NOTES FROM THE ELEVENTH INTERNATIONAL SweSAT CONFERENCE

Umeå, June 12–14, 2006

Christina Stage

EM No 58, 2006

ISSN 1103-2685

Abstract

The Swedish Scholastic Assessment Test (SweSAT), has been used for selection to higher education in Sweden since 1977, and it has by now become an integrated part of the Swedish Educational System. The International Scientific Advisory Board, was constituted in 1992, and met for the first time in Umeå in May 1993. The board met for the eleventh time in June 2006. A list of participants, and the program for the meeting are enclosed as appendices. In this report condensed summaries of the presentations and the discussions from the meeting are presented. The summaries of the presentations in this report are in the same order as in the program.

The SweSAT Program since June 2004

Christina Stage

At the last meeting the proposals from The Governmental Commission on rules for admission to higher education was presented. According to the decisions taken after that, the present situation is that:

- At a minimum 30 percent (and maximum 60 percent) of the study places should be allocated on the basis of grade from upper secondary school.
- At a minimum 30 percent (and maximum 60 percent of the study places should be allocated on the basis results on the SweSAT
- At a minimum 10 percent (and maximum 20 percent) of the study places should be allocated on the basis of selection models decided by the universities or colleges.

Hence, SweSAT is still one of the two most important selection instruments. The problem at present is insufficient finances for development of the test. It should be within the budget of the National Agency for Higher Education, and they are very worried about the losses this may cause. The fee for taking the test has been raised and it is now 350 SEK. But the number of test-takers has been decreasing at the last test occasions, in spite of predictions of the opposite. In Figure 1 the number of test-takers, in relation to number of 20-years old individuals may be seen. 'A' represents tests given in spring and, 'B' represents tests given in autumn.

The main reason for the decreased number of test-takers is probably the decreasing interest in higher education. The number of applicants for higher education has decreased still more than the number of test-takers.

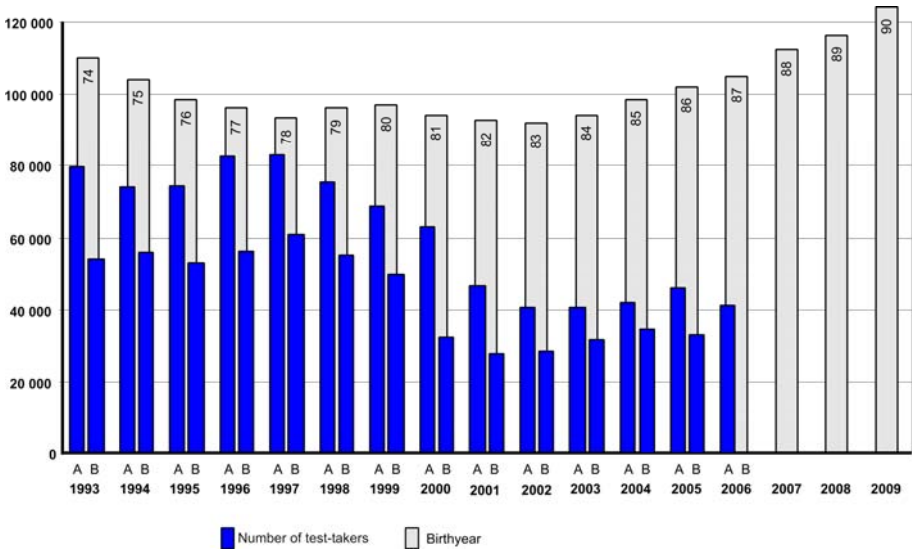


Figure 1. Number of test-takers in relation to the total number of 20-years old individuals.

We have been forewarned that in the end of June a public procurement for the construction of two of the subtests in SweSAT will be presented. This will be discussed later in this meeting, after introductions by Nils Olsson and Ingemar Wedman.

The VALUTA-project has been finished, since we last met, and a final report has been written. The project has resulted in two doctoral thesis, more than 20 published articles, more than 30 working papers, at least ten presentations at international conferences, and a vast number of presentations at Swedish conferences.

Some conclusions from the VALUTA-project:

The criterion referenced grades are inflated over years, are unequal between schools and between different study programs in upper secondary school. Christina Wikström will tell us more about these findings during the meeting.

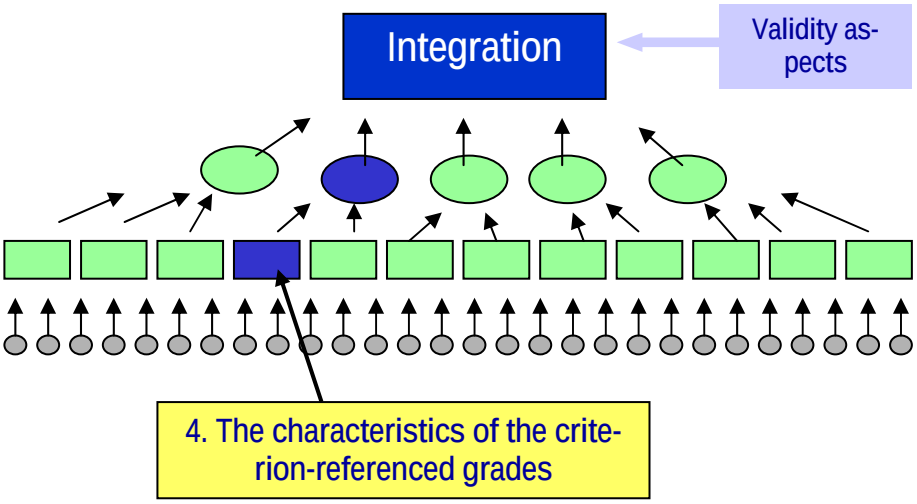
All the same, several studies have shown that the GPAs are better predictors of success in higher education than the SweSAT. Per-Erik Lyrén will talk about this paradox.

The National Agency for Higher Education has initiated a low budget change/development of the SweSAT, and the primary aim of the changes should be to improve the predictive validity. This will also be discussed during this meeting.

The grades as selection instrument for higher education

Christina Wikström

VALUTA "Validation of the University Entrance System"*



Are the criterion referenced upper secondary school grades reliable and valid instruments for

1. information?
2. selection?

Validity

- Content validity
- Criterion-related validity Concurrent & **predictive**
- Construct validity

* A joint project between Göteborg University and Umeå University, financed by the Bank of Sweden Tercentenary Foundation.

	Interpretation	*Use*
Evidential basis	Construct validity	Construct validity Relevance+ utility
Consequential basis	Value implications	Social consequences

"The facets of validity" according to Messick (1989)

Assessment and grading in Swedish schools

- Since 1994 Criterion referenced education/grading
- Classroom assessment – Limited control mechanisms

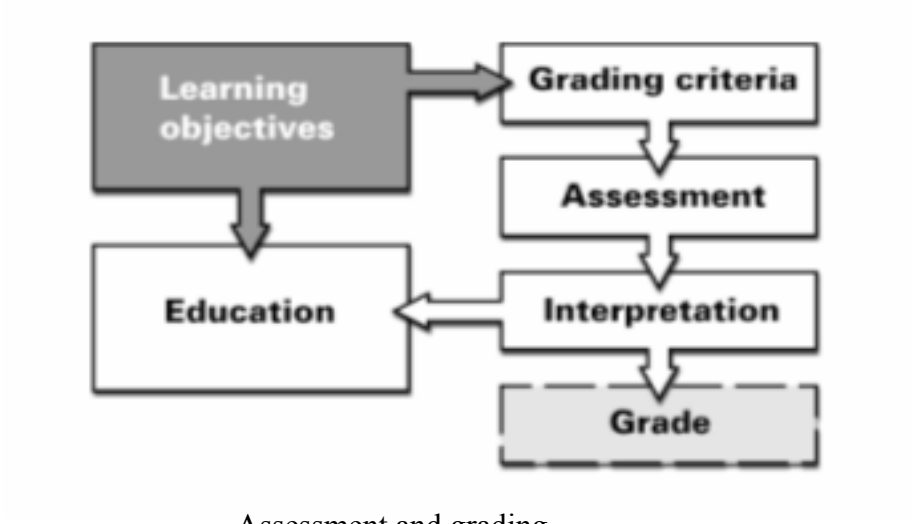
<u>The upper secondary school grades serve 3 purposes:</u>

1. motivation
2. information
3. selection (eligibility & ranking) ← *two instruments!*

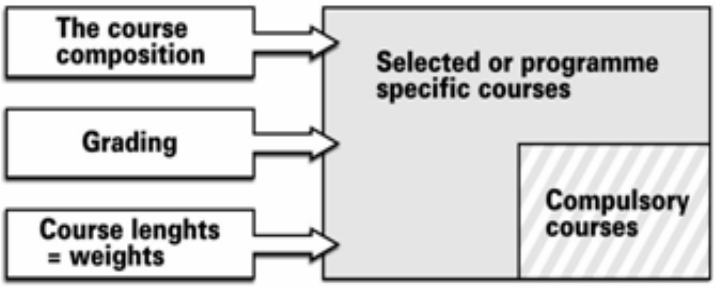
↑ grades — Absolute scale

↑ GPA — Relative scale

Threats to validity

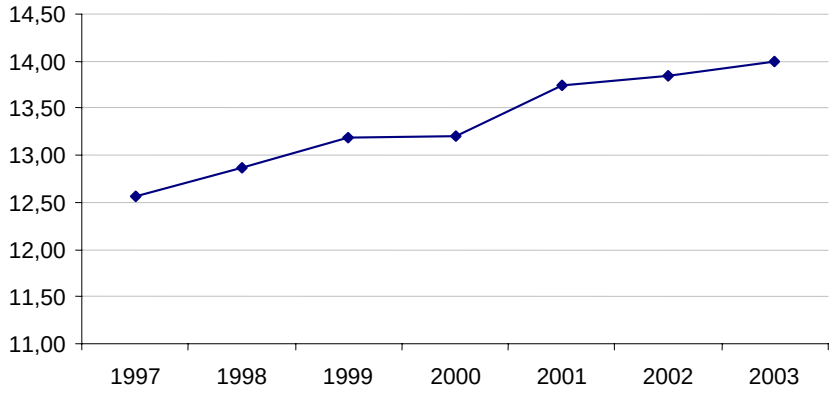


Assessment and grading

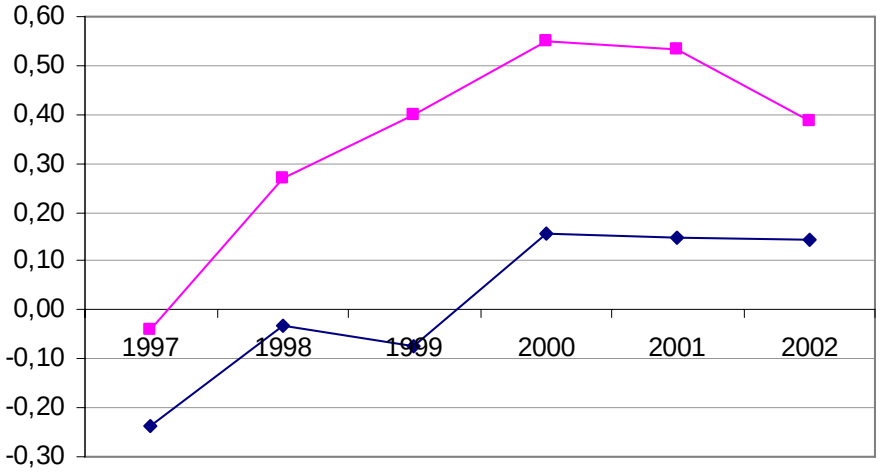


The GPA model

Upper secondary GPA 1997-2003 (N=all graduates, scale 0-20)

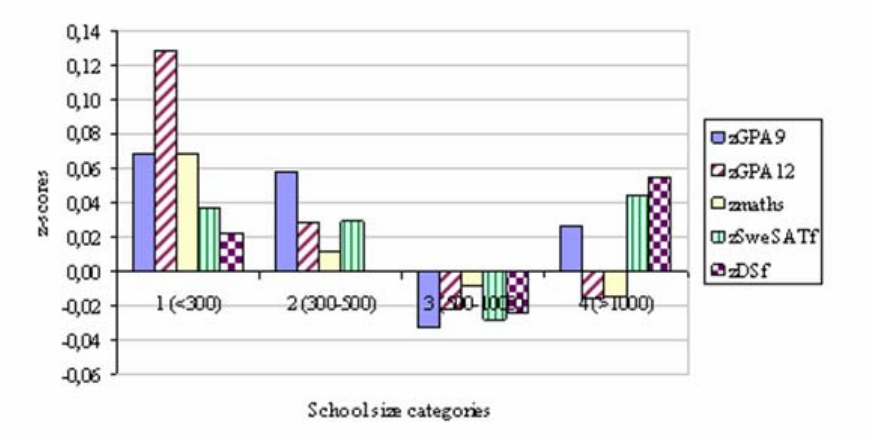


Public –private school rank differences (grades-test)

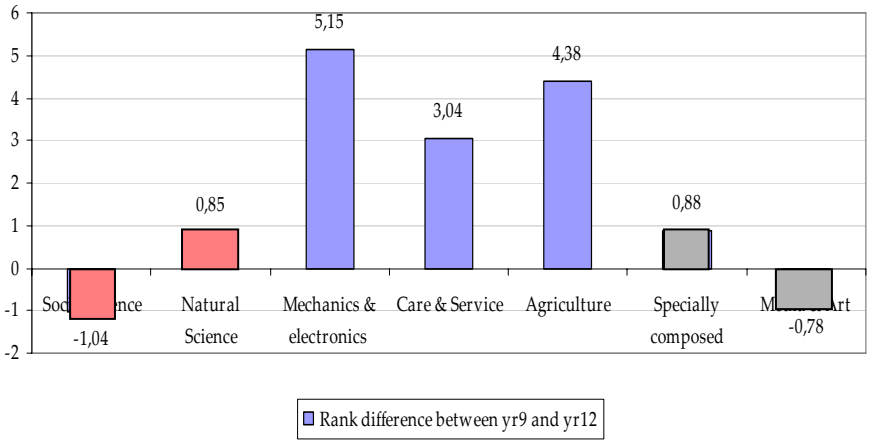


Diff maths grades – DS subtest (z-scale), Natural Science graduate students 1997-2002. (Grey lines = private schools, black lines = public schools)

School size, grades and test performance



Rank difference GPA 12 – GPA 9 (compulsory courses)



Why these variations in grading?

- Lack of calibration instruments (standardized tests etc.), unclear criteria, low skills in assessment and grading.
- The combination of criterion referenced grades and selection, or other forms of competition, are in conflict.
- The GPA model is leading to reliability and validity problems.

Empirical results

- Grade inflation! (high achieving students have gained the most)

- Grade outcome varies among schools of different type and size (schools exposed to competition are high grading)
- Students in vocationally oriented programs are slightly higher graded than students in academically oriented programs, and also benefit from the course composition in their programs
- Students (in particular those from high socioeconomic background) are strategic – (female) students supplement their grades and (male) students re-take the SweSAT.

But

- GPA predicts study success in higher education (in terms of credits and study rate) slightly better than the SweSAT.

Validity aspects

Consequences:

- No possibility to monitor educational outcome over time → incorrect evaluations of school performance.
- System more favourable for some categories → increased segregation.
- Strategic course choices → false positive students failing in higher education; increased expenses for universities accepting false positive students.
- Academic/advanced courses avoided.

The predictive validity of the SweSAT

Per-Erik Lyrén

Throughout the existence of the SweSAT there has been focus on its predictive validity. Lately, several researchers have presented studies in which the SweSAT is described as inferior to the GPA in terms of being able to predict academic performance. This is something that the Swedish Agency for Higher Education has seized upon and therefore they have put forward a demand/wish to improve the predictive validity of the SweSAT. This is in its essence a good thing. However, there are issues that need to be discussed before taking action towards any changes of the SweSAT. For example there are issues related to the studies on the SweSAT (and GPA) that need to be addressed. These issues are related not only to the predictor itself, but also to the criteria for academic performance (e.g., distribution of variable, relevance), the measure of predictor-criterion relationship (appropriateness of correlations and selection groups means respectively), the

sample on which the study is carried out (e.g., representative-ness, selectivity), and the procedure for selecting applicants to the programs (simple vs. compensatory procedure).

Development of the SweSAT

Christina Stage

The Demands upon the SweSAT

- The test should, as fairly as possible, rank the applicants with regard to expected success in higher education.
- The test should be in line with the goals and content of higher education, and be relevant to the entire sector of higher education.
- It should not be possible to improve individual results by mechanical practice or by learning certain strategies.
- The test-takers should regard the test as meaningful and suitable for selection to higher education
- The requirements for comprehensive recruitment shall be taken in consideration, so that nobody is treated unfairly due to sex or social background.
- It should be possible to mark the test quickly, cheaply, and objectively.

The Criterion – Study Success

The criterion, which the SweSAT should be able to predict, consists of about 170 different study programs, and the examination forms can vary from tests to oral exams, home exams, individual papers, group papers, discussions, compulsory attendance etc. The grades are failed, passed, and sometimes passed with distinction. Provided passed exams, 20 credit points per semester are given independently of study course. Failure in the studies is not necessarily caused by lack of talent or laziness, but can be due to job opportunity, or change of study program.

The main idea, at present, is to change the SweSAT into a similar format as SAT and PET. The test would then consist of two parts, which could be equated and scaled separately, and which could be used in different ways for different study programs. Since mathematical ability is supposed to be measured by the grades in upper secondary school, we have chosen to call the two parts verbal and analytical. At present we have three subtests (including ERC) for measuring verbal

ability with 80 items, but only two subtests, and 42 items, measuring analytical ability. Since it is not possible to equate and scale on basis of only 42 items, more analytical subtests would be needed. Also the WORD sub-test, which contains 40 of the 80 verbal items, has been criticized, for having too much weight on the final test result, and for being unfair towards immigrants. Hence, at least parts of the WORD sub-test would need to be substituted.

SAT Reasoning Test	
Critical reading	2 x 25 + 1 x 20 min = 70 min
Sentence completion	
Passage-based reading	
Math	2 x 25 + 1 x 20 min = 70 min
MC	
Grid in	
Writing	25 + 25 +10 min = 60 min
Total test time 3 h 45 min (including a 25 min pre-test section)	

PET Psychological Entrance Test	
Verbal Reasoning	27 items 25 min
Analogies	
Sentence completion	
Logic	
Reading comprehension	
Quantitative Reasoning	25 items 25 min
Questions and problems	
Graph or table comprehension	
Quantitative comparisons	
Number series	
English	27 items 25 min
Total test time 3 h 20 min (including 2 x 25 min pre-test sections)	

Verbal subtests

Ragnar Haake, Sandra Scott

At present Swedish verbal ability is measured by two subtests WORD and READ. The problems with the subtest WORD are:

1. the words are taken out of context

2. it contains 40 out of a total of 122 items, i.e. 1/3 of the total score
3. the time is only 15 minutes; sensitive to disturbances

The problems with the subtest READ are:

1. it is a very demanding test, which takes a lot of effort
2. it gives small credit; only 20 items In 50 minutes

Two types of verbal items have been tried this spring, not intended as replacements but complements to the WORD subtest:

1. sentence completion

This subtest intends to measure vocabulary in context, and also the test-takers sense of language style.

2. analogies

The intention of this subtest, as well, is to measure vocabulary. Therefore the difficulty in the items should not be in the relations but in the words.

Analytical sub-tests

Anders Lexelius, Gunilla Ögren

Analytical reasoning questions test the ability to understand a given structure of arbitrary relationships among fictitious persons, places, things or events, and to deduce new information from the relationships given.

- A set of related statements or conditions describing a structure of relationships
- Questions that test understanding of that structure and its implications

An analytical reasoning test (AR) was developed and tried out on a small scale in 2000.

Table 1. The test composition, and results

subtest	Items	minutes	Mean	sd	Males	Females	d
AR	16	50	9.19	3.11	9.24	9.15	.03
WORD	10	8	11.66	2.71	11.77	11.58	.07
DS	10	20	5.74	2.77	6.73	5.12	.58
READ	8	20	3.59	1.98	3.73	3.50	.12
Total	54	98	30.17	7.91	31.47	29.36	.27

Table 2. Reliabilities α (**bold**), inter-correlations, and inter-correlations corrected for attenuation (*italic*)

Subtest	AR	WORD	DS	READ	Total
AR	**.68**	.30	.52	.50	.82
WORD	*.48*	**.56**	.25	.43	.66
DS	*.72*	*.38*	**.77**	.49	.76
READ	*.86*	*.83*	*.70*	**.49**	.76

The subtests AR and DS measure very much the same ability, but the gender differences are considerably smaller on AR than on DS.

Analytical reasoning items have also been used in the DS-format, i.e. DS-items with no numbers, but only verbal relations.

Table 3. AR-items in DS, 03:A - 06:A

Test	p males	p females	r_{bis}
03:A	.44	.42	.47
03:B	.64	.59	.59
04:A	.69	.64	.53
04:B	.35	.34	.51
05:A	.72	.70	.58
05:B	.74	.64	.62
06:A	.37	.35	.49

With one exception (05:B) this type of items causes less gender differences than DS-items in general, where the average difference is .09.

Discussion

Analogy tests have been abolished in the USA, since there does not exist anything similar in the school work, and American teachers hate this test type.

Analogy tests are sensitive to coaching, but that could probably be avoided if the difficulty lies in the vocabulary and not in the relations.

Analogy tests could cause cultural problems, an alternative could be to include more analytical problems in the READ subtest. You could also increase the number of items to each text to five or six.

A possibility would be to give only half a score for each WORD item, and an alternative to that could be to give two points for each correct READ-item, which would give more reward for the big effort needed on this subtest.

The most important thing to do, before changing the test, is to ask university teachers which abilities they regard as important for success in higher education. In such an investigation it is very important to ask specific questions in order to get useful answers.

Construction of the SweSAT

Stig Eriksson

The 19 Steps in Test Construction

1. Planning & selection of texts/figures/topics/ words
2. Item-construction (internal & external
3. Editing: corrections, re-writings, additions, check-ups, etc.
4. Composing of try-out versions
5. Local review
6. Revision: new corrections, language, checking of facts, copyright, etc.

---

7. Printing (& checking "blue-prints")
8. National try-out (along with the regular test)
9. Analysis of the try-out results (HANALYS)
10. Archive & (new) planning
11. Selection of items for regular test-version (+ ev. New try-out versions)
12. National review
13. Revision: language, facts, communication with text-authors etc.

---

14. New national review
15. New revisions
16. Final check-up/all sub-tests: overlapping, p-values, prognosis etc.

---

17. Printing
18. Regular use of the test ("Day of the SweSAT").
19. "After test": complaints, analysis & evaluation (internal & external).

	06:A	06:B	07:A	07:B
Autumn 04	Steps 1-6			
Spring 05	Steps7-13	Steps 1-6		
Autumn 05	Steps 14-15	Steps 7-13	Steps 1-6	
Spring 06	Steps 16-18	Steps 14-15	Steps 7-13	Steps 1-6
Autumn 06		Steps 16-18	Steps 14-15	Steps 7-13
Spring 07			Steps 16-18	Steps 14-15
Autumn 07				Steps 16-18

Procurement of the SweSAT: The item-writer as both a handy craftsman and a psychometrician

Nils Olsson

The National Agency of Higher Education is planning to develop and field-test new sub-tests for the SweSAT. In parallel existing sub-tests (DS and DTM) will be subject of procurement.

In Sweden the Act (SFS 1992:1528) on public procurement states that "The award of contracts should be so arranged as to take advantage of existing competition and should also in other respects accord with the conventions of good business practice. No unwarranted considerations should affect the treatment of candidates, or tenders.

Thus one important issue in developing new sub-tests is to find people willing to construct the tests at the best price given quality demands. Given that test specifications have been defined, an important question emerges, when deciding what company (or academic institution) will be given the task of constructing the items: What qualifications are needed for being a competent item writer? Is good knowledge in the subject area sufficient? What (if any) psychometric competence is necessary among the item writers? Can the item writing process take place in isolation (i.e. individually by a consult) or is an academic environment, where exchange of ideas about psychometrics and other test related issues, of utmost importance?

The Advisory Council on Access to Higher Education

Ingemar Wedman

1 The future development of the SweSAT test battery. The National Agency of Higher Education is right now planning for a development bid concerning parts of the SweSAT test battery and would like to have a short discussion concerning international experiences from such bids. The coordination of the SweSAT test will furthermore, according to the plans, be a task for Department of Educational Measurement at Umeå University. Experiences concerning this last topic will also be discussed.

2. Presently, there is a discussion about cultural bias in the SweSAT program, implying that native students are favoured by the contents of the test battery. There are ideas about decreasing the impact of the subtest Word. The main issue is about making the test battery more internationally adjusted. What effect will such a change have on the results on the whole and what effect will such a change have on Swedish students with a foreign background?

3. Today there are three roads into higher education in Sweden. One of these roads go through the marking system applied at the upper secondary school. The other one is through The SweSAT program. A new road is opening up and concerns about 20% of the applicants that through a third road can get access to higher education. The third row is an option for each university to handle. However, there is an ongoing discussion about if the tests of knowledge and skills in the area of technique and caretaking (nursing) will a useful tool for universities in handling access to higher education. What consequences will such a third road have on opening up the access to universities?

Discussion

The discussion was concentrated to the first point *procurement of the SweSAT.*

In comparison with tests internationally the present costs for development of the SweSAT was regarded as very reasonable. There are also quite a lot of "hidden costs" in test construction.

There was a warning raised against private companies, since the total environment for test construction is important, and preferably should be academic.

It was also regarded as peculiar to start with two subtests for procurement. The normal thing, internationally, is to give item construction to external agents, but not test construction. This will cause a lot of responsibility and work to be transferred from Umeå to the National Agency. Split responsibility is not good for the test quality.
There was also a general discussion of the disadvantages for universities, in cases of procurement, because of their very high over-head costs in comparison with private agencies.

Five big challenges for educational assessment practices
Ronald K. Hambleton

In 1966 educational measurement was characterized of: 1) multiple choice tests, 2) relatively simple statistics, 3) routine psychometric studies (could be published) and 4) computer cards/tapes.
In 2006: 1) wide array of item types, 2) complex statistical modelling of data (IRT, GT, SEM), 3) Standard setting, DIF, CBT, CAT, performance testing, Automated scoring and test development, 4) Lap
Two goals of the presentation:

- Describe five measurement problems that are either understudied, or underappreciated in to-day's efforts to implement educational and psychological tests.
- Suggest some necessary research

1. Use of tests in international markets

- Major misunderstandings about the difficulties of translating and adapting tests from one language and culture to another.
  Example: "Out of sight, out of mind" (back translated from French) "Invisible, insane"

Common misunderstandings (in US):

- That most anyone who knows two languages can do the translation.
- That a backward translation design is sufficient.
- That translators, if they have the correct training, can produce a valid instrument in a second language and culture

What needs to be done?

- Hire qualified translators (and several of them)
- Use forward and backward designs to review the test items.
- Compile empirical evidence to address construct, method and item-bias.

- Follow the ITC guidelines.

Research: Integration of best practices and examples, to guide future test adaptation studies.

2. Advances in modelling of test data.
   - IRT models have become popular and for good reasons – lots of positive features.
   - With new item types come questions about dimensionality and local dependencies.

New IRT polytomous response models:

- Partial credit model
- Generalized partial credit model
- Graded response model
- Logistic multidimensional model
- Rating scale models
- Hundreds more models exist

Research:

- There are questions of model choice (fit, practicality), and calibration of items with small samples
- Identifying and handling dependencies in the data
- Handling outliers in equating

3. Generation of new item types

High fidelity simulations, item algorithms, item cloning, automated scoring of constructed response items (e.g. essays)

- Lots of "sizzle" here with simulations (e.g., virtual reality, performance tasks)
- Can new skills be measured?
- Can old skills be measured better?
- What's the value added versus the cost of development?

Research:

- An expanded commitment to validation initiatives of these new item types is needed.
- Face validity is important (but hardly sufficient). Much more evidence is needed to-day to support the use of new item types.

4. Advances with computer-based tests

Advantages are well-known:

- Flexibility in scheduling, potential
- Immediate score reporting.

- Capability to assess higher level skills
- Potential to shorten testing time
- Eliminate floor and ceiling effects.

BIG Challenge: item/problem exposure – when present, test score validity is lowered.

Research:

- How large item bank is needed?
- How can item exposure be detected?
- How much more vulnerable are performance based tasks?
- How can the tasks be disguised and/or cloned? Impact of even minor revisions in item statistics?

5. Improvements in score reporting

- Least studied topic today in assessment, and one of the most important
- Lots of evidence that score users are easily confused. (concept of measurement error is not understood, error bands are confusing)

Research:

- Can we find empirically-based principles available to assist in the design of meaningful and useful score scales and reports?
- How can diagnostic reports be enhanced?
- Evaluation of new methods for studying score reports: focus-groups, "think aloud" studies, experimental studies, field studies.

Tests for detecting answer copying[1]

Wim van der Linden

A statistical test for the detection of answer copying is presented. The test is based on the idea that the answers of examinees to test items may be the result of three possible processes: (1) knowing (2) guessing (3) copying, but that examinees, who do not have access to the answers of other examinees, can arrive at their answers only through the first two processes. This assumption leads to a distribution for the number of matched alternatives, between the examinee suspected of

[1] Van der Linden, W. J. & Sotaridona, L. (2004). A statistical test for detecting answer copying on multiple-choice tests. *Journal of Educational Measurement, 41*, 361-377.

copying and the examinee believed to be the source, that belong to family of "shifted binomials". Power functions for the tests for several sets of parameter values are analyzed. An extension of the test to include matched number of correct alternatives would lead to improper statistical hypotheses.

The g_2 index proposed by Frary, Tideman, and Watts (1977)[2] is an attempt to evaluate the number of matching alternatives between an examinee suspected to be a copier and another examinee believed to be the source against the expected number of matching alternatives. Two problems inherent in working with such an index are obtaining the distribution of the index under the null hypothesis of no copying, and evaluating the statistical power of the test based on it.

The *K* index (Holland, 1996[3]; Lewis & Thayer, 1998[4]) is another attempt to correct for the examinee's ability. The index focuses only on the number of matching alternatives on the items that were answered incorrectly by the source. The null model is a binomial with a success parameter that is obtained by piecewise linear regression of the proportion of matching incorrect alternatives on the proportion incorrect scores in a population of examinees. An alternative with quadratic regression is given by Sotaridona and Meijer (2002)[5].

The most elaborate null model for a test to detect copying is the one on which Wollack's ω index is based (Wollack, 1997[6]; Wollack & Cohen, 1998[7]). Like the *g* index, the ω index compares the observed number of matching alternatives against an estimate of the expected number.

In spite of attempts to condition on the examinee's ability, a fundamental feature of all three tests is their dependence on the distribution of the item scores in the population. In principle, such tests can result

---

[2] Frary, R., B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational Statistics, 2*, 235-256.

[3] Holland, P. W. (1996). *Assessing unusual agreement between incorrect answers of two examinees using the K-index. Statistical theory and empirical support*. (Technical Report No. 96- 4). Princeton, NJ: Educational Testing Service.

[4] Lewis, C., & Thayer, D. T. (1998). The power of the K-index (or PMIR) to detect copying. (Research Report RR-98-49) Princeton, NJ: Educational Testing Service.

[5] Sotaridona, L., S., & Meijer, R. R. (2002). Statistiacal properties of the K-index for detecting answer copying. *Journal of Educational Measurement, 39*, 115-132.

[6] Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21*, 307-320.

[7] Wollack, J. A. & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*, 144-152.

in statistically significant proof of answer copying for a pair of examinees in one population but acceptance of the null hypothesis of no copying if their vectors were included in the data set for another population.

We present a statistical test to detect answer copying that can be used when any reference to a population of examinees is undesirable. The only assumptions we make are about the response behaviour of the individual examinee suspected of copying.

Like the *K* index the test focuses on the items for which the source has an incorrect answer. The test is derived from the following assumptions about the behaviour of the copier on the items the source has answered incorrectly. If an examinee knows an item, he/she gives a correct answer. If an examinee does not know an item but has access to the source, he/she copies that. If an examinee does not know the answer, and does not have access to a source, he/she guesses blindly Thus for each item answered incorrectly by the source, the copier can be in one of three possible states, each characterized by a different probability of choosing the same alternative the source has chosen.

The hypothesis to be tested is that the examinee did not copy any of the items. We suggest testing this hypothesis against the alternative that for some of the items, that he/she did not know, the answers were copied.

The actual power of the test is a function of the unknown number of copied items. The shape of the power function depends on (1) the number of alternatives per item (2) the number of incorrect items (3) the significance level chosen for the test, and (4) the number of items the examinee knows.

Unidimensionality and interpretability of psychological instruments

Jan-Eric Gustafsson

Unitary or complex measures?

Much of our current thinking in measurement is based on the idea that each instrument should measure one "thing" only.

Heterogeneity of instruments is seen as a problem, while an ideal instrument is homogeneous in the sense that it measures one factor only.

Analytical techniques such as regression analysis assume each independent variable to measure one "thing" only. However, the dependent variable is understood in terms of multiple sources of variance.

Thus, while the independent variables are viewed as unitary, the dependent variable is viewed as complex.

Is heterogeneity a threat to interpretability?

The unidimensionality claim typically implies a focus on constructs with narrow referent generality, and it typically requires splintering broad constructs into more and more narrow constructs.

During a long period research on intelligence suffered from loss of focus on constructs with a broad referent generality (e. g., general intelligence, fluid ability, crystallized ability)

In many other fields too there is a need to measure broad constructs, such as depression, social support, or self-esteem.

A construct with broad referent generality has a "contingently clustered set of attributes that covary under mutual causation or share common causal mechanisms" (Lucke, 2005[8]).

The complexity of social behavioural phenomena may require tests to be heterogeneous to reflect broad constructs, and to be reliable and valid.

"A high level of homogeneity of items may, however, be a mixed blessing. While greater homogeneity will generally result in greater reliability, it may do so at the cost of validity. Just as low correlations among tests permit higher multiple correlation from a combination of those tests, so low correlation among items permits higher validity for the test composed of those items." (Thorndike, 1951[9]).

Hierarchical models

Hierarchical factor-analytical models allow identification of both broad and narrow sources of variance. Two types of models.

Higher-order models include first-order factors which account for performance on manifest variables, second-order factors which account for performance on first order factors, and so on.

Nested factor models include a general factor with relations to all manifest variables, along with one or more narrow factors with relations to subsets of the manifest variables. All latent variables are orthogonal.

---

[8] Lucke, J. F. (2005). The [Aplpha] and the [Omega] of congeneric test theory: an extension of reliability and internal consistency to heterogeneous tests. *Applied Journal of Psychological Measurement 29, 1.*

[9] Thorndike R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational Measurement,* Washington: American Council on Education.

The dimensionality of the SweSAT

Factor analysis at the sub-test level have identified two correlated factors: one analytical problem solving factor with relations to DS and DTM; and one verbal knowledge factor with relations to WORD, READ, GI and STECH.

Item-level analyses have revealed further factors. Some DTM items require mental arithmetic, and these items along with the DS test, may be used to define a quantitative factor. Some of the sub-tests (particularly READ and DTM) are mildly speeded, and for these sub-tests an "end-of-test" factor appears on which the last items in the test load, the size of the loading being a function of the position of the item.

The SweSAT thus is multidimensional, and involves several broad and narrow factors.

Table 1. Sources of variance in the SweSAT

	DTM	**DS**	**WORD**	**READ**	**STECH**	**GI**	**Swe-SAT**
Gen*	0.57	0.48	0.21	0.35	0.40	0.30	0.66
Quant*	0.06	0.09					0.01
End*	0.02						0.00
Knowl*			0.41	0.17	0.13	0.19	0.19
Resid*	0.35	0.44	0.38	0.49	0.46	0.50	0.13

*General factor, Quantitative factor, end of test factor, verbal knowledge factor, and residual variance.

The composite reliability of the SweSAT is .87

The construct reliability of the general factor is .66 for the SweSAT score.

The construct reliability of the Knowledge factor is .19

The total SweSAT score is to a larger extent dominated by the general factor than is any of the subtests.

Table 2. Sources of variance in the SweSAT, when combined with the Swedish Enlistment Battery in a nested-factor model.

	DTM	DS	WORD	READ	ERC	SweSAT
G (Gf)	0.31	0,36	0.10	0.13	0.10	0.33
Knowl	0.14	0.10	0.08	0.18	0.16	0.23
Voc	0,01		0.52	0.25	0.31	0.26
Quant	0,04	0.44				0.05
Resid	0.49	0.10	0.30	0.44	0.42	0.13

Conclusions

The previous Knowledge factor is split into a broad Knowledge (construct reliability = 0.23) factor (emphasis on reading skills) and a more narrow Vocabulary factor (construct reliability = 0.26) these factors represent Gc

The general SweSAT factor is split into a Gf factor (construct reliability = 0.33) and the broad Knowledge factor.

Thus the General SweSAT factor is a mixture of Gf and Gc, and overall the SweSAT scale score is dominated by Gc.

General conclusions

A single score based on a heterogeneous test tends to measure whatever is common among the items. This may be, for example, general abilities or method factors.

A single score based on a homogeneous test reflects both general abilities and specific abilities. If the purpose is to measure a specific ability the general abilities are sources of construct irrelevant variance. And if the purpose is to measure general ability the specific abilities are sources of irrelevant variance.

Thus, measurement needs to be multidimensional and hierarchical.

Who make use of the SweSAT? An investigation based on thirteen age cohorts

Allan Svensson

An investigation based on thirteen age-cohorts

The aim of the study is to answer the questions:

A How is the SweSAT used by those born between 1972 and 1984, and how does the number of test-takers vary between the age cohorts?

B What relations are there between test-taking, repeated test-taking, and some background variables?

C Which students gain most by taking the test?

The students will be divided by sex and social background. The social background ha been classified on the basis of information about the occupations of the parents. The following groups are distinguished:

Group I. Academic professions

Group II. Civil servants, and white-collar workers in lower management positions.

Group III. Skilled, and unskilled workers.

Table 1. The distribution of individuals according to year of birth, age in 2005, the size of each age cohort, and the number of students in each cohort who have taken the SweSAT.

Year of birth	Age in 2005	Size of the age cohort	Number of test-takers	Percent of test-takers
1972	33	113 483	48 329	41
1973	32	111 977	48 205	43
1974	31	112 948	48 894	43
1975	30	106 689	46 214	43
1976	29	101 775	144 130	43
1977	28	99 711	43 018	43
1978	27	97 613	43 199	44
1979	26	101 053	42 697	42
1980	25	102 599	39 869	39
1981	24	99 915	35 052	35
1982	23	99 666	31 137	31
1983	22	99 055	26 127	26
1984	21	101 663	22 840	22
Total		1 348 347	517 711	38

Table 2. The proportions of test-takers among all born 1972 -84, and among individuals up to 21 years of age according to sex and social background.

	Sex		Social	group	
	Women	Men	I	II	III
All	42	35	60	41	23
Up to 21	33	26	50	30	15

Table 3. The proportions among different categories who have taken the SweSAT 1, 2, 3, 4, 5, or more times

Number of tests	Sex Women	Men	Social I	group II	III
One	19	15	22	18	12
Two	12	9	17	12	6
Three	6	5	10	6	3
Four	3	3	6	3	1
> Five	2	3	5	2	1
Total	42	35	60	41	23

Summary
Of all individuals in the study 38 per cent has taken the SweSAT at least once from spring 1989 to spring 2005. In the eight oldest cohorts the percentage of test-takers is 40 per cent. Seven percentage points more women than men have taken the SweSAT, and the difference between social group I and II is 37 percentage points.

Repeated test-taking[10]

Birgitta Törnkvist, Widar Henriksson

Rules

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

- An obtained SweSAT 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.

The Model

	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, 1989, p.20[11])

Purpose

The main purpose of this study is to integrate and discuss results from studies with focus on the effects of repeated test-taking. Messick's

[10] Törnkvist, B. & Henriksson, W. (2006). Validity issues concerning repeated test-taking. (EM No 56). Umeå: Umeå University, Department of Educational Measurement.

[11] Messick, S. (1989). In R., L. Linn (Ed.) *Educational Measurement*. New York: American Council on Education & Macmillan.

four-faceted model of validity is used as an integrating and analytic tool.

I Construct validity

Messick points to two types of threats that can affect construct validity:

- Under-representation of the construct of interest. The instrument cannot cover all the important aspects and dimensions that the test intends to measure.
- Over-representation, i.e. when the instrument is also measuring irrelevant aspects.

Repeated test taking

A Main finding – the largest gain from repeated test taking occurs between the first and the second test occasion. A gain that is mainly due to increased test-wiseness (TW)

$$t = T + e$$

Assumption: A test taker must have a certain amount of TW to get a score that is a good estimate of his/her true score.

1. The obtained score at the first occasion is an underestimation of the test-taker's true score.
2. A test-taker must have a certain amount of TW in order to get a score that is a good estimate of his/her true score. Taking the first test gives a contribution to TW that can be used at the second test occasion. This contribution includes, in the first hand, an optimal time using strategy.
3. This also means that the proportion of construct relevant variance will be increased, as compared to the situation at the first test occasion.

Conclusion: The obtained score at the second test occasion is a better estimate of the true score, as compared to the first test occasion.

B Another finding – for many test-takers there is also a gain between the second and the third, and between the third and the fourth test occasion.

1. TW for taking the SweSAT is optimized. There is, on the whole, no further gain in TW (Henriksson, 1981).
2. Many repeaters are in educational settings during the period of repeated test taking, for example studies in upper secondary school (Hamrén, 2006)

Conclusion: the gain is (in many cases) a gain that is related to growth.

II Construct validity + relevance/utility

- All test scores contain random errors, which can be positive or negative.
- The direction of random errors is unknown, but the reliability coefficient provides an estimate of the proportion of variance in a test score that might be attributes to random errors.
- The reliability of SweSAT is about 0.92-0.94 (Stage & Ögren, 2005) and that implies that a rather small proportion of the score variance is random errors.
- The assumption is also that there is no relation between a test-taker's true score and the random errors.

When considering the concept of random error, it is also relevant to relate it to a certain rule for repeated test taking.

1. If a test taker has more than one valid SweSAT score, the best score will be used in the selection.
2. Thus, for some test takers, who have repeated the SweSAT, the selection will be based on a positive error, i.e. a score higher than the test taker's true score.

Conclusion: This is not fair from a strict perspective of measurement.

III Value implications

- A conception about a certain construct depends on ideas about the construct itself.
- If the conception is that the construct is constant and stable that influences the value implication.

The conception is that those who get high test scores also will have success in higher education. Value implications concerning repeated test taking is related to whether the test takers, and others involved change their opinion about the SweSAT when repeated test taking is allowed.

If a test allows strategies which can improve the score, without corresponding relation to ability, that is a reason for change of value implications.

Conclusion: There is no change in value implications

1. SweSAT is not susceptible to short time instruction (Henriksson, 1981).
2. All possible actions are taken to avoid undue score gains.
3. Repeated test taking incorporate the establishment of a rational time using strategy, and therefore allows a better estimate of the test takers´ true scores.

IV Social consequences

The last facet refers to consequences of the use of an instrument for individuals as well as other parties involved.

1. Test takers with high scores at the first test occasion repeat the test more often than those with low scores.
2. Young test takers repeat more often than old test takers.
3. Males repeat more often than females.
4. Social group I repeat more often than social group III.

How to reduce these unintended consequences?

- Motivate all test takers to repeat the test. A difference at the first test occasion is reduced as a function of number of tests taken.

Concluding remarks

It is very good to make a re-evaluation of the SweSAT. You should start to ask university professors, what they want their students to know, as Sten Henrysson did about 40 years ago. It is difficult to improve the predictive validity, if you do not know the criteria for study success. Study success is a very complicated construct. There will always be problems to define and measure study success, but you would get better results by asking professors, what they want the test to measure. In that way focus would be transferred to the construct validity of the test.

It is also a good idea to divide the test into one verbal and one analytical/quantitative part, and at the same time reduce the number of WORD-items. An analogy subtest works all right, but does not measure applied skills, and the items are too much an intelligence test. Analogy tests may also be sensitive for coaching.

There should be an evaluation of the information about the test. What is the opinion of the information? Does the information reach everybody? How does the information given in schools work? It would be very advantageous to have a test given, free of charge, to all students in upper secondary school. You should encourage practicing, and you should also encourage repeated test taking, and inform about the benefits of re-taking the test.

One way to reduce the expenses, would be to keep at least some test versions secret. Secret test versions would make it possible to build a substantial item-bank. In the US they keep most tests secret, not for cost reasons, but in order to keep or improve the quality of the tests.

The Swedish principle of access to official documents is a problem, however, for keeping test versions secret.

It would also be useful to look carefully into the security aspects in connection with the test-day. Cheating, in any form, gives the test a bad reputation, and reduces the credibility. Contact with Caveon Test Security company was recommended www.caveon.com (where our friend John Fremer is president).

Social and cultural bias on the test was also discussed. From that perspective, as well, a shorter WORD subtest would be desirable. In the US, in the development of SAT, they have test constructors from different cultures involved in the item writing, as well as in the review process. For the SweSAT there are people, with knowledge of differing cultural background, involved in the review, but all item writers have Swedish background. A survey in Sweden has shown that 50% of the immigrants want to study at university level, but only 28% of the Swedish born. The dilemma is that, even though, you have to give special considerations to immigrants, good knowledge of the Swedish language is a prerequisite qualification for university studies in Sweden.

Participants

Ronald K. Hambleton, USA
Wim van der Linden, The Netherlands (Monday and Tuesday)

Margaretha Hallgren, National Agency for Higher Education
Nils Olsson, National Agency for Higher Education
Ingemar Wedman, Advisory board, National Agency for Higher Education (Monday and Tuesday morning)

Jan-Eric Gustafsson, Göteborg University
Allan Svensson, Göteborg University

Widar Henriksson, Umeå University
Birgitta Törnkvist, Umeå University (Wednesday)
Christina Wikström, Umeå University

Axel Eklund, SweSAT
Stig Eriksson, SweSAT
Ragnar Haake, SweSAT
Mats Hamrén, SweSAT
Christina Jonsson, SweSAT
Ingegerd Jonsson, SweSAT
Anders Lexelius, SweSAT
Sandra Scott, SweSAT
Christina Stage, SweSAT
Gunilla Ögren, SweSAT

Program for the 11th SweSAT Meeting June 12 - 14

Monday June 12th

9.30 Welcome and opening address (Chistina Stage)
Coffee
The grades as selection instrument for higher education (Christina Wikström)
The predicitive validity of the SweSAT. (Per-Erik Lyrén)*

12.00 Lunch

13.00 Development of the SweSAT (Christina Stage)
Verbal subtests (Ragnar Haake, Sandra Scott)
Coffee
Analytical subtests (Anders Lexelius, Gunilla Ögren)*
Discussion

Tuesday June 13th

8.30 Construction of the SweSAT (Stig Eriksson)
Coffee

10.00 The Advisory Council on Access to Higher Education (Ingemar Wedman)*

12.00 Lunch

13.00 Five Big Challenges for Educational Assessment Practices. (Ron Hambleton)
Coffee
Tests for Detecting Answer Copying (Wim van der Linden)
Unidimensionality and Interpretability of Psychological Instruments (Jan-Eric Gustafsson)

Wednesday June 14th

8.30 WHO MAKE USE OF THE SweSAT? An investigation based on thirteen age cohorts. (Allan Svensson)
Repeated Test-taking (Birgitta Törnkvist, Widar Henriksson)
Coffee
Concluding Remarks

* Additional material to be sent later.

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

2006

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

Em No 56. VALIDITY ISSUES CONCERNING REPEATED TEST TAKING OF THE SWESAT. Birgitta Törnkvinst, Widar Henriksson

Em No 57. ECTS AND ASSESSMENT IN HIGHER EDUCATION. Conference Proceedings. Kent Löfgren