A Perfect Score
Validity Arguments for College Admission Tests

Per-Erik Lyrén

Academic Dissertations at the
Department of Educational Measurement
Umeå University
No. 4 · 2009
Abstract

College admission tests are of great importance for admissions systems in general and for candidates in particular. The SweSAT (Högskoleprovet in Swedish) has been used for college admission in Sweden for more than 30 years, and today it is alongside with the upper-secondary school GPA the most widely used instrument for selection of college applicants. Because of the importance that is placed on the SweSAT, it is essential that the scores are reliable and that the interpretations and uses of the scores are valid. The main purpose of this thesis was therefore to examine some assumptions that are of importance for the validity of the interpretation and use of SweSAT scores. The argument-based approach to validation was used as the framework for the evaluation of these assumptions.

The thesis consists of four papers and an extensive introduction with summaries of the papers. The first three papers examine assumptions that are relevant for the use of SweSAT scores for admission decisions, while the fourth paper examines an assumption that is relevant for the use of SweSAT scores for providing diagnostic information. The first paper is a review of predictive validity studies that have been performed on the SweSAT. The general conclusion from the review is that the predictive validity of SweSAT scores varies greatly among study programs, and that there are many problematic issues related to the methodology of the predictive validity studies. The second paper focuses on an assumption underlying the current SweSAT equating design, namely that the groups taking different forms of the test have equal abilities. The results show that this assumption is highly problematic, and consequently a more appropriate equating design should be applied when equating SweSAT scores. The third paper examines the effect of textual item revisions on item statistics and preequating outcomes, using data from the SweSAT data sufficiency subtest. The results show that most kinds of revisions have a significant effect on both \( p \)-values and point-biserial correlations, and as a consequence the preequating outcomes are affected negatively. The fourth paper examines whether there is added value in reporting subtest scores rather than just the total score to the test-takers. Using a method derived from classical test theory, the results show that all observed subscores are better predictors of the true subscores than is the observed total score, with the exception of the Swedish reading comprehension subtest. That is, the subscores contain information that the test-takers can use for remedial studies and hence there is added value in reporting the subscores. The general conclusion from the thesis as a whole is that the interpretations and use of SweSAT scores are based on several questionable assumptions, but also that the interpretations and uses are supported by a great deal of validity evidence.

Keywords: college admission tests; SweSAT; validity; interpretive arguments; predictive validity; equating; item revisions; subscores
Acknowledgements

Writing a thesis is a task that evokes both the joy and excitement in a toddler on a sugar rush and the pains and griefs in an executioner with a conscience. In other words, it is a mental roller coaster ride. The successful completion of a thesis is therefore not due to the author alone, and many people deserve credit for pushing and pulling me to where I am right now. First, I want to thank my advisors, Christina Stage, Simon Wolming, and Marie Wiberg, for guiding me on the path towards enlightenment. Although I am far, far away from that point, you all did a great job to get me as far as I am right now. Christina has an especially keen eye for linguistic and orthographical errors and typos, so she has been a VERY useful resource for me. Simon, string-bender extraordinaire, has tried to keep me structured, which must have seemed like an almost impossible task in his eyes. Marie, statistician extraordinaire, has made sure that all formulas are correct (which means that any errors that remain in the formulas are solely her responsibility). Here I would also like to extend my warmest thanks to Ron Hambleton for much support and dedication to my work and for being bold enough to actually write a paper with me. Many thanks to you all!

Gunilla Ögren deserves credit for having the guts to take me onboard the SweSAT test developing crew, a long, long time ago. Thanks to Anders Lexelius for helping out with data, for providing various twists and turns in discussions, and for the numerous uncalled-for visits to my workroom. Mats Hamrén has helped me with data collection and other stuff, and that I am grateful for. It would have sufficed to thank him for being a genuine norrbottning, or for making the best hallongrottor there are, but he did help me out with a lot of things too. Lotta Jarl has provided a lot of practical help during my years as a graduate student. Thanks Lotta! Many thanks also to the rest of the SweSAT crew and the rest of the department for inspiring discussions and great fika.

Many, many thanks to my “work-roomies”, who also happen to be former or current grad students: Christina Wikström, Anna Sundström, Peter Vestergren, and Tova Stenlund. Christina introduced me to the world of lost children, and she always has ideas about interesting studies. Christina, some day we will write that paper. Anna always has tips about all kinds of food-related stuff, which obviously is of great interest to me. In a near future, when my academic career has gone down the drain and Anna has gotten tired of me whining about it, we will start a bakery/boarding house/horse farm/measurement consulting firm together. Peter has an impressive knowledge of peculiar websites, measurement philosophy, tractors, back pain and other fun stuff. Tova, envelope-sealer extraordinaire, keeps me updated on things that I should know and books that I should read (such as
books that actually are about measurement). Many thanks also to the QME’sters at UC Berkeley (Mark Wilson, Heeju Jang, Jinnie Choi, Leah Walker, Amin Azzam, and others) and the folks at I-House (especially Karl, Wenke, and Jakob) for making my stay in Berkeley a pleasant experience.

Last and foremost, I want to thank my wonderful wife Gunilla for putting up with me and my work, for supporting me in everything I do, and for keeping me real. I am grateful to my daughter Sigrid for the power of her smile, because no matter how down in the dumps I’ve felt, she has always made me feel better. More importantly, she has taught me that there are greater things in life than writing excellent books and papers. I love you both more than words can say... Thanks also to all my family and friends for being who you are, and to all the test-takers out there for wanting to make a change.

Per-Erik Lyrén
Umeå in June 2009
A Perfect Score. Validity Arguments for College Admission Tests

This thesis is based on the following papers:


All referencing to the papers will be based on the enumeration used above.
# Contents

1 Introduction 1  
  1.1 Purpose 2  
  1.2 Disposition of the Thesis 2  
  1.3 A Note on Terminology and Statistical Notation 2  

2 Using Standardized Norm-Referenced Tests for College Admission 3  
  2.1 College Admission Testing Internationally 3  
    2.1.1 Historical Perspectives and Descriptions of Some Tests 3  
    2.1.2 Recent Developments 5  
  2.2 The Swedish Scholastic Assessment Test, SweSAT 7  
    2.2.1 A Brief History 7  
    2.2.2 Recent Developments and Current Status 10  

3 Validation 13  
  3.1 Traditional Perspectives 13  
  3.2 Recent Perspectives on Validation and Validity 16  
    3.2.1 A Unitary Conceptualization of Validity 16  
    3.2.2 Criticisms of Messick’s Validity Framework 17  
    3.2.3 An Argument-Based Approach to Validity 18  
    3.2.4 A Comment on the Argument-Based Approach to Validity 19  
  3.3 Gathering Validity Evidence for the SweSAT 20  
    3.3.1 Interpretive Arguments for the SweSAT 20  
    3.3.2 Validity Arguments for the SweSAT 30  

4 Methodological Issues 38  
  4.1 Test-Theoretical Approaches 38  
  4.2 Statistical Considerations 39  

5 Summaries of the Papers 39  
  5.1 Paper I 39  
  5.2 Paper II 40  
  5.3 Paper III 41  
  5.4 Paper IV 42  

6 Discussion 42  
  6.1 Main Results – Implications for the Validity Argument 42  
  6.2 Implications for Test Development 43  
  6.3 Implications for Test-Takers 44  
  6.4 Validation Issues 45  
  6.5 Limitations and Generalization 46  
  6.6 Suggestions for Further Research and Development 46  

References 48  
Appendix 55
1 Introduction

Education is one of the pivotal parts of society, and access to education is considered a fundamental human right by the United Nations, as established in the Universal Declaration of Human Rights. Specific elements of this right are also covered by the International Covenant on Economic, Social and Cultural Rights from 1966 and the Convention on the Rights of the Child from 1989. This right is in many cases exercised in elementary school and lower-secondary school, which are generally compulsory unlike upper-secondary school and tertiary education which are optional. However, changes in society over the last half century or so have led to extensive educational reforms and today, at least in Sweden, basically all children enter upper-secondary school and close to 50 percent of adolescents have entered higher education by the age of 24. This means that non-compulsory education has become a common issue for people, and many people most likely consider gaining the prerequisites required for higher education as a basic human right.

In an ideal world, at least from the prospective students’ point of view, all of those with a wish to study at college or university should have the opportunity to do so. However, to maintain a high standard of education certain eligibility requirements have to be placed on the candidates. Also, there are economical and practical factors that limit the number of available places in higher education. The issues of eligibility and selection are managed within an admission system. The Swedish higher education admission system is mainly based on grades from upper-secondary school and the Swedish Scholastic Assessment Test, SweSAT (Högskoleprovet in Swedish). The grades are mandatory since they are used both for eligibility and selection, while the SweSAT is optional and used for selection only. Because of the high stakes involved in gaining access to higher education, candidates and other users of the admission system want the measures used for eligibility and selection to be reliable and valid.

The SweSAT has been used for selection to higher education since 1977, and it has attracted more than one million unique test-takers to this date. The test is used in such a way that it gives eligible candidates a way into higher education without the need for a competitive upper-secondary school grade point average (GPA). Therefore, it is not an overstatement that the quality of the test in terms of its validity and reliability is crucial.

The title of this thesis, A Perfect Score, alludes to two things. First, it alludes to the fact that test-takers aim at achieving the highest score possible, and many of them work hard to reach their goal. Second, it alludes to the fact that those responsible for the test always aim at providing a test that yields reliable scores from which one can make valid interpretations and decisions.
Although this is the aim it is impossible to provide scores that are entirely perfect in this sense, which will become evident in this thesis.

1.1 Purpose

The purpose of this thesis is to examine some assumptions that are of importance for the validity of the interpretation and use of SweSAT scores. Specifically, the aim is to examine validity evidence for the SweSAT with regard to two uses of the test scores: a) admission decisions based on the total test score, and b) providing diagnostic information based on the subtest scores.

1.2 Disposition of the Thesis

The thesis consists of four papers and an extensive introduction and summary. Chapter 2 presents the origin of using standardized tests for admission to higher education. Also, it gives an overview of admission tests internationally and it presents the history and development of the SweSAT. Chapter 3 introduces the concepts of validation and validity. Here, I give an account of how different perspectives on validity have evolved over time as well as some criticisms towards commonly accepted validity perspectives. In this chapter I also present my view of what an interpretive argument for the SweSAT may look like and to what extent previous validation efforts support inferences in that argument. Chapter 4 deals with methodological issues, specifically test-theoretical issues and statistical considerations in the four studies. Chapter 5 summarizes the four papers, and in the last chapter (6) I discuss the findings in the studies and their implications for the validity arguments as well as practical implications for test development and test-takers. The last chapter also contains a discussion on limitations and generalizations, and some thoughts on further research. There is also an appendix with lists of statistical notation, acronyms, and abbreviations. Finally, the papers follow in numerical order.

1.3 A Note on Terminology and Statistical Notation

There is a wide array of terms and notations used in the field of educational and psychological measurement. In this thesis, differences in terms are related mainly to tests and the test-takers. Terms such as test, examination, and instrument interchangeably refer to any coherent collection of items that are used to measure some construct. Similarly, terms such as test-taker, examinee, and candidate refer to a person who is to be measured by a test/examination/instrument.
Some terms that are often used interchangeably are *try-out, field-test* and *pretest*. They all refer to the procedure where test items are administered to a sample of test-takers in order to get estimates of item statistics (e.g., difficulty and discrimination), which are then used in subsequent test construction. The term pretest is easy to inflect and combine with other words (pretesting, pretested, pretest item). Although the term is also used in experimental designs in the form of a pretest–posttest procedure and thus may cause confusion, this term was preferred in this thesis.

The breadth of the statistical notation used in the thesis is vast. Therefore, in the Appendix there is a list of the notation with references to the papers in which they occur. It should be noted that different notations may be used for the same statistical concept, and that a certain notation may stand for different statistical concepts. For example, both M and μ indicate a mean score, and r indicates a Pearson correlation in Paper I, a raw score in Paper II, and a regular-test score in Paper II and Paper III. In the Appendix there is also a list of acronyms and abbreviations used in this thesis.

2 Using Standardized Norm-Referenced Tests for College Admission

2.1 College Admission Testing Internationally

2.1.1 Historical Perspectives and Descriptions of Some Tests

The origin of university admission testing is somewhat debated, but it is generally agreed that it began in Europe. Specific times and places vary from 13th century or 18th century France to 16th century Spain (Zwick, 2006). However, most accounts agree that admission testing was introduced in England and Germany during the early 19th century. Even though admission testing began in Europe, it has been most prominent in the United States. In 1900, the College Entrance Examination Board was formed by representatives of the top northeastern universities in the United States. The College Board developed the Scholastic Assessment Test (SAT), which was first administered in 1926. The SAT was in many respects similar to the Army Alpha tests, which were used for selecting and assigning military recruits in World War I; these tests in turn were based on IQ tests. During World War II, both the College Board and the Iowa Testing Programs helped the military develop tests for screening individuals for military service and assigning them to jobs, which increased the interest in testing from educational institutions as well (Zwick, 2006). Also, large numbers of
returning WWII veterans were sent to college, which increased the interest in the multiple-choice and therefore efficient SAT.

In 1947, the College Board, the Carnegie Foundation for the Advancement of Teaching, and the American Council on Education merged their testing activities and founded the Educational Testing Service (ETS). ETS then took on the administration of the SAT under a contract with the College Board. For about a decade, ETS was basically the sole actor on the testing market until 1959 when the American College Testing Program, which came about from the Iowa Testing Programs, commenced. ACT, Inc., the testing company associated with American College Testing Program and the developer of the admission test ACT, was a spin-off of E.F. Lindquist’s Measurement Research Corporation and the Iowa Testing Programs. Interestingly, the establishment of both ETS and ACT, Inc. was largely associated with scoring issues and the development of the first test scoring machines.

ETS would become responsible for development and/or administration of several college admission tests. The Graduate Record Examination (GRE), which is used for admission to graduate school, was first administered in 1937 and was transferred to ETS in 1948. The Medical College Admission Test (MCAT) was first administered in 1946 and transferred to ETS along with the GRE in 1948. The first Law School Admission Test (LSAT) was administered in 1948 by ETS under a contract with the Law School Admission Council (LSAC), which itself took over the development of the test in 1979. In 1954, the first Graduate Management Admission Test (GMAT; then known as the Admission Test for Graduate Study in Business) was administered by ETS, which would be responsible for development and administration of the test until 2006 when the test development was taken over by ACT, Inc. and the administration was taken over by Pearson VUE (Zwick, 2006).

The SAT program consists of the SAT Reasoning Test (the original SAT test, referred to simply as the “SAT”) and the SAT Subject Tests. The SAT has changed considerably over the years and consists today of a critical reading section, a mathematics section, and a writing section. Most items are multiple-choice, but there are also constructed-response math items and an essay in the test. The writing section was added in 2005 as a result of discussions following the University of California (UC) controversy in 2001, when UC President Richard C. Atkinson recommended that the SAT Reasoning Test should be optional, not required, for admission to the UC universities. Atkinson argued that “students should be judged on the basis of their actual achievements, not on ill-defined notions of aptitude”, and consequently recommended that tests that are more closely linked to school curricula should be used instead of the SAT Reasoning Test (Atkinson, 2001). Intense discussions led to several significant changes to the test,
including the introduction of the writing section and more advanced math content, and the elimination of verbal analogy items and quantitative comparison items. Today, the SAT claims to “test[s] students’ basic knowledge of subjects they have learned in the classroom—such as reading, writing, and math—in addition to how students think, solve problems, and communicate” (College Board, 2008).

The development of the ACT was largely rooted in disappointments at the SAT. The test was considered to be too focused at the northeastern elite institutions, and its developers were viewed as resistant to change. The underlying construct of the ACT was also to be different. While the SAT consisted only of verbal and mathematical sections, the ACT consisted of four sections that resembled the school curriculum, namely English, mathematics, social studies reading, and natural science reading. Today, the ACT is based on content that is taught in grades 7 through 12, and the sections are now English, mathematics, reading and science. Also, since 2005 there is an optional writing section in the test. In addition to being more linked to school curricula than the SAT, the ACT also aims at facilitating course placement and academic planning.

Next to the tests in the USA, probably the most widely known college admission test is the Psychometric Entrance Test (PET), which is used for admission to Israeli universities. The test is developed and administrated by the National Institute for Testing and Evaluation (NITE), an independent organization founded in 1981 by the Israeli research universities (Beller, 2001). The PET was introduced in the early 1980s to be used in a manner similar to the SAT and the ACT. Prior to the introduction of the PET, the admission was based on the matriculation certificate (Bagrut). The fact that the Bagrut scores are not standardized (and therefore incomparable among schools) and that there was a desire to give a second chance to “late bloomers” led to the need for an additional admission instrument (Beller, 2001). According to Beller (1994), the PET “measures various cognitive and scholastic abilities to estimate future success in academic studies” (p. 13). Since 1990 the test “has become slightly more curricular-based” (Beller, 2001, p. 321), and now consists of three sections: verbal reasoning, quantitative reasoning, and English as a foreign language.

2.1.2 Recent Developments

Much of the recent developments in college admission testing have focused on the content changes to the SAT in the wake of the UC controversy and the “aptitude versus achievement” discussion in general. Other developments have involved the administration, scoring, and score reporting practices. For example, during the last few years several college admission tests have been computerized. The GRE was computerized as early as 1992, when a
computer-based test (CBT) form was administered alongside with the paper-and-pencil version (Schaeffer, Steffen, Golub-Smith, Mills, & Durso, 1995). The year after, a computerized adaptive test (CAT) version of the GRE was administered, and since 1999 the test has been administered exclusively as a CAT. However, there have been discussions about changing the test from a CAT to a linear (i.e., non-adaptive) CBT (Zwick, 2006). The GMAT has also been transformed into a CAT and has been administered as such since 1997 (The Graduate Management Admission Council [GMAC], n.d.), and the MCAT has been offered exclusively as a linear CBT since 2007 (Association for American Medical Colleges [AAMC], 2006). There are experimental CAT versions of the SAT, the ACT, and the LSAT, but these have not been used in the general test-taking population. Also, NITE has developed a CAT version of the PET to be administered to examinees with disabilities (Moshinsky & Kazin, 2005).

One of the main developments in scoring is the automated scoring of essays. Even though large theoretical and technical advances have been made in this area (see e.g., Dikli, 2006; Wang & Brown, 2007), only the GMAT and the GRE use automated essay scoring (both with the ETS e-rater® system) in addition to human scoring, while the SAT, the ACT, and the MCAT writing sections are scored only by humans (College Board, 2008; ACT, 2007; AAMC, 2009).

Score reporting issues have also received increased attention during the last decade or so. Traditionally, only the scores that are used for admission purposes are reported to the test-takers, and the information about the scores is rather scant. In recent years, however, there has been an increasing interest in providing more useful information to the test-takers, usually by reporting certain scores from subtests or other subsections of the test (see e.g., Monaghan, 2006). The motivation for reporting these subscores is that they provide diagnostic information to the test-takers. Based on these scores the test-takers can then take remedial action and thereby improve their scores. As an example, the SAT reports Sentence Completion and Passage-Based Reading scores from the Critical Reading section, and Numbers and Operations, Algebra and Functions, and two more scores from the Mathematics section. Similarly, the ACT reports Usage/Mechanics and Rhetorical Skills from the English section, and Pre-Algebra/Elementary Algebra, Algebra/Coordinate Geometry, and Plane Geometry/Trigonometry from the Mathematics section. In addition, the SAT has an online operational diagnostic program called the SAT Skills Insight (www.collegeboard.com/sat skillsinsight), in which test-takers can get more in-depth information about the skills associated with certain score levels. Another example of an operational diagnostic program is the online GMAT Focus™ system (www.gmatfocus.com), which “identifies quantitative strengths and abilities using real GMAT questions and a
computer adaptive process that mimics the GMAT exam”. Both systems are intended for test preparation; however, the SAT Skills Insight can be used both by first-time test-takers as well as re-takers, while the GMAT Focus system is intended mainly for first-time test-takers. In general, these new score reporting features are intended to make the scores more useful and valid for the test-takers.

2.2 The Swedish Scholastic Assessment Test, SweSAT

The SweSAT is a norm-referenced, paper-and-pencil, multiple-choice test with five subtests: Vocabulary (WORD; 40 items), Swedish Reading Comprehension (READ; 20 items), English Reading Comprehension (ERC; 20 items), Data Sufficiency (DS; 22 items), and Diagrams, Tables, and Maps (DTM; 20 items). Candidates may take the test an unlimited number of times, and it is their highest achieved score that is used in the selection process. Since 1991 SweSAT scores are valid for five years for college applications (before 1991 the scores were valid for two years). Further, the SweSAT reports only an overall composite score for use in the admission process while subtest scores are provided to examinees only. The test is administered twice a year, spring and fall, with approximately 30,000–50,000 examinees per administration. For future reference, the spring tests are labeled with an A, as in the 09A test (the 2009 spring test), and the fall tests are labeled with a B, as in the 08B test (the 2008 fall test).

2.2.1 A Brief History

The SweSAT was introduced for admission purposes in 1977, but several years of political discussions, developmental work, and research preceded its introduction. This section aims to give an account of this development.

During the 1950s and 1960s a number of school reforms took place in Sweden. These led to a substantial increase in the number of students in upper-secondary schools, which meant that the number of external examiners needed to assess the students’ final examinations had to increase considerably. This was not feasible and as a result the final examinations were abolished. Consequently, the basis for setting the grades was impaired, so in order to make the grades fair and comparable among schools across the nation, nationally developed tests and a school inspection system were introduced. This was particularly important since the upper-secondary school grades were the main instrument used for selection to higher education. The increase in student numbers also meant that the competition for admission to higher education increased. As a consequence, other kinds of instruments (e.g., aptitude tests) were considered for use in the admission process (Henrysson, 1994).
In 1965 a governmental commission was set up for the purpose of revising and simplifying the admission procedure (Marklund, Henrysson, & Paulin, 1968). The instructions to the commission included a number of requirements and requests that were to be taken into account, for example (a) the recruitment should be broadened (with respect to background variables such as social group, age, and sex), (b) the quality of the students should be maintained, (c) applicants were to be admitted on the basis of their suitability for the education applied to, (d) other selection instruments than grades should be used, among other things to reduce the biased distribution of social background variables of the students, (e) informal merits should be acknowledged, and (f) the selection regulations should be simple and comprehensible. These requirements led to the idea of using tests as a complement to, or instead of, the grades for selecting applicants. This became a key issue for the commission, mainly due to criticism of the dominant role the grades had in the admission procedure, and the aspiration to give applicants without a formal upper-secondary education access to higher education.

This aspiration resulted in the proposition to introduce a quota group system for selecting applicants. The four quota groups were

1. applicants with 3 or 4 years of upper-secondary school
2. applicants with 2 years of upper-secondary school
3. applicants with a completed folk high school education
4. applicants at least 25 years old and with 4 years of working experience (commonly referred to as the 25/4s)

In this system the applicants in a certain quota group were to compete only with those in the same quota group, not with other applicants. The number of applicants that were admitted from each quota group was in proportion to the number on applicants in the quota groups. For example, if 50 percent of the applicants were in quota group 1, then 50 percent of those who were admitted were selected from that group. When it came to selection instruments, the upper-secondary school grades were to be used as the selection instrument in quota groups 1 and 2, while the folk high school’s assessment of the applicant’s ability to profit from the higher education was used in quota group 3, and a nationally developed and standardized selection test was to be used in quota group 4 because the 25/4s did not have comparable grades. There were also discussions about whether such a test could be used for eligibility (in quota groups 3 & 4) and guidance, but these ideas were abolished at an early stage. Another idea was to use the test as a second chance for students who did not perform well in upper-secondary school and students that earned grades that did not reflect their true
knowledge and abilities (Stage, 2004a). In the end, it was decided that the test was to be used only for the 25/4s, mainly because there were great uncertainties in regard to what the test would look like, how it would work and how it would be received by the applicants (Henrysson, 1994). Then, through the 1977 revisions to the Higher Education Ordinance the quota group system was introduced and the test, which is now known as the SweSAT, was to be used for selecting applicants from quota group 4 (i.e., the 25/4s). Also, before the test’s inception it was decided that scores from different test administrations should be comparable (which is achieved through the statistical procedure called equating).

In 1983 another governmental commission (Tillträdesutredningen) was set up to further scrutinize the admission system. Again, the dominant role of the grades in the selection process and the aspiration to broaden recruitment were the reasons for setting up the commission. The commission expected test results and other qualifications to reduce the biased distribution of social groups, and it therefore proposed that the SweSAT should be used for selection for all groups of applicants (SOU 1985:57). Consequently, the SweSAT has been available for all candidates since 1991. Also, since then the selection of applicants has been based mainly on the upper-secondary school grade point average (GPA) or SweSAT scores. However, because one of the political objectives was to encourage older people to apply to higher education, those who had at least five years of work experience could add 0.5 score points to their SweSAT score (which ranges from 0.0 to 2.0 with 0.1 increments).

Not only the use of the SweSAT, and consequently its purpose, has changed over time. There have been several changes in the composition and administration of the test. When the test was introduced in 1977 it consisted of six subtests with a total of 150 items. Since then, one subtest has been added (ERC) and two have been removed (Study Technique, STECH, and General Information, GI). STECH was removed mainly because it was a very expensive test, while GI was removed mainly because there were always intense and unresolvable discussions in the test review groups about what was to be considered as general information. The subtests that have always been part of the SweSAT are WORD, READ, DS, and DTM. The number of items and time limits for each subtest are shown in Table 1.
Table 1. The number of items in the SweSAT subtests and the time allowed for each subtest

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>WORD</td>
<td>30 (15 min)</td>
<td>30 (15 min)</td>
<td>30 (15 min)</td>
<td>40 (15 min)</td>
</tr>
<tr>
<td>DS</td>
<td>20 (40 min)</td>
<td>20 (40 min)</td>
<td>20 (40 min)</td>
<td>22 (50 min)</td>
</tr>
<tr>
<td>READ</td>
<td>30 (50 min)</td>
<td>24 (50 min)</td>
<td>24 (50 min)</td>
<td>20 (50 min)</td>
</tr>
<tr>
<td>DTM</td>
<td>20 (50 min)</td>
<td>20 (50 min)</td>
<td>20 (50 min)</td>
<td>20 (50 min)</td>
</tr>
<tr>
<td>ERC</td>
<td>-</td>
<td>-</td>
<td>24 (50 min)</td>
<td>20 (35 min)</td>
</tr>
<tr>
<td>GI</td>
<td>30 (30 min)</td>
<td>30 (30 min)</td>
<td>30 (30 min)</td>
<td>-</td>
</tr>
<tr>
<td>STECH</td>
<td>20 (50 min)</td>
<td>20 (50 min)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>pretest</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>20–60 (50 min)</td>
</tr>
<tr>
<td>Total</td>
<td>150 (3 h 55 min)</td>
<td>144 (3 h 55 min)</td>
<td>148 (3 h 55 min)</td>
<td>122 (4 h 10 min)</td>
</tr>
</tbody>
</table>

Up to 1995 the item pretesting was carried out in upper-secondary schools with seniors at academically oriented programs. However, this procedure was associated with three problematic issues. First, the groups that the items were pretested in were too small for obtaining reliable item statistics. Second, because the pretesting situation was low-stakes to the students they probably lacked motivation for doing their best, which also affects the item statistics. Third, it became more difficult to get access to the students due to major reforms in the school system. As a consequence of these problems, the pretesting procedure changed in 1996 to be integrated into the regular test administration. This change involved dividing the testing time into five blocks of 50 minutes, where the subtests DS, READ, and DTM is administered in one block each, the subtests WORD and ERC are administered together in one block (15 + 35 minutes), and the fifth block is used for pretesting new items. The pretest block consists of a complete version of one or two of the other subtests (DS, READ, DTM, or WORD + ERC). The order of the blocks changes from one administration to another, and the idea behind this procedure is that test-takers will not be able to tell the pretest block from the regular blocks, and thus their motivation will be high when answering the pretest items. By having a large number of test-takers who are motivated when taking the pretest items the test developers can get reliable item statistics, which are essential when constructing parallel versions of the test.

2.2.2 Recent Developments and Current Status

There have not been any major changes to the test in recent years. However, there certainly have been discussions about further developments. The Swedish National Agency for Higher Education (NAHE), which is
responsible for the SweSAT, examined the test in 2000 and made some suggestions for how to further develop the test (Högskoleverket, 2000). The basis for the examination was to increase the predictive validity of the test and to reduce the impact of the biased distribution of social groups. The suggestions included (a) differential weighting of subtest scores or reporting section scores (a verbal score and a quantitative score), (b) developing domain-specific tests, (c) reducing the number of verbal items and/or increasing the number of problem solving items, (d) using a test-taker’s average SweSAT score instead of his or her highest SweSAT score in the selection process, and (e) reducing the significance of the additional credits for work experience.

The suggestion to reduce the number of verbal items was made because there were indications that the negative effects of repeated test-taking on the predictive validity were greater for these items. The suggestion to use the average score instead of the highest score was based on the understanding that repeated test-taking increases the differences in performance between social groups, and that using average scores would reduce the incentives for taking the test multiple times. The suggestion to reduce the significance of the additional credits for work experience was made simply because they do not add to the predictive validity of the test; this suggestion was realized in 2008.

Reporting section scores was discussed already in the 1983 commission, but back then it was concluded that the effect on admission would be insignificant. However, a NAHE report from 2001 (Svensson, Gustafsson, & Reuterberg, 2001) concluded that applying program-specific weights to the subtest scores could improve the prognostic value of the test. Yet, another NAHE report from 2002 (Högskoleverket, 2002a) concluded, again, that using a verbal score and a quantitative score in the selection of applicants would have a small effect on admission. Instead, it was suggested that domain-specific tests should be developed.

Domain-specific tests have also been a topic of discussion in the governmental commission. However, it was not until the recently mentioned NAHE reports (Högskoleverket, 2000, 2002a) and the report from the latest governmental commission on admission to higher education (SOU 2004:29) came out that this issue ended up on top of the agenda. This led to the development of a test for admission to engineering, and a test for admission to medicine and nursing. However, both tests were cancelled in 2008, before they made it to regular use; the engineering test because there was a lack of interest from the universities and the medicine/nursing test because the universities did not want to use a test that was not fully developed and thus lacked in quality (Högskoleverket, 2008).

In 2002 the SweSAT was evaluated by a group of internationally renowned psychometricians, John Fremer and David Lohman from the USA,
1. The SweSAT is a high quality test with a solid research program. They applaud past research and encourage continued research on the test.

2. Uses for which the test was originally designed have changed. Now the test is being used as an alternative selection mechanism for a substantial fraction of the applicant population. This requires a rethinking of the nature of the test and the ways in which it is used.

3. The decision rule for how SweSAT scores and grades are used should be reconsidered. In the current model, grades and test scores are considered separately. A selection model that simultaneously considers grades and test scores should be investigated.

4. Separate scores for components of SweSAT are recommended. Predictive validities of SweSAT for particular courses of study are likely to be higher if separate scores were available for at least verbal reasoning and quantitative reasoning.

5. Explore the possibilities of giving diagnostic feedback for counseling and individual score interpretation—at least on the profile of scores that are reported, and possibly on the underlying skill classifications.

6. Explore the use of computer-based test.

7. In conjunction with the previous point, consider adding a computer-scored measure of writing ability.

8. A cost-benefit analysis of the test should be conducted. Information about the contribution of the test to the overall efficiency and success of the educational system need to be part of the discussion about its value.

These recommendations led to the inclusion of subtest scores in the score reports to the test-takers, and an improved test preparation guide with in-depth descriptions of items and discussions about solutions (SOU 2004:29). Also, they had an impact on the NAHE’s current aspiration to computerize the test. Regarding point 8, the NAHE is currently developing new item types, both verbal and quantitative ones. It is interesting to note that, under the current propositions, the verbal section would become more context-based through a reduction of vocabulary items and the introduction of sentence completion items, while the quantitative section would become less context-based through the introduction of quantitative comparison items.
and basically context-free algebra, arithmetic, and geometry items. The development of new item types is also the result of discussions with the SweSAT International Advisory Board, which recommended that the test should be more time-efficient, that is, to increase the number of score points per minute of testing (Stage, 2008).

The cancellation of the domain-specific tests led to a renewed interest in reporting section scores. As a consequence, the work of developing new item types also involves an aspiration to develop two sections, a verbal and a quantitative, of approximately equal length that can be scored, scaled, and equated separately. The work with the new item types and sections commenced in 2006 and will continue during 2009 and probably 2010 as well. It deserves to be pointed out that even though there have been and still are many ideas for how to further develop the SweSAT, it is unclear which developments will potentially contribute the most to the validity of the interpretations and uses of the test scores.

3 Validation

The two related terms “validation” and “validity” are frequently used in discussions about measurement issues. According to the Standards for Educational and Psychological Testing (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999, p. 9), validity refers to “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests”, while validation is a process that “involves accumulating evidence to provide a sound scientific basis for the proposed score interpretation”. Inherent in these definitions is that one validates neither the test nor the test scores, but rather the proposed interpretations and uses of the test scores.

3.1 Traditional Perspectives

The early formal definitions of validity were mainly focused on the relationship between test scores and scores on some criterion. For example, Kelley (1927, p. 14) defined validity as the extent to which a test measures what it purports to measure. Similarly, Guilford (1946) argued that a test was valid for anything with which it correlated and Gulliksen (1950b) considered a test valid for any criterion for which it provided accurate estimates. These definitions reflect validity in terms of what was called criterion-related validity. Most validity studies of that time examined the predictive validity of tests, that is, the focus was on examining how well the test could predict
future achievement in some area. Another type of criterion-related validity that developed around this time was **concurrent validity**, which also refers to the extent to which the test scores correlate with some criterion. The difference between concurrent validity and predictive validity has to do with when the criterion scores are observed (concurrently with the test scores, or some time after the test scores). The biggest challenge involved in criterion-related validity studies was, and still is, the difficulty in obtaining a valid and reliable criterion (e.g., Wolming, 2000). The importance of this issue was addressed by Toops (1944), who maintained that

> Possibly as much time should be spent in devising the criterion as in constructing and perfecting the tests. This important part of a research seldom receives half the time or attention it requires or deserves. If the criterion is slighted the time spent on the tests is, by so much, largely wasted. (p. 290)

The problems faced by criterion-related validity studies led to the development of alternative approaches to validation. One approach focuses on interpretations of test scores based on a sample of performances in some area of activity as an estimate of overall level of skill in that activity (Kane, 2006). According to Rulon (1946), validation must include an assessment of the content of the test and its relation to the purpose of the measurement (Sireci, 1998). Like Rulon, both Mosier (1947) and Gulliksen (1950a; 1950b) acknowledged the importance of evaluating test content when performing validity studies. Shortly thereafter, Cureton (1951) introduced the term **content validity** to refer to the extent to which the test covers the relevant instructional or content domain. The emergence of content validity was also acknowledged in the *Technical Recommendations for Psychological Tests and Diagnostic Techniques: A Preliminary Proposal* (APA, 1952) put forth by the APA Committee on Test Standards, and in the subsequent *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (APA, 1954) put forth by a joint committee of the APA, the AERA, and the National Council on Measurements Used in Education (currently NCME). A few years later, Lennon (1956) defined content validity more formally as

> the extent to which a subject's responses to the items of a test may be considered to be a representative sample of his responses to a real or hypothesized universe of situations which together constitute the area of concern to the person interpreting the test. (p. 295)

The publication of the *Technical Recommendations* (APA, 1954) involved the presentation of a new “aspect” of validity, **construct validity** (referred to as **congruent validity** in the preliminary proposal version; APA, 1952). APA (1954) stated that “construct validity is evaluated by investigating what psychological qualities a test measures, i.e., by demonstrating that certain explanatory constructs account to some degree for performance on the test.”
The concept of construct validity was further developed by Cronbach and Meehl (1955), who claimed that construct validity is essential to validation whenever a test score is to be interpreted as a measure of some attribute that is not operationally defined. However, Cronbach and Meehl emphasized the importance of construct validation only in the absence of an acceptable criterion or an acceptable universe of content, which has been viewed as a depreciation of construct validity in relation to criterion-related validity and content validity (e.g., Shepard, 1993). Similarly, the successor to the Technical Recommendations, the Standards for Educational and Psychological Tests and Manuals (APA, AERA, & NCME, 1966) suggested that “construct validity is relevant when the tester accepts no existing measure as a definitive criterion” (p. 13). One of the key features of Cronbach and Meehl’s work was the concept of the “nomological network” (nomological meaning ‘law-like’), which shows the relationship between (a) different observable properties, (b) theoretical constructs and observable properties, and (c) different theoretical constructs. These relationships are then tested empirically through experimental and statistical procedures, for example by using structural equation modeling.

Though the concepts of criterion-related validity, content validity, and construct validity were identified as different “aspects” of validity in the 1954 Technical Recommendations (APA, 1954), over the years they came to be viewed more as distinct types of validity. Guion (1980) referred to this conception as “the trinitarian concept of validity” and the “holy trinity” in psychometric theology. On a critical note, Guion asserts that under this conception, those responsible for test validation get off easy because “if you cannot demonstrate one kind of validity, you have two more chances!” (p. 386). In practice, different types of validity evidence were used for different assessments. Typically, criterion-related evidence was used when validating selection and placement testing, evidence of content validity was used when validating achievement testing, and evidence of construct validity was used for theory-based explanations (Kane, 2006).

Loevinger (1957) criticized criterion-related validity and content validity as being ad hoc and gave prominence to construct validity, which she claimed to be “the whole of the subject from a systematic, scientific point of view” (p. 461, as cited by Kane, 2006). Basically, Loevinger viewed construct validity as an overarching concept that subsumed the other types of validity. Although Loevinger’s view did not gain recognition by the measurement community at the time of publication of her thesis, her argument prefigured the move toward a unitary conceptualization of validity (Moss, 2007).
### 3.2 Recent Perspectives on Validation and Validity

#### 3.2.1 A Unitary Conceptualization of Validity

During the 1970’s, validity theorists became more prone to support a more unified approach to validation. Messick (1975; 1980) argued strongly for a unitary conceptualization of validity. Like Loevinger (1957), Messick viewed construct validity as somewhat of a golden standard, asserting that “all measurement should be construct-referenced” (Messick, 1975, p. 957, italics in original) and that only construct validity should “bear the name ‘validity’ and … wear the mantle of all that name implies” (Messick, 1980, p. 1015); a position that was supported by Guion (1977a; 1977b) and Tenopyr (1977) among others. Messick argued that construct validity is a unifying concept that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships. This conceptualization was elaborated further and culminated in his seminal chapter in the third edition of *Educational Measurement* (Messick, 1989), in which validity was defined as

> an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. (p.13; emphasis in original)

Messick’s validity framework was summarized in his (in)famous two-by-two matrix, which described two interconnected “facets” of validity: the source of justification for testing (an evidential basis or a consequential basis) and the function or outcome of testing (interpretation or use). The cells in this matrix have the following functions. First, the evidential basis of test interpretation is construct validity. Second, the evidential basis of test use is also construct validity, but “as buttressed by evidence for the relevance of the test to the specific applied purpose and for the utility of the test in the applied setting” (Messick, 1989, p. 20). Third, the consequential basis of test interpretation is the appraisal of the value implications associated with the construct label. Fourth and finally, the consequential basis of test use is the appraisal of both potential and actual consequences, whether intended or not, of the applied testing. Although Cronbach (1971) already in the second edition of *Educational Measurement* gave significant attention to the consequences of testing, Messick went a step further and gave the consequential basis equal bearing with the evidential basis.
3.2.2 Criticisms of Messick’s Validity Framework

While Messick’s unified validity framework is widely endorsed (see e.g., Sireci, 2007), it has also been subject to much discussion and controversy. Those who disagree with Messick’s view of validity mainly fear that his framework does not help in the practical validation process. Markus (1998) agrees with the bulk of Messick’s arguments, but argues that the unified theory of validity is incomplete and that “there is a need to synthesize the evidential and consequential bases of test interpretation and use” (p. 31). Further, Sireci (2007) argues that although being theoretically sound, the unitary conceptualization of validity has a major shortcoming in being extremely difficult to describe to lay audiences. Sireci believes that policy makers and the general public can understand the concept of content validity more easily than the concept of construct validity, and that Messick “demoted content validity to a much lower status than construct validity” (p. 480). While not disagreeing with the substance of Messick’s argument, Shepard (1993) fears that the faceted presentation of his framework will lead to a new segmentation of validity requirements, and that the complexity of his argument will not help in identifying the validity questions essential to support a test use. Similarly, Kane (2008, p. 77) discusses the benefits and shortcomings of Messick’s framework in light of the problems associated with developing a validation strategy:

This framework is very elegant in many ways and highlights some issues that deserved more attention than they were getting at the time. ... However, this unitary framework may be more useful for thinking about fundamental issues in validity theory than it is for planning a validation effort ... Messick did not create this problem, but he may have made it worse by formulating his discussion of validity within an abstract, philosophical framework that made his discussion of validity remote from the interests of test developers and users.

Others have opposed to the unified framework more strongly. For example, Borsboom, Mellenbergh, and van Heerden (2004, p. 1061) claim that the unified theory “fails to serve either the theoretically oriented psychologist or the practically inclined tester”, and that “validity is not complex, faceted, or dependent on nomological networks and social consequences of testing”. Further, in an Educational Researcher special feature called “Dialogue on Validity”, Lissitz and Samuelsen (2007) reject the concept of a unitary definition of validity and proposes a change in the emphasis and vocabulary of validity theory. They argue that validity has to do with internal aspects of the technical evaluation of testing, namely latent process, content, and reliability, while external aspects such as criterion-related evidence, nomological networks, and consequences have nothing to do with the concept of validity. This view on validity was more or less
rejected by other validity theorists in the commentary articles (Embretson, 2007; Gorin, 2007; Mislevy, 2007; Moss, 2007; Sireci, 2007).

It seems reasonable to say that while some validity theorists have differing opinions, most agree that validity should be considered as a unitary concept and that consequences should be considered in validation efforts. Then one might ask, what is the reality? Do those who evaluate tests treat validity as a unitary concept, and what kind of validity evidence is reported? These questions were examined in a recent study (Cizek, Rosenberg, & Koons, 2008) where it was concluded that validity information is not routinely provided in terms of the unitary conceptualization of validity and that consequential evidence as well as other types of evidence are essentially ignored. More specifically, only 2.5 percent of the 283 test reviews that were examined had a unitary conceptualization of validity, and equally few reported validity evidence based on consequences. Also, only one out of four reviews specifically referred to validity as a characteristic of test score, inference, or interpretation. These results clearly indicate that there is a wide and disquieting gap between validity theory and practice.

3.2.3 An Argument-Based Approach to Validity

Cronbach (1988) proposed that validators should use evaluation argument when validating test score interpretations and uses, and that they should think of “validity argument” rather than “validation research”. Because validation speaks to a diverse audience, the argument must “link concepts, evidence, social and personal consequences, and values” (Cronbach, 1988, p. 4; italics in original). Kane (1992) further developed the concept of an argument-based approach to validity. He argued that validation should always begin with a specification of the proposed interpretations and uses of the scores, which is referred to as the interpretive argument. The latest edition of the Standards (AERA, APA, & NCME, 1999) adopted this approach, suggesting that “Validity logically begins with an explicit statement of the proposed interpretation of test scores along with a rationale for the relevance of the interpretation to the proposed use” (p. 9). The interpretive argument involves clear statements about the inferences related to the interpretation, and the assumptions associated with these inferences. Once the interpretive argument has been specified, the next step in the validation process is to evaluate the inferences and supporting assumptions in the interpretive argument using appropriate evidence; this evaluation is referred to as the validity argument (Kane, 2006).

Kane (2006) suggests the following strategy for validating a proposed interpretation and/or use of a test score. First, an interpretive argument is outlined along with a test development plan. Second, the test is developed. Developing the test consistently with the interpretive argument supports the
argument. Third, the inferences and assumptions in the interpretive argument are evaluated during test development to the extent possible and, if deemed necessary, changes are made to the interpretive argument. This is repeated until the test developers are satisfied with the fit between the test and the interpretive argument. Fourth, the overall plausibility of the interpretive argument is examined (under the assumption that the test is a finished product). This examination also involves a search for hidden assumptions and alternate possible interpretations. In general, studies of the most questionable assumptions are likely to be most informative. When evaluating the interpretive argument, it should be examined with respect to the clarity of argument, the coherence of the argument, and the plausibility of inferences and assumptions. The evidence relevant to a certain inference and its supporting assumptions can be of various kinds, including the results from previous research, empirical studies, expert judgments, and value judgments.

3.2.4 A Comment on the Argument-Based Approach to Validity

The idea of an argument-based approach to validity seems appealing, not least because it provides much needed guidance in allocating research efforts and in monitoring the validation effort. However, I believe that there is great risk that test developers will be confused by the abundance of concepts and terms used when describing this approach, and from the description of the approach itself. For one thing, the difference between interpretation and use is not always clear, at least not to someone who is a non-native English speaker. In some accounts, Kane (2006) talks about interpretation and uses, and in others he talks about interpretation and/or uses. For example, it is stated that “An interpretive argument specifies the proposed interpretations and uses of test results” (p. 23; italics in original), and “The test developers have some interpretations and/or uses in mind when they begin developing a test” (p. 25). Also, the term interpretive argument suggests that the argument deals with an interpretation. So, should a test developer also specify a utility argument? Or would a more appropriate name of the interpretive argument be an interpretive/utility argument? I would claim that it makes little sense to merely provide an interpretation of a score without using the score for some specific purpose. Conversely, it would be awkward to use a score if we had not provided some explanation of the meaning of the test score, that is, an interpretation of the score. Surely, a certain interpretation could be associated with different uses and a certain use could be associated with different interpretations. My point is that for all practical purposes you cannot have one without the other.

Another thing that may seem unclear to test developers is the difference between validating the interpretive argument and validating the proposed
interpretation. As I understand it, validation of a proposed interpretation involves specification and evaluation of the interpretive argument. Validation of the interpretive argument sounds more like the process of examining whether the inferences and assumptions make sense for the proposed interpretation. However, Kane’s description of the process suggests that the validation of the interpretive argument is the evaluation of the interpretive argument. In summary, the concept of the argument-based approach to validation should be appealing to test developers, but applying the approach in practice may not always be completely straightforward.

3.3 Gathering Validity Evidence for the SweSAT

In this section I will provide interpretive arguments and validity arguments for the SweSAT. I focus on two uses of SweSAT scores: the use of the scores for admission decisions, and the use of the scores for providing diagnostic information to the test-takers. The primary purpose of providing the interpretive arguments is to set a common basis for the papers in this thesis, and although they have some intrinsic value the interpretive arguments should be viewed as being preliminary. Similarly, the validity arguments should be viewed as a first attempt at providing a comprehensive evaluation of the assumptions underlying the interpretation and use of SweSAT scores.

3.3.1 Interpretive Arguments for the SweSAT

Drawing on the work of Kane (2006, 1992, 2002), I present interpretive arguments for the two uses of SweSAT scores mentioned earlier. There are four inferences that one wants to make. The first involves the scoring procedures, when one wants to make an inference from observed performance to observed score. The second inference is generalization, that is, an inference from observed score to true score. The third inference is extrapolation, that is, an inference from true score to the construct (general ability). The fourth inference is the decision about admission that is made on the basis of the level of ability. The first three inferences are referred to as semantic inferences and the decision inference is referred to as a policy inference; consequently, the assumptions involved with these inferences are referred to as semantic assumptions and policy assumptions respectively (Kane, 2002). Each of these inferences is based on assumptions that have to be met in order to argue that the inference is appropriate.
An Interpretive Argument for the Use of SweSAT Scores for Admission Decisions

Inference 1: Scoring – from observed performance to observed score

The scoring inference is based on assumptions about the scoring rule and the equating and scaling procedure. The scoring rule assigns a score to each test-taker’s performance on the items. Wainer and Thissen (2001) assert that “it is absolutely essential in any valid test that a specification of the score or code any answer is to receive is defined for every item” (p. 24). In the case of a multiple-choice test like the SweSAT, the scoring rule is simply the answer key. There are two assumptions related to the scoring rule. First, it should be appropriate, and second, it should be applied accurately and consistently. By “appropriate” I mean that there should be no apparent errors in the answer key, such as items having no correct answer or several correct answers. Of course, there are situations when an item can be keyed as having no or several correct answers without it being made by mistake, for example when an item is flawed. Further, the answer key should be applied accurately and consistently. That is, every test-taker that has answered a certain item correctly according to the answer key should be awarded a score for that item, and conversely, every test-taker that has answered a certain item incorrectly according to the answer key should not be awarded a score for that item. For tests using formula scoring (e.g., the SAT), incorrect answers should lead to the designated score reduction associated with the specific items.

The assumptions related to the scoring rule can also be applied to the equating and scaling procedures. In the case of the SweSAT, the equating and scaling might be placed under the same assumptions because these two procedures are carried out as one integrated procedure. Raw scores (observed number-correct scores, range 0–122) are transformed to normed scale scores (range 0.0–2.0), and in that transformation the scores from different administrations are equated as well; this procedure is described in detail in Paper II. However, because the equating part and the scaling part of the score transformation rely on separate assumptions they are treated as separate procedures. Also, in situations where raw scores are first equated and then transformed to scale scores it is definitely more appropriate to separate scaling assumptions from equating assumptions. The appropriate-ness of the equating procedure has to do with issues regarding parallelness of test forms, equating designs (i.e., how the data are collected) and equating methods. Equating is intended to be used for adjusting minor differences in difficulty between test forms. Therefore, it is important that the test forms to be equated are as equal as possible in terms of difficulty, and consequently that the test developers can as accurately as possible predict item and test
difficulty. The choice of equating design is dependent upon the test administration conditions and statistical properties of the test-taking groups (Kolen & Brennan, 2004). In other words, there are a number of assumptions associated with each equating design. The choice of equating method also involves assumptions. For example, the mean equating method assumes that two forms of a test differ in difficulty by a constant along the scale, and IRT equating methods involve the basic assumptions of the IRT models. If the assumptions associated with the equating design and equating method are not met then scores from different administrations will not be comparable, and that would threaten the validity of the test. In addition to the assumptions regarding specific designs and methods, there are also some general requirements that are often regarded as basic to all test equating (see e.g., Dorans & Holland, 2000):

1. **The Same Construct Requirement.** Tests that measure different constructs should not be equated.

2. **The Equal Reliability Requirement.** Tests that measure the same construct but differ in reliability should not be equated.

3. **The Symmetry Requirement.** The equating function for equating the scores of Y to those of X should be the inverse of the equating function for equating the scores of X to those of Y.

4. **The Equity Requirement.** Given that tests Y and X have been equated, it should not matter for a test-taker which test he or she is tested by.

5. **The Population Invariance Requirement.** It should not matter which (sub)population is used to compute the equating function between the scores of tests Y and X (i.e., the equating function should be population invariant).

Scaling is the process of associating numbers or other ordered indicators with the performance of examinees (Kolen & Brennan, 2004). As stated previously, the raw scores on the SweSAT are transformed to scale scores. This transformation should be carried out using a procedure that is appropriate for the intended purpose. In addition, this scaling procedure should be applied accurately and consistently. Petersen, Kolen, and Hoover (1989) stated that “the main purpose of scaling is to aid users in interpreting test results” (p. 156). This can be done by incorporating normative, score precision, or content information into the score scales. Thus, scaling procedures should be judged by how well they encourage accurate interpretations of test scores and discourage improper score interpretations (Kolen, 2006). If normative information is to be incorporated in the score
scale it is crucial that the norm group has been defined explicitly. When it comes to college admission tests, for example, it might be relevant to think about if the test-takers’ scores should be compared with whichever group that takes the test on a certain occasion or with a certain subgroup of test-takers. Further, because all scores contain measurement error, it seems reasonable to incorporate score precision information in the scale scores as well. Kolen (2006) points out that using too few scale score points leads to a loss of precision, while on the other hand, the use of too many scale score points might lead test users to attach significance to score differences that are small relative to the measurement error. Content information can be provided using procedures such as item mapping and standard setting. However, although content information can certainly be useful when interpreting scores, it is not as crucial for the interpretation of norm-referenced test scores as it is for the interpretation of criterion-referenced test scores.

Dorans (2002) suggests that a score scale should be well aligned with the intended uses of the scores. To achieve this, the score scale should possess six properties:

1. The scores of the norm group should be centered near the midpoint of the scale, and the average score in the norm group should be on or near the middle of the scale.
2. The distribution of scores for the norm group should be unimodal, and that mode should be near the midpoint of the scale.
3. The score distribution should be nearly symmetric about the average score.
4. The shape of the score distribution should follow a commonly recognized form, such as the normal curve.
5. The working range of scores should extend enough beyond the reported range of scores to permit shifts in population away from the scale midpoint without stressing the endpoints of the scale.
6. The number of scale units should not exceed the number of raw score points (this is similar to Kolen’s point about the number of scale score points).

In addition to the six properties, Dorans suggests that “a score scale should be viewed as infrastructure that is likely to require repair” (p.4), for example when the score distributions have moved far away from one of the endpoints or when norm groups have lost their relevance.
Inference 2: Generalization – from observed score to true score

In testing we expect our observations (e.g., test scores) to be generalizable over conditions of observation. In other words, we treat the observations as if they have been sampled from some universe of observations, involving different observers (e.g., raters), occasions, locations, and other conditions (Kane, 1992). The conditions of observation associated with the SweSAT include different locations (regions, places, premises, and rooms), time of observation, test forms, and testing proctors. In addition to the assumption that the observations can be generalized over testing conditions, the sample of observations is assumed to be large enough to control sampling error. That is, the number of items needs to be sufficiently large to produce reliable test scores.

The evidence needed to support the assumption for the generalization inference can be collected in reliability studies or generalizability studies (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 2001), or through multifaceted Rasch measurement (Xi, 2008). While reliability is not sufficient for validity, Kane (1992) points out that “reliability is a necessary condition for validity because generalization is a key inference in interpretive arguments” (p. 529). Judgments about the representativeness of the samples of observations in the test can also provide valuable information about generalizability (Kane, 2006).

Inference 3: Extrapolation – from true score to general ability

The extrapolation inference extends the interpretation from test performance to a claim about the abilities required to perform well in higher education (i.e., academic performance). For example, the READ subscore is interpreted as indicating the ability to comprehend a variety of texts and other written materials in a variety of contexts, even though the test consists of multiple-choice items that ask questions about excerpts of texts that are edited to a standardized format. Similarly, the DTM subscore is interpreted as indicating the ability to extract, use, and analyze information from diagrams, tables, maps, and other graphical representations of all shapes, sizes, and colors and from a variety of sources and contexts, but the graphical representations in the test are all grayscale and fitted to the A4 format. Also, scores on the SweSAT as a whole are interpreted as indicating the ability to extract, use, analyze, and comprehend information from a variety of sources (textbooks, reports, newspapers, spreadsheets, etc.) and contexts (lectures, self-instruction, laboratory work, etc.). While it is possible to sample information from a variety of sources, this information is always to some extent edited to fit the test format. Further, the testing context is very narrow compared to the possible contexts in which students are expected to
use their abilities. Therefore, we must extrapolate from test performance to claims about ability. Qualitative analyses of the relationship between test scores and non-test behavior can provide evidence to support the extrapolation inference (Kane, 1992).

Because the extrapolation involves claims about abilities required for college studies, it also involves expectations about performance in such studies. This can be supported by empirical evidence showing the relationship between test scores and performance in college studies (i.e., criterion-related/predictive validity evidence). The extrapolation inference is also based on the assumption that there are no construct-irrelevant sources of variability that seriously bias the score interpretation. For example, we assume that the background variables of the test-takers (e.g., sex and social class) have little impact on how the item functions. This can be examined through studies of differential item functioning (DIF; Xi, 2008). Also, we do not want the test-takers to be able to improve their scores through practice and coaching for the specific test without growth in the abilities measured by the test; this is examined in studies of repeated test-taking.

**Inference 4: Decision – from conclusion about level of ability to decision about admission**

All testing programs for admission to higher education, including the SweSAT, involve decisions in some form. The decision inference extends the interpretation of a test-taker’s ability to a decision about whether he or she is admitted or not. This inference is associated with assumptions about consequences and value judgments of using the test scores for admission (Kane, 1992, 2002).

One expected consequence of using SweSAT scores for admission that goes back to the introduction of the test, is that it would broaden recruitment to higher education. It was assumed that the introduction of the test would attract non-traditional student groups, specifically those from working class homes, and consequently that the score differences between different social groups would be small. This assumption is unrealistic, as pointed out by Henrysson (1994, pp. 12–13), initiator of the SweSAT program and former head of development of the test, in an account of a conversation he had back in the 1970s with the then minister of education:

*Minister: The SweSAT, does it yield less social group differences than the grades?*

*Henrysson: Hardly, both grades and tests register actual differences between the social groups. The test might yield somewhat less differences, but basically the same differences.*

*Minister: Then I'm not interested in your test.*
Henrysson: We have tried to adjust the test items to the working class, but that's only possible to a minor extent if the prognostic value of the test is to be maintained.

Minister: That doesn’t sound good.

Henrysson: If you want to favor the working class, I think you should suggest that you get extra credits if the father is a worker or has only attended elementary school.

Minister: We can’t do that, we would loose voters. I suggested drawing lots but I had to withdraw that. [Translated from Swedish]

Furthermore, the 2004 governmental commission on admission to higher education (SOU 2004:29) argue that the admission system is only one of ten different factors which a successful broadened recruitment is dependent upon and that other factors are much more important for achieving this goal. Therefore, it seems awkward to acknowledge this assumption in the interpretive argument. Although this assumption has not been officially discarded, it was not included in the interpretive argument.

In conjunction with the previous argumentation, I would argue that an important assumption is that using the SweSAT for admission will not lead to significantly lower levels of performance in higher education than had the test not been used. It can hardly be justified to use a test or any other instrument for admission decisions if it significantly lowers performance in higher education. Had the admission system been built on strict meritocratic principles it would be reasonable to assume (and expect) that using the test would significantly increase performance in higher education. However, because the introduction of the SweSAT was based much on sociopolitical grounds rather than meritocracy (Wikström, 2008), it makes sense to lower the expectations regarding performance in higher education. Furthermore, the benefits of using the test should be examined in relation to the costs involved in testing (Crouse & Brams, 1984; Kane, 2002; Phelps, 2002).

A requirement commonly referred to is that SweSAT scores should be accepted as meaningful for admission decisions by the users of the admission system, including the test-takers (SOU 2004:29; Stage, 2004a). This makes sense because if the test is regarded as not meaningful to the test-takers, then that may affect their motivation for taking the test and hence their performance. Also, if the colleges and universities using the test do not find it meaningful for admission decisions, then why should they support using the test in the first place?

In addition to examining the possible positive consequences of using the test scores for admission decisions it is also important to examine potential negative consequences resulting from testing (Kane, 2006). One such potential negative consequence is that testing has a negative impact on instruction in primary and secondary schools, such as disregarding parts of the curriculum to “teach to the test”. Therefore, it must be
assumed that using SweSAT scores for admission decisions will not have such negative consequences.

The policy assumptions accounted for above are all related to the use of SweSAT scores for admission decisions in general. If one were to account for policy assumptions related to the specific design of the admission system one

Table 2. An interpretive argument for the SweSAT when used for admission decisions

Inference 1: Scoring – from observed performance to observed score

| A1.1 | The scoring rule is appropriate |
| A1.2 | The scoring rule is applied accurately and consistently |
| A1.3 | The forms from different administrations are parallel |
| A1.4 | The equating procedure (design and method) is appropriate |
| A1.5 | The equating procedure is applied accurately and consistently |
| A1.6 | The scaling procedure for transforming raw scores to scale scores is appropriate |
| A1.7 | The scaling procedure for transforming raw scores to scale scores is applied accurately and consistently |

Inference 2: Generalization – from observed score to true score

| A2.1 | The observations made in testing are representative of the universe of observations defining the testing procedure |
| A2.2 | The sample of observations is large enough to control sampling error |

Inference 3: Extrapolation – from true score to general ability

| A3.1 | The test tasks tap the general abilities required in higher education studies |
| A3.2 | There are no construct-irrelevant sources of variability that seriously bias the interpretation of scores as measures of ability |
| A3.3 | Test-takers with a high level of ability perform better in higher education than test-takers with a low level of ability |

Inference 4: Decision – from conclusion about level of ability to decision about admission

| A4.1 | Using SweSAT scores for admission decisions will not lead to significantly lower levels of performance in higher education |
| A4.2 | SweSAT scores are accepted as meaningful for admission decisions by the users of the admission system, including the test-takers |
| A4.3 | Using SweSAT scores for admission decisions will not have a negative impact on instruction in primary and secondary school |
would also include policy assumptions about the other selection instruments, specifically the GPA. Given that the SweSAT is regarded as a second chance for candidates with a poor GPA, there is an underlying assumption that there are individuals with poor GPAs that are likely to perform well in higher education. However, because the interpretive argument is focused on the use of SweSAT scores in general, this assumption is not included in the argument. Table 2 shows an overview of the interpretive argument for the use of SweSAT scores for admission decisions.

**An Interpretive Argument for the Use of SweSAT Scores for Diagnostic Information**

The interpretive argument for the use of SweSAT scores for diagnostic information (Table 3) is based largely on the same assumptions as the interpretive argument for the use of SweSAT scores for admission decisions. The difference between the two lies mainly in fewer assumptions about the scoring inference, and other policy assumptions related to the decision inference.

When providing subscores (specifically subtest scores) to the test-takers there is no need to equate the scores from different test forms, because it is not important to make comparisons across forms and over time. The subscores could preferably be standardized or scaled in some other way to facilitate interpretations, but it is not done at present. Therefore, there are no assumptions regarding equating and scaling in this interpretive argument. Still, the scoring rule must be appropriate (A1.1) and applied accurately and consistently (A1.2). Also, the forms from different administrations must be parallel (A1.3) in order for the test-takers to make use of the remedial studies. The assumptions related to generalization (A2.1-2.2) are the same as for the admission decision argument, because the scores still have to be reliable and generalizable.

The assumptions related to extrapolation (A3.1r-3.2r) are basically the same as for the other interpretive argument, but they are defined at the subtest level. Assumptions A3.1r and A3.2r are then by and large the same as assumptions A3.1 and A3.2, respectively, and they are evaluated in the same way. Therefore, only the latter assumptions will be accounted for in the validity arguments. However, there is no need to make assumptions about the relationship between subtest scores and academic performance (see assumption A3.3), because that has no bearing on the test-takers’ possibilities of improving their scores. Furthermore, it could be argued that there should be an assumption stating that the subtest scores are indicators of the subtest constructs. The argument for this might be that it must be clear to the test-takers what the subtests measure in order for them to have a
Table 3. An interpretive argument for the SweSAT when used for providing diagnostic information

<table>
<thead>
<tr>
<th>Inference 1: Scoring – from observed performances to observed subscores</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1.1  The scoring rule is appropriate</td>
</tr>
<tr>
<td>A1.2  The scoring rule is applied accurately and consistently</td>
</tr>
<tr>
<td>A1.3  The forms from different administrations are parallel</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inference 2: Generalization – from observed subscores to true subscores</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2.1  The observations made in testing are representative of the universe of observations defining the testing procedure</td>
</tr>
<tr>
<td>A2.2  The sample of observations is large enough to control sampling error</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inference 3: Extrapolation – from true subscores to subtest abilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3.1r  The subtest tasks tap the abilities required in higher education studies</td>
</tr>
<tr>
<td>A3.2r  There are no construct-irrelevant sources of variability that seriously bias the interpretation of subscores as measures of ability</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Inference 4: Decision – from conclusion about levels of subtest abilities to decision about remedial studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>A4.1r  The subtest scores provide information that is useful for remediation and not already provided by the total score</td>
</tr>
<tr>
<td>A4.2r  The test-takers think the subtest scores provide valuable information</td>
</tr>
</tbody>
</table>

Note. The “r” after the assumption labels indicates that these assumptions are different from the assumptions in the other interpretive argument.

reasonable chance to actually improve their scores. Although this is desirable, it is not an absolute necessity because the test-takers have a great resource for practice in already administered tests, and they are readily available free of charge.

The policy assumptions regarding admission decisions are not applicable to this interpretive argument because the decision inference is not about admissions. Instead, the decision is at the individual’s own discretion whether or not to engage in remedial studies based on the subscore information. One assumption (A4.1r) is that the subscores provide information that is useful for remediation and not already provided by the total score. That is, the information provided by the subscores should be different from the information provided by the total score. Another assumption (A4.2r) underlying this decision inference is that the test-takers
find the provided subscores valuable as a basis for remedial studies. If the test-takers do not see any value in using the subscores, or if the subtest scores and the total score provide essentially the same information, then there is no point in reporting them. On the other hand, they can hardly do much harm if reported to the test-takers, other than possibly confusing them when interpreting the score reports.

3.3.2 Validity Arguments for the SweSAT

Under this heading, I will give a brief account of procedures and research that is relevant for the inferences and assumptions of the interpretive arguments. This should be considered as a first and preliminary attempt to provide comprehensive validity arguments for the different uses of the SweSAT.

A Validity Argument for the Use of SweSAT Scores for Admission Decisions

Inference 1: Scoring – from observed performance to observed score

Assumption 1.1: The scoring rule is appropriate. The answer key is established in cooperation with professional review boards, one for each subtest, which examine the final version of the subtests. The review boards might miss detecting flawed items, which would lead to a flawed answer key. However, because the items are made available to the public immediately after the test has been administered the test-takers or other people will almost always detect and report any potentially flawed items. Then, the test developer can decide on whether to change the answer key or, for example completely remove an item or provide an item with two correct answers. Fortunately, this has happened very seldom since the SweSAT was introduced.

Assumption 1.2: The scoring rule is applied accurately and consistently. The test is scored using state-of-the-art optical scanners that can distinguish markings of different depth, so that the boxes with the most lead or ink will be interpreted as the answer to the item rather than blurred markings caused by erasers. Any uncertainties in the markings as detected by the scanners are edited manually.

Assumption 1.3: The forms from different administrations are parallel. As stated previously, equating is to be used for adjusting only minor differences in test form difficulty and it is therefore crucial that test developers can accurately predict item and test difficulty in the regular test administrations. To aid in the construction of parallel test forms the test developers use
pretest data, and if these data are unreliable then it becomes difficult to make the test forms of essentially equal difficulty. One thing that can make the pretest data unreliable is if the group of test-takers taking the pretest items is very small or unrepresentative of the total group. However, the sample size is generally not a problem with at least 800–1,000 test-takers per item. Also, unrepresentativeness is usually controlled for by means of scores on the regular test (e.g., by using linear regression models). Another thing that threatens the reliability of the pretest data is if the items are revised before they are administered in a regular test, without an additional round of pretesting. Revisions can be made for different reasons, and if the test developers do not know what effect the revisions have on item difficulty it becomes harder to make test forms of equal difficulty. The effect of item revisions is the topic for Paper III, which will be discussed further in chapters 5 and 6.

Assumption 1.4: The equating procedure is appropriate. SweSAT scores are equated using the equipercentile equating method (see e.g., Braun & Holland, 1982) under the equivalent-groups design. One of the main advantages of equipercentile equating over mean equating and linear equating is that the equipercentile equated scores will always be within the range of possible scores, while mean equated and linear equated scores can be outside this range (unless the conversion is truncated at the highest and lowest scores). Regarding the choice of equating design, the single group design cannot be used because it is not the same group that takes the test forms. The nonequivalent groups anchor test (NEAT) design is preferable for many equating situations, because there are no strong assumptions involving the test-takers. The main advantage of the equivalent-groups design over the NEAT design is that a set of common items (an anchor test) does not have to be administered. However, using the equivalent-groups design relies on the assumption that the ability of the test-takers does not vary over time, and this is the topic of Paper II. Given that the assumption of equal ability is met, the equivalent-groups design is appropriate for equating SweSAT scores. If the assumption is not met, the NEAT design is preferable.

Assumption 1.5: The equating procedure is applied accurately and consistently. Once an appropriate equating method has been chosen it remains to apply it accurately and consistently. Kolen and Brennan (2004) suggest that analytical procedures are preferred over graphical procedures, and the SweSAT is equated using analytical procedures. Further, score distributions are often adjusted using a statistical procedure called smoothing (see e.g., Kolen & Brennan, 2004). This procedure is carried out to make the estimates of the empirical distributions and the equipercentile relationship from the sample more similar to what they would have been in
the population. Thus, it is hoped that the smoothed estimates will be more precise than the unsmoothed ones. There are two types of smoothing: smoothing of the score distributions (presmoothing) and smoothing of the equipercentile equivalents (postsmoothing). Another statistical adjustment of the score distributions is a procedure called *continuization* (Holland & Thayer, 1989; see also von Davier, Holland, & Thayer, 2004). Using equipercentile equating, at least as defined by Braun and Holland (1982), assumes that test scores are continuous random variables. However, test scores are usually discrete, such as the number-correct raw scores of the SweSAT. To circumvent this problem, the score distributions are continuized. SweSAT scores are neither smoothed nor continuized, and the potential effects of applying these procedures have not been examined.

**Assumption 1.6:** The scaling procedure for transforming raw scores to scale scores is appropriate. Normative information is included in the SweSAT score scale. For example, the scale scores are approximately normally distributed. Also, the norm group has been defined as the general population of test-takers. Score precision information is also included in the scale scores because each scale score corresponds to an interval of raw scores. The score scale also seems to possess most of the properties suggested by Dorans (2002): the average of the norm group is close to the midpoint of the scale, the distribution of scores for the norm group is unimodal and that mode is near the midpoint of the scale, the score distribution is nearly symmetric about the average score, the shape of the score distribution follows a commonly recognized form (i.e., a bell-shaped distribution), and the number of scale units does not exceed the number of raw score points (Stage & Ögren, 2006, 2007, 2008). However, it seems that the working range of scores does not extend enough beyond the reported range, because the last administrations have had much larger proportions of test-takers at the low end than at the high end, and if there occur more shifts in the population away from the scale midpoint then there will be too much stress on the low endpoint of the scale.

**Assumption 1.7:** The scaling procedure for transforming raw scores to scale scores is applied accurately and consistently. From looking at descriptions of how the procedure is carried out from administration to administration and from the outcome of the scaling, it seems that the procedure is carried out both accurately and consistently. For example, the range of scale scores is the same across administrations and the distributions have similar shapes and locations across administrations (Stage & Ögren, 2006, 2007, 2008). However, there are issues that should be further examined. In line with Dorans's (2002) suggestion that the score scale should be viewed as infrastructure it might be worth examining if the
score distribution has moved away enough from the high endpoint that it is in need of corrective action. Also, it should be examined whether the current norm group is still relevant or if there exist groups that are more relevant to compare test-takers with.

Inference 2: Generalization – from observed score to true score

Assumption 2.1: The observations made in testing are representative of the universe of observations defining the testing procedure. The conditions of testing that are examined regularly are the administrative regions and the test forms. The performance in the different regions is examined after each administration, and this information is used when estimating the total group performance on the pretest items. The equivalence of test forms is examined through the equating procedure, which has been described in more detail under assumptions 1.3–1.5. However, there are several conditions of testing that have not been examined, for example, the nature of the premises, group size, number of proctors, and time of administration (spring/fall).

Assumption 2.2: The sample of observations is large enough to control sampling error. This assumption is examined through reliability studies, including internal consistency and alternate forms reliability. The internal consistency (Cronbach’s $\alpha$) of the total score is usually around .92 (Stage & Ögren, 2004). Correlations between test-takers’ scores from two different administrations have also been estimated. Stage and Ögren (2004) examined the following pairs of administrations: 02A and 02B (correlation of .90), 03A and 03B (.90), 02A and 03A (.90), and 02B and 03B (.88). These correlations can be viewed as a combined test-retest and alternate forms reliability, because the forms are distinct and parallel, and the time between administrations is seven or twelve months. During the time between administrations the test-takers will progress (or regress) at different paces, which means that the estimated correlations are under-estimations of reliability.

Inference 3: Extrapolation – from true score to general ability

Assumption 3.1: The test tasks tap the abilities required in higher education studies. The SweSAT was developed with the basis in the skills and abilities considered important to perform well in higher education, such as reading comprehension ability, study technique, general knowledge, and some quantitative abilities. However, the test has changed since it was introduced, and direct measures of study technique and general knowledge are no longer part of the test. Today, two major parts of the test are knowledge of words out of context (WORD) and reading comprehension (READ and ERC).
Anyone would probably agree that reading comprehension is essential to all university studies, so having two reading comprehension subtests makes sense. However, it can be questioned whether the knowledge of words out of context is important in higher education. This is also one of the main reasons why it has been proposed that the weight of the WORD subtest should be reduced, and that items measuring knowledge of words in context should be introduced instead.

Andersson (2003) examined Social Welfare and Business Administration teachers’ view on which criteria are important for study success. The cognitive criteria that they agreed on as being important were among others writing skills and analytical skills. Åberg-Bengtsson (2005) found that SweSAT to some extent measures analytical skills. However, the assessment of writing ability/skills is not included in the SweSAT. Writing tests are usually considered expensive and complicated to score, but there are good reasons for considering adding a writing test. As mentioned in a previous chapter, an international evaluation of the SweSAT recommended adding a writing test. Adding a writing test would send an important signal to candidates and upper-secondary schools that writing ability is important. Also, it may result in a stronger relationship between test scores and academic performance. For example, an analysis of data from the University of California showed that of the various tests in the SAT I (the SAT Reasoning Test) and the SAT II (the SAT Subject Tests) examinations the best single predictor of student performance was the SAT II writing test (Atkinson, 2001). Consequently, it was concluded that “Given the importance of writing ability at the college level, it should not be surprising that a test of actual writing skills correlates strongly with freshman grades” (Atkinson, 2001, p. 5).

Assumption 3.2: There are no construct-irrelevant sources of variability that seriously bias the interpretation of scores as measures of general ability. DIF studies on the SweSAT have been performed mainly with regard to gender, but also social class. Stage (2004b; 2005) examined DIF with regard to social class and concluded that there are very few items that function differently for social groups. However, gender DIF has been more frequent. In the 92A test form, 11 percent of the items had severe DIF and 14 percent had moderate DIF (Stage, 2004b), while in the 02A test form 3 percent of the items had severe DIF and 14 percent had moderate DIF (Stage, 2005). In general, the WORD subtest seems to be the most problematic one. On the 92A form, where WORD accounted for 20 percent of the total number of items, 30 percent of the severe and moderate DIF items together were WORD items. The corresponding percentage on the 02A form, where WORD accounted for 33 percent of the total number of items, was 52 percent. The items favoring females are mostly related to art,
literature, care and nursing, culture, and home economics, while the items favoring males are related to sports, politics, technology, geography, and natural science. Given that most WORD items are specific to what can be considered as female or male areas, one might be compelled to question the content model that WORD is based on. However, this model specifies that items from several different areas of study should be chosen, such as culture and natural science. Consequently, if one wants to reduce DIF in the WORD subtest it is probably necessary to change the content model. The most troubling issue about the examination of DIF is that it does not seem to be carried out routinely as part of the item analysis.

The effects of practice and coaching have been examined in several studies. Henriksson (1981) reviewed the literature on practice and coaching and found that the effects of repeated test-taking in general are greater for speeded tests, and that this is due to the test-takers developing a strategy for how to use the allotted time. Henriksson also found that the effects are greater on non-verbal tests and on tests with complex item formats, and that the more able the test-takers are, the more they will benefit from unsupported practice. Törnqvist and Henriksson (2006) reviewed previous studies of repeated test-taking on the SweSAT and found that the largest increase in mean scores occurs between the first and the second test occasion, even after controlling for self-selection. Cliffordson (2004) found similar results, but she concluded that repeated test-taking involves growth effects in addition to practice effects. Törnqvist and Henriksson suggested that the score gain from the first to the second test is mainly due to test-wiseness (e.g., Rogers & Bateson, 1991). They argue that a test-taker’s observed score from the first test occasion is an underestimation of his or her true score and consequently that test-takers need a certain amount of test-wiseness in order to get a score that is a good estimate of their true score. One way to achieve this is to provide the test-takers with detailed information about the test, for example, item format and timing issues. Such information is readily available online and in information booklets published by the NAHE. Also, all tests are released to the public shortly after a test has been administered, which makes up an excellent source of practice material for potential test-takers. To conclude, there are practice effects on the SweSAT but measures are taken to make all test-takers sufficiently prepared before taking the test.

Assumption 3.3: Test-takers with a high level of ability perform better in higher education than test-takers with a low level of ability. This assumption is examined through criterion-related validity evidence. Typically, the correlation between test scores and criteria for academic performance is estimated to provide predictive validity evidence. Predictive validity studies are usually associated with a number of methodological
issues of which range restriction is one of the most common, as indicated by Wainer and Thissen (2001):

To understand how well a college admission test works in selecting students, one would have to admit some who are predicted to do very poorly. Because this is not ordinarily done, we must extrapolate from the usually narrow range of individuals who are admitted to the subjunctive case of what would have happened if admissions were carried out without the test. (p. 25)

Paper I provides a review of predictive validity studies of SweSAT scores and a discussion of methodological issues related to such studies.

Inference 4: Decision – from conclusion about level of ability to decision about admission

Assumption 4.1: Using SweSAT scores for admission decisions will not lead to significantly lower levels of performance in higher education. This issue does not seem to have been examined previously. If the SweSAT was the only instrument used for selection then one could simply look at the relationship between test scores and academic performance; if there is a null or positive relationship then the use of the test will not lead to lower performance as compared to simply selecting applicants at random or admitting all applicants. However, the SweSAT is one of two main selection instruments, which complicates things. Given that the upper-secondary school GPA is generally considered as being better at selecting good-performing applicants (see Paper I for more information on this issue) one might be tempted to conclude that better-performing applicants could be selected by using the GPA only. However, this issue needs to be examined more thoroughly before such conclusions can be drawn.

Assumption 4.2: SweSAT scores are accepted as meaningful for admission decisions by the users of the admission system, including the test-takers. The test-takers’ views on the SweSAT was examined on a regular basis during the 1980s and early 1990s, and the general understanding has been that the test-takers find the test to be meaningful for use as a selection instrument. In the most recent study (Eriksson, 2003), between 61 percent and 83 percent of the test-takers regarded the different subtests as relevant for selection (WORD 61%; DS 63%; DTM 67%; ERC 82%; READ 83%), which can be considered as fairly high numbers. The test-takers were also asked if they miss some kind of subtest in the SweSAT. About three fourths of them did not miss a subtest, while about one third of those who did wanted some kind of a general information test, which might be due to the test-takers’ view that such a test covers important knowledge (Eriksson, 2003).
Another aspect of this assumption is whether the selection procedure is perceived as fair. Wolming (2008) states that “If the applicants’ perceptions of the procedure are negative ... this could have serious implications for the validity of the selection procedure” (pp. 10–11). In this study, which examines test-taker perceptions of fairness in the selection procedure, Wolming concluded that the test-takers’ perceptions of procedural fairness were generally positive.

**Assumption 4.3:** Using SweSAT scores for admission decisions will not have a negative impact on instruction in primary and secondary schools. This assumption was partially examined through a survey aimed at English teachers in upper-secondary schools (Ohlander, 1999). The survey was conducted in 1998, six years after the ERC subtest was introduced. The results indicated that 87 percent of the teachers were at least somewhat familiar with the test, and 24 percent of the teachers stated that students at the courses English B and C occasionally or frequently wanted the teachers to prepare them for the ERC test. Further, 22 percent of the teachers stated that the ERC test to some or a high degree had affected their own and their students’ choice of texts to work with in class, and 19 percent of the teachers stated that the test had affected their way of teaching. However, in the cases where the ERC affected the choice of texts and the instruction it is not clear whether this was in a positive or a negative way. Either way, Ohlander concluded that the ERC test has affected English instruction in upper-secondary school to a relatively slight extent. What remain to be examined are how the WORD and READ subtests affect Swedish instruction and how the DS and DTM subtests affect mathematics instruction.

**A Validity Argument for the Use of SweSAT Scores for Diagnostic Information**

**Inference 4:** Decision – from conclusion about level of ability to decision about remedial studies

**Assumption 4.1r:** The subtest scores provide information that is useful for remediation and not already provided by the total score. To meet this assumption, the subscores should provide additional information beyond that already provided by the total score; this is what is examined in Paper IV.

**Assumption 4.2r:** The test-takers think the subtest scores provide valuable information. This assumption was examined in a pilot study of test-takers’ attitudes towards the SweSAT score report and the score scale (Lyrén, 2008). The test-takers were asked among other questions whether the information in the score report is sufficient to understand one’s strengths
and/or weaknesses. On a scale from 1 (No, not at all) to 5 (Yes, definitely) about 52 percent marked a 4 or a 5, while about 22 percent marked a 1 or 2. This indicates that for most test-takers the subtest scores provided in the score report can be useful for remedial action. Yet, this issue needs to be further examined before any definitive conclusion can be drawn.

4 Methodological Issues

4.1 Test-Theoretical Approaches

When dealing with test scores and item statistics one must choose a measurement model for how to score responses and estimate the item statistics. In general, two different approaches can be used. Classical test theory (CTT) is based on the notion of a true score, which is a variable that is not directly observable. It is assumed that each test-taker's observed score on the test consists of a true score and an error score. Consequently, if the measurement is error-free then the true score will be equal to the observed score. Some basic assumptions of CTT are that the expected value of the error score is zero and that true scores and error scores are perfectly uncorrelated.

One of the main shortcomings of CTT is that test-taker characteristics (e.g., ability) and item/test characteristics (e.g., difficulty) cannot be separated. Item response theory (IRT) deals with this problem through the use of probabilistic models, most commonly logistic models. When an IRT model fits the data, test-taker ability can be estimated from any set of items, and item difficulty can be estimated in any group of test-takers (Hambleton, Swaminathan, & Rogers, 1991). Some assumptions of regular unidimensional IRT models are unidimensionality (of course) and local independence.

IRT has become widely popular in psychometrics over the last decades, and one would have good reason to expect a thesis on educational measurement to capitalize on the advantages of this approach. However, IRT was not used in any of the four papers. The main reason is that the SweSAT is developed, scored and equated using CTT approaches, and it makes sense to examine the test scores using the same approaches as is done operationally. IRT has actually been used for several years in the equating of scores, but that is more on an experimental level and the equating procedure is not explicitly designed to facilitate IRT-based equating methods. Moreover, using IRT would not add much value to the papers. Paper I is not an empirical study at all, Paper II examines patterns of score differences (which are not likely to change when using IRT), Paper III examines item characteristics derived from CTT, and Paper IV examines the subtest scores,
which also are CTT-derived. In conclusion, given the data it makes more sense to use CTT methods rather than IRT methods.

4.2 Statistical Considerations

The empirical papers (II-IV) all use Pearson correlations to some extent. Pearson correlations presume normally distributed and continuous data, and the SweSAT scaled scores and number-correct raw scores can be considered normal but they are by definition not continuous. However, Pearson correlations are generally considered to be robust to violations of these assumptions and are therefore used in the studies. Furthermore, a level of significance of .05 was used for all statistical tests.

The preequating approach in Paper III is based on the binomial error model (Lord & Novick, 1968) and presumes that all items have equal difficulty, which is a strong assumption that is not very likely to hold for the SweSAT data. However, because it is a relatively straightforward approach that only needs two parameters to estimate score distributions and because these parameters are readily available in the item data this approach was considered the most appropriate one.

5 Summaries of the Papers

This chapter provides a summary of the four papers attached to this thesis. Papers I–III are related to assumptions concerning the use of SweSAT scores for admission decisions while Paper IV is related to an assumption concerning the use of SweSAT scores for diagnostic information.

5.1 Paper I. Prediction of Academic Performance by Means of the Swedish Scholastic Assessment Test

There is a widespread belief that the SweSAT is a poor predictor of academic performance. However, no study has summarized the findings from studies that examine evidence for the predictive validity of the test. This paper contains a review and discussion of such studies and the methodological issues related to them. The data consisted of ten studies, published as journal articles or reports, between 1994 and 2006. Only three of the studies report correlation coefficients while the other seven report the average of achieved credits for students in the GPA group and the SweSAT group respectively. The latter type of study only provides information about how the two selection instruments function relative to each other, and does not give any absolute measure of how the SweSAT in itself functions. The results
indicate that the predictive power of the test differs between programs, and that the SweSAT seems to be a slightly poorer predictor of academic performance, both in terms of correlations and the average number of course credits. However, there does not seem to be any differences in dropout rates or any significant prediction bias due to gender and social class. Given the results, it is suggested that the use of the test could be differentiated depending on the program applied to. It is also suggested that the construct of the test should be expanded to incorporate other skills and abilities required in college studies that are not currently measured by the SweSAT.

5.1.1 Errata for Paper I

- Page 567, paragraph 4, line 9: “HSPGAs” should be “HSGPAs”
- Page 570, Table 1, row 1, column 5: The last line that reads “educ. Y4–9” should be moved up one line, as below

| Civil engin. | 17,198 |
| Law | 2,378 |
| Teacher Y1–7 | 9,605 |

5.2 Paper II. Systematic Equating Error with the Randomly-Equivalent Groups Design: An Examination of the Equal Ability Distribution Assumption

The SweSAT, as well as many other high-stakes tests, is equated using the randomly-equivalent groups design. An underlying assumption of using this design is that those taking the different test forms have equal abilities. Because this is a very strong assumption the SweSAT forms are equated using not only the total group, but two assumingly homogeneous reference groups as well. However, the assumption of equal abilities still remains for the specific groups. Because of a variety of reasons, such as educational reforms taking place, there is reason to believe that the equal ability assumption might be violated. The purpose of this study was to examine this issue.

Using DS and WORD regular items and anchor items (22 DS anchor items and 15 WORD anchor items) the test-takers’ performance over 14 administrations was examined. Scores on the anchor items were estimated using the test-takers’ scores on the regular subtests. There were significant differences in mean WORD scores in the total group as well as the two reference groups, with an expected raw score difference of 2–3 points. For the DS anchor items, there were significant differences in mean scores in the
total group and reference group II (upper-secondary school seniors at academic programs) but not in reference group I (a proportionally stratified sample). Adding together the scores for the DS and WORD anchor items and assuming that the anchor items are representative of the total test and that the respective anchor test and regular subtest are parallel, the largest estimated total-score difference between two administrations would be more than 4 points. This implies quite severe potential implications for the test-takers, and it is concluded that the equal ability assumption seems to be highly problematic. Consequently, it is suggested that an equating design that is not dependent on this assumption should be used in the equating of scores from different forms of the SweSAT.

5.3 Paper III. The Effect of Item Revisions on Classical Item Statistics and Preequating Outcomes

The construction of parallel test forms is perhaps the most important test development procedure in college admission testing. This procedure is dependent on having reliable item statistics, which are obtained from pretestings of the items. The item statistics can change significantly if the items are revised after pretesting and in situations where it is not feasible to have revised items go through another round of pretesting it is important to know the effect of the revisions on the item statistics. The purpose of this paper was to examine the effect of textual item revisions on classical item statistics. The effect of the revisions on the adequacy of preequating outcomes is also examined in the paper, because using a preequating design on the SweSAT would have many potential benefits.

Three forms of the DS subtest, a total of 66 items, were examined. The items were divided into six categories after type of revision (0 = intact; 1 = linguistic/aesthetic revision; 2 = simplification; 3 = change of tense; 4 = single elucidation; 5 = multiple elucidations) and the differences in p-values and point-biserial correlations between regular test and pretest were averaged in each category. As expected, the mean of the difference in p-values for the 14 intact items was not significantly different from zero. However, the standard deviation was about .05, which means that intact items can also be expected to exhibit fluctuations in p-values. Because there were very few items in some categories, the six categories were collapsed into two categories: 0* (categories 0, 1, and 2) and 1* (categories 3, 4, and 5). The difference in both p-values and point-biserial correlations was not significantly different from zero in category 0* while it was significantly different from zero in category 1*. Consequently, it was concluded that textual revisions to items have a significant effect on item difficulty and discrimination. Regarding the preequating results, it was concluded that (a) the preequating procedure produced adequate results for the three test
forms, and (b) that making revisions to already pretested items leads to poorer preequating results.

5.4 Paper IV. Reporting Subscores from College Admission Tests

This paper focuses on the use of SweSAT subtest scores for diagnostic information. For some time now, there has been an increasing interest in the reporting of subscores for various tests, including college admission tests. In this study, a new CTT-derived method for determining the added value of reporting subscores was applied to the SweSAT subtest scores and section scores. The method is based on a comparison of the proportional reduction in mean squared error (PRMSE) of the observed subscore and the observed total score as predictors of the true subscore: the predictor that has the largest PRMSE is considered as the relatively better predictor. Further, if the observed subscore is the relatively better predictor it can be concluded that there is added value in reporting the subscore. The results showed that the observed subtest score PRMSEs were larger than the observed total score PRMSEs for four of the five subtests, which indicates that there is added value in reporting the subtest scores. There also seems to be added value in reporting section scores (Verbal + Quantitative), which gives support to the prospective use of section scores for admission decisions.

6 Discussion

The main purpose of this thesis was to examine some assumptions that are of importance for the validity of the interpretation and use of SweSAT scores. In Paper I, II and III, assumptions relevant for the use of SweSAT scores for admission decisions were examined, and in Paper IV an assumption relevant for the use of SweSAT scores as diagnostic information was examined. The basis for the studies was derived from interpretive arguments that specify the inferences and assumptions underlying the different interpretations and uses of the scores. In addition, preliminary validity arguments were provided that give an account of research that supports the interpretation and use of the scores.

6.1 Main Results – Implications for the Validity Argument

Do the results from the studies support the interpretive argument? For the use of SweSAT scores for admission decisions, the results from the studies both support and weaken the argument for the current use. First, the
SweSAT seems to be a reasonably good predictor of academic performance (see Assumption 3.2). However, there are large differences between study programs. Second, the assumption of equal abilities over time underlying the current equating procedure does not seem to be sufficiently met (see Assumption 1.4). Third, there are circumstances that complicate the construction of parallel test forms (see Assumption 1.3). For the use of SweSAT scores as a basis for diagnostic information, the results from the studies support this use because the subtest scores provide information that is unique from the total score. It should be pointed out that even though the research presented in these papers does not entirely support the arguments for the use of the scores for admission decisions, it was some of the most problematic assumptions that were examined. Also, as was shown in the validity argument there is appropriate backing for the majority of the other assumptions. In all, the SweSAT scores are perhaps not entirely perfect in the sense referred to in the introductory chapter, but they are not in any sense alarmingly far from being so.

6.2 Implications for Test Development

The results suggest that some parts of the test development procedure should be reconsidered. For example, a different equating design should be used that does not rely on the assumption of equal abilities. The most feasible alternative is to use some sort of NEAT design, which involves administration of anchor items/tests. How these would be administered is an issue that deserves the utmost attention. This issue is related to the issue of how to pretest items, and there are basically three approaches to item pretesting: (a) An embedded section within the test, (b) embedded items within a section of the test, and (c) external to the test itself (Wendler & Walker, 2006). Of course, the external approach is not feasible for administering anchor items because they have to be administered along with the tests that are to be equated. In addition, external item pretesting is usually associated with a number of problematic issues, for example, questionable motivation of the test-takers, and security issues. In addition, external pretesting is associated with high costs such as recruitment efforts, payment to examinees and administrators, etc. (Schmeiser & Welch, 2006). External pretesting was carried out on SweSAT items up until 1996 when it was changed to embedded sections pretesting, much because of the reasons just mentioned. Embedded sections pretesting and embedded items pretesting both have their advantages and disadvantages. If embedded sections pretesting is used then the test day must include a pretest section (also denoted variable section), as is done today. However, close to 60 percent of the test-takers think that the pretest section takes too much time and energy, and about 47 percent of the test-takers would prefer to have the
pretest items embedded within the regular test (Eriksson, 2003). Also, as is the case today there are large potential subtest order effects that might have an adverse impact on the reliability of the anchor and pretest item statistics. On the other hand, with complete tests item-level context effects are avoided. Having anchor items embedded within regular test forms introduces such context effects, which can be quite severe. Also, embedding items within a regular test section probably requires much more editorial time than embedding sections within the regular test as a whole.

In conjunction with the previous point, the procedures for reviewing and pretesting items should be reconsidered. Given the importance of having reliable item statistics and the effect of textual revisions on these statistics it seems that much can be gained from restricting the possibility of administering items that have been revised but not re-pretested. There are two approaches that can be considered to achieve this. First, when items are up for review for a future regular test, only revisions that eliminate actual or potential flaws should be carried out. Proposed revisions that are purely cosmetic by nature should not be carried out. Second, items that must be revised after pretest in order to avoid flaws should, to the greatest extent possible, be subjected to another round of pretesting. In the current SweSAT test development procedure the items are reviewed once before pretesting by a small group consisting of measurement experts, content experts, and language experts. Then, the items are pretested and, finally, the items that are chosen for a regular test are reviewed once again by a larger group of experts. If this procedure was expanded so that new items were subjected to more scrutiny through, for instance, double reviews before pretesting then the need to revise already pretested items would most likely be reduced and hence the item statistics would be more reliable. In summary, pretesting is one of the most important procedures for quality assurance, and the more rigorous the pretesting procedure the better the quality of the test.

6.3 Implications for Test-Takers

Given the results in Paper II, there is reason to believe that some test-takers have been wronged in that they have received lower scores than they should relative to test-takers taking other test forms. It is impossible to know exactly how many additional score points a certain test-taker should have received, and therefore the scores cannot be changed afterwards. Until a more appropriate equating design is being applied, the best advice to give to potential candidates is to take the test on every occasion possible, especially if they have an already valid score from several years ago. This is analogous to advising them to capitalize on measurement error. Every test score they receive contains measurement error, either positive or negative. Because the highest obtained score is used in the selection process, test-takers that take
the test multiple times will have a larger chance of maximizing their score by receiving a score with a (large) positive measurement error.

Another implication for the test-takers, which is derived from Paper IV, is that they can use the subtest scores for remedial studies with a good chance of improving their total score. However, to get a sense of their (Swedish) reading comprehension ability they should look at their total score rather than the READ subtest score.

6.4 Validation Issues

After having used the argument-based approach to validation in this thesis, I must conclude that it is an intuitive and powerful approach that should appeal to validators of all kinds of tests. Still, there are some issues to consider. For example, to what extent should subordinate assumptions be specified in the interpretive argument? In the interpretive arguments provided in this thesis, few subassumptions were explicitly specified. If important assumptions are not explicitly specified, such as the assumptions underlying the equating design, there is a risk that these assumptions will be overlooked in the validity argument. On the other hand, if very many assumptions are specified there is a risk that the not so important assumptions conceal the more important ones.

Another issue to consider is to what extent theory-based inferences and assumptions that are derived from nomological theories or process models should be considered (see e.g., Kane, 1992, 2006). No theory-based inferences or related assumptions were explicitly specified in the presented interpretive arguments, mainly because the extrapolation inference implicitly involves such inferences. For instance, if the total score is assumed to indicate some general ability that is of basically equal importance for academic performance in all areas of study, and if the score predicts academic performance in all areas of study equally well, then it is reasonable to conclude that the total score is an indicator of some general ability. However, if the total score predicts academic performance differently in different areas of study then it can be concluded that the total score contains subcomponents that cause this difference in prediction and, consequently, that the total score cannot be assumed to measure a truly general ability that is important for all areas of study. If section scores are to be reported and used differently for different areas of study then there is more reason to explicitly specify theory-based inferences. Also, if the proposed interpretation and use of the scores change in other ways then there is reason to examine the interpretive argument to see if there are inferences and assumptions that should be removed or added.
6.5 Limitations and Generalization

As is the case with basically all kinds of research, this thesis and its papers have limitations that are potential threats to generalizations. For one thing, the interpretive argument as presented in this thesis is my view of what such an argument might look like. Test owners, test developers, test users, test-takers, researchers, politicians, and others may have differing views of what should be included in the interpretive argument. Of course, in future work of establishing an interpretive argument these views should be considered. Further, the review in Paper I is based on a relatively small number of studies. In particular, there are few correlation studies. Yet, these are the studies that exist and the ones that can form the basis for a review. In Paper II, the main limitations are related to the WORD anchor items. These items are embedded within a pretest form, which means that context effects might influence the item statistics. Also, vocabulary use is anything but constant. Some words or phrases may be used less frequently and some may be used more frequently over time, but in a large sample of items these changes would probably cancel each other out. However, in small samples of items changes to specific items can have potentially significant effects.

Paper III is based on a relatively small sample of items, especially for this kind of study. Ideally, the sample of items would have been large enough to place at least 30 items in each revision category, or at least such that the number of intact items was about 30. Unfortunately, the DS anchor test has been administered during such a short period of time that it allowed for the examination of only these few DS test forms.

6.6 Suggestions for Further Research and Development

The current development work by the NAHE suggests that new item types should be introduced and that more coherent test sections should be formed. As part of this work, it is important to consider the potential impact on the validation of the test. For example, how are the inferences and assumptions in the interpretive argument affected? Which are currently the strongest assumptions that are least likely to be met, and will other questionable assumptions be introduced as a result of the current development work? As Kane (2006) points out, the most relevant kinds of validity evidence are those that support the most problematic inferences and assumptions.

Moreover, the equating issues must be examined more thoroughly. One of these issues is the potential benefit of applying a preequating design. If an appropriate model for estimating score distributions is used, then the problem of varying abilities over time can be reduced because the items in a regular test have been pretested at different administrations. Another issue to examine further is, as already suggested, how prospective anchor items
could be administered. In light of the apparent problems associated with equating, the alternative to equating the tests at all, which is to have the scores be valid for a single administration only, should also be considered.

Scaling is another scoring issue that deserves more attention. For example, the number of score points on the score scale should be reconsidered. Currently, the SweSAT score scale has 21 score points while the scale of the other selection instrument, the GPA, has 2000 score points in theory and about 1000 score points in practice (0.00–20.00 and 10.00–20.00, respectively). Given the tremendous amount of measurement precision that is attached to the GPA through the use of this many score points, it seems reasonable to attach more measurement precision to the SweSAT than is done today. Certainly, the GPA is a measure that is based on multiple assessments over a long period of time and thus it should be a more reliable measure than SweSAT scores (though, I am not saying it is), but the large difference in the number of score points can hardly be justified on that account. A related issue is what the labels of the score points should be. The current 0.0–2.0 scale must be far from optimal for facilitating appropriate interpretations. For one thing, receiving a score of 0.0 even though you actually got up to one fourth of the items right does not make sense. Why not have a lowest scale score that tells you something other than “You are not only a 0, but you are a 0.0, a zero to the first decimal”? If you look at the other major admission tests (e.g., the SAT, the ACT, and the PET) the score scales for these tests do not have a zero score, and they have more score points as well. Another thing is that the 0.0–2.0 scale is illusively similar to the GPA 0.00–20.00 scale, which can result in inappropriate comparisons between SweSAT scores and the GPA. To conclude, the scaling issues related to the SweSAT deserve further examinations.

Finally, the feasibility of using IRT in the development of the SweSAT should be examined further. If an appropriate item response model can be fitted to the SweSAT scores there are many potential benefits. Equating, scaling, DIF analyses, and score reporting are examples of procedures that can capitalize on the advantages of IRT. Other types of investigations that should be performed more frequently on the SweSAT are generalizability studies and DIF studies. Also, there is a need for predictive validity studies that address many of the methodological issues surrounding these studies, and that use data from areas of study that have not been examined previously. By continuously collecting validity evidence the test developers can hopefully maintain the test-takers’ right to receive reliable scores from which valid interpretations and uses can be derived.
References


54
Appendix

Statistical notation

Here follows a list of the statistical notation used in the papers. After each entry, it is indicated in which papers it occurs (I–IV = papers I–IV).

- \(a\) anchor-test score (II, III)
- \(d\) effect size (Cohen’s \(d\)) (II)
- \(df\) degrees of freedom (III)
- \(E\) estimated score (III), expectancy operator (IV)
- \(g\) test form (II)
- \(h\) negative hypergeometric distribution function operator (IV)
- \(i\) item (III)
- \(j\) raw score interval (II)
- \(k\) number of scale score points (III)
- \(M\) mean (II)
- \(n\) number of students (I)
- \(O\) observed score (III)
- \(p\) probability density function operator (II)
- \(P\) cumulative distribution function operator (II)
- \(q\) probability density function operator (II)
- \(Q\) cumulative distribution function operator (II)
- \(r\) Pearson correlation (I)
- \(r_{pb}\) point-biserial correlation (III)
- \(s\) scale score (II)
- \(S\) parameter for postsmoothing of equipercentile equivalents (III)
- \(t\) \(t\) statistic (in \(t\) tests) (III)
- \(X\) total score (IV)
- \(\bar{x}\) mean (of \(x\)) (III)
- \(\alpha_{21}\) Kuder-Richardson formula 21 reliability coefficient (III)
- \(\varepsilon\) effective weight of a test score (IV)
$\mu$ mean (II, III)

$\sigma$ standard deviation (II, III, IV)

$\rho$ correlation coefficient (IV)

$\rho^2(X_t, X)$ reliability (of a score $X$) (IV)

$\Psi$ proportional reduction in mean squared error (IV)

**Abbreviations and Acronyms**

AAMC Association for American Medical Colleges

ACT Formerly an acronym for American College Testing

AERA American Educational Research Association

APA American Psychological Association

CAT computerized adaptive test

CBT computer-based test

CTT classical test theory

DS Data Sufficiency (a SweSAT subtest; NOG in Swedish)

DTM Diagrams, Tables, and Maps (a SweSAT subtest; DTK in Swedish)

ERC English Reading Comprehension (a SweSAT subtest; ELF in Swedish)

ETS Educational Testing Service

GMAC Graduate Management Admission Council

GMAT Graduate Management Admission Test

GPA grade point average (it usually indicates college GPA, but here it is used to indicate upper-secondary school GPA)

GRE Graduate Record Examination

HSGPA high school grade point average

IRT item response theory
KR-20  Kuder-Richardson formula 20 reliability
LSAC  Law School Admission Council
LSAT  Law School Admission Test
MCAT  Medical College Admission Test
MSD  mean signed difference
MSE  mean squared error
NAHE  the Swedish National Agency for Higher Education (Högskoleverket in Swedish)
NCME  National Council on Measurement in Education
NEAT  nonequivalent groups anchor test equating design (also referred to as the common-item nonequivalent groups equating design)
NITE  National Institute for Testing and Evaluation
PET  Psychometric Entrance Test
PRMSE  proportional reduction in mean squared error
READ  Swedish Reading Comprehension (a SweSAT subtest; LÄS in Swedish)
RMSD  root mean square difference
S+WE  SweSAT score plus additional credits for work experience
SAT  Former acronym for Scholastic Aptitude Test. The SAT program consists of both the SAT Reasoning Test and the SAT Subject Tests, but in this thesis “SAT” refers to the SAT Reasoning Test.
SD  standard deviation
SweSAT  Swedish Scholastic Assessment Test (Högskoleprovet in Swedish)
UC  University of California
VALUTA  Swedish acronym for the research project "Validering av den högre utbildningens antagningssystem" (Validation of the higher education admission system)

WORD  Vocabulary (a SweSAT subtest; ORD in Swedish)