VALIDITY ISSUES CONCERNING REPEATED TEST TAKING OF THE SWESAT

Birgitta Törnkvist
Widar Henriksson

EM No 56, 2006

ISSN 1103-2685

ABSTRACT

The main purpose of this study is to integrate and discuss results from studies that focus on the effects of repeated test taking of the Swe-SAT. Messick's four-faceted model of validity is used as an integrating and an analytic tool. Another purpose is to use results from supplementary education as a reference in this integration.

The summarized conclusion is that the existing rule concerning repeated test taking for the Swedish Scholastic Assessment Test (Swe-SAT) is relevant. It is relevant as the test taker has a chance to obtain a good estimate of his or her knowledge and ability. The unintended social consequences are reduced if actions are taken to motivate test takers to repeat the test.

The conclusion is quite the opposite for supplementary education. A change in the rule that eliminates the usage of supplementary grades in the selection procedure will result in higher validity.

The conclusion is also that the Messick model is a very useful tool for validation. When applying the Messick model, the aim and the direction of the process of validation will be systemized and optimized as well as nuanced.

CONTENTS

INTRODUCTION

The quality of the selection to higher education has been evaluated mainly thorough studies of the predictive validity of the selection instruments, i.e. by examining the correlation between the admitted students' results on the chosen selection instrument and the indicator of the subsequent academic performance. It is important to observe that, from a perspective of the changed concept of validity (see e.g. Cronbach, 1988; Messick, 1989a; Shepard, 1993; Wolming, 2000; Nyström, 2004; Wikström, 2005b: Eklöf, 2006), the predictive validity of the selection instruments is a critical factor, but not in itself sufficient for a description of the validity of a certain instrument as well as the validity for the whole procedure for selection to higher education. Thus, the conclusion is that predictive validity is regarded as only one of several critical factors for the validity of a selection instrument. Other questions connected with the concept of validity must also be taken into consideration. For example; what are the rules for a certain instrument when it is used in the selection procedure? What is the influence of these rules on the behaviour of the applicant? What are the consequences of these rules in relation to a particular selection instrument? Are certain test takers favoured as a consequence of a combined effect of these rules? To sum up: these questions indicate that the concept of validity has become much more complex and, consequently, a multi-faceted concept.

Selection to higher education in Sweden

The main instruments in the present system for selection to higher education in Sweden are grades from upper secondary school, scores from the Swedish Scholastic Assessment Test (SweSAT) and, to a lesser extent, scores from various types of course-specific selection instruments for programmes like architecture, journalism and medicine (Högskoleverket, 1997a).

From a historical perspective the selection of students to universities and colleges in Sweden has, since the early 1940s, mainly been based on grade point average (GPA) obtained in upper secondary school. In the middle of the 1960s, however, a debate about the system of selection to higher education was initiated. A central ingredient in this debate was the ambition to broaden the population of students by in-

cluding new categories. To realise this ambition, it was proposed that a Swedish test battery, like e.g. the SAT in the US, should be developed. Hence applicants without an upper secondary school certificate would be able to qualify by way of a test (SOU 1968:25; SOU 1974:71; Stage, 2004).

Also discussed was the way in which an admission test should be used in the selection to higher education, i.e. whether it should be offered to all applicants or restricted to certain categories of students. When the first version of the SweSAT was administered in 1977, the decision was that the test could be taken and used only by applicants who were at least 25 years old and had at least four years of work experience. However, this restriction was dropped about 15 years later (1991). Since then, the SweSAT may be used by all applicants to Swedish universities and colleges (Henrysson, 1992; Wedman, 1992, 2002). The admission to higher education in Sweden has a number of goals to fulfil. Besides, as mentioned earlier, broadening the population of students by including new categories there is also an ambition not to treat any group unfairly, i.e. to not give any group of applicants advantages that others do not have (SOU 2004:29).

The present selection system, which is mainly based on grades from upper secondary school and scores from the SweSAT, is also influenced by certain rules that are valid for each selection instrument when it is used in the selection process. With reference to GPA, there is an opportunity for the students to increase their chances of being admitted to higher education by improving their upper secondary school-leaving grades by means of supplementary examination after leaving secondary school. With reference to the score from the SweSAT, there is also a possibility of raising the score by repeated test taking. Thus, from a validity perspective, it is necessary to analyze the effects of repeated testing as well as the effects of supplementary examination. The main focus of this article is on repeated test taking.

This article is structured as follows: first, the effects of repeated test taking are related, on a general level, to the test in itself and the test taker but also, on a more specific level, to the rules for selection. Analogously, the effects of supplementary examination of grades from upper secondary school are related to the student and the rules for selection. Second, the theoretical concept of validity will be addressed. This includes a brief description of Messick's (1989a) view of valid-

ity. Then the main results from studies of repeated test taking of the SweSAT will be summarized. As a basis for reference, these results are also related to the effects of supplementary examination of GPA. Finally, the results will be discussed and some conclusions will be drawn from a perspective of validity.

Effects of repeated test taking with reference to tests in general

In a literature review about practice and coaching, Henriksson (1981) summarized the main findings so far by concluding that the effects of repeated test taking (practice) are greater when a test has a speed component than when there is no time limit, i.e. the need to respond quickly is susceptible to practice. This means that the gain from repeated test taking on parallel test versions (cf. the SweSAT) is the establishment of a rational time-use strategy. Another result was that the effects of repeated test taking tend to be greater on non-verbal tests than on verbal tests such as vocabulary tests and verbal reasoning tests. Still another conclusion was that the gain from repeated test taking is greater on tests with a complex item format as compared to tests with a simple instruction and a simple format. It can also be stated that standardised tests, as well as the process of test construction for this type of tests, have been continuously developed on the basis of knowledge about these findings. In this context it is also relevant to mention that all possible actions are taken in order to avoid this kind of score gain for the SweSAT (Stage, 2004).

Effects of repeated test taking with reference to the student

Henriksson (1981) made a main distinction between repeated test taking with and without support. Test taking without support was called practice, while test taking with support was called instruction or coaching. In an actual situation there is of course no clear-cut border between the two, but the theoretical view is that practice implies no support, i.e. any special instruction regarding strategies for test taking or teacher support. Based on this distinction one main result supported by literature was that the more able a test taker is, the more he or she will gain and benefit from unsupported practice. Another main result was that the largest effects of practice occur when the test taker has little or no previous experience of test taking, i.e. when he or she is completely unfamiliar with tests and test situations in general. Knowl-

edge about these findings has of course influenced the potential test takers, and the consequence is that they nowadays usually are fairly familiar with the requirements of each test as well as the whole test situation, i.e. they are testwise in a general and practical sense (Millman, Bishop & Ebel, 1965; Rogers & Bateson, 1992; Gregory, 2004). These findings have also influenced those responsible for existing tests, test administration, and the use of test scores. Thus, different strategies are applied in order to reduce or eliminate the effect of deficiency in this respect. All students who enter for the SweSAT get an extensive and detailed description of the test as well as all other relevant circumstances in relation to the test administration (Högskoleverket, 2006). It can also be added that each version of the SweSAT is public as soon it has been administered, i.e. a student can examine and practice on earlier versions of the test. In spite of these circumstances repeated test taking is rather widespread. About one third of the test takers are repeaters (Stage & Ögren, 2005).

Repeated test taking with reference to the rules of selection

When SweSAT scores are being used in the process of selection to higher education in Sweden, certain rules apply. These are:

- An obtained SweSAT[1] score is valid for five years
- If a test taker has more than one valid SweSAT score, the best obtained score (normed score) is used in the selection procedure
- An applicant is selected *either* on the basis of the SweSAT score *or* on the basis of GPA
- If an applicant has *both* a valid SweSAT score *and* a valid GPA, the best result is used in the selection procedure

Supplementary examination of grade point average

A new grading system, a goal-related grading system, was introduced in upper secondary school in Sweden in 1994 and 1997 was the first time students had this type of grades as a basis for selection to universities and colleges. According to the rules for selection these students

[1] SweSAT of today consists of 122 items and the raw score distribution has a range 0-122. Raw score is, by equalization, transformed to a normed score with a range 0.0-2.0. It is the normed score that is used in the process of selection.

are allowed to raise their grades by means of supplementation (Högskoleverket, 1997b; Löfgren, 2003). Thus, the grades which are used for selection to higher education are either school-leaving grades or grades from subsequently supplemented grades. If the supplemented grades are higher than the school-leaving grades, it is the former grades that are valid in the selection procedure.

Analyses of the goal-related grading system in upper secondary school also indicates that there has been a relatively constant annual increase in GPA. This means that students with old school-leaving grades, compared with students with GPA at a later point of time, are disadvantaged in the process of selection to higher education (Cliffordson, 2004a). Based on data over a six-year period Wikström (2005a) came to the same conclusion, i.e., GPA has increased every year since the new grading system was introduced. The conclusion was also that this increase cannot be explained by improved performance, selection effects or strategic course choices. This fact, i.e. grade inflation, opens for a scenario of supplementary completion for students with an old GPA.

Effects of supplementary examination with reference to the student

In a study of the effects of supplementary examination of grades Löfgren (2004) focused on those students who had graduated from upper secondary education 1997-2001. He found that about 30 percent had studied in adult upper-secondary school. About one third of those had boosted their grades, and about 40 percent had applied to higher education. The results also indicated that there are more women than men among these applicants, and that they had parents with a high educational level. Thus, a comparatively large number of applicants have raised their grade by means of supplementation (Högskoleverket, 2004).

Cliffordson (2004b) investigated the impact of an increase of GPA by supplementary examination by relating the increase to an indicator of study success (credit points) during the first year of university engineering and medical programmes. The study included about 14,000 students and the design was based on a comparison of those students who had supplemented their GPA and those who had not. The main finding was that an observed grade increase did not correspond to an

increase in study success. Thus, the predictive validity of grade increases was zero.

Supplementary examination with reference to the rules of selection

When GPA obtained in upper secondary school is used in the process of selection to higher education in Sweden, certain rules apply. These are:

- An applicant is selected *either* on the basis of the SweSAT score *or* on the basis of GPA
- If an applicant have supplemented grades that are higher than the leaving grades, it is the former grades that are valid in the selection procedure
- If an applicant has *both* a valid SweSAT score *and* a valid GPA, the best result is used in the selection procedure

THE CONCEPT OF VALIDITY

The concept of validity has long been the subject of debate and change. From the mid-50s to the mid-80s, it was customary to divide validity into the following types depending on the purpose of the test: content validity, criterion-related validity, and construct validity (Anastasi, 1982). The choice of selection instruments has usually focused on the predictive validity of the measuring instrument, i.e. its ability to predict the academic success of the applicants. But, as Henriksson & Wedman (1992) stated: *predictive validity is certainly not a validity concept of its own (p 18)*. Rather, as Cronbach (1971, 1988) and more explicitly Messick (1987, 1989b, 1995) have made clear, it should be seen as one of many aspects of a more general concept of construct validity. Thus, the strict division into different types has now been replaced by the general opinion that validity cannot be divided into these categories. It is generally said that there are very few occasions when only one of these types is used, and that normally all aspects are present in a validation.

When focusing on instruments for selection to higher education it is also important to point out that validation is a process, starting from how the instrument is developed and constructed to how it is inter-

preted and used. And, as Stobart (2001) emphasized, all links in the chain must be validated in order to get indications of the quality of an instrument. It is also important to pay attention to the fact that validation, in the more comprehensive definition of this concept, is a never-ending process since society and the context in which the instrument is administered and used changes continuously. Therefore, the conclusion is that the more comprehensive definition of validity implies that validity is neither a fixed property nor definitively determined at a certain point of time.

Messick (1989a) described this comprehensive view of validity in his demonstration of how the concept of validity could be divided into two aspects. One constitutes the *source of justification of the testing* and the other aspect constitutes the *function or outcome of the testing*. The former aspect is based on either evidence or consequence and the latter aspect is based on either interpretation or use. When these aspects are related the result is a two-by-two matrix which, according to Messick, includes all aspects of validity. The conclusion is that a satisfactory validation should answer well to all four facets of validity. In spite of the theoretical idea that these two aspects could be considered to be unrelated, Messick also draws attention to the fact that they are not only interlinked but also overlapping. He expressed the advantages of his model in the following way:

> One advantage of this progressive matrix formulation is that construct validity appears in every cell, thereby highlighting its pervasive and overarching nature. Furthermore, evidence of the relevance and utility of test scores in specific applied settings, and evaluation of the social consequences of test use as well as of the value implications of test interpretation, all contribute in important ways to the construct validity of score meaning. This makes it clear that, in the generalized sense, construct validity may ultimately be taken as the whole of validity in the final analysis (p 21).

	Test interpretation	*Test use*
Evidential basis	*Construct validity*	*Construct validity + Relevance/utility*
Consequential basis	*Value implications*	*Social consequences*

Figure 1. Messick's facets of validity framework (Messick, 1989a, p 20).

The first facet (*Construct validity*) focuses on whether a test actually measures the quality or ability it is intended to measure. In this context Messick points to two types of "threats" that can affect construct validity. The first threat is under-representation of the construct of interest. The instrument cannot cover all the important aspects and dimensions of the quality or ability that the test intends to measure. The second threat refers to consequences of over-representation, i.e. when the instrument is also measuring irrelevant aspects. In this context, when the main focus is repeated test taking of SweSAT and supplementary education for GPA, the question of "threats" is related to the validity of gains, as a function of repeated test taking, and the validity of gains in GPA as a function of supplementary examination.

The second facet (*Construct validity + Relevance/utility*) focuses, not only on the construct validity of the instrument, but also on the evidence supporting a certain use of the instrument and the relevance and utility of certain rules that are related to this use. Thus, if an instrument that is used in a selection process has certain rules, the question of relevance and utility of these rules is a matter of validity.

The third facet (*Value implications*) points to the implied values that can be associated with the qualities and abilities that an instrument is intended to measure. According to Messick the designations of the variables that the instrument intends to measure are very important. He stated that when chosing a designation one should strive for logical consistency between the significance of the quality and the connotations that the designation could have for the interpretation of the result. Thus, the value implications of supplementary examination for GPA and repeated test taking of SweSAT is a matter of validity and must be examined.

The last facet (*Social consequences*) represents the potential consequences of the use of an instrument in a certain situation for all parties involved. The question whether the rules for an instrument, in this case supplementary examination for GPA and repeated test taking of SweSAT, should consider both the intended and the unintended consequences. If these rules for SweSAT and GPA indicate differences between certain groups, e.g. males/females or social groups, this may have social consequences for these groups. There may be many causes for these differences but Messick pointed out that the consequences for the groups may differ and therefore it is a question of validity. It is also worth mentioning that this last facet has caused some controversy, and there is no complete unanimity about the relevance of this aspect (Popham, 1997; Kane, 2004). But, given this fact, the conclusion is that the facet *social consequences* is a very important aspect of validity in this context. The main argument for this conclusion is based on the fact that the effects of certain rules for SweSAT and GPA, may differ between certain groups.

Repeated test taking of the SweSAT and supplementary examination of GPA from a perspective of validity

The Swedish model for selection to higher education can be described by a simple model where two different components are included. Applicants are ranked on the basis of their SweSAT scores and their GPA from upper secondary school. The main idea is that the students who are admitted on the basis of these instruments are those who will show good academic performance.

From a measurement perspective it is also of interest to state that scores from the SweSAT can be categorized as norm-referenced and that GPA can be categorized as criterion-referenced. The former measurement is an example of a relative approach, i.e., the test taker's standardized SweSAT-score is compared with the performance of other test takers in a norm group. The model for selection to higher education is based on ranking the applicants and this condition is in line with the idea of SweSAT as a norm-referenced measurement.

Grades in upper secondary school are an example of an absolute approach, i.e. grades are indicators of a student's performance in relation to a defined criterion. This means that the application of grade point average in the model for selection to higher education is a little more

problematic. Grades are, on the one hand, used for eligibility to higher education. This means that a certain type and level of previous knowledge is required for a certain study programme at the university. This application of grades is in accordance with grades as indicators of performance in relation to educational objectives. But, on the other hand, grades are also used as a basis for ranking of applicants and this application is not in accordance with the absolute meaning of a criterion-referenced measurement. However, in educational assessment it is not uncommon that a criterion-referenced measurement has multiple purposes, partly as information about how well a student has reached goals in upper secondary school, partly as a basis for ranking of applicants with reference to expected success in higher education (Payne, 1997).

Besides the fact that the SweSAT score is an example of a normed referenced measurement, and that GPA is also used as a normed referenced measurement, these two instruments for selection to higher education in Sweden also have another property in common. For the SweSAT this condition is the rule that allows for repeated test taking and the corresponding condition for the GPA is the rule that allows for supplementary examination.

The presence of these two conditions, and the fact that results from each condition has consequences for process of selection to higher education, makes repeated test taking and supplementary education a matter of validity.

PURPOSE

The main purpose of this study is to integrate and discuss results from studies with focus on the effects of repeated test taking of the SweSAT. Messick's four faceted model of validity is used as an integrating and an analytic tool. Another purpose is to use results from supplementary education as a reference in this integration.

STUDIES ABOUT THE EFFECT OF REPEATED TEST TAKING OF THE SWESAT

Main results from Swedish studies about the effect of repeated test taking are summarized below. The main part of these studies are based on a design that allows for describing longitudinal effects, i.e. in this context score changes over four test administrations. The reason for selecting a two-year-period is that it allows for a description of effects of repeated test taking (practice) at the same time as the possibilities of score changes as a function of true changes of test takers' ability are controlled for, or at least minimized.

These studies have been focused on the whole population of test takers at a certain test administration. The test takers are grouped according to prior experience from taking the SweSAT into four categories: one, two, three or four test administrations. These administrations are within a two-year-period.

The results from five studies (Henriksson, 1991; Henriksson & Wedman, 1993; Henriksson, 1995; Henriksson & Törnkvist, 2002; Törnkvist & Henriksson, 2004a) are summarized in Table 1.

The designations (2–4) in Table 1 refer to the number of SweSATs taken during the 2-year period. The figure 2 means, with reference for example to 86B[2] (Henriksson, 1990), that 86B was the *second* SweSAT for the test taker, 3 means that 86B was the *third* SweSAT, and 4 means that 86B was the *fourth* SweSAT. Test takers in that last-mentioned subpopulation hade taken all SweSATs that were administered during the observed 2-year period (86B, 86A, 85B, 85A). Analogously, the 2-year period for 93B is 93B, 93A, 92B, 92A (Henriksson, 1995) etc.

[2] SweSAT is administered twice a year, in spring and in autumn. The spring administration is labelled A and the autumn administration is labelled B. The number 86 refers to the year 1986.

Table 1. Mean difference (latest compared to the highest earlier) and standard deviation for differences in normed scores (M_d, s_d) for test takers who had taken the SweSAT 2, 3 or 4 times. Summary of five studies (86B, 91B, 93B, 97B and 02B).

Difference in normed score	Number of SweSATs taken					
	2		3		4	
	M_d	s_d	M_d	s_d	M_d	s_d
86B	0.055	0.20	-0.011	0.18	-0.036	0.21
91B	0.083	0.20	0.009	0.18	-0.034	0.17
93B	0.111	0.21	0.031	0.19	-0.023	0.18
97B	0.093	0.21	0.010	0.20	-0.028	0.19
02B	0.082	0.21	-0.002	0.19	-0.035	0.18

The summarized results from five studies of the effects of repeated test taking indicates that the highest gain of repeated test taking is from the first to the second test occasion (Table 1). The mean differences of normed SweSAT scores (M_d) shown in Table 1 are the mean differences between the second and the first score, between the third and the highest of the earlier scores and between the fourth score and the highest of the earlier scores.

The summarized conclusion concerning differences between males and females is that the mean differences in normed scores from the first to the second time are 0.08 - 0.09 for males and 0.08 - 0.10 for females, i.e. the gain is marginally larger for women. The observation is also that the mean normed score increases with the number of SweSATs, but the mean gain decreases and is about zero at the third SweSAT.

The main part of the studies about repeated test taking are based on a design that allows for describing longitudinal effects, i.e. in this context score changes over four test administrations. This design can also be used to describe the fact that an applicant, who has more than one valid SweSAT score, is allowed to use his or her highest score in the selection procedure, even though the obtained score at the latest SweSAT taken is lower. It is, according to the rules, the highest valid

score that is used in the selection process. This score, i.e. the score that is used in the selection, is labelled x_{max} for test takers who have more than one valid score. In Table 2 a distinction is made between test takers who have increased their score and those who have not (Törnkvist & Henriksson, 2004a). The label x_{max} in Table 2 also stands for those test takers who did not increase their score at the next test occasion.

Table 2. Mean (M), standard deviation (s), total number (N) and percentage (%) for the best previous obtained result (X_{max}) and the result at the latest test occasion (X_2, X_3, X_4) for test takers who have taken the SweSAT 2, 3 and 4 times respectively.

Variable	Number of SweSATs					
	2 (n=4,096)		3 (n=1,086)		4 (n=204)	
	X_{max}	X_2	X_{max}	X_3	X_{max}	X_4
M	0.94	0.92	1.07	1.03	1.12	1.09
s	0.40	0.42	0.36	0.40	0.40	0.43
N	1,818	4,096	664	1,086	148	204
%	44.4	100	61.1	100	72.6	100

From Table 2 we can see, for example, that 44.4% of the test takers, who had taken the test twice (two valid SweSAT scores), obtained a higher score on the first test occasion, i.e., 55.6% had a higher score (as compared to their first score) at the second test occasion. The mean score for the former group was 0.94 and the mean for the total group at the second test occasion was 0.92. Analogously, for those test takers who had four valid SweSAT scores, 27.4% of the test takers obtained a higher (and 72.6% a lower or equal) score at the fourth test occasion.

A more detailed description of the test takers in each category (2, 3 and 4 SweSATs) is presented in Table 3.

The relation between Table 2 and Table 3 is that the test takers in each category are divided into subgroups with reference to when they obtained their best result. For test takers in category “2 SweSATs” roughly the same information is presented in Table 2, i.e., 55.6% (n=2,278) obtained a higher score and 44.4% obtained the same or a lower score at the second test occasion. Analogously, test takers with three valid SweSAT scores are distributed in the following way:

20.5% obtained their best result at the first test occasion, 40.6% obtained their best result at the second test occasion and 38.9% obtained their best result at the third test occasion.

Table 3. Total number (N) and percentage (%) for the best obtained results, at a certain test occasion, for test takers who have taken the SweSAT 2, 3 and 4 times respectively.

	Number of SweSATs								
	2 (n=4,096)		3 (n=1,086)			4 (n=204)			
	Best result		Best result			Best result			
Variable	1	2	1	2	3	1	2	3	4
N	1,818	2,278	223	441	422	18	44	86	56
%	44.4	55.6	20.5	40.6	38.9	8.8	21.6	42.2	27.4

The overall score-gains as a function of repeated test taking can also be described more in detail with reference to increase and decrease between the latest score as compared to the best of earlier scores in terms of tenths of normed score. There are great variations in gains on the individual level which is described in Table 4 below (Henriksson & Törnkvist, 2002; Törnkvist & Henriksson, 2004a). The lines in this table divide gains and losses with reference to zero gain (=0). Thus, the difference in normed score for test takers who had taken the SweSAT 2, 3, or 4 times was divided into three categories (negative, zero and positive) difference and the percentage was calculated for each category. This operation is carried out mainly for descriptive reasons.

The observation that can be made from Table 4 is that 56-58% of the test takers increased their normed score if results from the first and second SweSAT are compared. The proportions of test takers that get a higher score at the third and forth test occasion are then diminishing gradually (39-42% and 27-34% respectively). Thus, the probability of increasing the score by repeating the SweSAT can be estimated from Table 4. For example, the estimated probability for increasing the score by repeating SweSAT a third time is about 0.27-0.34.

Table 4. The distribution of differences in normed scores (latest compared to the highest earlier) for test takers who have taken the SweSAT 2, 3 or 4 times. Frequencies and percentage (%) for 97B (N= 17,863) and 02B (N=5,386). The number of test takers in each score category for 02B are printed in italics and within parenthesis.

Difference in normed scores	Number of SweSATs taken					
	2		3		4	
-1.0 - -0.6	12 *(3)*		11 (*4*)		2 (2)	
-0.5	44 (19)		25 (*9*)		5 (2)	
-0.4	142 (*39*)	23.5% 26.5%	92 (31)	37.2% 39.4%	32 (*4*)	47.1% *46.1%*
-0.3	336 (148)		252 (*54*)		78 (9)	
-0.2	882 (337)		414 (117)		165 (*28*)	
-0.1	1,575 (*540*)		723 (213)		232 (*49*)	
0	2,364 (*732*)	18.6% *17.9%*	853 (236)	21.0% *21.7%*	208 (*54*)	19.1% *26.5%*
0.1	2,522 (*785*)		756 (194)		201 (*33*)	
0.2	2,113 (*660*)		550 (135)		86 (12)	
0.3	1,372 (*429*)		237 (*66*)		55 (*6*)	
0.4	804 (238)		109 (*21*)		23 (*3*)	
0.5	343 (108)	57.8% 55.6%	33 (*4*)	41.8% *38.9%*	1(-)	33.8% *27.4%*
.6	128 (*34*)		7 (1)		2 (-)	
0.7	47 (17)		2 (1)		1 (1)	
0.8 – 1.5	18 (*6*)		6 (-)		- (-)	
97B	12,702		4,070		1,091	
02B	4,096		1,086		204	

Repeated test taking is a matter of self-selection. A test taker who, for some reason, selects to repeat the SweSAT has a higher normed score on the first test occasion as compared to others who took the test on the same occasion.

When comparing mean normed score at the first test occasion for repeaters and non-repeaters the conclusion, in all studies reported on, is that the repeaters have a higher mean. This is illustrated in Table 5 with reference to Henriksson & Törnkvist (2002) and Törnkvist & Henriksson (2004a). Data from the latter study are printed in italics and within parenthesis.

Table 5. Mean normed score and standard deviation (M, s) on the 1st, 2nd, 3rd and 4th test occasion for test takers who had taken the SweSAT 1 (n=29,572, *n=14,959*), 2 (n=12,702, *n=4,096*), 3 (n=4,070, *n=1,086*), or 4 times (n=1,091, *n=204*).

Number of Swe-SATs	Test occasion							
	1		2		3		4	
	M	S	M	S	M	s	M	S
1	0.83 *(0.81)*	0.45 *(0.44)*						
2	0.86 *(0.84*)	0.43 *(0.43)*	0.95 *(0.92)*	0.42 *(0.42)*				
3	0.91 *(0.88)*	0.41 *(0.40)*	0.94 *(0.99)*	0.41 *(0.40)*	1.05 *(1.03)*	0.41 *(0.40)*		
4	0.94 *(0.87)*	0.40 *(0.42)*	1.07 *(1.00)*	0.38 *(0.42)*	1.13 *(1.08)*	0.40 *(0.42)*	1.16 *(1.09)*	0.40 *(0.43)*

The consequence is that the description of the effects of repeated test taking, based on the whole population, can be invalid because of self-selection. In order to control for the effects of self-selection the following strategy is used in the 97B (Henriksson & Törnkvist, 2002) and the 02B study (Törnkvist & Henriksson, 2004a). The strategy that is used is to select a sample from the population that has taken the SweSAT four times (population 4) with the same distribution of normed scores at the first test occasion as the population that has taken the SweSAT for the first time (population 1).

This is illustrated by reference to Henriksson & Törnkvist (2002). In order to separate the effect of self-selection from the effect of repeated test taking, a sample (n=600) was selected from population 4 with the same distribution of the normed scores at the first test occasion as in population 1 (Table 6).

Table 6. The distribution of normed scores, mean (M) and standard deviation (s) regarding the first SweSAT taken by population 1 (n=29,572), population 4 (n=1,091), and the sample (n=600).

Normed score	Population 1 (%)	Population 4 (%)	Sample	
			(%)	n
0.00 – 0.49	20.8	11.5	20.8	125
0.50 – 0.99	41.3	38.8	41.3	248
1.00 – 1.49	27.5	39.5	27.5	165
1.50 – 2.00	10.4	10.2	10.3	62
Total (%)	100	100	100	600
M	0.83	0.94	0.84	
s	0.45	0.40	0.43	

The normed scores for population 1 and the sample have approximately the same distribution, mean (M), and standard deviation (s). The mean and the standard deviation for population 4 on the first test occasion are 0.94 and 0.40 respectively, and the corresponding data of the sample are 0.84 and 0.43.

But, as a consequence of the sampling strategy the distribution of sex, age and education will be different for the sample compared to population 4. Therefore the analysis has to be based on a model that controls for these variables. A comparison between the effects of repeated test taking in the sample and in population 4 indicates that the tendencies are the same for males and females. However, there are significant differences within the sample between males' and females' mean normed scores, i.e. men have a higher score at all test administrations. The summarized conclusion is that, even when controlling for self-selection, the benefit from repeated test taking occurs between the first and the second test administration. Another observation is that the mean scores increase with the number of tests taken, but the increase gradually declines.

In order to describe the effects of repeated test taking for test takers from different social groups a study (Törnkvist & Henriksson, 2004b) has also been carried out. The design of this study was similar to the other studies reported, i.e. a 2-year period was analyzed. There is, however, a difference in the definition of population. Information about social group is not available in the databases for the SweSAT.

Therefore the data base for the VALUTA-project[3] was taken as a point of departure in this study. The variable social group, i.e. socio-economic background for the test takers, was categorized into three socio-economic groups on the basis of the parents' education and vocation: upper middle class (social group I), lower middle class (social group II) and working class (social group III). The number of test takers at the selected SweSAT administration (00B) was 20,415. This population was divided into four subpopulations labelled 1, 2, 3, and 4. The designations (1–4) refer to the number of consecutive SweSATs taken during the 2-year period (00B-99A).

The results in the Törnkvist & Henriksson (2004b) study indicated that the willingness to repeat the SweSAT differed between sexes and social groups (Table 7).

Table 7. The social group- and sex distribution for first timers (population 1) and repeaters (population 2-4) and for the total VALUTA- population, in per cent (%). Number of test takers in each population (N) and per cent of females (%).

Population	Social group I (%)	Social group II (%)	Social group III (%)	Number of test takers	Females (%)
1	30	49	21	14,780	55
2	35	48	17	3,874	51
3	41	44	15	1,339	43
4	43	45	12	422	40
Total (VALUTA)	21	48	31	1,266,598	49

In the total VALUTA-population, persons born in the period 1972 to 1984, and living in Sweden at the age of 16 years, the percentage of females was 49 and 21% of the population belonged to social group I, 48% to social group II and 31 % to social group III. Table 7 indicates that a higher proportion of females (55%) took one SweSAT, but on the other hand, also that males repeated SweSAT more often than females. A higher proportion of test takers from social group I, compared with social group III, took the SweSAT and they also repeated the SweSAT more often.

[3] The database within the VALUTA project consists of individuals born in 1972 to 1984 who lives in Sweden in the age of 16 years.

Given the fact that the repeaters have a higher mean than the non-repeaters the question is whether the high mean scores for repeaters can be explained by the fact that males have higher mean normed scores than females and that test takers from social group I have higher mean normed scores than social group III?

Törnkvist & Henriksson (2004b) focused on these questions and the obtained results indicated that the mean normed scores for repeaters at the first SweSAT are higher than for the non-repeaters, even when controlled for social group and sex. Thus, the conclusion is that social group and sex may only explain the self-selection procedure to a minor extent.

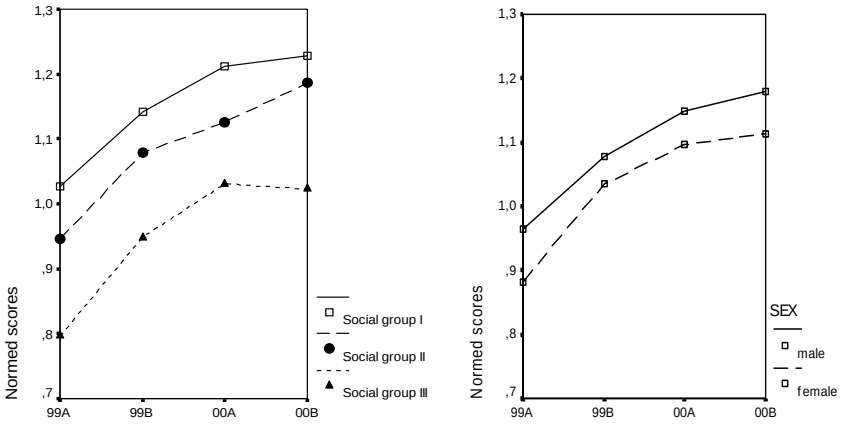

Figure 2. Mean normed SweSAT score for different social groups, when controlled for sex, and mean normed SweSAT score for males and females respectively, when controlled for social background. Evaluated at age 20 years.

When controlling for education and age (Henriksson & Törnkvist, 2002; Törnkvist & Henriksson, 2004a) or social group and age (Törnkvist & Henriksson, 2004b), the effects of repeated test taking were the same for females and males (no significant differences) and the gain in score is highest at the first repeat, i.e., the effect is about 0.08–0.17 scores, in average, from the first to the second test occasion (Figure 2 and Table 8).

For those test takers who repeated the SweSAT three times, the difference in mean normed score between social groups I and III was about 0.2 and between social groups I and II only about 0.1 scores, at

the first test occasion and when controlling for sex and age. For all social groups the highest gain was from the first to the second test occasion. Another finding was that the difference between social group I and III is higher than the difference between sexes.

Table 8. Mean normed scores for males (m) and females (f) for repeaters at each test occasion evaluated at the mean age when controlled for education or social group. Mean normed scores for social group I – III for each test occasion.

Test administration	First test occasion	Second test occasion	Third test occasion	Fourth test occasion
	m f	m f	m f	m f
96A – 97B	0.87, 0.73	1.02, 0.90	1.04, 0.93	1.10, 0.99
01B – 02B	0.90, 0.82	0.98, 0.91	1.03, 0.94	
99A - 00B	0.96, 0.88	1.08, 1.04	1.15, 1.10	1.18, 1.11
Social group (99A-00B)	All	All	All	All
I	1.03	1.14	1.21	1.23
II	0.95	1.08	1.16	1.19
III	0.80	0.95	1.03	1.02
Difference I – III	0.23	0.19	0.18	0.21

Cliffordson (2004b) examined the effects of repeated test taking of the SweSAT with focus on test takers from three cohorts with two and three test scores, i.e. test takers that have repeated the SweSAT once and twice. By using different regression models, a distinction was made between the effects of practice and the effects of growth. The effects of self-selection were also estimated and the conclusion was that test takers with higher grades tended to be younger when taking the first SweSAT. The main conclusions were, on the one hand that there are score gains as a function of practice from the first to the second SweSAT and, on the other hand, that there are effects of growth as well. Another observation was that the magnitude of gains related to growth was equal between the first and the second, and between the second and the third SweSAT taken.

The summarized conclusion, concerning the effects of repeated test taking of the SweSAT, is that the highest gain in score is from the first to the second test occasion. When controlling for age and level of education, males and females increase their average score with 0.1 normed score from the first to the second test occasion. The gain, as a

function of repeated test taking, is the same for males and females. When controlling for age and sex, the main finding is that the difference between social group I and III decreases as a function of the number of tests taken Törnkvist & Henriksson, 2004b). Another finding is that, when controlling for age and social group, the differences between male and female decreases as a function of the number of tests taken. The summarized conclusion, concerning self-selection, is that males repeat the SweSAT more often than females, young test takers more often than older test takers, social group I more often than social group III and test takers with high score more often that those with low score.

Subtests

Henriksson & Bränberg (1994) focused on the effect of repeated test taking, totally as well as on the level of subtests[4]. They studied score changes between the first and the second SweSAT taken, since the observation is that the largest score gain is obtained from the first to the second test occasion. They studied five consecutive populations (87A-89B) in order to control for the effect of self-selection, i.e. the fact that it is the test taker's own decision whether to repeat the SweSAT or not. A division into subgroups was made on the basis of the variables sex, age and educational background. The obtained results indicated that self-selection is a matter of constancy, i.e. the repeaters in the five populations had about the same distribution in the background variables. It is about the same categories of test takers who chooses to repeat the test. 52% of the repeaters were males, a major part (about 41%) being in the age category 25-29 years, and about 39% had graduated from the shorter (2-year) upper secondary school programme. Another observation was that the mean value for the repeaters at the first test occasion was consistently lower than the mean value for the non-repeaters. Another finding was that the standardized mean difference was higher at the second testing than at the first, and this was true for all five populations and all subtests. That taken into consideration, the largest average score gains appeared on the subtests DS, STECH, and DTM.

[4] The SweSAT in this study consisted of six subtests: WORD (Vocabulary), DS (Data sufficiency), READ (Reading comprehension), DTM (Interpretation of diagrams, tables and maps), GI (General information) and STECH (Study techniques).

When considering the main findings in this study by Henriksson & Bränberg (1994) it is relevant to take the following circumstances into consideration. Firstly, the population of test takers has changed, and secondly, the SweSAT has also changed as compared to the situation today. From 1991 and onwards the population of test takers consists mainly (about 90 percent) of students from upper secondary school. The SweSAT of today is also radically changed as compared to the SweSAT during the period 87A-89B. In the studies reported below (Henriksson & Törnkvist, 2002; Törnkvist & Henriksson, 2004a; Törnkvist & Henriksson, 2004b) the population of test takers consists mainly of students from upper secondary school, i.e. the population of test takers was not restricted to those who were 25 years old and had 4 years of work experience. In the studies reported the problem of comparing subtests[5], with different levels of difficulty, is controlled for by using calibrated scores. The reason for this is that the description of the effects of repeated test taking is complicated by the fact that the subtests are not being normed separately on the subtest level. The subtest scores of reference population 1 (Stage & Ögren, 2002) were used to calibrate the subtest scores.

There are significant differences in mean subtest scores between the first and the second test occasion for the subtest WORD, DS and ERC (Henriksson & Törnkvist, 2004a). For subtest DTM and READ there are significant differences from the second to the third test occasion (Table 9).

Table 9. Effects of repeated test taking in terms of mean score gain (M) between different test occasions.

Subtest	Score gain (M) $1 \rightarrow 2$	Score gain (M) $2 \rightarrow 3$
WORD	1.3*	0.1
DS	0.9*	0.0
DTM	0.2	1.0*
READ	0.4	1.0*
ERC	1.0*	0.1

* $p<0.05$

[5] The SweSAT in these studies consisted of five subtests: WORD, DS, READ, DTM and ERC (English reading comprehension).

For the subtests WORD and READ the differences between mean scores for males and females were not significant. For the subtests DTM and DS, males had significantly higher mean scores at every test occasion, when controlling for education and age. For subtest ERC there was no consistent results over the studies. In Henriksson & Törnkvist (2002) there were significant differences between males and females for ERC, but not in Törnkvist & Henriksson (2004a).

The differences between social group I and III decreased with the number of tests taken for most of the subtests except for the subtests WORD and READ (Figure 3). The lowest difference between social group I and III in relation to the maximum possible score was observed for the subtest DTM and the highest difference for the subtest ERC, at the first test occasion. At the fourth test occasion the subtest DS has the lowest difference in mean scores between social group I and III and READ the highest difference.

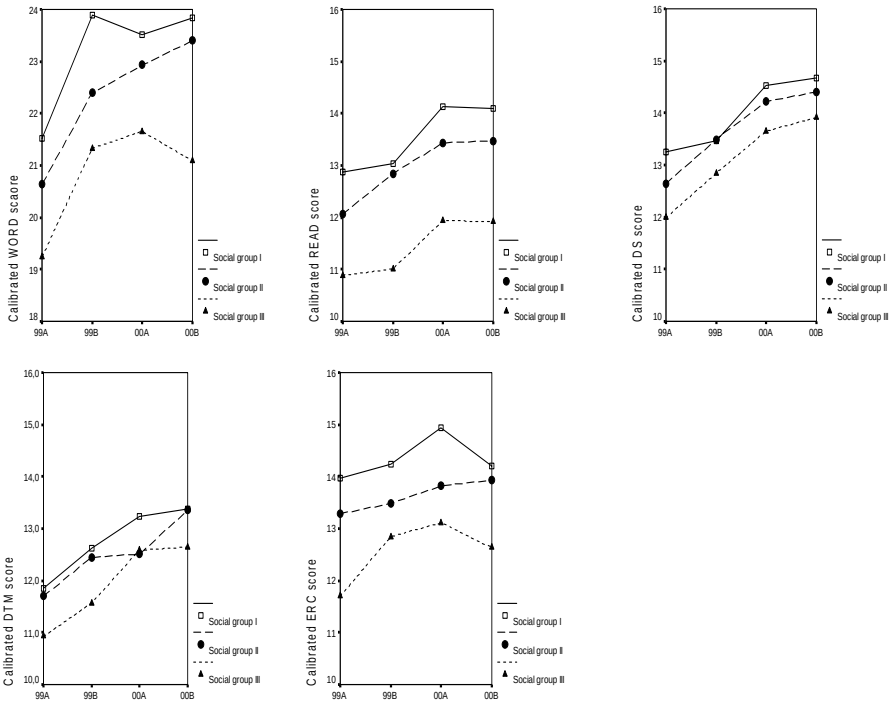

Figure 3. Effects of repeated test taking for social group I, II and III. Mean calibrated scores for the subtests WORD, READ, DS, DTM and ERC, evaluated at the age of 20.5 years.

The summarized finding, concerning the effects of repeated test taking for different subtests, is that the highest gain between the first and second test taking is observed for subtests WORD, DS and ERC. The highest gain between the second and third test occasion is observed for subtests DTM and READ. Another observation is that the differences between social group I and III decreased with the number of tests taken for the subtest DS, DTM and ERC. Still another observation is that the difference between social group I and III is higher than the corresponding difference between males and females for subtests WORD, READ and ERC at all test occasions (Törnkvist & Henriksson, 2004b).

Estimated consequences of changes in the structure of the SweSAT and its relation to repeated test taking

There is an ongoing debate about the structure of the SweSAT, i.e. the relative weight of each subtest, and its consequences for different subgroups of test takers. Different changes in the structure of the SweSAT and their consequences for the difference between social group I and III are illustrated in Table 10.

Table 10. Mean differences on subtests between social group I and III, and estimated (in brackets) mean score differences on total test, at first and fourth SweSAT, when controlled for sex and age.

Sub test	Maximum score	First SweSAT	Fourth SweSAT	Percent of the total test
WORD	40	2.28 (7.0)	2.76 (8.4)	32.8
READ	20	1.99 (12.1)	2.18 (13.3)	16.4
ERC	20	2.26 (13.8)	1.59 (9.7)	16.4
Sum(WORD+READ+ERC)	80	6.53 (10.0)	6.52 (9.9)	65.6
DS	22	1.26 (7.00)	0.77 (4.3)	18
DTM	20	0.92 (5.6)	0.74 (4.5)	16.4
Sum(DS+DTM)	42	2.18 (6.3)	1.51 (4.4)	34.4
Total score	122	8.72	8.03	100
Normed score	2.0	0.23	0.21	100

If all subtests had the same impact on the difference between social group I and III as the subtests WORD, READ and ERC taken together, the estimated difference in mean score between social group I

and III on the total test would be 10.0 at first test occasion instead of 8.7 and 9.9 scores instead of 8.0 at the fourth test occasion. This means that the higher weight for these subtests the higher difference between social group I and III.

If instead all subtests had the same impact as for the subtests DS and DTM taken together, the estimated difference in mean score between social group I and III would be 6.3 scores instead of 8.7 scores at the first and 4.4 scores instead of 8 scores at the fourth test occasion. This means that the higher the weight for these subtests the lower the difference between social group I and III and this difference is also lower for repeaters.

Thus, the summarized conclusion is that if the intensions are to decrease the differences between social groups the following line of action is to be considered – firstly to increase the weights of the subtests DS and DTM, secondly to motivate test takers from social group III to repeat the SweSAT.

Different changes in the structure of the SweSAT and their consequences for the difference between males and females are illustrated in Table 11.

Table 11. Mean differences on subtests between males and females, and estimated (in brackets) mean score differences on total test, at first and fourth SweSAT, when controlled for social group and age.

Subtest	Maximum score	First SweSAT	Fourth SweSAT	Percent of the total test
WORD	40	-0.75 (-2.3)	-0.36 (-1.1)	32.8
READ	20	-0.24 (-1.5)	-0.75 (-4.6)	16.4
ERC	20	0.86 (5.2)	1.24 (7.6)	16.4
Sum(WORD+READ+ERC)	80	-0.13 (-0.2)	0.13 (0.2)	65.6
DS	22	1.41 (7.8)	1.09 (6.1)	18.0
DTM	20	1.73 (10.5)	1.23 (7.5)	16.4
Sum(DS+DTM)	42	3.14 (9.1)	2.32 (6.7)	34.4
Total score	122	3.02	2.49	100
Normed score	2.0	0.08	0.07	100

If all subtests had the same impact on the difference between males and females as the subtests WORD, READ and ERC taken together, the estimated difference in mean score between males and females on the total test would be -0.2 at first test occasion instead of 3.0 and 0.2 scores instead of 2.5 at the fourth test occasion. This means that the higher the weight for these subtests the lower the difference between males and females.

If instead, all subtests had the same impact on the difference between males and females as for the subtests DS and DTM taken together, the estimated difference in mean score between males and females would be 9.1 scores instead of 3.0 scores at the first and 6.7 scores instead of 2.5 scores at the fourth test occasion. This means that the higher weight for these subtests the higher difference between males and females and this difference is also lower for repeaters.

Thus, the summarized conclusion is that if the intensions are to decrease the differences between males and females the following line of action is to bee considered - to increase the weights of the subtests WORD, READ and ERC.

If the importance of the numerical component of the SweSAT (DS and DTM) is increased the differences in mean scores between the social groups would decrease but at the same time the sex differences would increase.

If the importance of the verbal component of the SweSAT (WORD, READ and ERC) is increased the differences in mean scores between males and females would decrease but at the same time the differences between social groups would increase.

If social group III could be motivated to repeat the SweSAT the differences in mean scores between the social groups would decrease and the sex differences would be the same.

DISCUSSION

The main purpose of this discussion is to integrate and discuss presented results about the effects of repeated test taking of the SweSAT. Messick's four faceted model of validity will be used as an integrating and an analytic tool and the structure of the discussion follows Mes-

sick's 2x2-table. The presented results from supplementary education will also be used as a reference in this integration.

Construct validity

When discussing this facet it is relevant to mention that Messick points to two types of "threats" that can affect construct validity. The first threat is under-representation of the construct of interest. The instrument cannot cover all the important aspects and dimensions of the quality or ability that the test is intended to measure. The second threat refers to consequences of over-representation, i.e. when the instrument is also measuring irrelevant aspects.

Repeated test taking

A main finding is that the largest gain from repeated test taking occurs between the first and the second test occasion. This is interpreted as a gain that is mainly due to testwiseness (see for example Millman et al, 1965; Henriksson 1981). This means that the obtained score for a certain test taker at the first test occasion is an underestimation of the test taker's true score. A test taker must have a certain amount of testwiseness (TW) in order to get a score that is a good estimation of true score and taking the first test gives a contribution to TW that can be used at the second test occasion. This contribution includes, in the first place, an optimal time-use strategy. Thus, the conclusion is that the obtained score at the second test occasion is a better estimate of true score, as compared to the score obtained at the first test occasion. This also means that the proportion of construct relevant variance will be increased, as compared to the situation at the first test occasion.

However, for many test takers there is also a gain between the second and the third, and between the third and the forth test occasion. Given the assumption that testwiseness for taking the SweSAT is optimized, this gain is in many cases a function of growth. The reason for this is that many repeaters are in educational settings during the period of repeated test taking, for example studies in upper secondary school (Hamrén, 2006).

Supplementary examination

Supplementary examination implies that a student is permitted extended time for learning. The concept "time on task" is relevant in this situation (Smeets & Mooij, 1999; Spaulding & Dwyer, 2001; Nyroos, 2006). GPA is a criterion-related measurement and it is reasonable to assume that the level of knowledge and ability will be increased as a function of supplementary completion. But the problem is that when allowing for the condition of expended time for learning for certain persons the consequence is also that the GPA, in a perspective of selection to higher education, will include concept-irrelevant variance for those persons. With reference to selection another consequence is a restriction of variance in GPA.

Construct validity + relevance/utility

All test scores contain random error, which can be positive or negative and small or large. The direction of random error is unknown but the reliability coefficient provides an estimate of the proportion of variation in test score that might be attributed to random error. The reliability for SweSAT is about 0.94-0.96 (Stage & Ögren, 2005) and that implies that a rather small proportion (about 4-6%) of the variation in score is random error. The assumption is also that there is no relation between a test taker's true score and random error.

Repeated test taking

When considering the concept of random error it is also relevant to relate it to repeated test taking and one of the rules for the SweSAT when it is used in the process of selection to higher education. This rule is that if a test taker has more than one valid SweSAT score, the best obtained score will be used in the selection procedure. Thus, for some test takers, who have repeated the SweSAT, the selection will be based on a positive error of measurement, i.e. a score that is higher than the test takers true score.

This is not fair from a very strict perspective of measurement and the question is whether this rule should be replaced by some other rule? For example to base the selection on the latest obtained score, or on a mean of the two latest obtained scores or some other rule. Or - simply to state that repeated test taking is forbidden?

When discussing this question it is relevant to relate to Messick's two types of "threats" that can affect construct validity that are mentioned earlier. The second threat refers to consequences of overrepresentation, i.e. when the instrument is measuring irrelevant aspects. An example of an irrelevant aspect, or irrelevant variance, is when selection is based on a score that includes positive error of measurement. Thus, the rule that the best obtained SweSAT score is used in the selection procedure implies that the score that is used favours some repeaters as compared to non-repeaters.

But the weakness in this aspect must also be compared with the consequences of rules that influence the test taker's willingness to take the SweSAT. If the rule, for example, is that the latest obtained score, or a mean of the two latest obtained scores is valid for selection, the test taker will reflect on whether it is optimal to repeat or not. Thus, the relevance and utility of the rule that allows for repeated test taking is that the test taker is given a possibility of obtaining a good estimate of his or her knowledge and ability.

It is also relevant to mention that Svensson, Gustafsson & Reuterberg (2001) concluded that repeated test taking should be restricted and this conclusion was based on data about the relation between first and maximum SweSAT score and success in higher education during the first year. However, the obtained results were conflicting, for some educational programmes the relation between first and maximum SweSAT were lower, for others it was higher.

All rules that involve restriction will have consequences for the willingness to repeat the SweSAT. The existing rule is supported by two factors. The first factor can be illustrated by referring to correction for guessing on item level. Studies have indicated that the estimation of true score will be optimized if there is no correction for guessing. One conclusion is that a score that is obtained during such conditions also reflects partial knowledge (Henriksson, 1981). This problem can also be related to the first of Messick's "threats", i.e. to an underestimation of a test takers true score. These circumstances have also been the basis for not using correction for guessing for the SweSAT. The second factor is that the willingness to repeat the SweSAT would be lower if another rule than the existing rule was used. This second factor, i.e. conditions that are supposed to affect the willingness to repeat the SweSAT, can also be related to the desirable consequence

that the proportion of repeaters should be higher than it is today. It can also be related to the fact that many repeaters are in educational settings during the period of test taking.

Supplementary examination

Supplementary education implies expanded time for learning a defined educational context. From a theoretical perspective it is also reasonable to assume that concept irrelevant variance is incorporated when using GPA as a basis for selection for applicants with supplemented GPA. This conclusion is also supported by Wikström (2006). From an empirical perspective it is also in accordance with the conclusion drawn by Cliffordson (2004a). Her conclusion, based on empirical data, was that the increase in predictive validity that is caused by supplementary examination is zero.

Value implications

This facet in the Messick model refers to the fact that values are related to a certain construct and its label. To form an individual conception about a certain construct depends heavily on ideas about the construct itself. If the conception, for example, is that the construct is constant and stable, or the opposite - that it is fluctuating and unstable, then that is a factor that influences the value implication. Thus, value implications are related to the connotations that individuals, or a defined group of persons, have when they are confronted with a certain construct.

Repeated test taking

The intention with the SweSAT is to get an indication of study success in higher education. Thus, the assumption is that those test takers who obtain a high score will also succeed in higher education. The conception about the SweSAT must be in accordance with this assumption. Value implication concerning repeated test taking for the SweSAT is related to the question whether the test takers, and other persons involved, change their opinion about the quality of the SweSAT when repeated test taking is allowed? If a certain test allows for strategies that can be used to increase a test score, without the corresponding relation to higher ability, there is a scenario for a change of value im-

plication. Thus, there can be a change of the conception if the SweSAT is susceptible to short time instruction.

Studies have indicated that this is not the case for the SweSAT, i.e. the SweSAT is not susceptible to short time instruction (Henriksson, 1981). In this context it is also relevant to mention that all possible actions are taken in order to avoid undue score gains for the SweSAT (Stage, 2004).

The facet value implications for the SweSAT have also been illuminated by responses from test takers about the relevance of the SweSAT when used for selection to higher education. According to results from pilot studies with a random sample of test takers the summarized conclusion is that the test takers (Wester-Wedman, 1989; Eriksson, 2003) as well as personal from university education (Lyrén, 2003) considers the SweSAT to be relevant. This can be regarded as a summarized standpoint that also includes their opinion about repeated test taking. The fact that many test takers are in educational settings when taking the SweSAT can also be regarded as a contribution to the conception and value implications of repeated test taking.

Supplementary examination

The value implication of GPA, and the question of the role of supplementary examination for changing this value implication, focuses on the stability of the conception of GPA. It is reasonable to assume that the conception of GPA has changed, from being an indicator of performance in relation to criteria and standards in upper secondary school, to being a figure that can be manipulated. This change is, on the one hand, based on the fact that strategic selection of courses and educational programs occurs in upper secondary school. Advanced theoretical courses in basic subjects can for example be replaced with practical courses that are equal in length but for which obtaining a high grade is comparably easy. Courses that are equal in length are equivalent when it comes to sum for GPA (Wikström, 2006). On the other hand, the assumed change in value implication is most likely also based on the occurrence of supplementary education, i.e. when defined performance standards are obtained under expanded time conditions (Löfgren, 2004)

Supplementary examination implies that a student gets more time for acquiring a certain knowledge and ability. This increased time for learning implies in many cases a growth in knowledge but, at the same time, a decrease in the possibility of using GPA as a tool for prediction of success in higher education. Supplementary examination also leads to grade inflation. Grade inflation implies a decrease in variation and, as a consequence, also a decrease of the usefulness of GPA as an instrument in the process of selection to higher education.

Social consequences

This last facet refers to consequences of the use of an instrument in a certain situation, for individuals as well as all parties involved. The question whether the rules for an instrument, in this case supplementary examination for GPA and repeated test taking of SweSAT, should consider both the intended and the unintended consequences.

Repeated test taking

Social consequences can be regarded as a matter of selection. This means that test takers with a high score at the first test occasion repeat the SweSAT more often than those with a low score. It also means that young test takers repeat the SweSAT more often than old test takers, that males repeat more often than females and that test takers from social group I repeat more often than test takers from social group III. The summarized conclusion is also that these categories of test takers get a higher score as a function of repeated test taking. This can be considered as the unintended consequences of repeated test taking.

Then the question is - how to reduce these unintended consequences? The most important strategy is to motivate all test takers to repeat the test. This will reduce the differences between social groups and males and females. Thus, a difference that is observed at the first test occasion is reduced as a function of test taken. In this context it is also relevant to mention that Svensson & Nielsen (2005); Svensson (2006) made the opposite conclusion, i.e. that repeated test taking for the SweSAT should be restricted or even forbidden. The reason for this suggestion is that repeated test taking results in difference between Social group I and Social group III, due to self-selection. However, our point of view is that if we can motivate Social group III to repeat the SweSAT the differences between social groups will be reduced. It

is also important to point out that we, in our studies about differences between social groups, have controlled for the influence of sex and age. This is not the case in the studies reported on by Svensson & Nielsen (2005); Svensson (2006).

The intended consequences of allowing for repeated test taking is related to testwiseness, i.e. the opportunity to take the SweSAT more than once implies that the test taker will be familiar with the requirements of the test and the test situation. The obtained score, as an indicator of the test taker's true score, is also optimized if the test taker is test-wise. Thus, it is very important to motivate test takers from Social group III to repeat the SweSAT as they do not have the same opportunity to prepare for the SweSAT, or as Svensson (2006) state, that social group III does not have a supporting network.

Supplementary examination

The fact that GPA is used for selection to higher education, at the same time as supplementary education is permitted, leads to certain social consequences. From a perspective of society, and also from an economic macro-perspective, it is wastefulness with personal and material resources (Wikström, 2006). Students supplementing their grades imply that they extend their upper secondary education longer than stipulated and this is a matter of wastefulness.

Supplementary education is also a disadvantage for society since the prognostic power of GPA is undermined. Students that have supplemented their GPA get an unfair advantage, compared to those who have not, when it comes to selection to higher education. They are selected on a GPA that is not based on standardized conditions for acquiring a defined knowledge and ability. The consequence will also be that their success in higher education will be lower than expected.

Supplementary education is also a matter of concern from a perspective of social segregation. This means that certain categories of individuals utilize the possibility of supplementary education, for example high performing students and students from social group I (Svensson, 2006).

Conclusion

The summarized conclusion is that the existing rule for the SweSAT concerning repeated test taking is relevant. The rule that allows for repeated test taking is relevant as the test taker has an opportunity to obtain a good estimate of his or her knowledge and ability. The unintended social consequences are reduced if the test takers are motivated to repeat the test.

But, the conclusion is quite the opposite for supplementary education. A change in the rule that eliminates the usage of supplementary grades in the selection procedure will result in a higher validity.

The conclusion is also that the Messick model is a very useful tool for validation. When applying the Messick model, the aim and the direction of the process of validation will be systemized and optimal as well as nuanced. The model also implies a focus on traditional as well as non-traditional factors and consequences. Another advantage is that applying the Messick model also leads to a validation process that will not be fragmented since each facet in the model must be considered simultaneously.

But, one consequence of this strategy of simultaneously consideration can also be that the summarized evaluation and conclusion must be based on aspects with conflicting results. This was for example the case when the rule for repeated test taking on the one hand implies that the selection for some test takers is based on a positive error of measurement. On the other hand, the conclusion was that this negative aspect must be balanced against consequences of rules that affect the willingness to take the SweSAT. Thus, an application of a broadened validity perspective, by using the Messick model, also implies that the concept of validity must be balanced and related to overarching concerns that have to do with equality and fairness.

REFERENCES

Anastasi, A. (1982). *Psychological testing.* New York: MacMillan Publishing Co.

Cliffordson, C. (2004a). Betygsinflationen i de målrelaterade gymnasiebetygen [Inflation in goal-related grades from upper secondary school] *Pedagogisk Forskning i Sverige, 9*(1), 1-14.

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

Cronbach, L.J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (second edition). Washington, DC: American Council of Education.

Cronbach, L.J. (1988). Five perspectives on the validity argument. In: H. Wainer. & H. Braun (Eds.), *Test validity*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Eklöf, H. (2006). *Motivational beliefs in the TIMSS 2003 context: Theory, measurement, and relation to test performance* (Doctorial thesis). Umeå university.

Eriksson, S. (2003). *Vad tycker provdeltagarna om högskoleprovet? En pilotstudie* [What is the test taker's opinion about the SweSAT? A pilot study] (Arbetsrapport, Nr 2). Umeå universitet, Enheten för pedagogiska mätningar.

Gregory, R.J. (2004). *Psychological testing: history, principles and applications.* Needham Hights: Allyn and Bacon.

Hamrén, M. (2006). (Personal communication, June, 9, 2006).

Henriksson, W. (1981). *Effekter av övning och instruktion på testprestation. Några empiriska studier och analyser avseende övningens och instruktionens betydelse för testprestationen* [The effects of practice and coaching on test score] Doctorial thesis. Umeå university, Sweden.

Henriksson, W. (1991). *Effekter av upprepat provtagande* [The effects of repeated test taking] (PM, Nr 40). Avdelningen för pedagogiska mätningar, Pedagogiska institutionen, Umeå universitet.

Henriksson, W. (1993). *The problem of repeated test taking and the SweSAT.* (EM, No 5). Division of Educational Measurement, Department of Education, University of Umeå.

Henriksson, W. (1995). *Repeated test taking and the SweSAT* (EM, No 13). Division of Educational Measurement, Department of Education, University of Umeå.

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

Henriksson, W., & Törnkvist, B. (2002). *The effects of repeated test taking in relation to the test takers and the rules for selection to higher education in Sweden* (EM, No 41). Department of Educational Measurement, University of Umeå.

Henriksson, W., & Wedman, I. (1992). *Prediction of academic success in a perspective of criterion-related and construct validity* (EM, No 2). Division of Educational Measurement, Department of Education, University of Umeå.

Henriksson, W., & Wedman, I. (1993). *Effects of repeated test taking on the Swedish Scholastic Aptitude Test (SweSAT)* (EM, No 8). Division of Educational Measurement, Department of Education, University of Umeå.

Henrysson, S. (1992). *Högskoleprovets historia. Några bidrag* [The history of the SweSAT] (Pedagogiska Mätningar, Nr 91). Umeå universitet, Pedagogiska institutionen, Avdelningen för pedagogiska mätningar.

Högskoleverket. (1997a). *Examination vid universitet och högskolor - ur studentens synvinkel* [Examination at universities and in higher education - from the student's point of view] (Högskoleverkets skriftserie 1997:10 S). Stockholm: Högskoleverket.

Högskoleverket (1997b). *Tillträde till högre utbildning - en evighetsfråga* [Admission to higher education - a perennial problem] (Högskoleverkets skriftserie 1997:13 S). Stockholm: Högskoleverket.

Högskoleverket (2004). *Fortsatt hög andel av nybörjarna vid universitet och högskolor har studerat i kommunal vuxenutbildning (komvux)* [The proportion of novice students in higher education that have studied in adult upper-secondary education is still high] (Statistik och Analys, 2004-02-12). Stockholm: Högskoleverket.

Högskoleverket (2006). *Om högskoleprovet.* Stockholm: Högskoleverket.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement, 2*, 135-170.

Lyrén, P-E. (2003). Är *DTK-provet relevant enligt högskoleutbildningarna? En pilotstudie* [Is the DTM subtest relevant according to the university education? A pilot study] (Arbetsrapport, Nr 3). Umeå universitet, Enheten för pedagogiska mätningar.

Löfgren, K. (2003). *Enskild prövning och komplettering av betyg* [Individual examination and supplementary completion of grades] (PM, Nr 182). Umeå universitet, Enheten för pedagogiska mätningar.

Löfgren, K. (2004). *Utbyteskompletteringar bland dem som avslutade gymnasiet 1997-2001* [Supplementary upper secondary education among school-leavers 1997-2001] (PM, Nr 182). Umeå universitet, Institutionen för beteendevetenskapliga mätningar.

Messick, S. (1987). Large-scale educational research as policy research: Aspirations and limitations. *European Journal of Psychology of Education, 2*, 157-165.

Messick, S. (1988). The once and future issues of validity: assessing the meaning and consequences of measurement. In H. Wainer & H Brown (Eds.), *Test validity*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Messick, S. (1989a). Validity. In R. L. Linn (Ed.), *Educational Measurement* (Vol. 3, pp. 13-103). New York: Macmillan/American Educational Research Association.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5-11.

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

Millman, J., Bishop, H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*, 707–726.

Nyroos, M. (2006). *Tid till förfogande* [Available time]. (Doctorial thesis). Umeå: Umeå univerity.

Nyström, P. (2004). *Rätt mätt på prov: Om validering av bedömningar i skolan* [Validation of educational assessments]. (Doctorial thesis). Umeå: Umeå university.

Payne, D.A. (1997). *Applied Educational Assessment.* New York: Wadsworth Publishing Company.

Popham, W.J. (1997). Consequential validity: right concern – wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Rogers, W.T., & Bateson, D.J. (1992). The influence of test-wiseness on performance of high school seniors on school leaving examinations. *Applied Measurement in Education, 4*, 159–183.

Shepard, L.A. (1993). Evaluating test validity. In: L. Darling-Hammong (Ed.), *Review of research in education*. (pp. 405-450). Washington, DC: American Educational Research Association.

Smeets, E., & Mooij, T. (1999). Time on task, interaction, and information handling in multimedia learning environments. *Journal of Educational Computing Research, 21*(4), 487-502.

SOU 1968:25. *Studieprognos och studieframgång* [Predicting academic performance and academic success]. Stockholm: Utbildningsdepartementet.

SOU 1974:71. *Om behörighet och antagning till högskolan* [Qualification and admission to higher education]. Stockholm: Utbildningsdepartementet.

SOU 2004:29. *Tre vägar till den öppna högskolan* [Three routes to the open university]. Stockholm: Högskoleverket.

Spaulding, K., & Dwyer, F. (2001). The effect of time-on-task when using job aids as an instructional strategy. *International Journal of Instructional Media, 28*(4), 437-447.

Stage, C. (2004). *Entrance to higher education in Sweden* (EM, No 51). Umeå university, Department of Educational Measurement.

Stage, C., & Ögren, G. (2005). *Högskoleprovet våren och hösten 2005. Provdeltagargruppens sammansättning och resultat* [SweSAT, spring and autumn 2005. The test takers background and results] (BVM, Nr 17). Umeå universitet, Institutionen för beteendevetenskapliga mätningar.

Stobart, G. (2001). The validity of national curriculum assessment. *British Journal of Educational Studies, 49*(1), 26-39.

Svensson, A. (2006). Hur ska rekryteringen till högskolans mest eftersökta utbildningar breddas? [How to broaden the recruitment to the most attractive programmes in higher education?] *Pedagogisk Forskning i Sverige,11*(2), 116-133.

Svensson, A., Gustafsson, J-E., & Reuterberg, S-E. (2001). *Högskoleprovets prognosvärde. Samband mellan provresultat och framgång första studieåret vid civilingejörs- jurist- och grundskollärarutbildningarna* [The prognostic value of the SweSAT. Relation between test result and success during the freshman year for students at graduate engineer-, lawyer- and comprehensive school teacher education programmes] (Högskoleverkets rapport 2001:19R). Stockholm: National Agency of Higher Education.

Svensson, A., & Nielsen, B. (2005). *Toppresultat men ej antagen till högskolan. Studier av takeffekter hos gymnasiebetyg och högskoleprov* [Maximum result – but not admitted. Studies about ceiling-effects for GPA and SweSAT] (No 2005:08). Department of Education, University of Gothenburg.

Törnkvist, B., & Henriksson, W. (2004a). *SweSAT repeat* (EM, No 46). Department of Educational Measurement, University of Umeå.

Törnkvist, B., & Henriksson, W. (2004b). *Repeated test taking. Differences between social groups* (EM, No 47). Department of Educational Measurement, University of Umeå.

Wedman, I. (1992). *Selection to higher education in Sweden* (EM, No 1). Department of Education, Division of educational measurement, Umeå university.

Wedman, I. (2002). *Behörighet, rekrytering och urval. Om övergången från gymnasieskola till högskola* [Eligibility, recruitment and selection. About the transition from upper secondary school to higher education]. Stockholm: Högskoleverket.

Wester-Wedman, A. (1989). *Vad tycker provdeltagarna om högskoleprovet 1988-05-07?* [What is the test taker's opinion about the SweSAT 1988-05-07?] (PM, Nr 23). Avdelningen för pedagogiska mätningar, Pedagogiska institutionen, Umeå universitet.

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

Wikström, C. (2005b). *Criterion-referenced measurement for educational evaluation and selection.* (Doctorial thesis). Umeå university.

Wikström, C. (2006). Classroom assessment and grading – validity issues in the selection process to higher education. *Paper presented at the annual NCME-conference in San Francisco, 2006-04-08—10.*

Wikström, C., & Wikström, M. (2005). Grade inflation and school competition: an empirical analysis based on the Swedish upper secondary schools. *Economics of Educational Review, 24*, 309-322.

Wolming, S. (2000). *Validering av urval* [Validation of selection]. (Doctorial thesis). Umeå university.

Wolming, S. (2001). Att värdera urvalsinstrument. Några reflektioner över möjligheter och begränsningar [To validate selection instruments. Reflections on limitations and possibilities]. *Pedagogisk Forskning i Sverige, 6*(2), 122-130.

EDUCATIONAL MEASUREMENT

Reports already published in the series

EM No 1. SELECTION TO HIGHER EDUCATION IN SWEDEN. Ingemar Wedman

EM No 2. PREDICTION OF ACADEMIC SUCCESS IN A PERSPECTIVE OF CRITERION-RELATED AND CONSTRUCT VALIDITY. Widar Henriksson, Ingemar Wedman

EM No 3. ITEM BIAS WITH RESPECT TO GENDER INTERPRETED IN THE LIGHT OF PROBLEM-SOLVING STRATEGIES. Anita Wester

EM No 4. AVERAGE SCHOOL MARKS AND RESULTS ON THE SWESAT. Christina Stage

EM No 5. THE PROBLEM OF REPEATED TEST TAKING AND THE SweSAT. Widar Henriksson

EM No 6. COACHING FOR COMPLEX ITEM FORMATS IN THE SweSAT. Widar Henriksson

EM No 7. GENDER DIFFERENCES ON THE SweSAT. A Review of Studies since 1975. Christina Stage

EM No 8. EFFECTS OF REPEATED TEST TAKING ON THE SWEDISH SCHOLASTIC APTITUDE TEST (SweSAT). Widar Henriksson, Ingemar Wedman

1994

EM No 9. NOTES FROM THE FIRST INTERNATIONAL SweSAT CONFERENCE. May 23 - 25, 1993. Ingemar Wedman, Christina Stage

EM No 10. NOTES FROM THE SECOND INTERNATIONAL SweSAT CONFERENCE. New Orleans, April 2, 1994. Widar Henriksson, Sten Henrysson, Christina Stage, Ingemar Wedman and Anita Wester

EM No 11. USE OF ASSESSMENT OUTCOMES IN SELECTING CANDIDATES FOR SECONDARY AND TERTIARY EDUCATION: A COMPARISON. Christina Stage

EM No 12. GENDER DIFFERENCES IN TESTING. DIF analyses using the Mantel-Haenszel technique on three subtests in the Swedish SAT. Anita Wester

1995

EM No 13. REPEATED TEST TAKING AND THE SweSAT. Widar Henriksson

EM No 14. AMBITIONS AND ATTITUDES TOWARD STUDIES AND STUDY RESULTS. Interviews with students of the Business Administration study program in Umeå, Sweden. Anita Wester

EM No 15. EXPERIENCES WITH THE SWEDISH SCHOLASTIC APTITUDE TEST. Christina Stage

EM No 16. NOTES FROM THE THIRD INTERNATIONAL SweSAT CONFERENCE. Umeå, May 27-30, 1995. Christina Stage, Widar Henriksson

EM No 17. THE COMPLEXITY OF DATA SUFFICIENCY ITEMS. Widar Henriksson

EM No 18. STUDY SUCCESS IN HIGHER EDUCATION. A comparison of students admitted on the basis of GPA and SweSAT-scores with and without credits for work experience. Widar Henriksson, Simon Wolming

1996

EM No 19. AN ATTEMPT TO FIT IRT MODELS TO THE DS SUBTEST IN THE SweSAT. Christina Stage

EM No 20. NOTES FROM THE FOURTH INTERNATIONAL SweSAT CONFERENCE. New York, April 7, 1996. Christina Stage

1997

EM No 21. THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SWESAT. A study of the DTM subtest. Christina Stage

EM No 22. ITEM FORMAT AND GENDER DIFFERENCES IN MATHEMATICS AND SCIENCE. A study on item format and gender differences in performance based on TIMSS´data. Anita Wester, Widar Henriksson

EM No 23. DO MALES AND FEMALES WITH IDENTICAL TEST SCORES SOLVE TEST ITEMS IN THE SAME WAY? Christina Stage

EM No 24. THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SweSAT. A Study of the ERC Subtest. Christina Stage

EM No 25. THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SweSAT. A Study of the READ Subtest. Christina Stage

EM No 26. THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SweSAT. A Study of the WORD Subtest. Christina Stage

EM No 27. DIFFERENTIAL ITEM FUNCTIONING (DIF) IN RELATION TO ITEM CONTENT. A study of three subtests in the SweSAT with focus on gender. Anita Wester

EM No 28. NOTES FROM THE FIFTH INTERNATIONAL SWESAT CONFERENCE. Umeå, May 31 – June 2, 1997. Christina Stage

1998

EM No 29. A COMPARISON BETWEEN ITEM ANALYSIS BASED ON ITEM RESPONSE THEORY AND ON CLASSICAL TEST THEORY. A Study of the SweSAT Subtest WORD. Christina Stage

EM No 30. A COMPARISON BETWEEN ITEM ANALYSIS BASED ON ITEM RESPONSE THEORY AND ON CLASSICAL TEST THEORY. A Study of the SweSAT Subtest ERC. Christina Stage

EM No 31. NOTES FROM THE SIXTH INTERNATIONAL SWESAT CONFERENCE. San Diego, April 12, 1998. Christina Stage

1999

EM No 32. NONEQUIVALENT GROUPS IRT OBSERVED SCORE EQUATING. Its Applicability and Appropriateness for the Swedish Scholastic Aptitude Test. Wilco H.M. Emons

EM No 33. A COMPARISON BETWEEN ITEM ANALYSIS BASED ON ITEM RESPONSE THEORY AND ON CLASSICAL TEST THEORY. A Study of the SweSAT Subtest READ. Christina Stage

EM No 34. PREDICTING GENDER DIFFERENCES IN WORD ITEMS. A Comparison of Item Response Theory and Classical Test Theory. Christina Stage

EM No 35. NOTES FROM THE SEVENTH INTERNATIONAL SWESAT CONFERENCE. Umeå, June 3–5, 1999. Christina Stage

2000

EM No 36. TRENDS IN ASSESSMENT. Notes from the First International SweMaS Symposium Umeå, May 17, 2000. Jan-Olof Lindström (Ed)

EM No 37. NOTES FROM THE EIGHTH INTERNATIONAL SWESAT CONFERENCE. New Orleans, April 7, 2000. Christina Stage

2001

EM No 38. NOTES FROM THE SECOND INTERNATIONAL SWEMAS CONFERENCE, Umeå, May 15-16, 2001. Jan-Olof Lindström (Ed)

EM No 39. PERFORMANCE AND AUTHENTIC ASSESSMENT, REALISTIC AND REAL LIFE TASKS: A Conceptual Analysis of the Literature. Torulf Palm

EM No 40. NOTES FROM THE NINTH INTERNATIONAL SWESAT CONFERENCE. Umeå, June 4–6, 2001. Christina Stage

2002

EM No 41. THE EFFECTS OF REPEATED TEST TAKING IN RELATION TO THE TEST TAKER AND THE RULES FOR SELECTION TO HIGHER EDUCATION IN SWEDEN. Widar Henriksson, Birgitta Törnkvist

2003

EM No 42. CLASSICAL TEST THEORY OR ITEM RESPONSE THEORY: The Swedish Experience. Christina Stage

EM No 43. THE SWEDISH NATIONAL COURSE TESTS IN MATHEMATICS. Jan-Olof Lindström

EM No 44. CURRICULUM, DRIVER EDUCATION AND DRIVER TESTING. A comparative study of the driver education systems in some European countries. Henrik Jonsson, Anna Sundström, Widar Henriksson

2004

EM No 45. THE SWEDISH DRIVING-LICENSE TEST. A Summary of Studies from the Department of Educational Measurement, Umeå University. Widar Henriksson, Anna Sundström, Marie Wiberg

EM No 46. SweSAT REPEAT. Birgitta Törnkvist, Widar Henriksson

EM No 47. REPEATED TEST TAKING. Differences between social groups. Birgitta Törnkvist, Widar Henriksson

EM No 49. THE SWEDISH SCHOLASTIC ASSESSMENT TEST (SweSAT). Development, Results and Experiences. Christina Stage, Gunilla Ögren

EM No 50. CLASSICAL TEST THEORY VS. ITEM RESPONSE THEORY. An evaluation of the theory test in the Swedish driving-license test. Marie Wiberg

EM No 51. ENTRANCE TO HIGHER EDUCATION IN SWEDEN. Christina Stage

Em No 52. NOTES FROM THE TENTH INTERNATIONAL SWESAT CONFERENCE. Umeå, June 1–3, 2004. Christina Stage

2005

Em No 53. VALIDATION OF THE SWEDISH UNIVERSITY ENTRANCE SYSTEM. Selected results from the VALUTA-project 2001–2004. Kent Löfgren

Em No 54. SELF-ASSESSMENT OF KNOWLEDGE AND ABILITIES. A Litterature Study. Anna Sundström

Em No 55. BELIEFS ABOUT PERCEIVED COMPETENCE. A literature review. Anna Sundström