UMEÅ UNIVERSITY

Theory and Validity Evidence for a Large-Scale Test for Selection to Higher Education

Jonathan Wedman

Department of Applied Educational Science
Educational Measurement
No. 10, Umeå 2017

ISBN: 978-91-7601-732-6
ISSN: 1652-9650
Cover art and design by: Björn Sigurdsson
Electronic version available at: http://umu.diva-portal.org/
Printed by: UmU Printing Service, Umeå University
Umeå, Sweden 2017

De ser ut som bruna humlevingar
[They look like brown bumblebee wings]

/Lova, 25 months old, shares her observation of hamburgers

Table of Contents

Abstract

Validity is a crucial part of all forms of measurement, and especially in instruments that are high-stakes to the test takers. The aim of this thesis was to examine theory and validity evidence for a recently revised large-scale instrument used for selection to higher education in Sweden, the Swedish Scholastic Assessment Test (SweSAT), as well as identify threats to its validity. Previous versions of the SweSAT have been intensely studied but when it was revised in 2011, further research was needed to strengthen the validity arguments for the test. The validity approach suggested in the most recent version of the *Standards for education and psychological testing*, in which the theoretical basis and five sources of validity evidence are the key aspects of validity, was adopted in this thesis.

The four studies that are presented in this thesis focus on different aspects of the SweSAT, including theory, score reporting, item functioning and linking of test forms. These studies examine validity evidence from four of the five sources of validity: evidence based on test content, response processes, internal structure and consequences of testing.

The results from the thesis as a whole show that there is validity evidence that supports some of the validity arguments for the intended interpretations and uses of SweSAT scores, and that there are potential threats to validity that require further attention. Empirical evidence supports the two-dimensional structure of the construct *scholastic proficiency*, but the construct requires a more thorough definition in order to better examine validity evidence based on content and consequences for test takers. Section scores provide more information about test takers' strengths and weaknesses than what is already provided by the total score and can therefore be reported, but subtest scores do not provide additional information and should not be reported. All four quantitative subtests, as well as the Swedish reading comprehension subtest, are essentially free of differential item functioning (DIF) but there is moderate DIF that could be bias in two of the four verbal subtests. Finally, the equating procedure, although it appears to be appropriate, needs to be examined further in order to determine whether it is the best practice available or not for the SweSAT.

Some of the results in this thesis are specific to the SweSAT because only SweSAT data was used but the design of the studies and the methods that were applied serve as practical examples of validating a test and are therefore likely useful to different populations of people involved in test development, test use and psychometric research.

Suggestions for further research include: (1) a study to create a more clear and elaborate definition of the construct, *scholastic proficiency*; (2) a large and empirically focused study of subscore value in the SweSAT using repeat

test takers and applying Haberman's method along with recently proposed effect size measures; (3) a cross-validation DIF-study using more recently administered test forms; (4) a study that examines the causes for the recurring score differences between women and men on the SweSAT; and (5) a study that re-examines the best practice for equating the current version of the SweSAT, using simulated data in addition to empirical data.

Keywords: SweSAT, validity, theoretical model, score reporting, subscores, DIF, equating, linking.

Populärvetenskaplig sammanfattning

Ett prov, oavsett hur det används, bör hålla hög kvalitet för att kunna mäta det som provet verkligen avser att mäta. Kvaliteten för ett prov blir extra viktigt när konsekvenserna av att ta provet är stora. I Sverige är högskoleprovet ett välkänt exempel på ett prov med stora konsekvenser för provdeltagarna – ens högskoleprovspoäng kan avgöra om man kommer in eller inte på den utbildning man vill gå samt vid vilket lärosäte man kommer in. Det är därför mycket viktigt att högskoleprovet håller hög kvalitet, eller, för att använda en psykometrisk formulering, att validiteten i tolkningarna och användningarna av provpoängen är hög. Syftet med den här avhandlingen var att utvärdera högskoleprovets validitet, vilket behövs eftersom provet förändrades kraftigt hösten 2011, bland annat genom att antalet uppgifter och delprov blev fler, och genom att provet delades upp i två distinkta delar, en kvantitativ del och en verbal del. Forskningen som ligger till grund för denna avhandling täcker fyra områden vilka presenteras nedan.

Den teoretiska idén bakom högskoleprovet är att det ska mäta *studiefärdighet*. Genom en litteraturstudie över material som rör högskoleprovet skapades förslag till teoretiska modeller för ursprungsversionen av högskoleprovet (från 1977) och för dagens högskoleprov, vilka knyter samman teorin med provets faktiska innehåll. Begreppet *studiefärdighet* anses bestå av två komponenter, en kvantitativ och en verbal, och detta får stöd av litteraturstudien samt av statistiska analyser av närmare 100 000 provdeltagares resultat sett över två administrationer av högskoleprovet. Däremot framkom också att definitionen av *studiefärdighet* är relativt kortfattad med följden att högskoleprovet skulle gynnas av att *studiefärdighet* fick en utförligare beskrivning så att man vet mer om vad begreppet innefattar.

Tre typer av poäng på högskoleprovet rapporteras för närvarande till provdeltagare: totalpoängen, poängen på den kvantitativa och den verbala delen, samt poängen för varje delprov. Genom denna rapportering antyder man att delprovspoängen säger något om provdeltagarens styrkor och svagheter, som inte den kvantitativa eller verbala delen säger, och att poängen på den kvantitativa och verbala delen säger något som inte totalpoängen säger. En statistisk analys av närmare 100 000 provdeltagares resultat sett över två administrationer av högskoleprovet visar att både den kvantitativa och den verbala delen ger unik information om styrkor och svagheter och därmed kan rapporteras till provdeltagare. Däremot säger poängen på de flesta av de åtta delproven inte något utöver detta och bör därför inte rapporteras, eftersom rapporteringen inte ger den information som den anspelar på att göra.

Provdeltagare med samma poäng på högskoleprovet förväntas på gruppnivå att klara frågorna på provet ungefär lika bra. Den tredje studien bygger på statistiska analyser av provpoängen från cirka 250 000 provdeltagare över fem administrationer av högskoleprovet och har undersökt om kvinnor och män *med samma provpoäng* presterar lika bra på olika frågor. Resultatet visade att de presterar lika bra på de fyra delproven i den kvantitativa delen av provet men att det finns vissa skillnader inom den verbala delen, och speciellt inom två delprov. På ORD-provet, som mäter ordförståelse, verkar kvinnor gynnas på ungefär en femtedel av frågorna och på ELF-provet, som mäter engelsk läsförståelse, är resultatet precis det omvända: män verkar gynnas på ungefär en femtedel av frågorna. Detta resultat ska undersökas vidare med hjälp av fler administrationer av högskoleprovet, och om mönstret kvarstår bör dessa två delprov undersökas närmare för att se om några speciella åtgärder bör vidtas.

Högskoleprovsresultat är i dagsläget giltiga i fem år och det innebär att det är mycket viktigt att provresultat från olika tillfällen är jämförbara så att konkurrensen till program och lärosäten blir rättvis. Den sista studien i denna avhandling undersökte om den statistiska metod som används för att göra administrationerna jämförbara är lämplig och om den är den bästa möjliga metoden för ändamålet. Resultaten bygger på statistiska analyser av cirka 140 000 provdeltagare över tre administrationer av högskoleprovet och tyder på att nuvarande metod är lämplig men resultaten var något oklara om det var den bästa metoden eller inte. Studien behöver utvecklas vidare med fler administrationer och med så kallad simulerade data för att ge en tydligare bild av vilken den bästa metoden faktiskt är.

Sammanfattningsvis visar studierna i denna avhandling att det både finns validitetsbevis som stödjer högskoleprovets nuvarande utformning och andra fynd som tyder på att vissa delar inom högskoleprovet behöver undersökas närmare.

Acknowledgements

Now that my thesis is completed there are persons, near and far, that deserve my thanks and my gratitude. First, I want to thank my advisor Marie Wiberg, who has an immense knowledge of seemingly all aspects of research and the academic world, and who from the very start and all the way to the end has provided me with excellent counselling and exactly the type of guidance I needed to grow as a researcher. I also want to thank my assistant advisor Per-Erik Lyrén who has been a perfect complement to Marie and with his insight into the SweSAT project has been great in guiding me past the many, *many* pitfalls associated with the SweSAT data that I would likely have fallen into otherwise.

Furthermore I want to thank the two people who started their respective research paths at the same time I did. These are my former PhD colleagues Miguel Inzunza, with whom I have had countless interesting discussions and who has provided invaluable support during the hundreds of lunch talks we have had, and Eva Knekta, who has never let me get away with anything but instead always continues to ask "why?", forcing me to think considerably harder than I originally had intended. Although I appreciate having my own office now, I miss the many spontaneous discussions the three of us had when we all shared the same room; they were very rewarding.

I give my thanks to Björn Sigurdsson for creating the cover art and design for this thesis, to Luis Cobian who is the systems manager responsible for creating data sets that I used in three out of the four studies in this thesis, to Lotta Jarl for help with various practicalities, and to my other colleagues at BVM, who help create a fantastic work environment.

In 2013 I was fortunate enough to be allowed to spend a semester taking courses and getting to know the people at the School of Education at the University of Massachusetts. I would like to give my great thanks to Ron Hambleton, who was responsible for making this happen and who was kind enough to invite me and make me feel welcome during my entire stay in Amherst. Thank you also to the faculty and fellow doctoral students at the School of Education who made my visit both fun and educational.

Finally, I want to express my deepest gratitude to my friends and family for their relentless support during my work on this thesis. Without you this would not have been possible. My mother, you have not only been cheering me on and encouraging me through all of this, but have also provided valuable academic feedback on my work. My father, who was my main motivation to enter the world of psychometrics in the first place, you are always with me in spirit. My dear friend Iza, for the past 15 years we have shared the ups and downs of life together and you have supported me whole-heartedly when I have needed it the most. And last but not least (albeit the

youngest), my fantastic daughter Lova, now three and a half years old, you bring endless joy to my life and keep me grounded and sane – I am very fortunate to be your father.

Jonathan Wedman
Umeå, August 2017

Studies

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

I. Wedman, J. (2017). *From aptitude to proficiency: The theory behind the Swedish Scholastic Assessment Test*. Manuscript submitted for publication.

II. Wedman, J., & Lyrén, P.-E. (2015). Methods for examining the psychometric quality of subscores: A review and application. *Practical Assessment, Research & Evaluation*, *20*(21), 1–14. Available online: http://pareonline.net/getvn.asp?v=20&n=21.

III. Wedman, J. (in press). Reasons for gender-related differential item functioning in a college admissions test. *Scandinavian Journal of Educational Research*.

IV. Wedman, J., & Wiberg, M. (2017). *Equating challenges when revising large-scale tests: A comparison of different frameworks, methods and designs*. Manuscript in preparation.

[1] Two of the studies (II and IV) are co-authored and below, the authors' respective contributions are specified.
Study II: I was in charge of drafting most of the manuscript, namely past research, the discussion and all text in the introduction and method section that was related to Haberman's method, utility index, factor analysis, DIMTEST, DETECT, and augmented subscores. I also performed the data analyses for said methods and interpreted the results. PEL came up with the initial idea and design, which we developed further together, performed the multidimensional item response theory (MIRT) analyses and wrote the sections on MIRT in the introduction and in the method section. PEL also helped with the discussion and provided valuable revisions on the manuscript.
Study IV: I was in charge of drafting the manuscript, processed the data, performed all data analyses and interpreted the results. MW came up with the initial idea and design, wrote the equations in the study, and provided crucial support when I wrote the code in the statistical software *R version 3.3.2*. MW also helped with the discussion and provided valuable revisions on the manuscript.

1. Introduction

Prospective students can choose from two main paths when applying to higher education in Sweden, provided that they have already obtained eligibility for higher education studies. The first path involves competing using one's upper-secondary school grade point average (GPA), which is currently the "gold standard" of college admissions instruments in Sweden, while the second path involves competing using one's normed score on the Swedish Scholastic Assessment Test (SweSAT). This selection system alleviates the pressure on students and teachers alike. Students who, for one reason or another, lack a competitive GPA are provided with another opportunity to get into their higher education program of choice by taking the SweSAT. Swedish teachers are therefore not burdened with the task of solely determining students' futures.

The high-stakes setting of admissions to higher education requires the SweSAT to meet certain demanding quality expectations. These quality demands mainly originate from the Swedish Council for Higher Education (Universitets- och högskolerådet, 2016), but some also stem from the international psychometric community (e.g., American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). This thesis focuses on the most important aspect of measuring quality – validity.

1.1 Aims and research questions

The overarching aim of this thesis was to examine the theoretical basis and validity evidence for the current version of the SweSAT, which has been in use since the fall of 2011, as well as identify threats to its validity. The SweSAT has been intensely studied throughout its existence, but when it was revised and heavily altered, further research was needed to strengthen the validity arguments for the test and to determine how to improve areas in which threats to validity were found.

Specifically, the main research questions were:

1. What is the theoretical model of the current SweSAT, and how does it relate to the theoretical model of the original SweSAT as well as to the construct *scholastic proficiency*?
2. Is there any psychometric value in reporting the subscores (i.e., subtest and –section scores) to test takers, do different methods of examining subscore value reach the same conclusion, and can the information gathered from subscores be improved?
3. Does the probability of answering items correctly differ between women and men with equivalent quantitative or verbal ability? If so, what could be the cause of this differential item functioning (DIF)?

4. What is the best way to link the current SweSAT to the previous version of the SweSAT? Furthermore, what is the best way of equating the sections of the current SweSAT and how does the current equating method compare to the best equating method?

1.2 Terminology

This thesis builds upon research from four distinct areas that relate to the same topic, and as such, includes a wide array of terminology. The most central terms will be addressed here. The main term used throughout the text is *SweSAT*, but there are differences in how different versions of the SweSAT are described throughout the thesis. In this summary and Study I, on the theoretical model of the SweSAT, the term *SweSAT-11* is used to refer to the SweSAT version that has been in use since the fall of 2011. Similarly, the version used before SweSAT-11 is referred to as *SweSAT-96* because it was introduced in the spring of 1996. The versions introduced in 1977 and 1992 are referred to as *SweSAT-77* and *SweSAT-92,* respectively. It is important to note that this terminology is not consistent across all the studies included in this thesis. In Study II, on subscores, both SweSAT-96 and SweSAT-11 are anonymized and referred to only as *Test A* and *Test B,* respectively, with no mention of either SweSAT-77 or SweSAT-92. In Study III, on differential item functioning, SweSAT-11 is referred to as *the new version* whereas SweSAT-77, -92 and -96 are grouped together as *the older versions*. Finally, in Study IV, on equating, SweSAT-11 is referred to as *the revised SweSAT* and SweSAT-96 as *the old SweSAT*, with no mention of either SweSAT-77 or SweSAT-92. These variations in terminology are not ideal, and are the result of selecting terminology that was the most appropriate for the specific context of each paper.

A second central term is *subscore,* along with the accompanying *subscore value*. A subscore is any score that does not represent all of the items included in a test. Within this thesis, a subscore most often refers to the observed score on either the verbal or quantitative section of SweSAT-11, which ranges from 0–80, or to the score on any of the subtests in any version of the SweSAT. *Subscore value* refers to the information given by a subtest score in relation to the information provided by the section score, or to the information given by a section score in relation to the information provided by the total score, and is relevant only in the context of reporting subscores to test takers. A subscore is considered to have *adequate psychometric quality* when it adds value to the total score. This justifies reporting the subscore to test takers because the subscore will provide more information about their strengths and weaknesses than the total score alone.

The term *differential item functioning* (*DIF*) means that the probability of answering an item correctly differs between comparable individuals from different groups. Individuals are comparable if they have

similar ability, which is often estimated using a test score. Groups can be any constellation of individuals that are organized according to, for example, language, ethnicity, socioeconomic status or gender; in this thesis, gender is used as the grouping variable. It is important to note that the term *sex* could have been used instead of *gender,* and the accompanying term *gender-related DIF* (Study III), but gender was chosen as this term is more prevalent in research focusing on DIF (e.g., Chubbuck, Curley & King, 2016; Hidalgo, Gómez-Benito, & Zumbo, 2014; Wiberg, 2009).

The term *equating* refers to a statistical procedure that allows the scores from different forms of a test to be comparable, or interchangeable. In the context of the SweSAT, this means that scores from the most recently administered SweSAT are adjusted during the norming process using scores from a previously administered test form of the SweSAT. Equating is crucial because SweSAT scores are valid for five years; thus, the scores must be normed so that they are comparable regardless of which test form of the SweSAT the test score comes from. In other words, equating test scores from different administrations of the SweSAT allows the scores to be interchangeable.

Finally, ten abbreviations, which are the short-form names of SweSAT subtests (both historical and current), are used throughout this thesis. The names of the subtests, and their corresponding abbreviations, are presented in Table 1. Brief descriptions of the subtests are presented in section 2.2.3.

Table 1. Name, abbreviation and the ability that is measured by each subtest that is, or has been, a part of the SweSAT.

Abbreviation	**Name**	**Ability**[a]
DS	Data sufficiency	Quantitative
DTM	Diagrams, tables and maps	Quantitative
WORD	Vocabulary	Verbal
READ	Swedish reading comprehension	Verbal
GI[b]	General information	Verbal
STECH[b]	Study techniques	Verbal
ERC	English reading comprehension	Verbal
XYZ	Mathematical problem solving	Quantitative
QC	Quantitative comparisons	Quantitative
SEC	Sentence completion	Verbal

[a] The terminology used is from SweSAT-11, in which subtests are considered to measure either quantitative or verbal ability; it is not the original terminology from SweSAT-77, in which the subtests were considered to measure aptitude, ability or knowledge.

[b] Not part of SweSAT-11

1.3 Disposition of the thesis

This thesis will cover four original studies and begins with an introductory summary, which also serves to present the overall findings in a wider setting. Chapter 2 provides information on different college admissions tests, and chronicles how they have been used internationally. Chapter 2 also presents the SweSAT, outlining both its history and status today. Validity is discussed in Chapter 3 by considering the historical development of validity theory along with current validity propositions and supporting propositions for both uses of SweSAT-11. This chapter also includes a summary of how different aspects of validity has been addressed in previous SweSAT versions. Chapter 4 provides a brief overview of the methods that were used in the research underlying this thesis and highlights some of the similarities and differences between them. The four studies presented in this thesis are individually summarized in Chapter 5. The last chapter, Chapter 6, discusses the findings from the research underlying this thesis and specifically focuses on how the findings affect the SweSAT and its test takers in a broader sense than what is discussed in the studies themselves. Chapter 6 also provides a discussion on how the research presented in this thesis can be useful to a wider audience, along with a consideration of its limitations and generalizability, and concludes with suggestions for further research.

2. The use of college admissions tests

Tests for selection or admission to higher education are widely used and range from small-scale, local tests to large-scale tests that are administered nation-wide. In Sweden, the SweSAT is a nation-wide test for selection to higher education, and has been mainly influenced by the SAT from the United States. Internationally, three well-recognized selection tests are the SAT and ACT in the United States and the PET in Israel. However, other tests, such as the Prueba de Selección Universitaria in Chile and the General Certificate of Education Advance Level examinations in the United Kingdom, provide further examples of large-scale tests that play an important role in their respective nation.

2.1 Internationally

One of the most widely known large-scale college admissions tests is possibly the SAT (The College Board, 2017), which is used in the United States but also administered in other countries. The SAT has been used since the 1920s and currently includes a reading test, a writing and language test, a math test, and an optional essay component. The SAT includes 154 multiple-choice items and has a total testing time of three hours. The items are meant to reflect real-world contexts and the SAT is not aligned with the upper-

secondary school curriculum. However, the test aims to measure what a test taker has learned in upper-secondary school as well as what is needed to succeed in higher education. Observed scores are converted to a 121-step, standardized composite score ranging from 400 to 1600 (not including the essay scores) with intervals of 10. The SAT also has a separate series of 20 tests, called the SAT Subject Tests, which cover domains specific to five general subject areas: English; history; languages; mathematics; and science. These tests are closely linked to the upper-secondary school curriculum.

The main competitor to the SAT is the ACT (ACT Inc., 2017), which is a curriculum-based achievement test that aims to measure what test takers have learned in upper-secondary school. The ACT has been administered since 1959 and currently includes subject tests in English, mathematics, reading and science, as well as an elective writing test. The ACT comprises 215 multiple-choice items and has a total testing time of two hours and 55 minutes. The observed scores are converted into a 36-step, standardized composite score ranging from 1 to 36 (not including the essay scores) with intervals of 1. Both the ACT and SAT are considered widely accepted admissions tests by colleges and universities in the United States; SAT scores are accepted by all colleges (The College Board, 2017) and ACT scores are accepted by all four-year colleges and universities (ACT Inc., 2017).

Another large-scale test for admission to higher education is the Psychometric Entrance Test (National Institute for Testing and Evaluation [NITE], 2017) in Israel. This test aims to predict a test taker's academic performance during higher education. The Psychometric Entrance Test consists of three sections – quantitative, verbal and English – with a total of nine subtests of which eight are multiple-choice and one involves essay writing. Six of the multiple-choice subtests are used for calculating the test score while the remaining two are used to try out newly developed items and for equating the test. The observed scores are converted into a standardized score scale ranging from 200 to 800, which is then used for selection.

2.2 In Sweden

College admissions tests were introduced in Sweden during the early 1900s, but were discontinued in the 1930s when GPA, which was found to predict success in higher education equally well as admissions tests, replaced them (Wedman, 1983). The relatively low competition for higher education studies at the time made the rank-ordering of test takers less relevant than establishing eligibility.

However, discussions during later years of allowing a larger proportion of the population access to higher education were accompanied by a mandate to develop a new admissions instrument. The envisioned instrument was not intended to reinstate the admissions test of the early 1900s, but to rather

complement the GPA in the selection process (Kompetensutredningen, 1966).

2.2.1 The history of the SweSAT

In the 1960s, Kompetensutredningen (1966) suggested that another instrument should be used alongside GPA in selection to higher education. It was proposed that this instrument should measure something other than what was measured by GPA in order to create a new path to higher education. Several types of instruments were examined and disregarded, specifically those that aimed to measure study motivation, study habits, personality, interest and background data. The instrument that showed the most potential was a battery of tests measuring both knowledge and aptitude (Kompetensutredningen, 1968).

Several applications for the SweSAT were suggested, but in 1970 it was decided that the SweSAT should be constructed to rank-order all applicants to higher education who had academic eligibility (Kompetensutredningen, 1970). It was also suggested that the SweSAT should be used for the eligibility testing of applicants who lacked formal academic eligibility; however, this proposed use was later disregarded. Hence, the initial purpose of the SweSAT was to provide upper-secondary school students with an uncompetitive GPA a "second chance", and to act as a selection instrument for candidates with at least 25 years of age and four years of work experience (the 25/4:s) (Regeringens proposition 1972:84; Regeringens proposition 1975:9; Tillträdesutredningen, 1985a).

It was originally proposed that observed scores on the SweSAT would be transformed into a normally distributed, 41-step scale ranging from 1.0–5.0 with intervals of 0.1 (Kompetenskommittén, 1974). This was intentionally the same scale as what was used for the GPA at the time, and was deemed appropriate because it would enable the adequate differentiation of test takers so that selection to higher education would not be based on score differences that were meaningless in a practical sense (Kompetenskommittén, 1974). It was suggested that the normed SweSAT score and GPA would constitute a composite score in which both components received equal weight (Kompetenskommittén, 1974); but the composite score would only be used for selection if a candidate's SweSAT score was higher than the GPA; if the SweSAT score was lower, the GPA would be used alone. In 1975, it was decided that the SweSAT would be subjected to a trial period during which only the 25/4:s could use their scores for selection purposes (Regeringens proposition 1975:9). This also meant that the SweSAT's intended purpose to serve as a "second chance" for upper-secondary school students with an uncompetitive GPA was postponed. At this point, the SweSAT scale was changed from the previously suggested GPA scale into the work experience scale that the 25/4:s used: a

21-step scale ranging from 0.0-2.0 with intervals of 0.1 (Centrala organisationskommittén för högskolereformen, 1975; Skolöverstyrelsen, 1976). An important difference, however, was that the SweSAT scale was norm-related and scores were scaled to approximate a normal distribution with a mean of around 1.0 (Skolöverstyrelsen, 1976) whereas the scale used for the 25/4:s was criterion-related and had no pre-defined shape for its distribution.

The SweSAT's approximately normally distributed, 21-step scale is still used today to calculate the score of each SweSAT-11 section, and the total SweSAT-11 score is the mean of the two section scores. Interestingly, in contrast to the suggestion by Kompetenskommittén (1974), SweSAT scores have never been used in a composite score alongside any other scale.

The first version of the SweSAT was introduced in 1977, free of charge until 1982, and open for anyone to take (Henriksson & Wedman, 1993) although only the 25/4:s could use their scores for selection to higher education. During its first years, SweSAT-77 had 150 items, but READ proved too easy for test takers with an upper-secondary school education. As a result, in 1980 the True/False answer format used in READ was changed to the multiple-choice format that was already used in the other subtests (Lexelius & Wedman, 1980; 1981). This change also led to a reduction in the number of READ-items, from 30 to 24; therefore, SweSAT-77 had 144 items from 1980 onwards. During the first years of SweSAT administration, the vast majority of 25/4 test takers had at least nine years of work experience. This meant that they received the max score (2.0) on the work experience scale and work experience therefore played a relatively small role in selection to higher education when compared to the effect of SweSAT score (Lexelius & Wedman, 1978; Wedman, 1978).

After the trial period ended in 1991, the SweSAT could be used for selection purposes by all candidates to higher education. This meant that the primary purpose of the SweSAT shifted and the test became a "second chance" for students with uncompetitive GPAs, as previously intended. The decision to open the SweSAT to the general public was accompanied by anticipated problems with STECH, which comprised a cluster of 20 items, all related to a booklet of around 83 pages (Henrysson & Wedman, 1975), which needed to be tried out in full instead of item by item like the other subtests. SweSAT items were at this time still tried out on a separate occasion from the regular administration, using students in upper-secondary school. Exposing the entire STECH to that upper-secondary school students had not been an issue during the administrations of SweSAT-77 because that population had not been allowed to use their SweSAT-77 scores for selection to higher education. However, by allowing all test takers to use their scores for selection to higher education, starting in 1992, there was a real possibility of students participating in the try-out and then taking the regular SweSAT.

This meant that some students would have received the same STECH test both when it was tried out and when it was included in the regular SweSAT. This problem of exposing the entire STECH to parts of the test taking population before including it in the regular administration, in combination with the large cost of developing it (e.g., Wester-Wedman, 1990a) and the finding that removing it from the SweSAT only slightly affected the rank-order of test takers (Wester-Wedman, 1990a), led to the removal of STECH from the SweSAT prior to the regular administration in the spring of 1992. Simultaneously, ERC was added to the SweSAT, marking the introduction of SweSAT-92.

SweSAT-92 was short-lived. This was partly because soon after its introduction, the SweSAT's international expert committee[2] suggested that GI should be removed from the test (Wedman & Stage, 1994). Also, in 1996, the routines for trying out items changed. Instead of trying out items with upper-secondary school students on a separate occasion from when the test was administered, the items would now be tried out as part of the regular administration. Thus, the SweSAT now included more items, but because the test-time could not be lengthened, several time-saving changes were made to the test, with the main change being the removal of GI (Stage & Ögren, 2001). GI had also been plagued by intense discussions within test review groups concerning what general information actually was, and these conflicts likely facilitated its removal (Lyrén, 2009a). The next version was SweSAT-96.

SweSAT-96 was in use for almost 15 years without any major changes even though the preliminary work for SweSAT-11, during which three international experts in the field of testing jointly suggested that separate scores should be available for at least quantitative and verbal reasoning, had already begun in 2002 (Högskoleverket, 2002).

2.2.2 Current status of the SweSAT

Following several years of development, a revised form of SweSAT-96 was introduced in 2011 as SweSAT-11. The primary change was a formal division of the test into two independent sections – a quantitative and a verbal section. During the development process, three new subtests were added and the total number of items in the test increased from 122 to 160, with 80 items in the quantitative section and 80 in the verbal section. The number of items in some of the existing subtests was also adjusted, with the largest change in WORD, which was reduced from 40 items to 20. Another change

[2] The committee was formed in 1992 and initially met every other year to discuss questions regarding the SweSAT. In 1994 the committee consisted of Dr. Michal Beller, Israel, professor Ronald Hambleton, USA, professor Wim van der Linden, the Netherlands, professor Jan-Eric Gustafsson, Sweden, professor Allan Svensson, Sweden, professor Ingvar Lundberg, Sweden, professor Sten Henrysson, Sweden, Dr. Günter Trost, Germany, and professor Ingemar Wedman, Sweden.

was that the SweSAT was no longer administered subtest by subtest, which had been the protocol in all previous versions, but in blocks. The current version includes four blocks, each with 40 items. Two of these blocks are quantitative (including XYZ, QC, DS and DTM) and two of the blocks are verbal (including WORD, READ, SEC and ERC). The test takers also take a fifth "try-out" block, which is either quantitative or verbal and contains newly developed items that do not count toward the test score. Test takers do not know which block is the try-out block. The observed scores from the two quantitative blocks are summed into a quantitative section score from 0–80, which is then transformed into an approximately normally distributed, 21-step scale score ranging from 0.0–2.0 with intervals of 0.1. The same process is performed for the observed score from the verbal blocks. The two resulting normed section scores are then averaged into the final score, which is reported on a 41-step scale ranging from 0.00–2.00 with intervals of 0.05.

2.2.3 Past and present subtests of the SweSAT

It is possible that as many as 50 subtests have been tried out during the continuous development of the SweSAT since the late 1960s, yet many failed due to sensitivity to coaching (Henrysson, 1994). Of these 50, evidence that 36 subtests were developed at some point has been found during the research described in this thesis. Ten of these subtests (explained in more detail below) have been part of the SweSAT at one time or another (Table 2), and four of the original subtests are still in use today.

Data sufficiency (DS) is a quantitative subtest in which each item consists of a stem and a question related to the stem. In addition, there are two statements that provide information about the answer to the question. The object is to assess whether the question can be answered by both statements on their own, by the two statements together or not at all.

Diagrams, tables and maps (DTM) is a quantitative subtest consisting of visual stimuli and clusters of items relating to each stimulus or group of stimuli. It tests the test taker's ability to read and interpret different types of diagrams, tables and maps.

Vocabulary (WORD) is a verbal subtest in which the item stem is a word that has a synonym or best match among the alternatives. It tests Swedish vocabulary and contains Swedish and foreign words from different content domains.

Swedish reading comprehension (READ) is a verbal subtest that consists of long texts with two or more accompanying items. The object of each item is to measure a test taker's understanding of the text rather than the ability to memorize the content.

General information (GI) was a verbal subtest that measured general knowledge of humanities, society, nature and culture in a broad sense.

Study techniques (STECH) was a verbal test in which the items asked for one or several pieces of information that could be found in an accompanying pamphlet complete with a table of contents and registers. The pamphlet was around 83 pages long; therefore, it was not possible to read through in its entirety and test takers had to use one or more registers to find the requested information. It was the subtest that most closely reflected the complex study situations students would face upon admission to higher education (Henrysson & Wedman, 1975)

English reading comprehension (ERC) consists of long and short texts, as well as sentence completion texts. The long texts and sentence completion texts have several accompanying items whereas short texts include one item each. As with READ, the object of each item related to a long or short text is to measure the test taker's understanding of the text. The object of the sentence completion items is, as in SEC, to choose which of the given words is most suitable for the blank in the text.

Mathematical problem-solving (XYZ) is a quantitative subtest that assesses a test taker's ability to solve mathematical problems within the domains of arithmetic, algebra, geometry, functions, and statistics.

Quantitative comparisons (QC) is a quantitative subtest that assesses a test taker's ability to make comparisons between two numerical quantities within the same domains as XYZ.

Sentence completion (SEC) is a verbal subtest in which the items include one or a few sentences with either one, two or three words omitted. The object is to choose which of the given words is most suitable for the blank, or blanks, in the text.

Table 2. The subtests and the number of items they have contained in various SweSAT versions.

Subtest	**SweSAT-77**	**SweSAT-92**	**SweSAT-96**	**SweSAT-11**
DS	20	20	22	12
DTM	20	20	20	24
WORD	30	30	40	20
READ	30/24[a]	24	20	20
GI	30	30	-	-
STECH	20	-	-	-
ERC	-	24	20	20
XYZ	-	-	-	24
QC	-	-	-	20
SEC	-	-	-	20
Items total	150/144[a]	148	122	160

[a] READ was reduced from 30 to 24 items in 1980. SweSAT-77 thus contained only 144 items from 1980 to 1991.

Note. The time allotted for each subtest has also changed over the years. This is not included in the table.

3. Validity

Validity is, in layman's terms, best described as when an instrument measures what it is meant to measure. However, validity theory has a far broader scope, and views on validity and the validity framework have undergone several changes during the 1900s and early 2000s. As validity is such an extensive topic, this chapter provides a mere sample of the origins of the current approach to validity as presented in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). The *Standards*' approach to validity is used in this thesis to provide a contemporary approach to the validity discussions regarding the SweSAT.

3.1 History and development

Validity was introduced in the 1920s and initially focused solely on what later became known as criterion validity. One of the earliest formal descriptions of validity was presented by Cureton (1951), who proposed that validity is linked to a particular purpose for which a test score is used, instead of being a static attribute of a test. Cureton also discussed an aspect of validity which he called "relevance", and this concept was similar to what later became known as construct validity. However, Cureton did not endorse relevance, but instead dismissed it due to concerns that the inherent subjective nature of the term could have a negative effect on validity, and focused instead on criterion validity.

Thurstone (1955) labeled criterion validity as obsolete. He suggested that criterion validity should be replaced with internal consistency reliability in validity studies, at least those used to assess personality tests, and thus proposed that reliability should be incorporated into validity. Thurstone's idea, however, did not receive much support, and reliability and validity thus remained separate concepts.

The 1954 *Technical recommendations for psychological tests and diagnostic techniques* (APA, AERA, & National Council on Measurements Used in Education, 1954) presented four types of validity – predictive, concurrent, content, and construct validity. Categorizing validity was partly a response to increasing criticism about criterion validity being the gold standard of validity theory and partly a way to validate theoretical attributes, which had previously had very limited validation models (Kane, 2006). Among these four types of validity, Cronbach and Meehl (1955) stressed that construct validity, which was a relatively new concept at the time, is a valuable complement to the other three types, especially when no criterion is accepted as adequate to define what one aims to measure.

The *Technical recommendations* (APA, AERA, & National Council on Measurements Used in Education, 1954) described construct validity as the

extent to which a test measures the psychological qualities that are linked to achieving certain aims, and recommended using both logical and empirical approaches when investigating construct validity. It should be investigated, as also stated in Cronbach and Meehl (1955), when no definitive criterion was available and indirect measures have to be used instead. This was restated in the second version of *Technical recommendations,* renamed *Standards for Educational and Psychological Tests and Manuals* (APA, AERA, & NCME, 1966), and also relates to Ebel (1961), who raised the question of how the criterion itself should be validated when using criterion validity.

The *Standards for Educational and Psychological Testing* (APA, AERA, & NCME, 1985), hereafter called *Standards* for the sake of brevity, treated validity as a unitary concept instead of four types, but added that different types of evidence are needed when making different types of interpretations. These types of evidence relate to the four validity types – criterion-related evidence (concurrent and predictive), content evidence and construct evidence.

Messick (1989; 1995) also advocated that validity should be considered as a unitary concept, yet one that is made up of four aspects based on a cross-tabulation of the two facets that, according to him, make up validity: (1) the source of justification; and (2) the outcome of testing. Messick divided the source of justification into *evidence* and *consequence*, and the outcome of testing into *test interpretation* and *test use*. Three of the four aspects of validity in Messick's framework are subsumed under construct validity, with the exception of social consequences.

Messick's (1989) idea of including *social consequences* in the validity framework became a subject of debate. Mehrens (1997) opposed the inclusion on the basis that it would overburden the already complex validity concept. Mehrens argued that the concept of validity should be narrowed instead of expanded. Messick (1995) responded to the critique by stating that social consequences were not added to, but intrinsic to, validity, and are thus part of validity by default. Shepard (1997) arrived at the same conclusion, stating that consequences are a logical part of test use. Linn (1997) agreed with Messick (1995) and Shepard (1997), and voiced concern that consequences could be relegated to a lower priority if they were excluded from the validity framework, as validity was (and still is) considered to be crucial in test evaluation.

Kane (1992; 2006) suggested an argument-based approach in which validation is a process. Kane's approach, which requires a clear declaration of the interpretations and usage of a test, reflects the essence of construct validity, but with less focus on formal theories. Two types of arguments are needed in Kane's approach, the interpretive argument, which specifies the proposed interpretations and uses of test results, and the validity argument,

which is an evaluation of the interpretive argument. The criteria for evaluating interpretive arguments are clarity and coherence of the argument, along with plausibility of the inferences and assumptions (Kane, 2006).

Standards (AERA, APA & NCME, 1999) also promoted the ongoing process of validation in relation to the interpretation and uses of test scores rather than a focus on distinct types of validity. *Standards* describes validity as a unitary concept, which linked back to Messick (1989) and an earlier version of *Standards* (APA, AERA, & NCME, 1985), and it comprises theory and five sources of validity evidence: Test content, response processes, internal structure, relations to other variables and consequences of testing. Strands of evidence based on these sources are accumulated and integrated into a coherent validity argument, which purports to support the intended interpretation of test scores for specific uses.

All of these developmental stages of validity have paved the way for a current view of the validity framework, which is described in the validity chapter in the most recent version of *Standards* (AERA, APA, & NCME, 2014), which is the theoretical approach to validity that is adopted in this thesis. The *Standards* is not the only work available that contains guidelines for proper test use and that addresses validity; other works, such as the International Test Commission (ITC) Guidelines on Test Use (ITC, 2013), are also available. However, the ITC Guidelines are partly based on previous versions of *Standards* (AERA, APA, & NCME, 1999), and do not address validity further than an endorsement of its importance in various areas. Other options are, for example, to approach validity based on the suggestions by Kane in the validation chapter in the latest edition of the *Educational Measurement* series (Kane, 2006), or based on the suggestions by Messick in the validity chapter in the previous edition (Messick, 1989). However, because (a) the validity chapter in the latest version of *Standards* is the most up-to-date validity approach available, (b) it is created and supported by AERA, APA and NCME, and (c) it provides a comprehensive and structured set of guidelines, concerning validity evidence and other technical quality aspects of testing, it is the approach that is followed in this thesis.

3.2 Present view

The current definition of validity, according to the most recent version of *Standards* (AERA, APA, & NCME, 2014), is "validity is the degree to which evidence and theory support the interpretations of test scores and proposed uses for tests" (p. 11). The validity argument is described as a process and builds on the idea presented in the previous version of *Standards* (AERA, APA, & NCME, 1999) that validity relates to the interpretations of test scores for proposed uses rather than the test itself, a viewpoint that links all the way back to Cureton (1951). Theory and validity evidence has to be provided for

each test score interpretation as well as for each proposed use, and five sources of evidence are required in the validation process. These sources relate back to Messick (1995) and *Technical recommendations* (APA, AERA, & National Council on Measurements Used in Education, 1954), as they include evidence based on test content, on the relations to other variables, and on consequences of testing as well as evidence on internal structure and on response processes. Other available evidence, related to the technical aspects of testing, such as reliability, scoring, scaling and equating, should also be integrated into the validity argument. The five sources of validity evidence are described in further detail below.

Validity evidence based on test content can be accumulated by analyzing test content and its relationship to the construct that the test aims to measure. Content evidence can come from, for example, expert judgments or empirical analyses of the extent to which the test content represents the content domain. It is especially important to assess the suitability of the existing content domain when a new purpose is added to a test.

Validity evidence based on response processes is required when a construct entails an assumption about one or more cognitive processes, such as *reasoning*, engaged in by test takers. Evidence often comes from analyses of individual responses but can also come from analyses of different subgroups. Such evidence can also provide information about construct-irrelevant sources of variance, in which abilities or other characteristics that are irrelevant to the construct affect test performance.

Validity evidence based on internal structure provides information on how relationships among test components are true to the construct that the test aims to measure. Dimensionality analyses and analyses of differential item functioning can provide evidence of internal structure, which can then be analyzed for conformity to the construct.

Validity evidence based on relations to other variables is gathered by analyzing the relationship between test scores and one or more variables that are external to the test. Relationships with external variables can provide either convergent or discriminant evidence, and can take the form of either a predictive or concurrent study.

Finally, validity evidence based on consequences of testing concerns both intended and unintended consequences, within and beyond the interpretation of test scores proposed by the test developer. An unintended consequence of testing could be admitting fever test takers from a particular group into a certain class or higher education program due to construct underrepresentation or construct-irrelevant sources of variance. Analyses of consequences are also important when a test is used for purposes other than that for which it was developed.

The research underlying this thesis investigated validity theory and four of the five sources of validity evidence in the context of SweSAT-11 to examine

different aspects of, and threats to, validity. Study I, which focused on the theoretical model of the SweSAT, described the underlying theory and examined validity evidence based on test content by studying construct representation. Study II, on subscores, examined validity evidence based on internal structure and on consequences of testing by studying subscore reporting and the relationship between the two SweSAT sections. Study III, which covered DIF, investigated validity evidence based on response processes, internal structure and consequences of testing through studying gender-related DIF and its causes. Equating was the focus of Study IV, and, as such, the work did not directly address validity but rather evaluated the interchangeability of test scores by studying the suitability of the current equating framework, design and method.

A study that examined the fifth source of validity evidence, evidence based on relations to other variables, in SweSAT-11 was initially supposed to be included in this thesis but was instead investigated by Lyrén, Rolfsman, Wedman, Wikström and Wikström (2014). Our findings showed that SweSAT-11 was a better predictor of success in higher education than SweSAT-96, but, as expected, not as robust of a predictor as GPA. We also found that the different weighting of SweSAT-11 sections for different programs improved prediction marginally, but this was only noted for certain programs.

3.3 Validity in a SweSAT context

The following descriptions of how validity has been addressed in a SweSAT context need to be understood as my own interpretations, which are based on the validity framework described in *Standards (*AERA, APA, & NCME, 2014). The descriptions are not absolute and undisputable but would, on the contrary, be interesting to develop further using discussions from different validity frameworks.

The examination of validity in a SweSAT context has remained relatively unchanged from the initial SweSAT-77 to today's SweSAT-11. The focus has been on the five sources of validity evidence: relations to other variables (correlation with success in higher education); test content (the relationship between content and construct); response processes (DIF); internal structure (dimensionality analyses and DIF); and the consequences of testing (group differences and DIF).

The most notable change has been the addition of validity arguments relating to using SweSAT scores for providing diagnostic information to test takers in the form of subscores (Lyrén, 2009). This application of the SweSAT became the second use of the SweSAT scores (Tillträdesutredningen, 2004) following recommendations by John Fremer, David F. Lohman and Werner W. Wittmann in 2002 (Högskoleverket, 2002), and needed to be validated as well.

Another change is how subtests have been allowed to intercorrelate. In SweSAT-77, -92 and 96, all of which produced a single total score and contained no sections, subtests were not allowed to correlate highly with one another as this would narrow the part of the construct *scholastic aptitude/-proficiency* that was being measured by SweSAT and thus make the test less relevant from a scholastic aptitude/-proficiency perspective (Kompetensutredningen, 1968; Stage & Jonsson, 1995). In contrast, the subtests in SweSAT-11 are supposed to intercorrelate highly with subtests in their respective section and show fairly low correlations with subtests in the other section to "strengthen" each section (Stage & Ögren, 2010). The means by which validity has been addressed in a SweSAT context are described further below.

3.3.1 Validity in previous versions of the SweSAT

The SweSAT initially had a single use – selection to higher education – and thus only that use needed to be validated. The first source of validity evidence mentioned for the SweSAT was relations to other variables, more specifically, the ability to predict success in higher education (Kompetensutredningen, 1968). This was before the SweSAT development had started, when the search for an alternative to GPA as the second best predictor of success in higher education was underway. Theory and validity evidence based on test content became a large part of SweSAT development when the construct *scholastic aptitude* was defined in Skolöverstyrelsen (1976), meaning that the SweSAT would aim to measure as much as possible of what higher education institutions in Sweden demanded of admitted students. Validity evidence based on internal structure was examined using dimensionality analyses. Initially a single factor was found (Tillträdesutredningen, 1985b) and later two factors were found (Gustafsson, Wedman & Westerlund, 1992). The two-dimensional structure received further support in Svensson, Gustafsson and Reuterberg (2001). Validity evidence based on the consequences of testing was also considered; for example, the use of test scores from SweSAT-77 for selection to higher education was not allowed to adversely affect primary or secondary education (Skolöverstyrelsen, 1976). Finally, evidence based on response processes and, again, internal structure and consequences of testing, was investigated because the SweSAT was not allowed to give test takers an unfair advantage based on the belonging to a certain group, where grouping categories can be, for example, gender or socioeconomic status (Skolöverstyrelsen, 1976). This was measured in terms of score differences (e.g., Stage 1988) and DIF (e.g., Wester, 1994; 1997).

3.3.2 Validity in SweSAT-11

There are currently two uses of the SweSAT – selection to higher education and providing diagnostic information to test takers – which means that both of these uses should be validated. The validity model, comprising validity propositions and supporting propositions, that is presented in this thesis for SweSAT-11 is based on, and resembles, the validity model for SweSAT-96 presented by Lyrén (2009), but is structured differently. The main differences between the proposed validity model for SweSAT-11, presented here, and the proposed validity model for SweSAT-96, presented in Lyrén (2009), are: (1) the addition of propositions A1.6, A1.7, A3.3, A3.5, A3.6, A3.7, B1.1, B1.5, B1.6, and B1.9r; and (2) many of the arguments in section 3.3.3, which are intended to support the propositions in Tables 3 and 4.

Tables 3 and 4 list the validity propositions for both uses of SweSAT-11. Supporting propositions on reliability, scoring, scaling and linking are also listed because they are relevant to the technical quality of SweSAT-11 and therefore affect the validity of the intended interpretations and uses of its scores (AERA, APA, & NCME, 2014). Tables 3 and 4 also include an acceptance proposition because public acceptance of SweSAT-11 is a quality demand made by the Swedish Council for Higher Education (Universitets- och högskolerådet, 2016).

Each proposition is explained in further detail in sections 3.3.3 and 3.3.4. The validity propositions aim to serve as both a summative evaluation of SweSAT-11 (what propositions were adequately fulfilled in these past administrations?) and as a formative evaluation (what should be done, and how, in the present and future to maintain highly valid interpretations and uses of SweSAT scores?)

Table 3. Validity and supporting propositions for SweSAT-11 when used for admissions decisions.

Validity propositions	
A1.1	The SweSAT measures the general skills required in higher education.
A1.2	The measurements obtained during testing are representative of the universe of measurements that defines the testing procedure.
A1.3	The test forms from different administrations of the SweSAT are parallel.
A1.4	There are no construct-irrelevant sources of variability that seriously bias the interpretation of the total score as a measure of *scholastic proficiency* or the section scores as measures of quantitative and verbal ability.
A1.5	Test-takers with a high level of *scholastic proficiency* perform better in higher education than test takers with a low level of *scholastic proficiency*.

A1.6	Test-takers with a high level of quantitative ability perform better in higher education with a quantitative/numerical orientation than test takers with a low level of quantitative ability.
A1.7	Test-takers with a high level of verbal ability perform better in higher education with a verbal orientation than test takers with a low level of verbal ability.
A1.8	Using SweSAT scores for admission decisions will not lead to significantly lower levels of performance in higher education than when GPA is used for the same decisions.
A1.9	Using SweSAT scores for admission decisions will not have a negative impact on instruction in primary and secondary school.
Reliability proposition	
A2.1	The sample of measurements is large enough to keep sampling error at an acceptable level.
Scoring, scaling and linking propositions	
A3.1	The scoring protocol is appropriate.
A3.2	The scoring protocol is applied accurately and consistently.
A3.3	The normed scales used for the section scores and the total score are appropriate.
A3.4	The scaling procedure for transforming observed scores to scale scores is appropriate.
A3.5	The scaling procedure for transforming observed scores to scale scores is applied accurately and consistently.
A3.6	The procedure for transforming the two separate scale scores into a single scale score is appropriate.
A3.7	The procedure for transforming the two separate scale scores into a single scale score is applied accurately and consistently.
A3.8	The equating procedure (framework, design and method) is appropriate.
A3.9	The equating procedure is applied accurately and consistently.
Acceptance proposition	
A4.1	SweSAT scores are accepted as meaningful for admission decisions by the users of the admission system, including test takers.

Table 4. Validity and supporting propositions for SweSAT-11 when used for providing diagnostic information to test takers.

Validity propositions	
B1.1	Both sections measure the general skills required in higher education.

B1.2	The measurements obtained during testing are representative of the universe of measurements that define the testing procedure.
B1.3	The test forms from different administrations of the SweSAT are parallel.
B1.4	There are no construct-irrelevant sources of variability that seriously bias the interpretation of the total score as a measure of *scholastic proficiency* or the section scores as measures of quantitative and verbal ability.
B1.5	Test-takers with a high level of quantitative ability perform better in higher education with a quantitative/numerical orientation than test takers with a low level of quantitative ability.
B1.6	Test-takers with a high level of verbal ability perform better in higher education with a verbal orientation than test takers with a low level of verbal ability.
B1.7r	The quantitative subtest scores provide information that is useful for remediation and that is not already provided by the quantitative section score.
B1.8r	The verbal subtest scores provide information that is useful for remediation and that is not already provided by the verbal section score.
B1.9r	The section scores provide information that is useful for remediation and that is not already provided by the total score.
Reliability proposition	
B2.1	The sample of measurements is large enough to keep sampling error at an acceptable level.
Scoring, scaling and linking propositions	
B3.1	The scoring protocol is appropriate.
B3.2	The scoring protocol is applied accurately and consistently.
B3.3	The normed scale used for the section scores is appropriate.
B3.4	The scaling procedure for transforming observed scores to scale scores is appropriate.
B3.5	The scaling procedure for transforming observed scores to scale scores is applied accurately and consistently.
Acceptance proposition	
B4.4r	Subtest scores are accepted as meaningful by test takers

Note. The "r" after propositions B1.7, B1.8, B1.9 and B4.4 indicates that these are new propositions to the diagnostic information use and therefore will be explained in further detail in section 3.3.4. The other propositions are identical or near identical to those already presented in Table 3 and are therefore explained in section 3.3.3.

3.3.3 Propositions for SweSAT-11 when used for admissions decisions

Validity propositions

A1.1 The SweSAT measures the general skills required in higher education. Historically, each subtest chosen for SweSAT-77 was tried out to determine whether it adequately measured a part of the construct *scholastic aptitude*, and thus whether or not it reflected the demands that higher education studies placed on admitted students. (Skolöverstyrelsen, 1976). It has been shown that the construct *scholastic proficiency* for SweSAT-11 needs a more thorough definition (See Study I, on the theoretical model of the SweSAT), such as the definition of *scholastic aptitude*, in order to better provide evidence for or against this proposition. That is, the general skills that are currently required in higher education need to be more precisely defined.

A study by Lyrén et al. (2014) that investigated the relationship between SweSAT-11 and an external criterion, success in higher education[3], found that SweSAT-11, when the quantitative and verbal sections were weighted equally, predicted success in higher education in varying degrees (*r*-values ranging from 0.16–0.37) across eight of the eleven university programs included in the study. This result demonstrates that despite the desire for a more elaborate definition of *scholastic proficiency*, SweSAT-11 still successfully measures what is required to succeed in higher education for most of the evaluated study programs. However, the results also showed that SweSAT-11 was not correlated with academic performance in three study programs.

A1.2 The measurements obtained during testing are representative of the universe of measurements that defines the testing procedure. The universe of measurements for each subtest is governed by the item construction guidelines specific for that subtest. Each topic in the guidelines includes a continuum of item difficulty, discrimination and correlation to the subtest score. The addition of pretest data and review panels allow test developers to ensure that the items, and thus the measurements obtained during testing, are representative of the universe of possible measurements according to the guidelines.

A1.3 The test forms from different administrations of the SweSAT are parallel. Tests forms are assembled using information gathered from the pre-testing of items. The pre-testing is performed when the SweSAT is

[3] "Success in higher education" was defined as the quotient between the number of ECTS credits that a student obtained during the first two semesters in higher education and the number of ECTS credits that the student was registered for during that time.

administered, and the test takers do not know which blocks are "live" and which block is the pre-test. Therefore, we can assume that test takers try equally hard on the pre-test items and "live" items. In this way, data gathered from pre-testing are representative of the test taker population. The pre-test block does not count towards the SweSAT score.

The pre-test data allow test developers to assemble subtests with similar content, difficulty and discrimination for different administrations, where each item correlates positively – within established intervals – to the respective subtest score. Each subtest also has separate item content guidelines, and this, together with pre-test data and an appropriate equating method, produces tests that are comparable in terms of content and difficulty. The test scores from different administrations are therefore interchangeable.

A1.4 There are no construct-irrelevant sources of variability that seriously bias the interpretation of the total score as a measure of scholastic proficiency or the section scores as measures of quantitative and verbal ability. Study III, which studied DIF in both sections of SweSAT-11, revealed that the WORD and ELF subtests may include construct-irrelevant variability associated with item content and item format, respectively. A confirmation of these results in a cross-validation study of later administrations of SweSAT-11 would suggest that means of reducing this potential bias should be considered. All four quantitative subtests, as well as READ, are essentially free of DIF, and SEC contains relatively little DIF, which suggests that there is no serious bias in these six subtests.

Another possible source of construct-irrelevant variability is repeated test taking, the effects of which have been examined in several studies (e.g., Cliffordson, 2004; Henriksson, 1993; Henriksson & Bränberg 1994; Henriksson & Wedman, 1993; Törnkvist & Henriksson, 2004). Findings have shown that, in general, test takers who repeat the test increase their scores in subsequent administrations, and that the greatest increase is between the first and second administration. However, the results showed high variability, which means that several test takers had *decreased* scores in subsequent administrations. The observed increase in score over test repetitions is believed to stem from practice effects, test-wiseness/test-familiarity, self-selection factors, and intellectual growth.

Furthermore, construct-irrelevant variability could result from undue practice, also called mechanical practice or coaching, through which test takers are able to solve items correctly based on special principles or solution strategies rather than ability (Skolöverstyrelsen, 1976; Högskoleverket, 2004; UHR, 2016). This is not likely to have been a problem in previous versions of the SweSAT, as any potential subtests that were found to be sensitive to coaching were eliminated already during the "try-out" of that

subtest (Henrysson, 1994), nor is it likely to be a problem in SweSAT-11 because during its development the only subtest – *verbal analytical reasoning* – that was believed to be sensitive to coaching was excluded at an early stage of development (Ögren & Stage, 2009). No further problems with coaching or mechanical practice have been reported for any of the other SweSAT-11 subtests.

A1.5 Test-takers with a high level of scholastic proficiency perform better in higher education than test takers with a low level of scholastic proficiency. The predictive validity study by Lyrén et al. (2014) showed that there was a positive correlation between normed score on SweSAT-11 and success in higher education for eight out of the eleven examined programs. For the remaining three programs, the correlation between these two variables was not significant. Thus, it seems as though test takers with a higher level of ability will perform better in their higher education studies than test takers with a lower level of ability.

A1.6 Test-takers with a high level of quantitative ability perform better in higher education with a quantitative/numerical orientation than test takers with a low level of quantitative ability. The findings from Lyrén et al. (2014) showed that placing more weight on the quantitative section of SweSAT-11 than the verbal section translated to a small increase in predictive ability for three of the four studied technical programs (Master in Engineering, Bachelor in Engineering and Master in Business and Economics) whereas the predictive ability decreased for Biomedical Laboratory Science. These findings show that, in general, test takers with a high quantitative ability perform better in higher education with a quantitative/numerical orientation that test takers with a low ability, and support the existence of an independent quantitative section, although the decision to weight the sections differently is questionable due to the marginally positive effects of doing so.

A1.7 Test-takers with a high level of verbal ability perform better in higher education with a verbal orientation than test takers with a low level of verbal ability. The findings from Lyrén et al. (2014) showed that placing more weight on the verbal section of SweSAT-11 than the quantitative section would only increase predictive ability for one of the seven assessed verbally-oriented programs (Subject matter teaching) whereas the predictive remained the same for the other six verbally-oriented study programs (Agriculture; Early Childhood Education; Medicine; Pharmacy; Nursing; and Social Work). These results show that, in general, test takers with a high verbal ability do not perform better in higher education with a verbal

orientation than test takers with a low verbal ability, and do not support a future decision to weight the sections differently.

A1.8 Using SweSAT scores for admission decisions will not lead to significantly lower levels of performance in higher education than when GPA is used for the same decisions. A review of international research concerning instruments for predicting success in higher education in the 1960s found GPA to be the best predictor of success in higher education, followed by aptitude tests and intelligence tests (Kompetensutredningen, 1968). This meant that from the very start of the SweSAT development process, the developers were aware that using the SweSAT for selection to higher education would lead to lower values for whatever criterion was used to represent "performance in higher education". Thus, researchers and policy-makers have always understood that the use of any general selection instrument other than GPA would result in slightly lower performance in higher education. However, SweSAT-11 could never lead to *significantly* lower levels of performance in higher education because it is a selection instrument rather than an eligibility instrument. Hence, in order to be accepted to higher education, even with a SweSAT-score, a student still needs to be eligible, which means that the student must meet prerequisites believed to be sufficient for participating in the higher education program.

A1.9 Using SweSAT scores for admission decisions will not have a negative impact on instruction in primary and secondary school. This topic has received limited research attention, with only one study investigating the effect of ERC on the subject English in Swedish upper-secondary schools. The study (Ohlander, 1999) included a survey that was sent out to 3,069 English teachers. Although the answer rate was low, 15%, the answers collected did not indicate that ERC negatively impacted how English was taught. Only a single teacher out of the 470 who had answered the survey stated that ERC affected their teaching, but they did not state whether this was in a negative or positive way. Thus, there is nothing in the scarce evidence that suggests that ERC negatively impacts instruction in primary and secondary school. More research is required on the remaining subtests, and newer research concerning ERC would also be beneficial.

The SweSAT is not expected to have a negative impact on prior education because it has always been constructed to measure something other than the secondary school curriculum (e.g., Kompetensutredningen, 1966). Tillträdesutredningen (2017) suggested that there should be an age limit of 19 years (with some exceptions) for taking SweSAT-11, which today has no age limit. However, this proposition was not explicitly motivated by concerns of a negative impact on instruction in upper-secondary school, but because of a belief that it is important to send upper-secondary school pupils a

message that their studies should be prioritized while they are in still in school (Tillträdesutredningen, 2017). Furthermore, one effect of using the SweSAT as a "second chance" is that it can make upper-secondary school studies more relaxed for students (Regeringens proposition 1972:84). Thus, using SweSAT as an admissions instrument involves the delicate task of allowing studies to become more relaxed, but not *too* relaxed.

Reliability proposition

A2.1 The sample of measurements is large enough to keep sampling error at an acceptable level. This proposition can be evaluated using reliability measures from different administrations. SweSAT-11 has 160 items and there have been between 40,431–76,094 test takers in each of its twelve administrations since the fall of 2011. The reliability, measured by coefficient alpha, has ranged from 0.93–0.94 for the whole test, 0.89–0.92 for the quantitative section and 0.90–0.92 for the verbal section. Thus, the reliability of the measures has been large and consistent over time, effectively minimizing sampling error. The reliability of the total test has also been consistently larger than 0.90, which was suggested by Nunnally and Bernstein (1994) to be the minimum acceptable requirement for high-stakes tests in which the scores are used for decision-making on an individual level.

Scoring, scaling and norming propositions

A3.1 The scoring protocol is appropriate. The scoring protocol for SweSAT-11, as well as its earlier versions, is the answer key because all the items in SweSAT-11 are multiple-choice. The items, along with the answer keys, are examined by professional review boards, and this is done separately for each subtest. Each item, along with its answer key, is reviewed three times before the item is tried out. After an item is tried out, the item and answer key are examined again if necessary. This ensures that the scoring protocol is appropriate.

A3.2 The scoring protocol is applied accurately and consistently. Answers to SweSAT test items are marked on a separate answer sheet, which is then scanned using three dedicated Canon DR-X10C optical scanners and scored using software that is able to differentiate between non-answers, answers and smudges or erased answers. All of the scanned sheets are saved as two copies, one in black-and-white and one in color. In the case of double markings or uncertainty, the answer sheet is flagged and edited manually. This procedure guarantees that the scoring protocol is applied consistently.

A3.3 The normed scales used for the section scores and for the total score are appropriate. The scale used for the SweSAT-11 is a normally distributed, 41-point scale ranging from 0.00–2.00 with intervals of 0.05. The scale is the arithmetic mean of the quantitative and verbal section scores, which each are reported on a 21-step scale ranging from 0.0–2.0 with intervals of 0.1. Kolen (2006) suggested that test score scales should have neither too many nor too few scale steps; too many steps would imply that practically meaningless score differences could still be differentiated whereas too few steps would lead to a loss of score precision. Kolen (2006) identifies two rules of thumb that can be used to determine the appropriate number of scale steps for a test based on its reliability: The Iowa Tests of Educational Development (ITED) rule and Kelly's rule, where Kelly's rule generally suggests approximately twice as many scale steps as the ITED rule. When the number of scale steps for the SweSAT is analyzed using reliability coefficients of 0.93 and 0.94 (see *A2.1* for details), the ITED rule suggests 23–24 scale steps while Kelley's rule suggests 68–73 scale steps. As the SweSAT's 41 scale steps lie between those proposed by the ITED and Kelley's rule, the scale arguably contains neither too many nor too few scale steps and is therefore appropriate to use.

Each section of SweSAT-11 also has normed scales. An analysis of the scale steps for the quantitative section using reliability coefficients of 0.89 and 0.92 (see *A2.1* for details) results in 18–21 scale steps (ITED rule) or 54–64 scale steps (Kelley's rule). For the verbal section, the same analysis using reliability coefficients of 0.90 and 0.92 (see *A2.1* for details) proposes either 19–21 scale steps (ITED rule) or 57–64 scale steps (Kelley's rule). As each section contains 21 scale steps, both scales, even though on the lower end, are appropriate to use according to the presented rules of thumb. The section scales, as well as the total score scale, are also in line with the seven properties that a well-aligned score scale should possess, as suggested by Dorans (2002).

A3.4 The scaling procedure for transforming observed scores to scale scores is appropriate. The SweSAT scaling procedure involves three steps and is conducted separately for each section. The results from the equating procedure (described in *A3.8*), which includes all test takers, are compared to results from the equating procedures under the equivalent groups design, which is applied to two reference groups. The reference groups are similar from year to year. Reference group I consists of a randomly selected, stratified sample of 20,000 test takers that stays constant across administrations in terms of the gender, education and age distribution. The second reference group comprises all the test takers that are upper-secondary school students, in a three-year upper-secondary school program, and between the ages of 18–20 (spring administration) or 19–21 (fall

administration). The results from these three procedures are then compared and the observed scores are converted into the two normed scores (one for each section), each on a normally distributed, 21-point scale ranging from 0.0–2.0 in steps of 0.1.

The appropriateness of the scaling procedure depends on the appropriateness of the reference groups. Two large reference groups based on user norms (test takers) can facilitate score interpretation as they tend to remain fairly stable over short periods of time (Kolen, 2006). Short-term stability is sufficient because each reference group is only used for one administration. Therefore, it can be argued that the scaling procedure is appropriate because it is based on all test takers as well as two appropriate reference groups.

A3.5 The scaling procedure for transforming observed scores to scale scores is applied accurately and consistently. The observed scores are transformed to scale scores with an algorithm that is identical for all test takers. The algorithm is based on the outcome of *A3.4*. This ensures that the algorithm is correct and that it is the same for all test takers, which thus means that it is applied accurately and consistently.

A3.6 The procedure for transforming the two separate scale scores into a single scale score is appropriate. This procedure involves the calculation of a simple arithmetic average from the two scale scores, which produces an approximately normally distributed, 41-point scale ranging from 0.00–2.00 in steps of 0.05.

A3.7 The procedure for transforming the two separate scale scores into a single scale score is applied accurately and consistently. Statistical software that applies the same algorithm to each test taker is used to average the two scale scores.

A3.8 The equating procedure (framework, design and method) is appropriate. This proposition was investigated in Study IV, on equating, and the findings did not suggest the current equating procedure to be inappropriate according to any of the three evaluation criteria included in the study. The current equating procedure applies traditional equipercentile equating with a non-equivalent groups with anchor test design and utilizes a post-stratification equating method. However, this topic should be researched further, and a study with simulated data in addition to empirical data would be the optimal method for gaining definitive information on the appropriateness of the current equating procedure.

Among the three commonly used designs – single group, equivalent groups and non-equivalent groups with anchor test – the single group design

is inappropriate because test takers differ between administrations and the equivalent groups design is inappropriate because in introduces systematic bias into tests such as the SweSAT, which cannot be reused for security reasons (von Davier, Holland, & Thayer, 2004). This means that out of the three commonly used designs, the non-equivalent groups with anchor test design, which is currently used to equate SweSAT-11, is the most appropriate. A recently developed design, non-equivalent groups with covariates (Wiberg & Bränberg, 2015), is potentially interesting but must be further tested using SweSAT data. Previous research suggests that among the two frameworks, traditional equating and kernel equating, neither appears indisputably superior over the other (von Davier et al., 2006), suggesting that either could be used to equate SweSAT-11. This could not be confirmed in Study IV because of the considerable differences between the methods in terms of equated scores and therefore needs to be tested further.

A3.9 The equating procedure is applied accurately and consistently. The equating procedure (e.g., Lyrén, Cobian, & Bardt, 2017) is performed using *Common Item Program for Equating* software (Kolen, 2004) that includes all test takers with valid scores and two reference groups to ensure accuracy. In each SweSAT administration, the scores of all test takers are converted using the same equating function. Moreover, the process is identical over different administrations. Both of these practices ensure consistency.

Acceptance proposition

A4.1 SweSAT scores are accepted as meaningful for admission decisions by the users of the admission system, including test takers. Eklöf, Lyrén and Lindberg (2013) examined test takers' perceptions of SweSAT-11. They found that 61% of test takers perceived the verbal section to be relevant or highly relevant for all areas of higher education. A further 30% perceived the SweSAT to have moderate relevance and 7% perceived little or no relevance (2% non-responses). The quantitative section received somewhat less support, with 45% of test takers reporting that the quantitative section was relevant or highly relevant for all areas of higher education, while 35% perceived moderate relevance and 17% perceived little or no relevance (3% non-responses). Men were more likely to perceive the quantitative section as being relevant (and easier) than women, which partially explains why the quantitative section received less support than the verbal section, as 62% of the respondents (n=1,778) were female. Perception of relevance was positively correlated with performance; that is, test takers with higher scores tended to view SweSAT-11 as more relevant to various aspects of higher education than test takers with lower scores. The Swedish and English

reading comprehension subtests were perceived as the two most relevant subtests in SweSAT-11.

Although this study found support for the perceived relevance of the SweSAT, it would be interesting to replicate this study using new data, as organized cheating has received much attention during recent years and may thus have negatively affected the public opinion of SweSAT-11 (Tillträdesutredningen, 2017). If public acceptance of SweSAT-11 has decreased significantly since the findings presented by Eklöf, Lyrén and Lindberg (2013) it would be important to take measures to restore confidence in using SweSAT as a selection instrument to higher education.

3.3.4 Propositions for SweSAT-11 when used for providing diagnostic information

Validity propositions

B1.7r The quantitative subtest scores provide information that is useful for remediation and that is not already provided by the quantitative section score. The findings in Wedman & Lyrén (2015) show that among the quantitative subtests, DTM provides information about test takers' strengths and weaknesses that is not already provided by the quantitative section score and that XYZ lies on the border of providing additional information as it sometimes does and sometimes does not provide additional information due to sampling variability. QC and DS do not provide additional information. This suggests that the QC and DS scores lack adequate psychometric quality and should not be used for remedial decisions. The DTM subtest scores may be used for remedial decisions and it is plausible that XYZ scores can be used for remedial decisions as the subtest showed reliability (coefficient alpha) of 0.80 and therefore, as suggested by Haberman (2008) concerning "rather reliable" subscores, can be used to produce a relatively accurate approximation of the true subscore.

B1.8r The verbal subtest scores provide information that is useful for remediation and that is not already provided by the verbal section score. The findings in Wedman & Lyrén (2015) show that among the verbal subtests, only WORD likely provides any additional information to what is already provided by the verbal section score when using the argument by Haberman (2008) that was applied to XYZ (see *B1.7r*). READ, SEC and ERC do not provide additional information. This indicates that the READ, SEC and ERC scores lack adequate psychometric quality and should not be used for remedial decisions. The WORD subtest scores can likely be used for remedial decisions.

B1.9r The section scores provide information that is useful for remediation and that is not already provided by the total score. The findings in Wedman & Lyrén (2015) show that the scores of both sections – verbal and quantitative – provide additional information to what is already provided by the total score. This indicates that the section scores have adequate psychometric quality and that they may be used for remedial decisions.

Acceptance propositions

B4.4r Subtest scores are accepted as meaningful by test takers. No studies have yet investigated test takers' opinions of reported subscores from SweSAT-11, but the results of a pilot study by Lyrén (2008), which covered the SweSAT-96 score reports, found that a slight majority of test takers felt that the SweSAT score reports provided information on their strengths and weaknesses. Furthermore, a slight majority of test takers believed that SweSAT test scores could be informative from a remedial perspective, and thus, the subtest scores were accepted as meaningful. However, research that investigates the perceived value of section scores and subtest scores in SweSAT-11 is still needed.

4. Methods

The research underlying this dissertation used different methods to analyze quantitative data. Most of these methods are based on classical test theory (CTT) (e.g., Lord & Novick, 1968; Crocker & Algina, 1990), also called true score theory, because this theory has always been used when constructing and assembling the SweSAT. Other methods applied item response theory (IRT) (e.g., Lord & Novick, 1968; Hambleton & Swaminathan, 1985), either unidimensional or multidimensional, which can provide useful information due to the sheer size of the SweSAT datasets.

Study I is a literature review and therefore does not contain any statistical methods. Study II, which focuses on subscores, utilized several statistical methods. Although these methods were used to assess whether subtest scores provide additional value to section scores and whether section scores provide additional value to the total score, they can nevertheless be divided into groups. Haberman's method and utility index are fully interchangeable methods that rely heavily on information about internal consistency reliability and a subscore's correlation with the total score. Factor analysis and DETECT are exploratory methods that assess optimal dimensionality by reducing the number of variables into considerably fewer factors/clusters. DIMTEST employs significance testing to detect differences between subtest scores within the same section. Haberman's method, utility index and factor

analysis are based on CTT, whereas DIMTEST and DETECT are based on IRT. Two other methods used in Study II, subscore augmentation and multidimensional IRT-estimates, which are based on CTT and IRT, respectively, were employed in an attempt to add information to the subscores instead of assessing the psychometric quality of the already provided subscore information. Study II also applied more conventional methods, such as Pearson correlation, reliability coefficient alpha and correction for attenuation.

Two CTT-based methods were used in Study III, on gender-related DIF, with the objective that both methods would produce the same results through different statistical procedures. The first method was the Mantel-Haenszel procedure, which compares, on an item level, differences in the odds-ratios of giving a correct response between women and men, conditional on ability. This method can detect uniform and ordinal non-uniform DIF, which, however, becomes classified as uniform DIF. The second method was logistic regression, which compares the coefficient of determination (explained variance: R^2) between separate regression models for women and men, also conditional on ability. This method can detect uniform, ordinal non-uniform, and disordinal non-uniform DIF.

Study IV, which investigated equating, used both CTT and IRT methods to equate a given test form to a previous test form. Equating employs different designs, methods and frameworks based on which are most applicable to specific conditions. The most appropriate design, given the data used in the research, was non-equivalent groups with anchor test design (NEAT). Under the NEAT design, equipercentile equating was performed using: (a) the chained equating method, both as a CTT version and as an IRT version; and (b) the similar, CTT-based, post-stratification equating method. Equipercentile equating was also performed under the single group and equivalent groups designs using both the traditional framework and the kernel equating framework.

5. Summary of studies

The four studies included in this thesis relate to different aspects of validity in SweSAT-11. These studies are summarized below.

5.1 Study I – Theoretical model

From aptitude to proficiency: The theory behind the SweSAT

The theoretical framework of a test is an important part of its validity arguments, which concern the interpretation and uses of its scores.

Surprisingly, no theoretical model has been formulated for the SweSAT since its introduction 40 years ago.

The purpose of this study was to, for the first time, formulate and present a theoretical model of the original SweSAT-77 and the currently used SweSAT-11. The purpose also entails an adequate description of the construct *scholastic aptitude* and its successor *scholastic proficiency*.

The study, which took the form of a literature review, investigated peer-reviewed research studies, reports from the Swedish School Board, Official Reports of the Swedish Government, Government propositions, as well as memos and reports from the different departments at Umeå University that have been responsible for developing most of the subtests in the SweSAT.

The construct *scholastic aptitude* was defined by Skolöverstyrelsen (1976) and included four domains: Aptitude; abilities; knowledge; and personality traits. Personality traits had shown practically no correlation with success in higher education, and practical limitations restricted the content of the other three domains. In the end, SweSAT-77 measured the entire knowledge domain and only parts of the aptitude and abilities domains.

SweSAT-77 originally measured *scholastic aptitude* but, in 1985, *scholastic proficiency* replaced *scholastic aptitude* as the construct that SweSAT-77 was intended to measure. However, the domains of the construct *scholastic proficiency* were initially identical to that of its predecessor, *scholastic aptitude* (Tillträdesutredningen, 1985a). A suggestion by the SweSAT's international expert committee in 1994, along with dimensionality analyses based on empirical SweSAT data (e.g., Gustafsson et al., 1992; Svensson et al., 2001) eventually led to the SweSAT adopting a two-dimensional structure that measured quantitative and verbal abilities instead of the original three-dimensional structure that aimed to measure aptitude, abilities and knowledge. Despite these changes in construct and structure, the core purpose of the SweSAT remained very similar: to reflect what is required to succeed in higher education (Andersson, 1999, p. 7).

The initial goal of SweSAT-11 development was to create two self-sustained sections, one quantitative and one verbal. New subtests were tried out by Stage and Ögren (2007) and Ögren, Lexelius and Stage (2008), leading to the inclusion of three new subtests that, together with the other subtests, made up two distinct sections, quantitative and verbal. This was in agreement with previous research and suggestions by the SweSAT's international expert committee.

The theoretical models of SweSAT-77 and SweSAT-11 are markedly different from each other. The three-dimensional model of SweSAT-77 is theoretically based and more elaborate, whereas the two-dimensional model of SweSAT-11 is empirically based and less theoretically elaborate. However, some of the changes between models were purely superficial. For example, certain subtests remained the same between versions but were defined as

measuring different contents as the theoretical model of the SweSAT developed.

The principal reason behind changes to the SweSAT and the construct it intends to measure was the belief that separate scoring of a quantitative and verbal section, with the option of weighting the scores differently, would increase the SweSAT's predictive validity (Högskoleverket, 2002). This was partially supported by findings from Lyrén et al. (2014), who reported that the predictive validity of SweSAT-11 was higher than that of its predecessor, SweSAT-96. However, they also reported that differential weighting of the sections would lead to only a trivial increase in predictive validity for the majority of higher education programs.

5.2 Study II – Subscores

Methods for examining the psychometric quality of subscores: A review and application

If subscores are to be reported to test takers, they must tell test takers more about their strengths and weaknesses than what is already provided by the total score. Subscores that accomplish this are considered to have adequate psychometric quality over the total score whereas subscores that fail to accomplish this are considered to lack adequate psychometric quality and should not be reported to test takers. The SweSAT reports subscores to test takers on two levels: the section level and the subtest level, and according to *Standards* (AERA, APA, & NCME, 2014), the appropriateness of reporting these subscores must be evaluated (Standard 1.14, p. 27; Comment to Standard 1.14, p. 27).

The purpose of this study was to review different methods that are used to assess subscore value and to apply each method to SweSAT data in order to examine the value of the five subtest scores in SweSAT-96 along with the eight subtest and two section scores in SweSAT-11. The objective also included an assessment of whether more information than what was provided by the observed subscore could be added to the subscore.

There are several methods for examining subscore value, and this study reviewed seven such methods and then applied them to SweSAT data. Four of the methods were based on classical test theory and the remaining three were based on item response theory. The methods reviewed were Haberman's method, Utility index, factor analysis, subscore augmentation, DIMTEST, DETECT, and multidimensional item response theory. Subscore augmentation and multidimensional item response theory are not subscore value methods *per se*, but rather focus on the possibility of *improving* subscore information by using information from other subtests and by applying models based on multidimensional item response theory.

Data from four administrations of the SweSAT were analyzed using each of the seven methods. Two of the test forms came from SweSAT-96,

administered in the fall of 2010 and the spring of 2011, and two came from SweSAT-11, administered in the fall of 2011 and the spring 2012. The data comprised test takers' scores on each item.

The results showed large variation in whether or not subtest scores were considered to have added value, but all of the seven subscore value methods showed that both SweSAT sections provide added value to the total score. Among the two methods that focus on the possibility of improving subscore value, subscore augmentation did not add information to the sections while multidimensional item response theory did.

The overall recommendation was that Haberman's method should be used as it likely fulfills the needs of most test developers and users. The results from Haberman's method indicated that both section scores have adequate psychometric quality and can be reported to test takers while most SweSAT-11 subtest scores should not be reported to test takers. The study also indicated that future research should investigate the implications of using subscores that lack adequate psychometric quality for remediative purposes.

5.3 Study III – Differential item functioning

Reasons for gender-related differential item functioning in a college admissions test

Test fairness is a key issue in any testing situation. One tenet of the concept of fairness is that every test taker at a certain ability level should have the same probability of answering an item correctly. If members of different groups of test takers, given that they are equally proficient, have the same probability of answering an item correctly then that item is free from DIF. On the other hand, if women, for example, have a lower probability of answering an item correctly than men even though they are equally proficient, then this item displays DIF.

The purpose of this study was to examine the presence and causes of gender-related DIF in SweSAT-11, and there were five hypotheses. The first and second hypothesis were that there was a moderate amount of DIF present in both sections of the SweSAT. The third and fourth hypothesis were that DIF in the quantitative section was linked to the mathematical operations required to solve an item, but not to its content domains. The final hypothesis was that DIF in the verbal section was linked to content domains.

Five administrations of SweSAT-11, with approximately 250,000 test takers and a total of 800 items, were included in this study. The data comprised the test takers' scores on each item. Two methods – Mantel-Haenszel and logistic regression - were used to examine DIF. Mantel-Haenszel can detect both uniform DIF, in which the difference in proportion correct answers between women and men stays constant across all ability

levels, and ordinal non-uniform DIF (which, however, becomes classified as uniform DIF), in which the difference in proportion correct answers between women and men differ across ability levels (Swaminathan & Rogers, 1990). Thus, in ordinal non-uniform DIF the difference in proportion correct answers between women and men can gradually increase or decrease across ability levels. Logistic regression, on the other hand, can detect uniform and ordinal non-uniform DIF (which is correctly classified as non-uniform DIF), as well as disordinal non-uniform DIF. Disordinal non-uniform DIF exists when the difference in proportion correct answers between women and men differs across all ability levels (Swaminathan & Rogers, 1990), and where these differences will cause an item, for example, to benefit low-ability men as well as high-ability women, or vice versa.

Mantel-Haenszel found more DIF than logistic regression, and the results showed that there was practically no DIF in the quantitative section of SweSAT-11. This rejected the first hypothesis, yet the results revealed a moderate amount of DIF in the verbal section, confirming the second hypothesis. DIF was found in all of the subtests in the verbal section, with the most occurring in WORD and ERC. Most DIF items in WORD favored women whereas all of the DIF items in ERC favored men. The lack of DIF in the quantitative section meant that the third and fourth hypotheses could neither be confirmed nor rejected. The final hypothesis, that DIF would be linked to content domains in the verbal section, was confirmed as almost all of the DIF items in WORD favored women and appeared to stem from traditionally female domains. The DIF in ERC was found to be related to the item format "sentence completion", which suggests that this item format measures something different from what is measured by the "reading comprehension" item format.

The findings suggest that SweSAT-11 includes certain gender-related bias, and that this bias favors women in WORD and men in ERC. A cross-validation of this study that would use data from more administrations of SweSAT-11 should be conducted in the future. If this study confirms the earlier DIF results then measures can be taken to reduce bias and thus increase the validity and test fairness of SweSAT-11.

5.4 Study IV – Equating

Equating challenges when revising large-scale tests: A comparison of different frameworks, methods and designs

When the results of a certain test are valid for several years, and when test takers with valid scores from different administrations compete with one another for admission to higher education, it is crucial that scores from different administrations are equivalent. The process of ensuring the equivalence of test scores is called equating, or, in a wider context, linking.

This study had two distinct purposes. First, the study aimed to determine the best way to link SweSAT-11 to SweSAT-96 (with fewer items and subtests) while maintaining the validity of the test scores. The second purpose was to find a best practice, in terms of framework, method and design, for equating SweSAT-11.

Two frameworks, kernel equating and traditional equating, were included in the study, and the hypothesis was that the differences between the frameworks would be negligible. Several data collection designs – single group, equivalent groups, and nonequivalent groups with anchor test as well as nonequivalent groups with covariates – were considered for each of the frameworks. Moreover, the focus was on equipercentile equating using kernel equating methods, especially chained and post-stratification equating for the NEAT design and kernel equating for the NEC design.

The study included data from four administrations of the SweSAT, with one test form from SweSAT-96 (spring of 2011) and three from SweSAT-11 (fall of 2011, spring of 2012 and fall of 2012). Several samples were used depending on the test form and equating design, with sample sizes ranging from 561 to 59,332 test takers. The data comprised the test takers' scores on each item although some designs only used section scores and anchor test total scores. The evaluation criteria were percent relative error, the standard error of equating, and difference that matters.

The results showed that kernel equating was preferable to traditional equating under the single group design when SweSAT-11 was linked to SweSAT-96. However, it was unclear whether the kernel equating or traditional framework was advantageous when different test forms of SweSAT-11 were equated. Therefore, the recommendation was that SweSAT-11 equating should continue with traditional equating methods but the possibility of switching to kernel equating in the future should not be overlooked, but rather researched further.

6. Discussion

The four studies presented in this thesis have examined the theoretical basis for the validity arguments, along with different sources of validity evidence and threats to validity, in the context of SweSAT-11. The research identified the strengths of the SweSAT-11, from the perspective of validity, to be high quality and overall fairness, while the weaknesses were partial unfairness, manifested through DIF, and theoretical gaps.

6.1 Main results – implications for the SweSAT and its test takers

Overall, the findings from the research underlying this thesis provide some evidence in favor of the validity of SweSAT and some evidence that identifies gaps in the validity arguments.

The theoretical framework for the SweSAT is well-motivated and has a strong scientific base. However, the construct *scholastic proficiency*, which is key to SweSAT-11, would benefit from a more detailed description of what this construct contains. One strategy would be to follow current research and the suggestions of representatives from higher education institutions, which was the procedure used to define *scholastic aptitude* (Skolöverstyrelsen, 1976). Validity evidence based on test content found that SweSAT-77, -92 and -96 all suffered from construct underrepresentation that was mainly caused by the demand for fast, inexpensive and objective scoring. Whether this underrepresentation is, or is not, present in SweSAT-11 depends on the definition of the construct.

The validity evidence based on internal structure and consequences of testing concerned the decision to report subscores to test takers as well as the presence and causes of gender-related DIF. The research found statistical evidence for the practice of reporting section scores, but did not identify sufficient evidence for the reporting of subtest scores. The research also found empirical support for the two separate sections – quantitative and verbal – in the theoretical model of SweSAT-11. The lack of DIF in the quantitative section provided evidence that this section has valid internal structure, but the presence of moderate DIF in the verbal section, caused by item content (which favored women) and item format (which favored men), highlighted the need for possible improvement in this section. In the same study, the validity evidence based on response processes were in accordance with what was found for internal structure. Furthermore, there was no indication that the operational equating method of SweSAT-11 is inappropriate, but a full understanding of this topic will require more research.

6.1.1 Study I – Theoretical model

The theoretical model was compiled to provide a theoretical basis for the validity arguments for SweSAT-11, as this is a requirement, alongside an investigation of all five sources of validity evidence, in the validity framework presented in *Standards* (AERA, APA, & NCME, 2014). This study also provided validity evidence based on test content.

Validity evidence based on test content found that SweSAT-77, -92 and -96 all suffered from construct underrepresentation that was largely caused by the demand for fast, inexpensive and objective scoring. This demand eliminated the possibility to measure a test taker's ability to (1) give oral or

written statements, (2) critically scrutinize and present their opinion on given information, and (3) perform quantitative computations and compilations, which were all part of the construct *scholastic aptitude* (Skolöverstyrelsen, 1976). Excluding these measures could wrongly affect the rank-ordering of a group of test takers who performed relatively better or worse on any of these eliminated ability measures than another group of test takers, and could thus alter who is accepted to higher education and who is not.

The results also uncovered a misconception concerning the original intended test-taking population. The misconception was that the SweSAT had been originally constructed with the intent of being used only by the 25/4:s and not by upper-secondary school students (e.g., Högskoleverket, 2002). This was not the case, as the SweSAT was constructed for, and available to, all potential test takers (Kompetensutredningen, 1970) although the 25/4:s were the only ones that were allowed to use their scores for selection to higher education during the initial administrative trial period of the test (Regeringens proposition 1975:9). The critique surrounding the perceived change in test-taking populations (Högskoleverket, 2002) was thereby invalidated.

The main provision of the study is that SweSAT-11 gained a long-needed theoretical framework that can be traced back all the way to SweSAT-77, and which could support claims of validity for the interpretations of the proposed uses of SweSAT-11 test scores. The second major finding was that the SweSAT-11 framework could benefit from a more elaborate definition of its construct, *scholastic proficiency*, which could then lend greater support for the validity of said interpretations.

A practical implication for test takers would be that they can get a complete and coherent description of what the SweSAT intends to measure as well as information about the studies that motivated the construction and assembly of the test. This would allow for a deeper understanding of the relevance of SweSAT-11 and thereby possibly increase public support for its existence and continued use.

6.1.2 Study II – Subscores

A study of the value of subscores in SweSAT-11 revealed that the quantitative and verbal section scores are distinct from each other, and thus, measure two different aspects of the construct *scholastic proficiency*. This provides empirical support for the decision to divide the SweSAT into two sections and to score, norm and equate them separately. It is also empirical support for the potential decision to weight the scores differently when they are used for selection to different higher education programs.

The finding that most subtest scores lack adequate psychometric quality suggests that reporting them to test takers is incorrect and should be

discontinued. This is based on Standard 1.14 in *Standards* (AERA, APA, & NCME, 2014), which states that there must be evidence of the distinctiveness of subscores when they are reported to test takers. Furthermore, subscores that are not distinct from the total score, by definition, provide the same amount of, or less, information than the total score, but with lower reliability. This implies that, in such cases, the total score is a better predictor of future subtest performance than the subscore itself (Monaghan, 2006). A lack of adequate psychometric quality can become a problem when test takers that repeat a test base their remedial decisions on previous subscores. This could cause them to allocate their study time less effectively than if the allocation was based on the total score.

This finding does not impact SweSAT-11 itself, first-time test takers, or repeat test takers that have not made active remedial study decisions, but the finding has the potential to adversely affect repeat test takers that have made remedial study decisions based on previous subtest scores. When remedial studies are based on a subtest score that lacks adequate psychometric quality and are allocated towards improving proficiency on content from that specific subtest instead of the overall content of the particular section, then there is a theoretical possibility that the improvement in the particular subtest score will be less than it could have been. For example, ERC measures English reading comprehension and the subtest score lacks adequate psychometric quality. If remedial study time with the goal of improving the ERC-score is spent only on tasks relating to English reading comprehension instead of also spending time on Swedish reading comprehension and vocabulary (the other parts of the verbal section), then the ERC score might increase less that if time had been spent on both English and Swedish.

Theoretically, a repeat test taker with a low score on ERC should spend 75% of their remedial study time on Swedish vocabulary and Swedish reading comprehension (60 out of 80 verbal items), and only 25% of their time studying English reading comprehension (20 out of 80 verbal items items). This should increase their ERC score more than if all of their study time dedicated to ERC was spent studying English reading comprehension.

6.1.3 Study III – Differential item functioning

The results from this study suggest that there is no DIF in the quantitative section but that, in the verbal section, content causes DIF in WORD and item format causes DIF in ERC. This provides validity evidence in support of the internal structure of the quantitative section of the SweSAT as it does not contain gender-related bias in terms of DIF. The results also provide validity evidence for READ, which was not found to contain DIF. The validity evidence gathered from WORD and ERC suggests that these subtests may need to be revised.

These results are conditional on using section scores as the matching variables. If other matching variables, such as subtest scores, the total score or some external measure of ability, had been used, then the results would have been different, as shown by Wester (1994). However, considering the results in Wedman and Lyrén (2015), and that the subtest scores were not constructed to be used on their own as measures of ability, the most appropriate choice of matching variable was the section scores.

The findings could have a large impact on the SweSAT, because there would be reasonable cause to modify the test in order to eliminate DIF if these results are replicated in a cross-validation study. For WORD, such a change could mean the exclusion of words that can be argued to come from a traditionally female or male domain. What constitutes female and male domains, and more importantly, what constitutes a neutral domain, could be ascertained in part from the results of this study, the planned cross-validation study, and previous research.

The impact on test takers would be that they may thus face a different SweSAT in the future if the presented findings are reproduced in a cross-validation study. In WORD, there may be fewer words that are tied to such domains as food, food preparation, feelings and horses. On the other hand, ERC may no longer contain items with the format "sentence completion", but rather include more items that measure reading comprehension. The changes could lead to more accurate and valid test scores as well as a SweSAT with less gender-related bias as long as the changes do not introduce new gender bias into the test.

6.1.4 Study IV – Equating

The current operational equating design and framework in SweSAT-11 is the non-equivalent groups with an anchor test design that applies an equipercentile, post-stratification method of equating. The results indicated that this method appears to be a suitable method for equating the SweSAT but were unclear regarding whether the equating procedure could be improved further by using the kernel equating framework. In the context of SweSAT-11, this means that the current operational equating method appears to be appropriate, but there is insufficient evidence to determine whether or not it is the best method available. This, in turn, suggests that test scores from different SweSAT administrations can be used interchangeably for selection to higher education. For test takers this is a partial confirmation of test fairness between administrations.

6.2 Usefulness of the research

The relevance of the research presented in this thesis can be divided into two distinct categories: the usefulness of the thesis as a whole and the usefulness of each separate study. This distinction is important because specific

findings from the studies may be relevant for different populations or recipients.

6.2.1 Usefulness of the thesis as a whole

The SweSAT is a fairly distinct large-scale assessment and, like any test, is dependent on the context in which it is used. However, the process of validating a test, which is shown in practice in this thesis, is intended to be similar across many testing situations. Therefore, the thesis as a whole can be considered to be useful as it is a practical example of test validation for test owners and test developers who have already constructed a test and who are in the process of validating it. The presented research is also applicable to those who are in the process of constructing or developing a test and are interested in figuring out how to build a theoretical model and what methods can be used to examine different types of validity evidence. In sum, this thesis, as a whole, may be relevant for a large part of the testing community.

On a more local level, this thesis, with its focus on theory, validity evidence, and threats to validity in SweSAT-11, will be of interest to the test owners, test developers, and test takers of the SweSAT. The test owners, the Swedish Council for Higher Education, can gain knowledge about the SweSAT-11, which could be useful for decisions regarding future developmental steps. The test developers, located at Umeå University and University of Gothenburg, can benefit by receiving information about which parts of the developmental process that are working well in practice. Finally, test takers can educate themselves on what the test measures and why, gain some reassurance about the interchangeability of different administrations, and understand issues of content, item format and the lack of subscore value.

6.2.2 Usefulness of each study

The generalizability of each study will be discussed in further detail in section 6.3, but in general, different populations can benefit from different parts of this thesis. Study I, which focuses on the theoretical model of the SweSAT, is likely to benefit both those who are in the process of developing a test of their own and those who have already developed a test but are lacking a proper theoretical model for it. They could gain insight into the process of developing a theoretical model for a test, as the study covers the limitations of the process as well as how the theoretical model can change over time, shifting from external sources of validity to using data from previous administrations to develop what the test measures. This is useful when developing both educational and psychological tests.

Study II provides a comprehensive review of several methods for examining subscore value, and contributes information on the benefits and drawbacks of each method as well as a recommendation of which method is

most suitable for practitioners that are reporting a test's subscores (in addition to the total score) to test takers. In this way, these results are likely to be useful for the individuals who are in charge of making decisions on whether or not to report subscores.

Study III, with its focus on DIF, is useful for test developers and test constructors who are responsible for constructing items or assembling tests, as well as to researchers who conduct DIF-studies. Knowledge of the effects of content and item format can provide tools for constructing tests that are more fair to test takers and items that require less revisions before being used (given that they are tried out first). Furthermore, the study provided insight into how different methods for evaluating DIF can give varying results. This finding is especially useful for researchers who are conducting DIF studies and sheds light on the potential problems that can arise from logistic regression and the presence (or absence) of disordinal non-uniform DIF.

Finally, the usefulness of Study IV, which investigated equating, to a wider audience is less clear as there were discrepancies between the comparison of traditional and kernel frameworks in this study and those in previous research (e.g., von Davier et al., 2006). Earlier research suggested that there are little differences between the frameworks, whereas the findings from Study IV revealed relatively large differences. The utility of this research would be more tangible if the discrepancy to prior research could be explained.

6.3 Limitations and generalizability

The data used in the research underlying this thesis were exclusively SweSAT data, and no cross-validation using data from other large-scale tests was performed. The included studies vary in terms of the generalizability of their results. The theoretical model of SweSAT-11 (Study I) is, by definition, not generalizable, but is nevertheless important as it constitutes a theoretical underpinning of the validation arguments for the SweSAT. As stated in section 6.2, it also provides other researchers and test developers insight into the process of developing a theoretical model, although their end product will be different.

The research on subscores (Study II) has elements which are generalizable, such as the problem of sample size when using Haberman's method. This should be a problem for any test as it depends only on reliability and correlation. On the other hand, applying the dimensionality methods factor analysis and DETECT on tests that have been developed using factor analysis (or some equivalent method of dimension reduction) would likely produce different results from when these methods are applied to a test like the SweSAT, which is developed using other premises.

The findings from research on gender-related DIF (Study III) concerning how content explains DIF in the vocabulary subtests could be highly generalizable. If one group of test takers has certain interests or areas with which they are more familiar than another group, then, independent of culture or nationality, it is reasonable for them to score higher on items sampled from those domains when controlling for overall ability. This claim is supported by findings from Carlton and Harris (1992). The DIF identified for the English reading comprehension, which was connected to item format, could be generalizable to other reading comprehension tests of non-native languages, although this is somewhat uncertain and requires further research.

The generalizability of results from Study IV, which investigated equating, to other tests is questionable due to the discrepancy of the results of this study and the results from previous equating studies (e.g., von Davier et al., 2006). More definitive conclusions on the generalizability of the results could be drawn once the cause of this discrepancy is identified.

The generalizability of results from Studies II, III and IV to future administrations of the SweSAT is likely high because the results from analyzes of different administrations were similar. This is not surprising because the SweSAT is constructed to be interchangeable across administrations and the sample of test takers for each administration is very large and relatively similar. The findings from Study II were stable between administrations of SweSAT-96 and SweSAT-11, and previous research on subscores in SweSAT-96 (Lyrén, 2009b) confirms that subscore value remained stable across administrations within the same SweSAT version. Similarly, Study III revealed that the amount of DIF has been fairly stable across administrations, with the exception of the first test form of SweSAT-11, 11B, which contained more DIF than the test forms that came after it. The results presented in Study IV were also stable across administrations, and this is why the results from only one pair of equated test forms are presented in that study.

6.4 Suggestions for further research

There are several interesting lines of research that can be pursued as a result of the research presented in this thesis. The presented theoretical model makes it clear that SweSAT-11 would benefit from a clear and elaborate definition of its construct, *scholastic proficiency*. A methodology similar to the one that led to the definition of the original construct, *scholastic aptitude*, could be pursued, but it should also include input from the SweSAT's international expert committee and empirical data gathered from previous SweSAT-11 administrations.

Results concerning subscore reporting, along with the subsequent discussion, indicate that subscore research should shift from the theoretical

field into the empirical field. A good starting point would be the empirical validation of subscore methods by, for example, assessing repeat SweSAT test takers. Because subtest scores that lack adequate psychometric quality are considered worse predictors of future administrations of themselves than the corresponding section scores, a simple correlational study using repeat test takers could be conducted to investigate whether this holds true in practice and, thereby, validate the claim made by the subscore methods.

Furthermore, it would be interesting to study how test takers who conduct remedial studies based on the subscores that lack adequate psychometric quality actually fare on the test. For example, would test takers who base their remedial study decisions on a subtest score with a correlation of r=0.76, with a future administration of itself, score worse on a future administration of that subtest than test takers who base their remedial study decisions on a subtest score with a correlation of r=0.78? And if so, how much worse? In relation to this point, it would be interesting to examine how a test taker plans their remedial studies when they know and when they do not know if a subtest has added value.

Potential future research on subscore value could answer two central questions: (1) To what extent does remedial studying that is based on subscores that lack added value adversely affect a repeat test taker's future subscore and (2) how many repeat test takers actually use remedial strategies that would subject them to these types of effects? Both of these potential studies should also be conducted in light of the findings in Feinberg and Jurich (2017), which indicate that examining subscore value using Haberman's method with an added effect size could be more informative than using Haberman's method by itself. This finding should be considered in future research that concerns subscores.

The DIF study (Study III) found no DIF, and thus no gender-related bias, in the quantitative section of the SweSAT. This suggests that the section is free from gender-related bias even though there are recurring large score differences between men and women in this section that cannot be explained. The results should be cross-validated with later administrations of the SweSAT and the generalizability should be examined by studying the presence and causes of DIF in other large-scale assessments. Also, all of the non-uniform DIF identified in this study was found to be ordinal non-uniform DIF. Hence, the DIF items contained both uniform and non-uniform DIF, which means that the non-uniform DIF was also detected by the Mantel-Haenszel method, but classified as uniform DIF. This raises the question of whether there is an actual need to examine the presence of non-uniform DIF in the future or whether detecting uniform DIF is sufficient. This question should be pursued in future DIF studies.

The score difference in the quantitative section is, on average, 6.5 observed score points, or about 0.2–0.3 normed score points, to the men's

advantage. The score difference in the verbal section is smaller, but the average is still 2.5 observed score points, or 0.0–0.1 normed score points, to the men's advantage. The potential causes for the score differences between women and men in the SweSAT, especially in the quantitative section, requires more attention.

In terms of equating, a large-scale equating study using a simulated pseudo-test with characteristics resembling those of the SweSAT should be conducted. Such a study could replicate the equating study presented in this thesis but the simulated data would allow comparisons of the outcomes of various equating frameworks and methods under controlled conditions. This could clarify the appropriateness of the operational equating method of SweSAT-11 and whether it is advantageous when compared to different methods, such as the kernel equating framework.

References

ACT Inc. (2017, June 20). Helps and faq's [Information on a website]. Retrieved from http://www.act.org/content/act/en/products-and-services/the-act/help.html.

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (1999). *Standards for educational and psychological testing*. Washington, D.C.: AERA.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, D.C.: AERA.

APA, AERA, & NCME. (1966). *Standards for educational and psychological tests and manuals*. Washington, D.C.: APA.

APA, AERA, & NCME. (1974). *Standards for educational and psychological testing*. Washington, D.C.: APA.

APA, AERA, & NCME. (1985). *Standards for educational and psychological testing*. Washington, D.C.: APA.

APA, AERA, & National Council on Measurements Used in Education. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, D.C.: APA.

Andersson, K. (ed.) (1999). *Högskoleprovet: Konstruktion, resultat och erfarenheter* [The SweSAT: Construction, results and experiences]. (Report No. PM nr 153). Umeå, Sweden: Umeå University.

Carlton, S. T., & Harris, A. M. (1992). *Characteristics associated with differential items functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons*. (Report no. ETS-RR-92-64). Princeton, NJ: Educational Testing Service.

Centrala organisationskommittén för högskolereformen. (1975). *Högskolelag och andra författningar för högskolan* [Laws and other constitutions for higher education]. Stockholm, Sweden: Utbildningsdepartementet.

Chubbuck, K., Curley, W. E., & King, T. C. (2016). *Who's on first? Gender differences in performance on the SAT test on critical reading items with sports and science content*. (Report No. RR-16-26). Princeton, NJ: Educational Testing Service.

Cliffordson, C. (2004). Effects of practice and intellectual growth on performance on the Swedish Scholastic Aptitude Test (SweSAT). *European Journal of Psychological Assessment*, *20*(3), 192–204.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart and Winston.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

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

Dorans, N. J. (2002). *The recentering of SAT scales and its effetcs on score distributions and score interpretations*. (Report no. ETS RR-02-04). Retrieved from https://www.ets.org/Media/Research/pdf/RR-02-04-Dorans.pdf.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist*, *16*, 640–647.

Eklöf, H., Lyrén, P.-E., & Lindberg, J. (2013, June). *Test-takers' perceptions of the SweSAT*. Paper presented at the 14th SweSAT Conference, Umeå, Sweden.

Feinberg, R. A., & Jurich, D. P. (2017). Guidelines for interpreting and reporting subscores. *Educational Measurement: Issues and Practice, 36*(1), pp. 5–13.

Gustafsson, J.-E., Wedman, I., & Westerlund, A. (1992). The dimensionality of the Swedish Scholastic Aptitude Test. *Scandinavian Journal of Educational Research*, *36*(1), 21–39.

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston, MA: Kluwer-Nijhoff.

Henriksson, W. (1993). *The problem of repeated test taking and the SweSAT*. (Report no. Em nr 5). Umeå, Sweden: Umeå University.

Henriksson, W., & Bränberg, K. (1994). The effects of practice on the Swedish Scholastic Aptitude Test (SweSAT). *Scandinavian Journal of Educational Research*, *38*(2), 129–148.

Henriksson, W., & Wedman, I. (1993). *Effects of repeated test taking on the Swedish Scholastic Aptitude Test (SweSAT)*. (Report no. Em nr 8). Umeå, Sweden: Umeå University.

Henrysson, S. (1994) *Högskoleprovets historia: Några bidrag* [The history of the SweSAT: Some contributions]. (Report No. PM nr 91). Umeå, Sweden: Umeå University.

Henrysson, S., & Wedman, I. (1975). *The Contents of the Scholastic Aptitude Test*. (Report No. Spånor från Spint Nr 3). Umeå, Sweden: Umeå University.

Hidalgo, M. D., Gómez-Benito, J., & Zumbo, B. D. (2014). Binary logistic regression analysis for detecting differential item functioning: Effectiveness of R2 and delta log odds ratio effect size measures. *Educational and Psychological Measurement*, *74*(6), 927–949.

Högskoleverket. (2002). *The Swedish national aptitude test: A 25-year testing program, current status and future development*. (Report No. SOU 2004:29). Stockholm, Sweden: Högskoleverket.
Högskoleverket. (2004). *Tre vägar till den öppna högskolan* [Three roads to open higher education]. Retrieved from http://www.regeringen.se/rattsdokument/statens-offentliga-utredningar/2004/03/sou-200429/.
International Test Commission [ITC]. (2013). *ITC guidelines on test use*. Retrieved from https://www.intestcom.org/files/guideline_test_use.pdf.
Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. doi:10.1037/0033-2909.112.3.52
Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: Praeger Publishers.
Kolen, M. J. (2004). Common Item Program for Equating (CIPE) version 2.0 [computer program]. Available online: https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs#equatinglinking.
Kolen, M. J. (2006). Scaling and Norming. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 155–186). Westport, CT: Praeger Publishers.
Kompetenskommittén. (1974). *Om behörighet och antagning till högskolan* [On eligibility and admisssion to higher education]. Stockholm, Sweden: Utbildningsdepartementet.
Kompetensutredningen. (1966). *Tillträde till postgymnasala studier: Förslag till provisoriska bestämmelser* [Access to post upper-secondary school education: Suggestions for provisional regulations]. Stockholm, Sweden: Ecklesiastikdepartementet.
Kompetensutredningen. (1968). *Studieprognos och studieframgång* [Prediction of education and educational success]. Stockholm, Sweden: Ecklesiastikdepartementet.
Kompetensutredningen. (1970). *Behörighet, meritvärdering, studieprognos: specialundersökningar av kompetensfrågor* [Eligibility, assessment of merits, prediction of education]. Stockholm, Sweden: Ecklesiastikdepartementet.
Lexelius, A., & Wedman, I. (1978). *Resultat från prövningar med högskoleprovet 1978* [Results from administrations of the SweSAT 1978]. (Report No. Spånor från Spint Nr 11). Umeå, Sweden: Umeå University.
Lexelius, A., & Wedman, I. (1980). *Mätteknisk beskrivning av högskoleprovet 1977–79* [Technical description of the SweSAT 1977–1979]. (Report No. Spånor från Spint Nr 19). Umeå, Sweden: Umeå University.

Lexelius, A., & Wedman, I. (1981). *Resultat från prövningar med högskoleprovet 1980* [Results from administrations of the SweSAT 1980]. (Report No. Spånor från Spint Nr 20). Umeå, Sweden: Umeå University.

Linn, R. L. (1997). Evaluating the Validity of Assessments: The Consequences of Use. *Educational Measurement: Issues and Practice*, *16*(2), 14–16. doi: 10.1111/j.1745-3992.1997.tb00587.x

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

Lyrén, P.-E. (2009a). *A Perfect Score – Validity Arguments For College Admission Tests* (Doctoral dissertation). Umeå, Sweden: Umeå University. Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-25433.

Lyrén, P.-E. (2009b). Reporting subscores from college admission tests. *Practical Assessment, Research, & Evaluation, 14(*4), 1–10. Available online: http://pareonline.net/pdf/v14n4.pdf.

Lyrén, P.-E., Cobian, L., & Bardt, S. (2017). *Normering av högskoleprovet hösten 2016* [Norming of the SweSAT in the fall of 2016]. (Report No. Arbetsrapport nr 58). Umeå, Sweden: Umeå University.

Lyrén, P.-E., Rolfsman, E., Wedman, J., Wikström, C., & Wikström, M. (2014). *Det nya högskoleprovet – samband mellan provresultat och prestation i högskolan* [The new SweSAT – correlations between test results and performance in higher education]. (Report No. Arbetsrapport nr 52). Umeå, Sweden: Umeå University.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, *16*(2), 16–18. doi: 10.1111/j.1745-3992.1997.tb00588.x

Messick, S. J. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–104). New York, NY: Macmillan Publishing Company.

Messick, S. J. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749.

Monaghan, W. (2006). The facts about subscores (Report No. RDC-04). Princeton, NJ: Educational Testing Service.

National Institute for Testing & Evaluation [NITE]. (2017, June 20). Psychometric Entrance Test (PET) [Information on a website]. Retrieved from https://www.nite.org.il/index.php/en/tests/psychometric.html.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Ohlander, S. (1999). Påverkas engelsk-undervisningen av ELF-provet? [Is English education affected by the ERC subtest?] In Högskoleverket (Ed.) Fokus på högskoleprovet [Focus on the SweSAT]. Stockholm, Sweden: Högskoleverket. Retrieved from

http://www.uka.se/download/18.12f25798156a345894e2d5d/1487841933041/9906S.pdf.
Regeringens proposition 1972:84 [The Swedish Government's proposition No. 1972:84] (1972).
Regeringens proposition 1975:9 [The Swedish Government's proposition No. 1975:9] (1975).
Shepard, L. A. (1997). The Centrality of Test Use and Consequences for Test Validity. *Educational Measurement: Issues and Practice*, *16*(2), 5–24. Doi: 10.1111/j.1745-3992.1997.tb00585.x
Skolöverstyrelsen (1976). *Prov för urval till högre studier* [Test for selection to higher education]. Stockholm, Sweden: Skolöverstyrelsen.
Stage, C. (1988). Gender differences in test results. *Scandinavian Journal of Educational Research*, *32*(3), 101–111.
Stage, C., & Jonsson, I. (1995). *Luckprov med flervalsuppgifter: Ett alternativt ORD-prov?* [A sentence completion test with multiple-choice items: An alternative to the WORD subtest?]. (Report No. PM nr 96). Umeå, Sweden: Umeå University.
Stage, C. & Ögren, G. (2001). *Högskoleprovets utveckling under åren 1977–2000: Provets sammansättning och provdeltagargruppens sammansättning och resultat* [The development of the SweSAT during the years 1977–2000: The composition of the test and the composition of the test taking group and their results]. (Report No. PM nr 169). Umeå, Sweden: Umeå University.
Stage, C., & Ögren, G. (2007). *Olika sätt att mäta ordförståelse* [Different ways of measuring vocabulary]. (Report No. Arbetsrapport nr 17). Umeå, Sweden: Umeå University.
Stage, C., & Ögren, G. (2010). Ett nytt högskoleprov [A new SweSAT]. (Report No. BVM 42:2010). Umeå, Sweden: Umeå University.
Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.
Svensson, A., Gustafsson, J.-E., & Reuterberg, S.-E. (2001). *Högskoleprovets prognosvärde – Samband mellan provresultat och framgång första året vid civilingenjörs-, jurist-, och grundskollärarutbildningarna* [The prognostic value of the SweSAT – Correlations between test score and success during the first year of the master in engineering, law, and primary school teacher education programs] (Report no. Högskoleverkets rapportserie 2001:19 R). Stockholm, Sweden: National Agency for Higher Education.
The College Board. (2017, June 20). Inside the test [Information on a website]. Retrieved from https://collegereadiness.collegeboard.org/sat/inside-the-test.

Thurstone, L. L. (1955). The criterion problem in personality research. *Educational and Psychological Measurement*, *15*, 353–361.

Tillträdesutredningen. (1985a). *Tillträde till högskolan: Betänkande* [Admission to higher education: A report]. Stockholm, Sweden: Liber.

Tillträdesutredningen. (1985b). *Prov för urval till högskolan: Rapport* [Test for selection to higher education: A report]. Stockholm, Sweden: Liber.

Tillträdesutredningen. (2004). *Tre vägar till den öppna högskolan* [Three roads to open higher education]. Retrieved from http://www.regeringen.se/49b71e/contentassets/3b95ce95164f4a5b9310af78408b7aeb/sou-200429b.

Tillträdesutredningen. (2017). *Tillträde för nybörjare – Ett öppnare och enklare system för tillträde till högskoleutbidning*. [Admission for beginners – A simpler and more open system for admission to higher education]. Retrieved from http://www.regeringen.se/rattsdokument/statens-offentliga-utredningar/2017/03/sou-201720/.

Törnkvist, B., & Henriksson, W. (2004). *Repeated test taking: Differences between social groups*. (Report no EM no 47). Umeå, Sweden: Umeå University.

Universitets- och högskolerådet. (2016). *Överenskommelse mellan Universitets- och högskolerådet och Umeå universitet om konstruktion av högskoleprovet m.m.* [Agreement between the Swedish Council for Higher Education and Umeå University on the construction of the SweSAT etc.]. (Diarie no. 3.5.1-986.2015). Stockholm, Sweden: Archives of the Swedish Council for Higher Education.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer-Verlag.

von Davier, A. A., Holland, P. W., Livingston, S. A., Casabianca, J., Grant, M. C., & Martin, K. (2006). *An evaluation of the kernel equating method: A special study with pseudotests constructed from real test data*. (Report No. ETS-RR-06-02). Princeton, NJ: Educational Testing Service.

Wedman, I. (1978). *Resultat från prövningar med högskoleprovet 1977* [Results from administrations of the SweSAT 1977]. (Report No. Spånor från Spint Nr 10). Umeå, Sweden: Umeå University.

Wedman, I. (1983). *Den eviga betygsfrågan* [The never-ending issue of grades]. (Report no. FoU Rapport 48) Stockholm, Sweden: Skolöverstyrelsen.

Wedman, I., & Stage, C. (1994). *Notes from the first international SweSAT conference May 23–25, 1993*. (Report No. EM nr 9). Umeå, Sweden: Umeå University.

Wedman, J., & Lyrén, P.-E. (2015). Methods for examining the psychometric quality of subscores: A review and application. *Practical Assessment,*

Research & Evaluation, *20*(21), 1–14. Available online: http://pareonline.net/getvn.asp?v=20&n=21.

Wester, A. (1994). *Gender differences in testing: DIF analyses using the Mantel-Haenszel technique on three subtests in the Swedish SAT*. (Report No. EM 12). Umeå, Sweden: Umeå University.

Wester, A. (1997). *Differential item functioning (DIF) in relation to item content: A study of three subtests in the SweSAT with focus on gender*. (Report No. EM 27). Umeå, Sweden: Umeå University.

Wester-Wedman, A. (1990a). *Studiefärdighetsprovets (STUF) betydelse i högskoleprovet – En studie av simulerat utfall med och utan STUF-provet* [The significance of the study techniques (STECH) subtest in the SweSAT – A study on a simulated outcome with and without the STECH subtest]. (Report No. PM nr 32). Umeå, Sweden: Umeå University.

Wester-Wedman, A. (1990b). *Vad tycker provdeltagarna om högskoleprovet 1990-05-05?* [What are the test takers' opinions of the SweSAT 1990-05-05?]. (Report No. PM nr 36). Umeå, Sweden: Umeå University.

Wiberg, M. (2009). Differential item functioning in mastery tests: A comparison of three methods using real data. *International Journal of Testing*, *9*(1), 41–59. doi: 10.1080/15305050902733455

Wiberg, M., & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement*, *39*(5), 349–361. doi: 10.1177/0146621614567939

Ögren, G., Lexelius, A., & Stage, C. (2008). *Fortsatt utvecklingsarbete av ett prov som avser att mäta förmågan till kvantitativa jämförelser* [Continued development of a test that is intended to measure the ability to make quantitative comparisons]. (Report No. Arbetsrapport nr 24). Umeå, Sweden: Umeå University.

Ögren, G., & Stage, C. (2009). *Högskoleprovet: Utveckling mot en ny utformning* [The SweSAT: Development toward a new format]. (Report no. Arbetsrapport nr 31). Umeå, Sweden: Umeå University.