerged.

During this time, a discussion about content emerged and the requirement that a test not only correlates with some criterion but also measures some content of interest (Kane, 2006). For example, a test in mathematics should reasonably contain items that reflect the curriculum. In other words, *content validity* also became important.

The third type of validity, *construct validity*, was introduced by Cronbach and Meehl (1955), and refers to a situation where there is no obvious content to assess or criterion to relate to but rather qualities that are not direct measurable. Even if they did not propose that construct validity was better than criterion and content validity they argued that all tests should have construct validation and thus that construct validity should be regarded as the most important form of validity. In his chapter in the second edition of *Educational Measurement* Cronbach (1971) emphasized the importance of construct validation for all types of tests.

Into the 1970s, the three types of validity were regarded as distinct aspects of tests. However, the 1974 *Standards* (American Psychological Association, American Educational Research Association, & National council on Measurement in Education, 1974) argued that the three aspects were interrelated, although they also said they could be treated independently for the sake of convenience.

4.2 Messick and the unification of validity

The years from the middle of the 1970s to the millennium can be called the "Messick years", due to the influential work of Samuel Messick. Even if much has been attributed to Messick, he was not the only person working on validity during this period. Cronbach (1971), Linn (1980), Hambleton (1980) and Haertel (1985), to name a few, all contributed to the discussion. However it was Messick's seminal chapter on validity in the third edition of *Educational Measurement* (Messick,

1989) that has shaped modern discussions around validity. In the first paragraph in the chapter Messick begins with:

> Validity is an integrated judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on the test scores or other modes of assessment. (p. 13, Italics in original)

The quote highlights that it is not the test itself that is to be validated but rather the inferences from the test scores, a view initially proposed by Cureton (1951) in the first edition of *Educational Measurement*. It is also clear that it is not only tests with test scores that are to be validated. In addition, the quality of all measurement activities that serve as a basis for decisions have to be evaluated. However, even though the focus has moved from the test itself to the interpretation of the test results, this does not imply that the quality of the test is unimportant. It is still necessary, albeit insufficient, that the test instrument is of high quality.

In his chapter, Messick (1989) argues and concludes that construct validity is an overarching concept and that both content and criterion referenced validity, as well as reliability, are parts of construct validity. In this definition of validity, Messick included an ethical aspect arguing that unintended uses of test scores also need to be investigated.

Messick (1995) divides the sources of invalidity, or threats, into two main groups; construct underrepresentation and construct-irrelevant variance. Construct underrepresentation appears if the test is to narrow, and does not include all important parts of the construct. This is a threat apparent in most tests since it is usually necessary to limit what is being tested and only include a selection of items. Therefore, the selection has to be well-grounded. The second threat, construct-irrelevant variance arises if the test results are influenced by irrelevant factors, for example if the language is so difficult that the test result mirrors reading ability rather than for example mathematical knowledge.

The 1999 *Standards* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), included Messick's work on validity, which can be seen as a recognition of his work. Even if Messick's view on validity has been adopted and widely used by the measurement community, it has also been criticized. One of the perceived problems with Messick's work on validity is that it is highly theoretical and rather complicated to grasp and understand, partly due to the complex language it employs. His approach has also been criticized for including social consequences (See e.g. Borsboom, Mellenbergh, & van Heerden, 2004; Lissitz & Samuelsen, 2007). The main criticism is that an already multifaceted concept has been made even more complex by the inclusion of ethics and consequences. However, at the same time there was an emerging discussion of the negative effects of

standardized testing and the misuse of test results (Gipps, 1994), a discussion supporting the importance of including the uses of the test when considering validity. My interpretation of this is that Messick's idea of including social consequences in the validation process was intended to suggest that the validation is not finished when the test results are interpreted, as in the traditional view of validity, and that is necessary to continue the validation process. Making social consequences part of the concept of validity is beneficial both for those developing the tests and for the test takers, who are affected by the consequences, since it demands a reflection on what the test scores actually mean and what conclusions from those scores are valid. Even if the Messick's unified theory of validity is accepted there has been and continues to be a lot of discussion on how the theory could be applied in practice and how validation of a particular test could be carried out in practice (See e.g. Cizek, Rosenberg, & Koons, 2008).

4.3 An argument-based approach to validity

As a response to the criticism regarding the applicability of Messick's theory of validity, the fourth edition of *Educational Measurement* contained a chapter called *Validation* written by Kane (2006). The main theme in Kane's work is how the validation can be divided into smaller parts and be implemented practically. Even if this chapter can be seen as Kane's seminal paper, much of his early work is also widely cited (See e.g. Kane, 1992, 2001a, 2002) and he has continued to elaborate on his model for argument-based validation (See e.g. Kane, 2008, 2013, 2016).

In argument-based validation, two types of argument are needed - the interpretive argument and the validity argument. The interpretive arguments specify the uses of the test results and the proposed interpretations. The validity arguments are validity indicators determined by evaluations of the interpretive arguments. Kane (2006) also argues that different interpretive arguments may have to be set out, depending on the purposes of the test that is to be validated. With this approach it becomes necessary to formulate the interpretive arguments, which will reveal what is important to investigate.

4.4 The validation approach in this thesis

Adopting the modern view of validity, the validation process needs to include a range of aspects in order to make conclusions about the overall quality. However, it is necessary to have some sort of framework as a basis for the validation, in order to avoid potential confusion. Messick (1989) proposed a model having six aspects to be taken into account when investigating construct validity: Content, Substantive, Structural, Consequential, Generalizability and External. However, as Kane (2006) concludes, one problem with this model is that it does not include instructions on how to draw conclusions about the validity evidence. In order to

facilitate the validation and make the process tangible, Crooks, Kane and Cohen (1996) suggested a structured model with clearly defined validation criteria. The model builds on an idea where the assessment process is described by eight distinct stages characterized as a linked chain, see Figure 2, and the validation process is about elaborating each of these links. The inferences from the validation process with this model can be compared to a chain. Just as a chain can never be stronger than the weakest link, overall validity cannot be higher than the lowest validity associated with any one of the links.

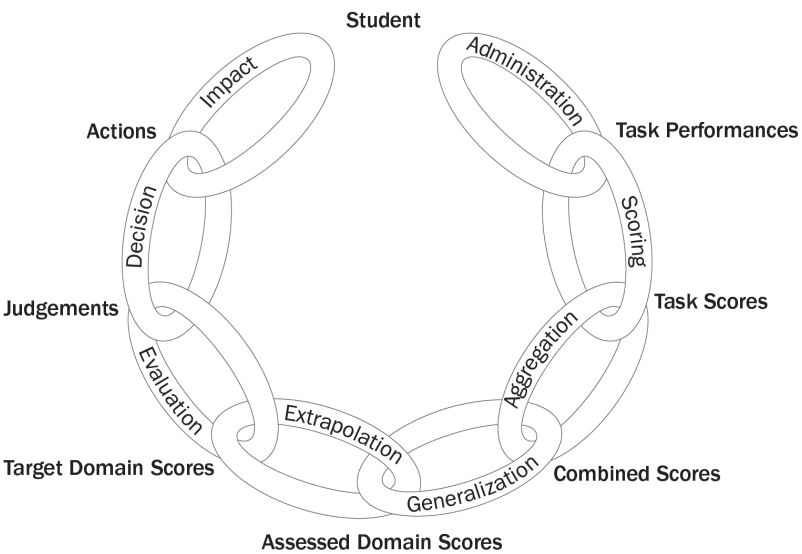

Figure 2. A model of educational assessment to use in the validation and planning of assessments proposed by Crooks, Kane and Cohen, 1996.

This model might be regarded as rather complex due to the many links. Indeed Crooks *et al.* (1996) note that some of the links in their model are combined into one link in some other validation models: scoring, aggregation and generalization, for example, are often merged into a generalization link. However, they argue that such a combination of links might make it more difficult to detect some of the threats, since they might be hidden in the aggregated link. On the other hand, for anyone who has been involved in tests from assessment to decision, the model seems very reasonable since the links follow a recognizable scheme.

Crooks *et al.* (1996) point out the need to elaborate threats in each of the links in relation to the specific purposes that are to be validated. They also identify several threats that might appear in each link, although they note that the list of threats associated with each link should not be regarded as comprehensive. The authors also emphasize the importance of focusing on the threats to validity

that really matter in relation to the purposes. Kane, Crooks and Cohen (1999) argue that for performance assessments, that is assessments where answers need to be evaluated by a human being, special attention should be given to the generalizability of the results over different administrations, scorers and tasks included.

The general framework for the Swedish national tests (Skolverket, 2017b) uses the validation process proposed by Crooks, Kane and Cohen (1996) for the national tests. For this reason, it seems reasonable to also use their model in the work described in this thesis.

This model has also been used to analyze the validity of the National Curriculum assessments in England (Stobart, 2001). In the study by Stobart, the specified threats in Crooks *et al.* (1996) are the threats investigated in relation to the National Curriculum assessments. One of the conclusions in the paper is that even if the eight-link model facilitates a structured validation, it still becomes complicated. Stobart (2009) also concludes that one factor that might complicate the validation is that the National Curriculum assessments consists of two parts, teacher assessments and a curriculum test, and it is only the latter which can be analyzed. The National curriculum assessments also have several purposes and that makes the validation even more difficult (Stobart, 2009).

Even if Stobart (2009) proposes a slightly simpler framework as a basis for the validation of the National Curriculum assessment, it is probably the many purposes, not the framework itself, that makes the validation complex. Therefore, I stand with the choice of using the chain model by Crooks *et al.* as validation framework. In order to identify and visualize possible threats it becomes necessary to evaluate each of the links in relation to the features of the national tests and their purposes, as described in the following section.

4.5 The chain model in relation to the national tests in mathematics

As stated before, the focus of this thesis is the national tests in mathematics for upper secondary school. Since all the tests from the second course and upwards are developed with similar processes and have similar appearances much of the benefits and the threats are the same regardless of course. The validation will be done in relation to the purpose that the national tests are to support fair and equal assessment and grading. In the following part of the chapter, each of the eight links is analyzed in relation to this purpose in the context of the Swedish national tests in mathematics.

4.5.1 Administration

For this link Crooks *et al.* highlight the threat of *inappropriate assessment conditions* in relation to the national tests. This threat might arise if testing time, tools or other conditions differ between test takers. For the national tests there is

a given test date and it is not allowed to administer the test before that date. However, the school principal can decide, if there are special reasons like accidents or major diseases, to postpone the test for all students at the school. However, it is rather uncommon, especially in mathematics, since these tests have test dates late in the semester.

Having a system where everything surrounding the administration of the tests has been delegated to the schools raises concerns about equality and fairness. Today the national tests have a specified test day and since some years ago the national tests have to start no later than 9 am, which prevents schools from starting the test late, thereby preventing students from obtaining information about the test. Schools are also given comprehensive guidelines about how to administer the national tests, before, during and after the test session. The SSI have undertaken inspections to check that schools, in this case compulsory schools, are following the instructions, which a vast majority of the schools seem to do (Skolinspektionen, 2017). Nevertheless, several of the national tests in mathematics have been disclosed before the test day. Thus the administration of the tests by schools is a potential threat to validity since trust in the national test system might decrease and therefore may not be used to support decision-making, something that needs be studied further.

Another issue connected to inappropriate assessment conditions, is investigated in the first article in this thesis and is concerned with students taking the test under different conditions (Lind Pantzare, 2012). The specific threat to validity is how the use of a calculator with CAS influences the administration of the national tests in mathematics. As mentioned before, from 2007 it is permissible, but not necessary, to use CAS or tools with dynamic geometry (such as GeoGebra), when solving parts of the national tests where digital tools are allowed. However, these tools are rather expensive and are thus tools only for those who could afford them. Since these kind of tools are able to solve complicated equations their use by comparatively wealthier students could affect the test results, and hence the fairness of those results.

An additional threat connected to administration is *assessment anxiety*. The rise in the stakes may also increase test anxiety, which could have a negative impact on the effort students put in when sitting the test. There are few studies investigating anxiety for Swedish students at upper secondary school but a recently published thesis focusing on motivation shows that students are more motivated but experience higher anxiety when sitting a regular national test than a field trial (Knekta, 2017). However, the results show differences between the groups indicating that anxiety is not only connected to the stakes of the test. One can assume that external tests, like the national tests, will probably raise the anxiety, for both teachers and students. The differences between the groups could be a result of how the teachers view the national tests. Since the test results can be used to evaluate teaching, the comparisons of test results and grades are important to teachers and the source of considerable anxiety to them.

4.5.2 Scoring

The risk that the *scoring fails to capture important qualities of task performance* might be a problem if there are several paths to a correct solution or several correct answers but the scoring guide is too general or only shows one of them. This is something that should be investigated during the test development process. The scoring guide for the national tests is developed together with the items, which makes it possible to also evaluate how they work via the field trials. At least, this makes it possible to detect whether or not the teachers who are scoring the answers understand how to apply the scoring guide.

If the scoring guide does not function as intended, there is an obvious risk of *lacking interrater reliability*, which means that different raters come to different conclusions about the correctness of an answer. The second study in this thesis investigates interrater reliability in the mathematics tests (Lind Pantzare, 2015). When a test includes items that need to be scored manually it is important that it does not matter who does the scoring. In systems like the Swedish system where it is common for a teacher to score her or his own students and where there is no system of control such as training and monitoring, this is a threat that needs to be investigated.

4.5.3 Aggregation

After scoring, the individual scores are often aggregated, most commonly by summing them in order to generate a total score. The threat associated with aggregation concerns the *diversity of tasks* where the aggregated sum might hide information, thereby limiting the generalizability and influencing the interpretations of the aggregated score. The results from the national tests are aggregated and reported as a test grade. The test grades have the same notation as the grades in the syllabuses and are decided via cut scores. This is probably the most visible support for fair and equal grading since the teachers will be provided with a summary of the results that have similar denotation as the grades. This is, simultaneously, one of the most problematic parts of the test development process since the cut scores for each new test form are set before the test administration. The standard setting is performed using a modified Angoff method (Angoff, 1971; Hambleton & Plake, 1995) and as mentioned earlier the Swedish national tests in mathematics are not regularly equated. One threat to validity is whether the cut scores are equally demanding in different test forms or not. In study III (Lind Pantzare, 2017), a judgmental standard setting with the Angoff method is compared to a statistical method.

4.5.4 Generalization

After aggregating the scores, a natural step is to investigate the generalizability or the reliability of the test results. Since the aggregated score is used to make conclusions that go beyond what is actually measured in the specific test form it is necessary that the scores are reliable. Reliability is connected to the number of

test items: the inclusion of more items will increase the reliability since a test with more items will generate more information. However, a test can normally never include more than a sample of all the possible tasks that can be included, and therefore it is important that the tasks in a test form assess the knowledge as reliably as the tasks in another test form. In tests like the national tests one threat to generalizability is *too few tasks*. When few tasks are included in a test form there is a risk that the total score is affected by chance since each task will have a large impact. However, the reliabilities for the national test forms in mathematics, measured using Coefficient alpha (Cronbach, 1951), are normally at least .90, which is the recommended level for test results that are to be used for decisions for individual test takers.

Even if the reliability is high on the aggregated level, the threat of having too few items in the generalization becomes even more pronounced if subscores, based on subsets of tasks, are reported. The interest in reporting subscores is also connected to the grading support. Since the goals in the mathematics syllabus are based on competencies, it may be interesting to consider the results associated with tasks evaluating each of these competencies. In the fourth study (Lind Pantzare & Wikström, 2018) the aim is to investigate if there is added value in reporting subscores in the national tests. This study deals with the generalization link but the aggregation, decision and impact links might also be affected.
Other threats connected to generalization are that *conditions of assessment are too variable* and *inconsistency in scoring criteria for different tasks*, and these threats are partly investigated in studies I and II.

4.5.5 Extrapolation

Extrapolation is concerned with drawing conclusions about features that go beyond the actual test. For example, if the syllabus includes goals on oral communication and there is no oral part, or if knowledge of all parts of the problem solving process is a goal but the tests only include multiple-choice items, then extrapolation can be problematic. The highlighted threats connected to this link are that *conditions of assessment are too constrained*, i.e. only one kind of item type or constraints in time, and *parts of the target domain are not assessed or given little weight,* which is especially problematic if the parts not assessed are weakly correlated to the parts included in the assessment.

Since it is not possible to assess everything in different ways with several item types there will always be a tradeoff between testing time, exhaustion and a variety of items. The Swedish national tests include a wide range of item types chosen to best suit what is being tested. However, choosing to use mainly constructed response items reduces the number of items and therefore the possibility to cover the target domain.

4.5.6 Evaluation

Evaluation is about putting a meaning to the scores. The threats to the evaluation link are especially apparent in the context of external tests, since those who use information from the tests are not always fully aware of how the results can and cannot be used. There is an apparent risk that those evaluating the tests have a *poor grasp of assessment information and its limitations*. In addition, teachers often have a lot of information about each student before they take the national tests so there is a risk for *biased interpretations or explanations* where the teachers evaluate the results in relation to what the student has achieved before.

The transformation of the result from a total score to a test grade is something that also could be associated with this link. The test grades are supposed to communicate that, on the basis of the test result, the student meets the knowledge requirements for the grade. Therefore both study III and study IV contribute validity evidence to this link.

As a side effect in study II, it is possible to obtain information relating to the threat of biased interpretations since the original scoring is compared to the re-scoring in the study.

4.5.7 Decision

The next step is to decide how to act based on the results. The Swedish national tests are not decisive when it comes to the final grading but are supposed to support assessment and grading. If the results from the tests are not used at all, or used differently by different teachers, there could also be a threat to validity. In addition, if teachers do not fully understand the composition of the national tests there is a risk for differences in the interpretations and the risk is then that the tests contributes to *poor pedagogical decisions*. Studies II, III and IV all provide information to use concerning the validity in this link.

4.5.8 Impact

The final link concerns the potential consequences of a test. The major threats are connected to the risk that *positive consequences are not achieved* and that *negative impact occurs*.

Since there are no guidelines on how supportive the national tests should be in the grading, it is possible that some teachers follow the results strictly and others do not take them into consideration at all. In both cases there is a risk for negative impact and that the national tests do not function as intended. There is also a risk that the results are used in a way that is not appropriate which can have undesirable consequences. The results from all four studies contribute to the validation in the impact link.

5 Materials and Methods

Writing a dissertation is an education in becoming a researcher and one ambition in my own education was to work with a wide range of methods, both qualitative and quantitative. It is necessary to choose methods that are suitable for the question being investigated. With quantitative methods it is possible to analyze large datasets, for example results from a test administration, and, depending on the quality of data, it is possible to draw general conclusions. However, in order to get more information of, for example how students think when they solve a test item, it might be necessary to also collect qualitative information as a complement to the quantitative data. In the validation of a test it is often necessary to use a mixed method approach to obtain a more comprehensive picture. Using a combination of qualitative and quantitative data has become more common since the combination of analyses takes the best from each of the approaches and gives a broad, deep and hopefully better understanding of the phenomenon investigated (Johnson & Onwuegbuzie, 2004).

5.1 Data collections and participants

The work in the thesis is based on two main types of data. The regular reporting from the national tests is one type and serves as the basis for study II, III and IV in this thesis. In study IV only data from the regular reporting of the national tests were used. In studies I-III specific data was required for the analyses as described below.

5.1.1 Think aloud study

In study I two types of data were collected. The main data consisted of video recordings of how eight students used an advanced calculator when working with mathematics items. All students were in their final year of upper secondary school (18-19 years old). The items were chosen among the released items from TIMSS advanced 2008 (Arora, Foy, Mullis, & Martin, 2009). This choice of data might be seen as rather strange when the aim is connected to the impact on the national tests. However, the question of particular interest in the study was to see if and how students use the calculator for solving equations, derive and calculate integrals. It was seen as problematic to use the types of construct-response items that are included in the national tests since they often require that the student has to grasp the problem and state the equation to be solved. The risk with these items is that the problem either is so easily solved by hand that the calculator is redundant or so complicated that the students never realize what to calculate and therefore never obtain a solution. The idea was to present tasks where it is obvious what to calculate but where the equations are a little bit complicated so the calculator would be useful. The national tests did not include such items at the

time of the study. Therefore, it seemed reasonable to choose TIMSS items since they are high quality multiple-choice items that the students had not seen before. The video only focused on the booklet containing the items, the calculator and what the students said when they were working with the items. This approach made it possible to get information about the solution strategies and how the students were using the calculator, information that could not have been obtained by only looking at the answers in the booklets.

The data in this study also included a questionnaire given to all students in the two school classes from where the eight students were chosen. The questionnaire was used for acquiring explanatory information, and focused on the students' opinion of mathematics in general and the work with calculators in particular.

5.1.2 Inter rater reliability study

In study II booklets from the regular reporting of the national tests were used. In connection to each test session, the schools are supposed to send copies of booklets for students born on certain dates to the test developing departments. A random sample of 100 booklets from one of the upper secondary school mathematics courses were used in this study. Five teachers, each with at least ten years of teaching experience, rescored the 100 booklets and all these scores together with the original scoring became the data in the study.

5.1.3 Validating standard setting

In study III two test forms and two types of data were used. The data for the first test form was the regular collection of data connected to the national tests. The data for the second test was from a group of students who sat a national test and an anchor test that contained items from the first test form.

5.1.4 Investigating subscores

In study IV, results from the regular collection of data connected to the national tests were used. Teachers are supposed to report results for each item, for students born on certain dates to the test development departments. All items are also categorized by the competence they are intended to measure which makes it possible to combine items measuring the same competence into a subscore. In this study, the subscores for six different national tests are investigated.

5.2 Methodological issues

Within the educational measurement field, the two main approaches used when analyzing tests, estimating item statistics and test taker characteristics are Classical Test Theory (CTT) and Item Response Theory (IRT). Classical test theory (CTT) is, as the name suggests, the more traditional approach (Crocker & Algina, 1986). The basic assumption of CTT is that an observed score, the result from a test, consists of a true score, which is not directly observable, and an error.

A fundamental problem with CTT is that the characteristics of an item, for example difficulty, depend on the ability of the test taker group. This drawback of CTT results in a situation where an item administered in a high ability group will seem to be easier than if it had been administered in a low ability group.

With Item response theory (IRT), or modern test theory as it sometimes is called, it is possible to separate test taker ability from item characteristics. This means that it is possible to decide test taker ability from any set of items and also decide item difficulty using the results from any group of test takers, at least if the model fits the data and the assumptions are fulfilled (Hambleton, Swaminathan, & Rogers, 1991; Lord & Novick, 1968). IRT has over the years become very popular and the default choice when analyzing test items, especially tests with items where there is only a right or wrong answer, such as multiple choice.

Traditionally, test developers have only used CTT-methods to analyze the Swedish national tests in mathematics (Lindström, 2003). There are several reasons for this choice but the most obvious is that the characteristics of the items, where most of the items are partial credit items and very few are multiple choice, means that the items have to be analyzed with a polytomous IRT model (Muraki, 1997; Samejima, 1969). These kinds of models are rather complex to handle and for a number of the mathematics courses it is problematic to collect the volumes of data needed in the analyses.

For the papers in the thesis it would probably not have made any difference if IRT methods had been used instead of CTT methods. In study I, none of these methods were applicable. In study II, interrater reliability was analyzed with percent correct and Cohen's kappa (Cohen, 1960). In study III, the equating was done using a CTT-method mainly because of the rather small sample of data, which is a shortcoming even with the CTT method. In study IV, the subscore analysis, only basic statistics like reliabilities and correlations were needed.

Irrespective of which methods are used, one challenge with the analyses is that for several of the courses the data is rather skewed. The skewness arises mainly because of the connection between the different mathematics courses and the types of programs that the students follow. Therefore, some of the courses have almost no students reaching the highest grades while in other courses the results are spread over the entire range. Another complexity in the analysis is that there are items in the tests that are very easy, while other items are rather difficult. Since the courses are criterion-referenced there will be items in the tests that have to be included to ensure construct representation and either almost all students or hardly any students manage to answer them. I have made some attempts to analyze the items using IRT but the results show that it is problematic to fit the IRT models to the items.

5.3 Ethical considerations

Being part of the working group that develops the national tests, while at the same time conducting research on those tests, raises certain ethical concerns. Scrutinizing material with which one is very familiar can of course be problematic, while, at the same time, it might be necessary to have quite detailed knowledge about the process and the structure of the tests in order to conduct the studies. However, there is an obvious risk that I am unable to be as critical as I could have been if the research had been on tests developed somewhere else. To make it clear, I have not been responsible for the everyday work of developing, analyzing and reviewing the items in any of the analyzed tests in the studies. I have only been part of the more general discussion on principles for the development process.

Investigating the quality of the national tests is an essential part of the development process. Being the project leader for the national tests in mathematics, I have, in the work on the studies in this thesis been aware of my involvement in the test development. However, as the project leader it is also my mission to develop assessments that are as valid as possible and therefore it is necessary to study what is functioning well and what has to be developed further. It is of course an advantage if the analyzes show that the tests are of high quality and generate valid interpretations. On the other hand, it is important, and a part of the commission, to detect flaws that have to be addressed. It is also necessary to conduct studies of those parts that are more complicated to investigate even though the results might indicate shortcomings, since they are more useful when refining the development process.

The choice of themes in the studies also indicate that I have not been afraid of investigating parts of the development process for which there may be doubts. For example the interrater reliability and standard setting are two such themes that had never been studied before and for which the results could have been problematic from a project leader point of view, still I chose to include them as a part of the thesis.

6 Summary of the papers

This chapter provides a summary of the four papers attached to this thesis. The four studies relate to different aspects of validity for the Swedish upper secondary school national tests in mathematics.

6.1 Paper I

Advanced calculators with, for example, computer algebra systems (CAS) are powerful tools when solving mathematics tasks since they can handle algebraic expressions. Much of the research associated with such calculators has focused on the pedagogical benefits of using calculators in the classroom (See e.g. Guin & Trouche, 1999; Kendal & Stacey, 2002; Kieran & Drijvers, 2006). However, in an assessment situation when some students have such devices and others do not, the use of the calculator can introduce bias. In this case, differences in students' performance might not be due to differences in ability but rather differences in the availability of advanced calculators. In addition, the use of the calculator can make it difficult to draw valid conclusions from the results due to differences in solution strategies. Even if previous research has identified the pedagogical advantages of using advanced calculators, no one has investigated the effect of such calculators on student performance in an assessment situation.

The aim of the study was to investigate how the use of CAS calculators influences the validity of an assessment by studying what students actually do if CAS calculators may be used in the work with mathematics assessment tasks. In this study, the main approach was a think aloud part where eight students from two classes in the third year of upper secondary school, who had been working with a CAS calculator for nearly two years, participated. In addition, all students in the two classes answered a questionnaire.

In the think aloud part, the students worked with five released items from TIMSS advanced 2008. The items chosen included different mathematical content and could be solved by hand. However, using the CAS calculator was believed to reduce the risk of making errors in four of the items. In addition, four of the items were multiple choice or short answer items in order to make it obvious what to calculate.

The think aloud study showed that all of the students tried to solve the items by hand to start with and few used the features of CAS at all. Out of 32 possibilities (eight students and four tasks) where the CAS calculator could be used only five correct answers were reached with the calculator. The think aloud part also revealed that it was mainly the high achieving students who were able to use the CAS calculator in those cases where they could not solve the item by hand. This was a somewhat surprising result since the hypothesis was that low-achieving

students could be expected to benefit more from using the calculator since they are more likely to make errors when calculating by hand.

The questionnaire was only used to gather some background information that could support or refute the explanations of the results. More than half of the students stated that mathematical solutions have to be calculated by hand even if use of a calculator is permitted. The questionnaire responses also showed that many of the students never use the calculator to solve simple equations but may use one to solve integrals or trigonometric equations.

The results indicate that in this group, the rank ordering of the students is not changed when CAS calculators are used, but the difference between high-achieving and low-achieving students was increased. However, those students who are mainly high-achieving and use a CAS calculator gain an additional advantage indicating therefore that validity can be threatened.

6.2 Paper II

One important part of the validity of an assessment is that the results are reliable and that it should not matter who did the scoring. In many assessment systems a wide range of mechanisms are put in to place in order to guarantee the rater reliability. Normally the scoring is centralized where expert scorers receive training and the scorers are then monitored and moderated if needed (Arora et al., 2009; Baird, Greatorex, & Bell, 2004). However, there are systems where the teachers have full responsibility to score the answers with only the help of a scoring guide and no other training. The aim of this study was to empirically examine the interrater reliability of teachers' ratings of students' performance on a Swedish national test in mathematics for upper secondary school where there are no external quality controls. These tests mainly consist of constructed response items that require manual scoring.

Since there is no regular control of the interrater reliability connected to the Swedish national tests there was a need to set up a special study by using booklets collected when the national test was administered. Five upper-secondary teachers, each with more than ten years of teaching experience, were commissioned to rescore a random sample of 99 student booklets. Initially there were 100 booklets but one booklet was not rescored since some tasks were missing. The booklets were scanned and the original scoring and teacher comments were removed in order to avoid influences on the rescoring. The scorers in the study were given a copy of the test and the scoring guide together with the student booklets. This is the same information as teachers normally receive when scoring the national tests.

The interrater reliability was analyzed for the whole test as well as item per item with respect to Stemler's (2004) categorization of consensus, consistency and measurement estimates. The results show that the interrater reliability is acceptable and even fairly high according to the recommendations in the

literature (See e.g. Landis & Koch, 1977). The study also showed that items with more than one possible solution strategy, where the scoring guide only can handle one of these strategies, are often more difficult to score. In addition, the analysis gave information about problematic scoring guides for specific items, information used in the development process for future tests.

The conclusion from this study was that it is probably not worth the money or the effort to introduce a control system with training and monitoring, at least not for the tests in mathematics. However, for complex solutions in mathematics it seems to be more difficult to achieve high interrater reliability, which suggests that training might be necessary for items with complex answers, not only in mathematics but also in other subjects.

6.3 Paper III

The programs in Swedish upper secondary school are built up by courses, and depending on what program the students follow the courses could be finished in one or two semesters. This leads to a situation where students taking the same course but not at the same time sit different forms of the national tests. Even if the ambition is to develop valid and equivalent test forms, it is difficult to accomplish in practice especially since a majority of the items are constructed-response. In addition, the test grades for each test form are defined by cut scores, which are set individually for each test form and before administration. The aim of this study was to investigate how a validation of the cut scores for parallel forms of national tests in mathematics can be done.

The cut scores for the Swedish national tests in mathematics are decided via a standard setting procedure using a method developed by Hambleton and Plake (1995). This method is a modification of the well-known and commonly used Angoff method (1971). However, since the standard setting for each test form is done on a per-test basis and since the standard setting method relies on a panel of judges estimating the difficulty of each item it is not possible to know if the cut scores for each test form are equally demanding. Even if the standard setting procedure follows the recommendations in the literature (See e.g. Cizek & Bunch, 2007), there is no regular control of the equivalence in the cut scores between test forms. In order to validate the standard setting Kane (2001b) recommends a validation including three parts. First, follow the recommended procedure in order to have procedural validity. Second, evaluate the consistency of standard setting and, third, evaluate the standard setting towards an external criterion. The first and second parts are normally done since the standard setting procedure generates data that can be used in the analysis. In order to accomplish the third part it is necessary to collect other information. In this study the judgmental linking, that is the Angoff standard setting procedure for two parallel test forms, was compared to a statistical linking using a non-equivalent group anchor test (NEAT) design.

In the study, two consecutive spring test forms were included. The anchor test contained items included in the first test form. The students who sat the new test form (the second) were also administered the anchor test in the subsequent mathematics lesson. Since the group who were administered the new test form and the anchor separately was rather small it was only possible to use mean and linear equating methods (Kolen & Brennan, 2004).

The results show that the judgmental standard setting and the statistical linking yielded almost the same results for the cut scores for the higher grades C and A but the statistical linking recommended that the cut score for the grade E should be one or two points below the standard setting result. However, there are some limitations of the study that should be mentioned. The anchor was internal for one group and external for the other, which could have influenced the results. The students who received the anchor externally were told it was a test where a good result could raise their grades, but still they knew it was not the national test and therefore the stakes were not as high. Also the second group was rather small, which made it impossible to use equipercentile equating, something that could have been beneficial since the results are slightly skewed to the left.

6.3.1 Errata paper III:

Page 156, Table 9.4, column 1, row 2 it should be “Equated score linear equating, statistical linking”.

6.4 Paper IV

A significant threat to validity arises if the proposed interpretations of the results are based on assessments that are not developed for those purposes (American Educational Research Association et al., 2014). There has been increased interest in analyzing and reporting subscores (Sinharay, Puhan, & Haberman, 2011). One reason for this might be an increasing demand to not only report a total score or a test grade, which is the normal case for summative assessments, but also give formative information to the students and the instructor (Haladyna & Kramer, 2004). However, even if the validity of the interpretations for the total score is investigated carefully, the quality of the subscores might not be studied at all. The purpose of the study was to investigate if it is psychometrically sound to present subscores for the Swedish national tests in mathematics.

Taking inspiration from NCTM standards (National Council of Teachers of Mathematics, 2000), the aim in the Swedish syllabus for upper secondary school mathematics is that the teaching should give students the opportunity to develop a number of competencies. In addition, the knowledge requirements are based on the competencies together with value expressions to illustrate the progressions among knowledge levels (Skolverket, 2011). Consequently, the national tests in mathematics have adopted a categorization connected to these competencies and

it has been suggested that the results from the tests should also be reported on a per-competency basis.

Competence-based subscores for six test forms were analysed using a method developed by Haberman (2008). In the study, the spring test forms from 2015 and 2016 for three different courses at upper secondary school – 2b, 3c and 4 – were analyzed.

The results show that none of the subscores have any added value and therefore do not need to be reported. However, when it comes to test development it is necessary to monitor and verify that each new test form follows the blueprint and the categorization of the competencies is a necessary part of this process.

7 Discussion

The overarching aim of this thesis was to examine evidence for the validity of the Swedish national tests in upper secondary school mathematics, thereby identifying potential threats to validity that may affect the interpretations of the results, rendering invalid conclusions. Validation involves an extensive process of accumulating evidence supporting the proposed interpretation and use of an instrument or assessment system. As stated in chapter 4, the validation process in this thesis was guided by the *Standards* (American Educational Research Association et al., 2014) and the validation chain framework developed by Crooks, Kane and Cohen (1996). The validation was made in relation to the purposes for the national tests, which is to support fair and equal assessment and grading. More specifically, the focus was to investigate how differences arising from the use of digital tools, different scorers and the standard setting process affects the results, and also to investigate if subscores can be used when interpreting the results. The thesis is based on four empirical studies connected to different parts of the validation chain.

The purposes for the national tests have changed over time, from a more diagnostic and formative approach towards an approach focused on fair and equal grading. Until about 2008 the debates and discussions on the national tests were based on a belief that only by providing national tests could teachers understand what to teach and how do develop fair and equal tests by themselves and consequently grade their students in a fair and equal manner. However, different studies have showed that there has been a certain amount of grade inflation also in subjects with national tests, and this has led to new regulations stating that results from the national tests are to be taken into special account when grading. With this new view on the tests and their importance much more emphasis has been put on reliability issues, leading to standardization of, for example, the date and time to start the test, objective scoring and the comparison of test results from year to year. Still, there is an ambition to include innovative and exemplary items and the wide range of competencies tested. All of these ambitions for the national test system together with the increased importance of the tests for students and teachers were the basis for my interest in investigating issues associated with the validity of the national tests in mathematics.

This concluding chapter begins with a summary of the main findings from the studies. I then discuss the research questions in relation to the validation framework. Finally, I discuss the limitations of this work, present ideas for future studies, and make some concluding remarks.

7.1 Main findings

This thesis contributes to an understanding of the quality of the national tests in mathematics, something rarely investigated before. By relating the validation of the national tests in mathematics to the chain model of Crooks, Kane and Cohen (1996), it is apparent that some parts of the validation must be addressed during the development process, while other parts have to be handled by regulations or by teachers at the schools. Relating to the first research question connected to fairness in relation to the use of digital tools, different scorers and cut scores, the results from paper I, II and III indicate that the results are not affected by these issues. However, the number of test takers in study I was small and it is necessary to repeat the study with more test takers in order to draw more firm conclusions. In addition, the revised syllabus in mathematics is supposed to increase the knowledge and use of digital tools and with that, the impact on the test results will probably change. At the same time, since it is clearly written in the syllabus and the tools have become more available as computer programs or applets, more students will have the opportunity to work with these tools and be able to use them when they take the test, which one hopes will make the conditions for test takers more equal.

According to the results in study II, the scoring of the mathematics tests seems to work well without any control, and it is questionable whether it is necessary or worth the money to introduce a central scoring system. It is my firm conviction that having teachers scoring the national tests is the largest regular in-service training for teachers in Sweden. Through the scoring process, teachers have the opportunity to discuss what kind of items can be used to assess different parts of the syllabus, what to value in the answers and what it takes to reach each test grade.

The results from paper III show that the judgmental standard setting yields almost the same results as if an objective statistical procedure is used. However, this study only reports a comparison of two consecutive test forms and highlights the need to develop a process where these comparisons can be made on a regular basis in order to ensure the consistency of the cut scores. Since the national tests only are supportive, the teachers may and do use other information when grading their students. The variation in the grading between years is smaller than the variation in the test grades (Skolverket, 2009).

Relating to the second research question, the investigation of the subscores in study IV show that they need not be reported since they give no added value. Today the cut scores are only dependent on the total score and should not be changed as long as the tests retain their current form. However, the teachers receive information connected to the competencies, mainly to help them to understand the connection between the items and the competencies in the syllabus. This is something that should be taken into consideration if this information is likely to mislead the teachers, especially if they believe that the

subscores are reliable and that it is possible to draw conclusions about the students' strengths and weaknesses in relation to the competencies. One important implication of this study is that it might be necessary to better inform and educate the users of the national tests about what the tests can and, more importantly, what they cannot be used for.

The main conclusion, then, is that the results from the four studies suggest the results from the national tests in mathematics can be used to support fair and equal grading. Nevertheless, there are threats that have to be taken into consideration and require further investigation. No matter how well the scoring is done or how stable the cut scores are, as long as teachers can use the results as they want when grading their students, the interpretations and uses of the test results may differ and might have consequences for the students both in the classroom and when, for example, they apply to tertiary education.

Below, these main findings are discussed in more detail in relation to the research questions and the purpose of the thesis.

7.2 Validity evidence and potential threats to validity in Swedish national tests

The validation chain contains eight interrelated links, so the final decision link is affected by all the previous links in the chain. In this following part, the validity evidence and possible threats connected to each of the links in relation to the national tests in mathematics will be discussed in more detail.

7.2.1 Administration

The specific issue investigated connected to this link is the question of how differences in the kind of digital tool that students use during the test session affect the test result. Students are required to have a graphical tool when they sit the national test for the highest course but, with the change in the regulations in 2007, students may also use other types of digital tools. From the beginning, the intention was that students with CAS calculators should not have any advantage by using the tool when taking the national tests. The reason for this was that CAS calculators were, and still are, quite expensive and very few schools provided them to the students since they were not mandatory. Performance in the national test in mathematics should not become a question of whether or not the students could afford to buy CAS calculators. The results in study I showed that the students did not use the tool as much as one could have imagined, especially not the weak students who could gain the most by avoiding algebraic mistakes. The study cannot give any clear answers why the students do not use the tools as much as they could have. However, this result is consistent with the findings of other studies (Weigand, 2017). It is still not so common to use CAS calculators in the mathematics classroom internationally. One of the reasons highlighted is that the

teachers have to rearrange the teaching and both teachers and students have to learn to use the tools.

The results of study I provide evidence supporting the view that the test results are or at least were at the time of the study not largely affected by the use of advanced digital tools.

7.2.2 Scoring

Quality of scoring is known to be a problem in large-scale testing, especially when scoring constructed-response items, for which extensive procedures involving training, monitoring and moderation may be required (See e.g. Arora et al., 2009; Baird et al., 2004). Since the scoring procedure of the Swedish national tests does not include any of these controls it seemed reasonable to investigate the quality of the scoring. Although earlier research has concluded that mathematics items are the easiest items to score reliably (Murphy, 1982; Newton, 1996), these studies were done in the context of training and monitoring the scorers.

The results in study II showed that the interrater reliability is acceptable and for many of the items fairly high, based on the recommendations by Landis and Koch (1977). The items having lower interrater reliability are items for which there are several correct answers or several different solution strategies that each leads to the correct answer. Since it is recommended (and usually only feasible) to have one scoring guide (Ahmed & Pollitt, 2011) there is a risk that the scorers may not recognize a solution as correct if it does not follow the guide. Also, solutions in which erroneous calculations lead to a correct answer are difficult to score reliably. Based on the results from this study, the scoring guides were revised by clarifying some of the general instructions and including more benchmarks for the items with more than one solution strategy. The interrater reliability for the tests developed after the revision of the syllabuses 2011 have been analyzed and the results show even higher agreement (Lind Pantzare, 2016). These results indicate that the scoring guide can account for the possible solution strategies that are present and that mathematics teachers are rather good at scoring the test items.

The studies of interrater reliability show that, at least in mathematics with the kind of scoring guides that are used, there is no significant risk that the solutions will be scored differently by different scorers and there are no strong indications that the teachers are biased in relation to their own students. Also, based on the results from study II, the scoring guide has been extended, rather than removing items of a problematic nature. However, it is not possible to conclude that all types of scoring guides work well. If the scoring guides are changed or other kinds of items and scoring guides are introduced it will be necessary to repeat the study.

7.2.3 Aggregation

Aggregation is an issue in all forms of item-based assessments aggregation since the information from each specific item becomes a part of the aggregated sum.

On the one hand, the aggregated sum is less affected by errors in the specific items but on the other hand the aggregated sum can hide specific shortcomings in the composition of the test form or shortcomings in student knowledge. In a criterion-referenced system like the one used in Sweden where the results are reported as test grades defined via cut scores it is important that the aggregation is made the same way for every new test form.

In study III, the process of defining the cut scores was investigated by comparing the judgmental standard setting with a statistical linking procedure. The results from the study show that the two methods provide approximately the same cut scores and the study supports the view that at least these cut scores are valid for the test forms considered. However, the study only considered two consecutive test forms, which is a shortcoming, and more studies are needed in order to draw conclusions that are more robust. In addition, the nature of the national test system, where it is almost impossible to include anchor items in the regular tests and where the student groups for some of the tests are skewed, contributes to the complexity of implementing these types of studies. Even if there exists literature on the best practice of equating (See e.g. Kolen & Brennan, 2004), there is a need to develop procedures that can be implemented as part of the regular test development procedure and not only in a special case like in study III.

7.2.4 Generalization

Many of the issues in the previous three links also affect this fourth link, especially the question of aggregation and cut scores. In addition, since the knowledge requirements are written in relation to the competencies and the regulation that the criteria for each grade have to be fulfilled as a whole, aggregation and generalization could hide shortcomings in the knowledge. Therefore, the issue of subscore reporting has emerged. In paper IV, the added value of reporting subscores is investigated and the conclusion is that with the kind of tests we have today the subscore results should not be reported, a finding consistent with other studies where the added value of subscores has been investigated (See e.g. Haberman, Sinharay, & Puhan, 2009; Puhan, Sinharay, Haberman, & Larkin, 2008). However, in order to be able to develop tests for which the aggregated results are valid, and where generalizations could be made in a valid manner it is necessary to ensure that the tests have a combination of items from all parts of the domain that is to be assessed. In the test development process it is important to ensure that all the different categories of items are included. Since items are categorized it has, as a result of the formative role of the tests, been seen as natural to also inform the teachers about the test model. As long as the results are only reported at the total score level the subscores do not influence the validity, but it is important that the information given to those using the test results is clear so no one draws erroneous conclusions from those subscores.

I also see the recent focus on reliability as a potential problem. It is of course, as discussed earlier, necessary to have high reliability in order to have validity but as I see it, the reliability has to be not too high and not too low, or "lagom" as we say in Sweden. Raising the reliability further could introduce threats like construct underrepresentation since some content and competencies in mathematics demand more complex items and replacing them with short answer or multiple-choice items might lead to that not all competencies will be assessed. Making the tests even longer, by increasing the number of items, will probably introduce an unacceptable burden on the test takers.

7.2.5 Extrapolation

As in all tests it is only possible to include a selection of items among all the items that theoretically could be included. Lack of validity connected to extrapolation is normally about the risk of including items that are too homogeneous leading to a narrowing of the curriculum. In standardized testing, it is common to use the multiple-choice format, yet conclusions are still drawn about the test takers' ability to communicate, solve problems, or reason (Kane et al., 1999). Ever since the introduction of the national tests in the middle of the 1999s there has been an ambition of "not only making the easy assessable what is assessed" (Erickson, 2017). Even if the Swedish national tests in mathematics consists of rather few items, these items are of different kinds, including oral parts, extensive problem solving and modelling items, and items demanding mathematical reasoning. It is not possible in a single test to include all kind of items in relation to all content and all competencies, but in the review process when the test forms are developed, much effort is put into selecting items that are a good representation of the domain as a whole. Studies analyzing the composition of the national tests with respect to the competencies included show that the tests are representative of the syllabuses (Boesen, 2006; Boesen et al., 2018). Also in relation to this link, the demand of high reliability might be a problem if items, that measure an essential part of the construct, are excluded because they reduce reliability. Such reduction might lead to construct underrepresentation and a reduction of the validity connected to extrapolation.

7.2.6 Evaluation

The threats connected to evaluation are not investigated as a special aim in any of the studies in the thesis. However, one important evaluation is the transition from total scores to test grades. Test grades are supposed to communicate how the results on the national tests correspond to the knowledge requirements. Since the national tests are course tests with the purpose of supporting the grading it is important that the composition of the tests together with the cut scores reflect the knowledge requirements accurately. The thorough qualitative review during the test development process together with the standard setting process are important mechanisms for ensuring that the evaluations of the test results are

valid. The rather small standard deviations of the cut scores in study III is one indication that the panelists agree on what is required for each grade.

In addition, in study II, on interrater reliability, the comparison between the original scoring and the rescoring shows a high agreement indicating that the students' own teachers are following the scoring guide and not using their additional knowledge of the students in an inappropriate, biased way.

The results in study IV also influence the evaluation link. It is problematic from a validity perspective if teachers use the results connected to the competencies when evaluating the level of student knowledge, since the competence subscores have no added value. It might even be deceptive since the low reliability, due to few items in the subscores, might introduce differences in the student profiles that are not real differences.

7.2.7 Decision

The decision link is problematic in relation to the Swedish national tests since it is not defined how the results from the tests are to be used in the grading. Some teachers follows the results strictly for all students and some teachers do not follow the test grades at all (Gustafsson et al., 2014). This is problematic from a validity point of view since, even if the national tests are of high quality and every link from administration to extrapolation is valid, when the results are used differently by the teachers the whole idea of the national tests is lost.

One indication of how well the results from the national tests are consistent with the teachers' perceptions of the students' levels of knowledge is to compare the course grades with the test grades. For the tests in mathematics about 80% of the students get the same grade as the test grade (Skolverket, 2018). As long as the tests only are supportive and when there are no regulations on how supportive the test should be, the agreement is unlikely to be 100%. Even if the Swedish system, where the national tests are supportive, might introduce differences in the decisions taken, I think it would also be problematic to have a system where the national tests are decisive at the student level. In such systems, all the work in the classroom becomes uninteresting and not worth anything and the grades only depends on how you perform on the test day. There have been discussions on the possibility of reintroducing a system where the national tests are decisive at a group level as it was with the central tests (SOU 2016:25). That might be one of few options available to limit the differences between test results and final grades without making the tests crucial for each individual student.

7.2.8 Impact

The impact link is affected by the shortcomings in all other links and there are many threats concerning impact. The most prominent threat is about misuse of the results. The results from the studies and the regular evaluations of the tests show that the teachers are in general pleased with the tests and that the results are in line with the information teachers had before the tests (Arbetsgruppen för

nationella prov i matematik, 2018). However, in the competition for students in the Swedish school market, national test results are used in the marketing of the school. With that competition for students, there is a prominent risk that teachers and principals are raising the stakes even more by trying to increase the results in order to be able to show that the school is performing well (Vlachos, 2018).

A threat to the impact link, often discussed in other countries, concerns the problems of teaching to the test and narrowing the curriculum. This is of course a problematic impact if the test is very different from the classroom activities. However, there have not been such discussions concerning the national tests in upper secondary school. In terms of the mathematics test, the teachers are pleased with the material and verify that the tests are in line with what is done in the classroom (Arbetsgruppen för nationella prov i matematik, 2018).

One positive impact and one of the strengths with a system where the teachers are involved in the scoring procedure of the national tests is the opportunity to discuss assessments in general and relate the structure of the national tests to the assessment practice at the school. The kind of items and the scoring guides that are obviously working well could perhaps inspire and support teachers when they are developing assessment material by themselves.

The proposal of introducing external scoring where teachers no longer score their own students' tests or only scoring some of the items is problematic when it comes to impact. If the teachers only see a subset of the items, their understanding of the composition of the tests might be even lower than it is today. That might result in a situation where some teachers trust the results without questioning anything and other teachers do not use them at all since they cannot understand how they are derived.

7.3 Implications for test development

The rigorous test development process where items and scoring guides are reviewed and field trialed several times seems to be working rather well and is essential for the production of high-quality tests. The quality of the scoring guides is especially important. However, the studies show that there is a need to consider what information about the categorization of the items should be included in the material provided to the teachers. The test developers have to use the categorizations in order to develop parallel test forms that are a good representation of the construct. Nevertheless, there is a risk that teachers will misuse the test results due to a lack of understanding of how the categorization is done.

Another result affecting the test development is the question of standard setting and equivalence. There is a need to continue the work of analyzing the cut scores and finding methods where it is possible to equate them and make them equally demanding from year to year. This is a significant challenge in the

decentralized system we have today, especially since the cut scores are decided before the test is sat by students.

7.4 Limitations and generalizations

The major limitation of the studies in the thesis is connected to the fact that Sweden is quite unusual when it comes to assessment and grading. Many of the threats to validity are related to the decentralized Swedish education system where the teachers have considerable responsibility and a lot of freedom.

One issue connected to the data from the regular reporting used in studies III and IV is that teachers and schools should report results item by item for students born on certain dates, to the test development departments, which is a prerequisite for getting a random sample. However, the test development departments do not know which schools are administering which tests and therefore it is not possible for them to know which schools have reported their results and not. In addition, there is no student identification, which makes it impossible to connect any background variables other than those reported, gender and preliminary grade. In order to validate the data reported to the department, the distributions of test grades are compared with the results collected by NAE, which is the test grade for all students taking each test. Generally, these distributions are very similar.

A more general limitation is that only some test forms in some courses in mathematics have been investigated. However, many of the threats analyzed in relation to the tests in mathematics are the same, irrespective of subject. The scoring guides have to work well in relation to the question asked, the cut scores have to be stable and how the results are used is important. Even if the results from the analyses of the mathematics tests do not give any information about the other national test subjects, I think that the structural validation procedure followed in the thesis could usefully be applied to other subjects.

One limitation with the validation model used in the thesis is what happens before test administration. Even if Crooks, Kane and Cohen argue that the validation chain can be used "backwards", starting at implications, in the test development process, it is my opinion that it is rather abstract to relate each of the links to something concrete to be done when developing the tests.

7.5 Suggestions for future studies

It would have been interesting to expand and repeat study 1 again, since the study was small and much has happened since then. There are two major changes that might affect the results obtained in this study: the revision of the syllabus in mathematics has led to a greater focus on digital competence; and many students now have access to better and more user-friendly digital tools. Both of these changes presumably contribute to a greater use of the tools in the classroom, and

students are like to be able to handle the tools more fluently. The possibility of using computerized tests, techniques where it is possible to include eye-tracking studies together with information about the work with the digital tool will also make it possible to include more students in such a study.

There is a need for further development of an equating procedure that is sustainable over time and also a further investigation on how well the statistical methods available are working in the Swedish national test context. One part of this work would be to introduce IRT in the development and analysis of the items in national tests in mathematics.

It would also be interesting to investigate how the national tests affect assessment materials developed by teachers, both positive and negative impact. The idea that has dominated thinking in this area from the beginning is that if the national tests included exemplary test items the teachers would also include such items in their own teaching and their assessments. It has also been suggested that a visual categorization of, say, content and competencies, would help teachers interpret the syllabuses. The question then is whether teachers would interpret and use this information in a consistent way.

The national tests are to be digitalized by 2022 according to current directives. The digitalization introduce both benefits and obstacles. In a system like the one operated in Sweden where many items require a constructed response and where students produce solutions that include algebraic expressions and pictures, digital tests may be problematic. The test systems are not built to handle mathematics items other than multiple-choice and other items with automatic scoring. In addition, neither students nor teachers are used to writing solutions on a computer and it might be difficult to develop items that cover the domain to be assessed. On the other hand, digitalization might make it easier to connect results from the field trials with those from the regular tests, which could facilitate item analysis and equating. This is an area where several validity studies have to be conducted.

7.6 Concluding remarks

The overall aim with the thesis was to examine validity evidence for the Swedish national tests in upper secondary school mathematics, thereby identifying potential threats to validity in relation to the purpose that may affect the interpretations of the results, resulting in invalid conclusions. I think that my work has contributed to an understanding of how the validation of a national test could be done, and also highlighted potential threats in relation to the national tests in mathematics in Sweden.

I hope that the work in this thesis can be an inspiration for other researchers who are interested in how to perform structured validation of tests.

8 Acknowledgement

Finally, I can see the finish line. I have been a doctoral student for a long time, at least when counting from when I started until today. However, during long periods, I have had other obligations, putting work on the thesis in stand-by mode. Nevertheless, in one way or the other, I have managed to return to the studies. Of course, I could not have come this far all by myself and there are some people that deserves to be mentioned. First, I want to thank my current supervisors, Christina and Pecke, you have been so supportive. Well aware of my situation you have pushed me, not too much and not too little. Thank you for gently forcing me to prioritize the work with the thesis. In our discussions, both in the formal supervising and during coffee breaks, you have helped me to see beyond all the details and instead focus on the "big picture". You have also given good advice concerning methodological issues, structure and language, thus making the work so much better. I also want to thank my first supervisors, Peter and Widar: you made me curious about research and gave me a solid 'measurement' ground to stand on.

Furthermore, I want to thank all colleagues in the national test project group, no one mentioned and no one forgotten. You are all doing incredible work developing high-quality national tests in mathematics and in the science subjects. You are a fantastic team to work with and one of the reasons why I have managed to continue the work with the thesis. Nevertheless, in the project group some people deserve to be mentioned in particular. Peder for always being ready to take care of all the messy issues I hand over to you. Lena, Jennie and Helen for being so well organized project administrators who have everything under control, and reminding us all of what needs to be done next so nothing is forgotten, and Maria, who has been a fantastic support in all kinds of personnel issues. You all make my life so much easier. I also want to mention the former project administrator Monika, I would not have managed my first years as a project leader without you.

I also want to thank Tova for valuable comments at the final seminar and for the inspiration to figure 1 and Björn for helping me with the cover art and everything needed for the printing. In addition, Lotta deserves a special thank you since she, as always, has taken care of all practicalities.

Outside the department I should mention Umeå Mathematical Education Research Center (UMERC). Via this research group I have had the opportunity to attend and give seminars in a friendly environment where my assessment focus has been scrutinized by researchers who have a teaching and learning focus in mathematics, something that has been fruitful for me and hopefully for all of you. In this group, I want to mention Ewa who has followed me all the way and I really value our lunch discussions over the years.

Finally, I want to thank my family. My parents, Karin and Torgny, who have always supported me, especially when it comes to education. I am so thankful for

everything you taught me when I grew up. My beloved husband Ulf, without you it would not have been possible to do this. I really appreciate that you continually remind me it has to be fun to go to work and that life contain other things than work. Thank you for suggesting various activities like going to concerts or the cinema, inviting people for dinner, riding motorcycles in the summer or just taking a walk. I am lousy at suggesting things but I rarely hesitate to come along. Last but no longer the least, or at least no longer the shortest, my daughters Ida and Agnes. You have become used to the fact that mom is often sitting at the kitchen table and working. However, helping you with the homework and following you to football and basketball games have been really good for realizing that there are other important things in the world.

Anna Lind Pantzare
Umeå, October 2018

9 References

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice, 18*(3), 259-278. doi:https://doi.org/10.1080/0969594X.2010.546775

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1974). *Standards for Educational and Psychological Tests*. Washington, DC: American Psychological Association.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-601). Washington, DC: American Council of Education.

Arbetsgruppen för nationella prov i matematik. (2018). Resultat kursprov [Results from the national tests]. Retrieved from http://www.edusci.umu.se/np/np-2-4/resultat/

Arora, A., Foy, P., Mullis, I., & Martin, M. (Eds.). (2009). *TIMSS Advanced 2008 Technical report*: TIMSS & PIRLS International Study Center. Boston College, Chestnut Hill, MA.

Bagger, A. (2015). *Prövningen av en skola för alla: nationella provet i matematik i det tredje skolåret [Trial of a school for all: the national test in mathematics for school year three.].* (Doctoral dissertation), Umeå Universitet, Umeå, Sweden. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:855578

Baird, J.-A., Greatorex, J., & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy and Practice, 11*(3), 331-348. doi:https://doi.org/10.1080/0969594042000304627

Baird, J.-A., Isaacs, T., Opposs, D., & Gray, L. (Eds.). (2018). *Examination Standards. How measures and meanings differ around the world*. London, United Kingdom: UCL IOE Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7-73. doi:https://doi.org/10.1080/0969595980050102

Blanchenay, P., Burns, T., & Köster, F. (2014). *Shifting responsibilities - 20 years of education devoution in Sweden: A governing complex education systems case study*. Paris, France: OECD Publishing.

Boesen, J. (2006). *Assessing mathematical creativity. Comparing national and teacher-made tests, explaining differences and examining impact*. (Doctoral dissertation), Umeå universitet, Umeå, Sweden. Retrieved from http://www.diva-portal.org/smash/get/diva2:144670/FULLTEXT01.pdf

Boesen, J., Lithner, J., & Palm, T. (2018). Assessing mathematical competencies: an analysis of Swedish national mathematics tests. *Scandinavian Journal of Educational Research, 62*(1), 109-124. doi:https://doi.org/10.1080/00313831.2016.1212256

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061. doi:http://psycnet.apa.org/doi/10.1037/0033-295X.111.4.1061

Bunar, N. (2010). Choosing for quality or inequality: current perspectives on the implementation of school choice policy in Sweden. *Journal of Education Policy, 25*(1), 1-18. doi:https://doi.org/10.1080/02680930903377415

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: SAGE Publications.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412. doi:https://doi.org/10.1177/0013164407310130

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46. doi:https://doi.org/10.1177/001316446002000104

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Group.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi:https://doi.org/10.1007/BF02310555

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443-507). Washington, D.C.: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281. doi:http://dx.doi.org/10.1037/h0040957

Crooks, T., Kane, M. T., & Cohen, A. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice, 3*(3), 265-286. doi:https://doi.org/10.1080/0969594960030302

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 621-694). Washington, DC: American Council on Education.

Danish Ministry of Education. (2017). Love og regler for de gymnasiale uddannelser [Laws and legislation for upper secondary education programmes.]. Retrieved from https://www.uvm.dk/gymnasiale-uddannelser/love-og-regle

DS 1990:60. *Betygens effekter på undervisningen [The effects of grades on the teaching]*. Stockholm, Sverige: Utbildningsdepartementet.

Dufaux, S. (2012). *Assessment for qualification and certification in upper secondary education: A review of country practices and research evidence.* Retrieved from https://www.oecd-ilibrary.org/docserver/5k92zp1cshvb-en.pdf?expires=1535288082&id=id&accname=guest&checksum=79124D6B94CFE02247F71587B900E722

Eklöf, H., Andersson, E., & Wikström, C. (2009). The Concept of Accountability in Education: Does the Swedish School System Apply? *Cadmo, 2*. doi:https://doi.org/10.3280/CAD2009-002006

Erickson, G. (2017). Experiences with standards and criteria in Sweden. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard Setting in Education* (pp. 143-160). New York, NY: Springer.

Eurydice. (2018a). Norway: Assessment in general upper secondary education. Retrieved from https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-39_en

Eurydice. (2018b). Iceland: Assessment in general upper secondary education. Retrieved from https://eacea.ec.europa.eu/national-policies/eurydice/content/assessment-general-upper-secondary-education-24_en

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 579-621). Westport, CT: ACE/Praeger.

Finnish parliament. (1998). *General Upper Secondary Schools Act (629/1998)*. Retrieved from http://www.finlex.fi/sv/laki/ajantasa/1998/19980629#L4P18.

Gipps, C. (1994). *Beyond Testing: Towards a Theory of Educational Assessment*. London, United Kingdom: Falmer Press.

Gipps, C., & Stobart, G. (2009). Fairness in assessment. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational Assessment in the 21st Century* (pp. 105-118). Dordrecht: Springer.

Gregory, R. J. (2004). *Psycological testing. History, principles and applications* (2 ed.). Needham Heigts, MA: Allyn and Bacon.

Gronlund, N. E., & Waugh, C. (2009). *Assessment of student achievement* (9th. ed.). Upper Saddle River, NJ: Pearson.

Gruber, K. H. (2006). The German 'PISA-Shock': some aspects of the extraordinary impact of the OECD's PISA study on the German education system. In H. Ertl (Ed.), *Cross-national attraction in education: Accounts from England and Germany*. Oxford, United Kingdom: Symposium Books Ltd.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*(4), 427-438. doi:https://doi.org/10.1177/001316446600600401

Guin, D., & Trouche, L. (1999). The complex process of converting tools into mathematical instruments: The case of calculators. *International Journal of Computers for Mathematical Learning, 3*(3), 195-227. doi:https://doi.org/10.1023/A:1009892720043

Gustafsson, J.-E. (2006). Indikatorer på kvalitet [Quality indicators]. In J.-E. Gustafsson (Ed.), *Barns utbildningssituation - bidrag till ett kommunalt barnindex*. (pp. kap. 4 s.40-62). Stockholm, Sverige: Rädda barnen.

Gustafsson, J.-E., Cliffordson, C., & Erickson, G. (2014). *Likvärdig kunskapsbedömning i och av den svenska skolan-problem och möjligheter [Equal assessment in and of the Swedish school - problems and possibilities]*. Stockholm, Sverige: SNS-förlag.

Gustafsson, J.-E., & Erickson, G. (2013). To trust or not to trust? - teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability, 25*(1), 69-87. doi:https://doi.org/10.1007/s11092-013-9158-x

Haberman, S. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204-229. doi:https://doi.org/10.3102/1076998607302636

Haberman, S., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*(1), 79-95. doi:https://doi.org/10.1348/000711007X248875

Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research, 55*(1), 23-46. doi:https://doi.org/10.3102/00346543055001023

Haladyna, T. M., & Downing, S. M. (2011). *Handbook of Test Development*. Abingdon, United Kingdom: Routledge.

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation & the Health Professions, 27*(4), 349-368. doi:https://doi.org/10.1177/0163278704270010

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 80-123). Baltimore, MD: Johns Hopkins University Press.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 41-55. doi:https://doi.org/10.1207/s15324818ame0801_4

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park/London/New Delhi: Sage Publications.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26. doi:https://doi.org/10.3102%2F0013189X033007014

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527. doi:https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2001a). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319-342. doi:https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

Kane, M. T. (2001b). So much remains the same: Conception and status of validation in standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53-89). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31-41. doi:https://doi.org/10.1111/j.1745-3992.2002.tb00083.x

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: ACE/Praeger.

Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher, 37*(2), 76-82. doi:https://doi.org/10.3102%2F0013189X08315390

Kane, M. T. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement, 50*(1), 115-122. doi:https://doi.org/10.1111/jedm.12007

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198-211. doi:https://doi.org/10.1080/0969594X.2015.1060192

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17. doi:https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Kendal, M., & Stacey, K. (2002). Teachers in transition: Moving towards CAS-supported classrooms. *ZDM, 34*(5), 196-203. doi:https://doi.org/10.1007/BF02655822

Kieran, C., & Drijvers, P. (2006). The co-emergence of machine techniques, paper-and-pencil techniques, and theoretical reflection: A study of CAS use in secondary school algebra. *International Journal of Computers for Mathematical Learning, 11*(2), 205-263. doi:https://doi.org/10.1007/s10758-006-0006-7

Knekta, E. (2017). *Motivational aspects of test-taking: measuring test-taking motivation in Swedish national test contexts*. (Doctoral dissertation), Umeå University, Umeå. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:1071134

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer Publishing.

Korp, H. (2006). *Lika chanser i gymnasiet? En studie om betyg, nationella prov och social reproduktion [Same chanses in upper secondary school? A study of grades, national tests and social reproduction.].* (Doctoral dissertation), Malmö högskola, Lärarutbildningen, Malmö, Sverige. Retrieved from http://muep.mau.se/bitstream/handle/2043/7717/HelenaKorpFINAL.pdf?seq

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159. doi:https://doi.org/10.2307/2529310

Lind Pantzare, A. (2012). Students' use of CAS calculators : effects on the trustworthiness and fairness of mathematics assessments. *International Journal of Mathematical Education in Science and Technology, 43*(7), 843-861. doi:https://doi.org/10.1080/0020739X.2012.662289

Lind Pantzare, A. (2015). Interrater reliability in large-scale assessments – Can teachers score national tests reliably without external controls? *Practical Assessment, Research & Evaluation, 20*(9). Retrieved from https://pareonline.net/getvn.asp?v=20&n=9

Lind Pantzare, A. (2016). *Bedömaröverensstämmelse på skriftliga delprov. En studie av interbedömarreliabiliteten vid bedömning av skriftliga nationella prov i matematik i gymnasieskolan. [Rater agreement in the witten parts. A study of the interrater reliability in the scoring of written national tests in mathematics for upper secondary school.* Tillämpad utbildningsvetenskap. Unpublished.

Lind Pantzare, A. (2017). Validating Standard Setting: Comparing Judgmental and Statistical Linking. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard Setting in Education: The Nordic Countries in an International Perspective* (pp. 143-160): Springer.

Lind Pantzare, A., & Wikström, C. (2018). Using summative tests for formative purposes. An analysis of the added value of subscores. *Manuscript submitted for publication*.

Lindström, J.-O. (2003). *The Swedish national course tests in Mathematics*. Umeå, Sweden: Umeå university, Department of Educational Measurement.

Linn, R. L. (1980). Issues of validity for criterion-referenced measures. *Applied Psychological Measurement, 4*(4), 547-561. doi:https://doi.org/10.1177%2F014662168000400407

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448. doi:https://doi.org/10.3102%2F0013189X07311286

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lundahl, C. (2006). *Viljan att veta vad andra vet. Kunskapsbedömning i tidigmodern, modern och senmodern skola. [The wish to know what others know. Knowledge assessment in the early modern, modern and late modern school.].* (Doctoral dissertation), Uppsala universitet, Uppsala, Sverige. Retrieved from http://journals.lub.lu.se/index.php/aio/article/view/16844/15222

Lundahl, C. (2016). Nationella prov - ett redskap med tvetydiga syften [National tests - a tool with ambiguous purposes]. In C. Lundahl & M. Folke-Fichtelius (Eds.), *Bedömning i och av skolan - praktik, principer, politik*. Lund: Studentlitteratur.

Lundahl, C., & Tveit, S. (2014). Att legitimera nationella prov i Sverige och Norge-En fråga om profession och tradition. [To legitimize national tests in Sweden and Norway-A question of profession and tradition.]. *Pedagogisk Forskning i Sverige, 19*(4-5), 297-323. doi:https://open.lnu.se/index.php/PFS/article/view/1397/1241

Lundahl, L. (2002). Sweden: decentralization, deregulation, quasi-markets-and then what? *Journal of Education Policy, 17*(6), 687-697. doi:https://doi.org/10.1080/0268093022000032328

Lundgren, U. P., Säljö, R., & Liberg, C. (Eds.). (2017). *Lärande, skola, bildning: Grundbok för lärare. Fjärde utgåvan [Learning, school, education: Book for teachers. Fourth edition]*. Stockholm: Natur och kultur.

Lyrén, P.-E. (2009). *A perfect score. Validity arguments for college admission tests.* (Doctoral dissertation), Umeå university, Umeå. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:231769

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education & Macmillan.

Messick, S. (1995). Validity of psychological assessment. Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. doi:https://doi.org/10.1037/0003-066X.50.9.741

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153-164). New York, NY: Springer.

Murphy, R. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology, 52*(1), 58-63. doi:https://doi.org/10.1111/j.2044-8279.1982.tb02503.x

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.

Newton, P. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, 405-420. doi:https://doi.org/10.1080/0141192960220403

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London, United Kingdom: Sage Publications.

Niss, M., & Højgaard, T. (2011). *Competencies and Mathematical Learning: Ideas and inspiration for the development of teaching and learning in Denmark (IMFUFA tekst)*. Roskilde, Denmark: Roskilde University.

Nyström, P. (2004). Reliability of educational assessments: The case of classification accuracy. *Scandinavian Journal of Educational Research, 48*(4), 427-440. doi:https://doi.org/10.1080/00313830420000245816

Olovsson, T. G. (2015). *Det kontrollera(n)de klassrummet. Bedömningsprocessen i svensk grundskolpraktik i relation till införandet av nationella skolreformer. [The controlling classroom. The assessment process in Swedish compulsory school in relation to the introduction of national school reforms.].* (Doctoral dissertation), Umeå universitet, Umeå. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:811225

Pettersson, A. (2004). *The national tests and national assessment in Sweden*. Retrieved from https://www.su.se/polopoly_fs/1.133072.1366873372!/menu/standard/file/Sw_test_ICME.pdf

Popham, J. W. (1999). *Classroom assessment. What teachers need to know* (2nd ed.). Needham Hights, MA: Allyn & Bacon.

Popham, J. W. (2003). *Test better, teach better. The instructional role of assessment*. Alexandria, VA: Association for supervision and curriculum development (ASCD).

Prop 1992/93:250. (1993). *En ny läroplan och ett nytt betygssystem för gymnasieskolan, komvux; gymnasiesärskolan och särvux. [A new curriculum and a new grading system for upper secondary school, adult education; upper secondary school for disabled and school for disabled adults.].* Stockholm, Sverige: Utbildningsdepartementet.

Prop 1997/98:169. (1998). *Gymnasieskola i utveckling - kvalitet och likvärdighet [Upper secondary school in development - quality and equality]* Stockholm, Sverige: Utbildningsdepartementet.

Prop 2017/18:14. (2017). *Nationella prov-rättvisa, likvärdiga, digitala [National tests-fair, equal, digital].* Stockholm, Sverige: Utbildningsdepartementet.

Prop 2017/2018:9. (2017). *Skolstart vid sex års ålder [School start at the age of six].* Stockholm, Sverige: Utbildningsdepartementet.

Puhan, G., Sinharay, S., Haberman, S., & Larkin, K. (2008). Comparison of subscores based on classical test theory methods. *ETS Research Report Series, 2008*(2), i-23.

Rambøll. (2013). *Evaluering af de nationale test i Folkeskolen*. Retrieved from https://dk.ramboll.com/-/media/files/rm/rapporter/nationale-test-2013.pdf?la=da

Ramstedt, K. (1996). *Elektriska flickor och mekaniska pojkar. Om gruppskillnader på prov - en metodutveckling och studie av skillnader mellan flickor och pojkar på centrala prov i fysik. [Electrical girls and mechanical boys. On group differences in tests - a method development and a study of differences between girls and boys in national tests in physics.].* (Doctoral dissertation Academic thesis), Umeå universitet, Umeå, Sverige. Retrieved from http://www.diva-portal.org/smash/get/diva2:156255/FULLTEXT02

Ruch, G. M. (1924). *The improvement of the written examination*. Chicago, IL: Scott, Foresman and Company.

Samejima, F. (1969). *Estimation of ability using a response pattern of grades scores (Psychometric Monograph No. 17)*. Iowa City, IA: Psychometric Society.

Shepard, L. A. (1993). Chapter 9: Evaluating Test Validity. *Review of Research in Education, 19*(1), 405-450. doi:https://doi.org/10.3102%2F0091732X019001405

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30*(3), 29-40. doi:https://doi.org/10.1111/j.1745-3992.2011.00208.x

Skolinspektionen. (2011). *Lika eller olika? Omrättning av nationella prov i grundskolan och gymnasieskolan - Redovisning av regeringsuppdrag Dnr. U2009/4877/G. [Equal or not? Rescoring of national tests in compulsory and upper secondary school.]* Retrieved from https://www.skolinspektionen.se/globalassets/publikationssok/granskningsrapporter/omrattning/2011/omratt2011-slutrapport.pdf

Skolinspektionen. (2012). *Riktad tillsyn av bedömning och betygssättning hos skolor med stora avvikelser vid omrättning av nationella prov [Special inspection on assessment and grading for schools with large deviances in the rescoring].* Retrieved from https://www.skolinspektionen.se/globalassets/publikationssok/granskningsrapporter/riktade-tillsyner/2012/nationella-prov/riktad-nat-slutrapport.pdf

Skolinspektionen. (2017). *Skolors hantering och förberedelse av nationella prov [School's management and preparation of national tests.].* Retrieved from https://www.skolinspektionen.se/sv/Beslut-och-rapporter/Publikationer/Granskningsrapport/Flygande-inspektion/hantering-och-forberedelse-av-nationella-prov/

Skolverket. (1995). *Information till lärare våren 1995. [Information to teachers spring 1995.].* Stockholm, Sverige: Skolverket.

Skolverket. (2000). Matematik [Mathematics syllabus]. Retrieved from https://www.skolverket.se/laroplaner-amnen-och-kurser/gymnasieutbildning/gymnasieskola/kursplaner-fore-2011/subjectKursinfo.htm?subjectCode=MA2000&courseCode=MA1203&lang=sv&tos=gy2000#anchor_MA1203

Skolverket. (2007). *PISA 2006 - 15-åringars förmåga att förstå, tolka och reflektera - naturvetenskap, matematik och läsförståelse [PISA 2006 - 15 year olds ability to understand, interpret and reflect - science, mathematics and reading comprehension].* Stockholm, Sverige: Skolverket.

Skolverket. (2009). *Likvärdig betygssättning i gymnasieskolan? En analys av sambandet mellan nationella prov och kursbetyg [Equal assessment in upper secondary school? An analysis of the relation between national tests and course grades.].* Stockholm, Sverige: Skolverket.

Skolverket. (2010). *Rustad att möta framtiden? PISA 2009 om 15-åringars läsförståelse och kunskaper i matematik och naturvetenskap [Ready to meet the future? PISA 2009 on 15-year olds reading comprehension and knowledge in mathematics and science.].* Stockholm, Sverige: Skolverket.

Skolverket. (2011). Mathematics. *Syllabus for upper secondary school.* Retrieved from http://www.skolverket.se/polopoly_fs/1.174554!/Menu/article/attachment/Mathematics.pdf

Skolverket. (2012). *Lärarinformation för samtliga kursprov. Matematik [Teacher information for all course tests. Mathematics.]* Stockholm, Sverige: Skolverket.

Skolverket. (2016). *Utvärdering av den nya betygsskalan samt kunskapskravens utformning [Evaluation of the new grading scale and the knowledge requirements].* Stockholm, Sverige: Skolverket.

Skolverket. (2017a). *Skillnader mellan provresultat och betyg i gymnasieskolan 2016 [Differences between test results and grades in upper secondary school 2016].* Stockholm, Sverige: Skolverket.

Skolverket. (2017b). *Skolverkets systemramverk för nationella prov [The Swedish national agency for education system framework for national tests]* Stockholm, Sverige: Skolverket.

Skolverket. (2018). Siris Retrieved from https://www.skolverket.se/skolutveckling/statistik/sok-statistik-om-forskola-skola-och-vuxenutbildning

Skolöverstyrelsen. (1970a). *Läroplan för gymnasieskolan Lgy 70, Allmän del [Curriculum for upper secondary school Lgy 70, General part].* Stockholm, Sverige: Skolöverstyrelsen.

Skolöverstyrelsen. (1970b). *Läroplan för gymnasieskolan Lgy 70, Planeringssupplement Naturorienterande och Tekniska ämnen [Curriculum for upper secondary school Lgy 70, Planning part, Natural sciences and technical subjects]*. Stockholm, Sverige: Skolöverstyrelsen.

SOU 1942:11. *Betänkande med utredning och förslag angående betygssättningen i Folkskolan. [Report with investigation and propositions about the grading in the school.]*. Stockholm, Sverige: Ecklestiastikdepartementet.

SOU 1992:86. *Ett nytt betygssystem [A new grading system]*. Stockholm, Sverige: Utbildningsdepartementet.

SOU 1992:94. *Skola för bildning [School for education]*. Stockholm, Sverige: Utbildningsdepartementet.

SOU 2007:28. *Tydliga mål och kunskapskrav i grundskolan - Förslag till nytt mål- och uppföljningssystem [Clear goals and grading criteria in compulsory school - Suggestions for a new goal and monitoring system]*. Stockholm, Sverige: Utbildningsdepartementet.

SOU 2008:27. (2008). *Framtidsvägen-en reformerad gymnasieskola [A reformed upper secondary school in the future]*. Stockholm, Sverige: Utbildningsdepartementet.

SOU 2016:25. (2016). *Likvärdigt, rättssäkert och effektivt – ett nytt nationellt system för kunskapsbedömning [A new national system for assessing knowledge.]*. Stockholm, Sverige: Utbildningsdepartementet.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability [Electronic Version]. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved from http://pareonline.net/getvn.asp?v=9&n=4

Stobart, G. (2001). The validity of national curriculum assessment. *British Journal of Educational Studies, 49*(1), 26-39. doi:https://doi.org/10.1111/1467-8527.t01-1-00161

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research, 51*(2), 161-179. doi:https://doi.org/10.1080/00131880902891305

The Swedish National Agency for Education. (2012a). Mathematics. *http://www.skolverket.se/polopoly_fs/1.174554!/Menu/article/attachment/Mathematics.pdf*.

The Swedish National Agency for Education. (2012b). *Upper secondary school 2011*. Stockholm, Sweden: Fritzes.

Tholin, J. (2006). *Att kunna klara sig i ökänd natur: en studie av betyg och betygskriterier-historiska betingelser och implementering av ett nytt system [Managing in disreputable nature: a study of grades and grading criteria-history and implementation of a new system.]*. (Doctoral dissertation), Högskolan i Borås, Borås, Sverige. Retrieved from http://www.diva-portal.org/smash/get/diva2:876774/FULLTEXT01.pdf

Tveit, S. (2014). Educational assessment in Norway. *Assessment in Education: Principles, Policy & Practice, 21*(2), 221-237. doi:https://doi.org/10.1080/0969594X.2013.830079

Tveit, S. (2018). Ambitious and ambiguous: shifting purposes of national testing in the legitimation of assessment policies in Norway and Sweden (2000–2017). *Assessment in Education: Principles, Policy & Practice*, 1-25. doi:https://doi.org/10.1080/0969594X.2017.1421522

Utbildningsdepartementet. (1994a). *Dnr U94/1031/Gru. Ett uppdrag till Statens skolverk om att utarbeta och tillhandahålla nationella prov [A commission to the Swedish national agency of education to develop and providing national tests.].* Stockholm, Sverige: Utbildningsdepartementet.

Utbildningsdepartementet. (1994b). *Läroplan för de frivilliga skolformerna, Lpf 94 [Curriculum for the upper secondary schools, Lpf 94]* Stockholm, Sverige: Utbildningsdepartementet.

Utbildningsdepartementet. (1999). *Dnr U1999/3290/S. Ett uppdrag till Statens skolverk om ett nationellt provsystem. [A commission to the Swedish national agency for education concerning a national test system.].* Stockholm, Sverige: Utbildningsdepartementet.

Utbildningsdepartementet. (2004). *Dnr U2004/5293/S. Uppdrag till Statens skolverk avseende det nationella provsystemet. [Commission to the Swedish national agency for education concering the national test system.].* Stockholm, Sverige: Utbildningsdepartementet.

Weigand, H.-G. (2017). What is or what might be the benefit of using computer algebra systems in the learning and teaching of calculus? In E. Faggiano, F. Ferrara, & A. Montone (Eds.), *Innovation and Technology Enhancing Mathematics Education* (pp. 161-193). Cham, Switzerland: Springer International Publishing.

Wikström, C. (2005). Grade stability in a criterion-referenced grading system: the Swedish example. *Assessment in Education: Principles, Policy & Practice, 12*(2), 125-144. doi:https://doi.org/10.1080/09695940500143811

Wikström, C. (2006). Education and assessment in Sweden. *Assessment in Education: Principles, Policy and Practice, 13*(1), 113-128. doi:https://doi.org/10.1080/09695940600563470

Wikström, C., & Wikström, M. (2005). Grade inflation and school competition: an empirical analysis based on the Swedish upper secondary schools. *Economics of Education Review, 24*(3), 309-322. doi:https://doi.org/10.1016/j.econedurev.2004.04.010

Vlachos, J. (2018). *Trust-based evaluation in a market-oriented school system* (Vol. 1217). Stockholm, Sweden: Research Institute of Industrial Economics.

Yang Hansen, K., & Gustafsson, J.-E. (2016). Causes of educational segregation in sweden–school choice or residential segregation. *Educational Research and Evaluation, 22*(1-2), 23-44. doi:https://doi.org/10.1080/13803611.2016.1178589

Zwick, R. (2006). Higher education admissions testing. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., Vol. 4, pp. 18-64). Westport, CT: ACE/Praeger.