NOTES FROM THE TENTH INTERNATIONAL SweSAT CONFERENCE

Umeå, June 1–3, 2004

Christina Stage

EM No 52, 2004

ISSN 1103-2685

Preface

The SweSAT has been used for selection to higher education since 1977, and it has by now become an integrated and generally accepted part of the Swedish educational system. An International Scientific Advisory Board was constituted in 1992, and up to 2001 the board met once a year, every other year in Sweden and the other year in connection with the AERA/NCME annual meeting. The first meeting was held in Umeå in May 1993 (Wedman & Stage, 1994). For two years, 2002 and 2003 the meeting had to be cancelled, but in 2004 the tenth meeting was held in Umeå.

This report gives a condensed summary of the presentations at the tenth meeting of the scientific advisory board. A list of the participants, and the program of the meeting are enclosed as appendices. The summaries of the presentations in this report are in the same order as in the program, and some of the presentations are followed by comments, which summarize the discussions.

The studies presented by: Ewa Andersson, Kent Löfgren, Christina Stage, Allan Svensson, Birgitta Törnkvist, and Christina Wikström were financed by The Bank of Sweden Tercentenary Foundation, and are part of a research program (VALUTA) with the aim of validating the system of admittance to higher education in Sweden.

The SweSAT Program during the last years

Christina Stage

Today it is 11 years since the first meeting with this group. In the notes from this first meeting, which took place in May 1993, the following conclusions were drawn:

- The present item specifications for the SweSAT should be made more precise.
- The present procedure with internal item-writers should be maintained.
- There is every reason to change the present manner of field testing new items. Hidden sections or items interspersed in the regular administration of the SweSAT should be used.
- Item writing information should be exchanged with item developers in other countries.
- More mathematical content should be added to the test
- The legitimacy of the sub-test GI (General Information) was questioned by several members of the board.
- A warning was issued against the problem of coaching.
- It was suggested that we should continue to use the highest obtained score, even though this is combined with some statistical problems.
- The situation of the handicapped groups and the SweSAT must be incorporated in the research agenda of the SweSAT program.

Most of the recommendations from the first meeting have been followed, the exception being that no more mathematical content has been added so far.

In the annual report from 1993 (Wedman & Stage) a report on the staff, working in the SweSAT-program at that time can be found. In Table 1 the number of staff in 1993 is compared with the number of staff working in the program in 2003.

Table 1. The SweSAT staff in 1993 and 2003.

	1993	**2003**
Project coordinator:	0.50	0.50
Researchers:	2.50	0.75
System analysts:	2	1.75
Superintendent:	1	0.80
Test developers:	5	5
Clerical officers:	3.5	2
Research assistants:	2	none
Doctoral students:	none	1

As may be seen in Table 1, the decrease in number of staff during the ten years is substantial, especially with regard to research staff.

In Figure 1 is shown that there has also been a considerable decrease in the number of test-takers during the same years. Unfortunately the number of test-takers influences the income of the test, while the development costs of the test are independent of the number who takes the test. In the background of Figure 1 the number of nineteen years old in the country is shown for each year.

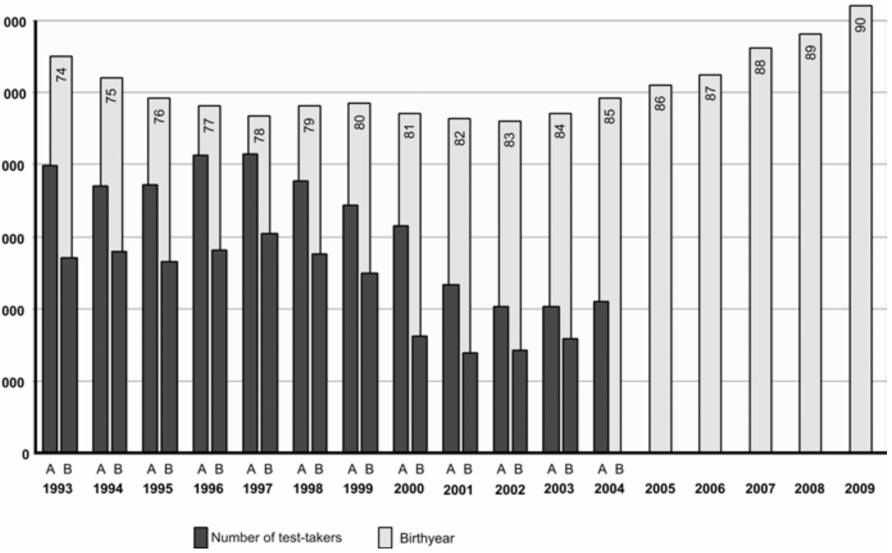


Figure 1. Number of test-takers from 1993 until spring 2004; (A= spring test, B= autumn test). In the back-ground the number of nineteen years old is shown.

In 1996 the test was changed, and therefore a lot about the test was written in the papers, which is the most probable reason why the number of test-takers rose in fall 1996. In 1997 a new grading system hade been introduced, and also a new rule, that allowed you to improve your grades after you have finished school (this will be discussed later in connection with the report of the government commission). The effects of this rule on the number of test-takers can be noticed already 1998 and the effect (i.e. the decrease of the number of test-takers) has increased further after that.

As you all know SweSAT was introduced in 1977, and in 2002 the 25-year birthday was celebrated by The National Agency for Higher Education by inviting an international evaluation group consisting of John Fremer, ETS (at that time) David Lohman from IOWA, and Werner Wittman from Mannheim, Germany. Their general findings and common recommendations were:

1. The SweSAT is a high quality test with a solid research program. We applaud past research and encourage continued research on the test.

2. Uses for which the test was originally designed have changed. The initial purpose of the SweSAT was to provide an estimate of academic potential for young adults who had been in the workforce for several years, and whose academic grades were thought to underestimate the likelihood of their success in university study. Now the test is being used as an alternative selection mechanism for a substantial fraction of the applicant population. This requires a re-thinking of the test and the ways in which it is used.

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

4. Grade-point average should be decomposed. The predictive validity of secondary school grades is likely to be higher if universities took into account the subjects in which the grades were earned, possibly using different weightings of grades to predict success in different institutions or programs (e.g., giving more weight to math and science grades for predicting performance in science or engineering programs).

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

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

7. Consider strategies to establish better proportions of males and females (and minorities) in all areas by using profile information of grades for different secondary school subjects (or clusters of subjects) and the scores for components of SweSAT. This includes investigating how well the university curricula in the different areas could be redesigned to accommodate students with different ability profiles.

8. Investigate the possibility of adding more context-based measures of reasoning (i.e., achievement-oriented tests). Both context-reduced and context-rich measures of reasoning abilities can contribute usefully to decisions about the likelihood of future academic success. Assessing both is an especially attractive possibility if the test can be administered adaptively.

9. Explore the use of computer-based test.

10. In conjunction with (9), consider adding computer-scored measure of writing ability.

11. Anticipate the need to make accommodations for testing disabled students.

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

One big issue during last year was that a new government commission (headed by Lars Lustig, who is deputy secretary of the Student Services Center at Umeå university) was set up to revise the admission rules for higher education. Ewa will give us a summary of the content in the commission report.

The Government Commission on Admission to Higher Education: "Three routes to the open university"

Ewa Andersson

The point of departure was that the Swedish government stated that:

- Direct transition from upper secondary school to higher education should be increased.
- The reward of the efforts of upper secondary school students should be changed.
- The incentive to improve grades in adult upper secondary school, in order to increase competitiveness, should be reduced.
- Recruitment to higher education should be broadened.

"Admission rules have few, if any, friends.Almost everyone is dissatisfied with them almost all the time."

The admission system is supposed to:

- Predict study success
- Highlight desirable behaviour in students
- Help to achieve educational policy goals
- Be straightforward
- Be comprehensible
- Be stable
- Be fair
- Be legally secure
- Be cost effective
- Be legitimate

The commission highlights that we actually have two educational systems in one:

- 10 % of the study places are at highly selective institutions or programs. > 30 % of the applications refer to those programs
- 60 % of the study places are at institutions or programs characterized by a great demand for recruitment efforts. > 40 % of the applications refer to those programs.

The Commission´s conclusions about SweSAT are that:

- On the whole the test works well and in accordance with the original intentions, according to current research
- Over the years the test has got a positive reception, and also today there is great confidence in the test.
- Even in the future the test should be an important selection instrument

However, the commission highlights that:

- Research and evaluations conducted over the years show that some aspects of the test should be scrutinized.
- SweSAT should be adapted to developments on the educational field and within society as a whole.
- In future development of the test, the international development (especially the American) with testing should be followed.
- Other selection instruments, besides SweSAT and grades, should be developed.

Proposals for new admission rules:

- SweSAT should also in the future be a general selection instrument to higher education
- At least 30 % of the study places should be distributed on the basis of SweSAT results
- The “selection group” SweSAT in combination with credits for work experience should be abolished
- Consider to allow all students in upper secondary school to take one SweSAT free of charge during their last semester
- SweSAT should be developed in accordance with knowledge within the field of test and measurement
- SweSAT should be adapted to higher education of today and the applicants´ experiences

Tests adapted to persons with special needs

Margaretha Hallgren

Since autumn 1999 there is a special test for people with defective vision. The test is transferred to braille and cassette. This test is secret and exists in only two versions, and therefore this group of test takers are allowed to take the test only twice, once in spring and once in autumn. There are less than 10 test takers each year, which is the reason for only two opportunities. The test consists of 112 items distributed on four of the five subtests in SweSAT. Since it is not possible to transfer the subtest DTM to neither Braille nor cassette, this subtest is excluded, and instead the number of items in the DS subtest is increased to 30. The READ subtest has been decreased to 18 items (to four texts), while the subtest ERC is increased to 24 items, and the WORD subtest is the same number as in the regular test, i.e.40. The time for each subtest is twice as long as for the regular test-takers.

People with certified dyslexia (i.e. they must have an accepted certificate of their handicap) can make the regular test under special conditions. The time for each subtest is 50 per cent longer, and they do not have to do the pre-test block. The total testing time for this group is 5 hours, while the regular test takes 4 hours and 10 minutes. Since this opportunity was introduced, in 1999, there has been on average somewhat more than 400 test-takers a year. Since quite a few of these test-takers have also done the regular test, it is possible to compare the results; 75 – 85 per cent of these test-takers have got better results on the adapted test version.

Since the regular test is always given on a Saturday, there is once a year the possibility to do the test on a weekday. This is to make it possible to take the test for some groups who for religious reasons do not want to do it on a Saturday. This possibility is used by between three and ten people a year.

Comments

A composite score (GPA + SweSAT-score) would probably improve the prediction but it could also give unintended side-effects, which would favour the coaching companies. Furthermore if SweSAT were compulsory it would be still more high-stake, and hence further favour the coaching companies. In the US there is a correlation of .40 to .60 between SAT-score and first year performance, while the GPAs from different schools are not comparable.

SAT is not diagnostic but there is an effort to get more informative reporting scales; to give a profile of strengths and weaknesses of the test-takers. If the test is to be used diagnostically this should be built into the test, but that is very difficult and could even weaken the original aim of the test.

The wishes for more authentic measurement should be taken seriously. But the first step is to find a good definition. Experiences have shown that it has not always been an advantage to make tests more authentic, if they do not measure what they should, or do not categorize in the right way. In the US authentic tests have caused enlarged differences between ethnical groups. The development towards authentic measurements is a political goal, like measuring writing ability – this will not increase the predictive validity but only face validity.

The WORD sub-test. What does it measure?

Sandra Scott

The subtest is short:

- 40 items in 15 minutes
- words and short phrases or expressions
- Swedish words and a few loan words
- synonyms or hyponyms

Item format:

Word

A distractor

B distractor

C distractor

D correct answer

E distractor

Content from: (1) technical and natural science (2) administration, economy and society (3) medicine (4) culture and information (5) education

Criticism of the test:

- Isolated words, taken out of their context → no connection to reality or real life situations
- Discriminates students for other countries
- Stressful – 40 items in 15 minutes
- Too big place in the total test

Word comprehension is a complex mental process!

Mean of factor loadings for the SweSAT subtests administered between fall 1996 and spring 2002:

	WORD	DS	READ	DTM	ERC
F1	0.892	0.195	0.780	0.263	0.790
F2	0.073	0.889	0.356	0.855	0.377

What is the subtest WORD measuring?

A) Intelligence? Vocabulary tests often occurs in IQ-tests

B) Verbal abilities, like general reasoning ability?

C) General knowledge?

Or a little bit of everything?

Comments

Discrete words measure verbal ability as well as anything else, but have been abolished as item type in the US. If the test is very high-stake vocabulary tests tend to become memory-tests. Some types of words, however, may measure general knowledge.

The experience of writing/essay tests at ETS

Michal Beller

Some form of writing assessment is now included in every test at ETS. The format varies between: one essay (i.e. in SAT II), two essays (i.e. in GRE), three essays (i.e. AP English), Multple-choice and one essay (i.e. New SAT), and two or more writing tasks (i.e.NAEP).

The skills measured by the prompts given are:

- Articulation of complex ideas clearly and effectively
- Evaluation of claims and accompanying evidence
- Support of ideas with relevant reasons and examples
- Coherence and focus of discussion
- Control of the elements of standard written English

Multiple-choice prompts (for New SAT):

- Identifying sentence errors
- Improving sentences
- Improving paragraphs

Scoring rubrics:

- Holistic scoring (used by all writing assessments, Scale 1 – 6 with 1 point increment, score based on overall impression of quality of writing)
- Analytical scoring (E-rater)
- Human online scoring (OSN)

ETS Automated scoring capabilities:

- e-rater
- Critique^{SM} Writing Analysis Tools
- Grammar, Usage, and mechanics score
- C- rater^{TM}

Psychometrics of essays vs. multiple choice writing assessments:

Reliability typically lower for essays than multiple choice writing test, *but* essays are more predictive of writing skills used in real-life contexts than multiple choice writing tests *and* generally yield smaller group differences than multiple choice writing and verbal ability tests.

Demonstration of the OSN-system

Michal Beller & Sandy Abraham

The Online Scoring Network (OSN) system, which has been used at ETS since 1997, was demonstrated via internet by Sandy Abraham in situ at ETS and Michal Beller in Umea.

OSN Features:

- Scalable
- Reader diversity
- Home scorers
- Cost-efficient
- Quality-control built-in
- Certification/Calibration

- Monitoring (imbedded pre-scored papers)
- Scoring Leader Confirm/Override

- On-demand statistics

Test score reporting

Ronald Hambleton

1. Considerable investment of time and money has been made to address technical problems in large scale assessments – test specifications, IRT modelling and equating, reliability assessment, DIF analyses, and validation studies.

2. Surprisingly, given importance, test score reporting attracts very little research attention. – And yet, without clear and meaningful reporting of information, the other steps are of little value!

If educational testing were ice-hockey:

We would say the hockey player has all the moves but no finish – i.e. can´t score goals.

This is unfortunate because:

- Reporting scales are confusing for many persons (e.g. percent, percentile, IQ, SAT vs. ACT etc.)
- Quantitative literacy is not high among policy makers and the public
- Growing body of evidence highlights reporting problems

Sources of confusion:

- Proficiency scale vc proficiency category
- Standard errors
- < and > signs
- at or above
- footnotes
- meaning of the numbers

What does statistically significant mean?

- more than a couple of percentage points
- at least a 5-point increase
- statisticians decide using certain criteria
- when the results are important
- big differences and important

Conclusions from empirical studies:

- design data displays to address the questions the report was prepared to answer
- panellists preferences for displays are not always consistent with displays on which their performance is highest!

Recommendations:

- report results in multiple ways (e.g., using numbers, graphics, and narrative texts)
- highlight main findings from the assessment
- in student score reports, include all information essential to interpretation

Domain specific tests

Ingemar Wedman

The advisory council on access to higher education gives advice to The National Agency for Higher Education, concerning many aspects connected to admission to higher education. One of the aims is to follow the use of the SweSAT program. The following issues are presently discussed:

1. Diagnostic information from large scale testing like the SweSAT program (or similar programs).
2. Computerized versions of the SweSAT program.
3. Using written composition in large scale testing.
4. Is a country specific test always country specific?

Comments

Diagnostic tests should give hints on how to improve; give feed back to the students. The diagnostic information should be built into the test from start. However, it is possible to get some diagnostic information also from an existing test, i.e. by describing what the individual items are measuring, and advice the students what to do in order to improve. There must be enough items in order to build a reliable profile, and instructions about what to do must be specific. When building a diagnostic test you should answer the questions:

What do we want to know about a person?

What kind of evidence would give us what we want to know?

What kind of items would give us this evidence?

The advantages would be the possibility to add new types of items; shorter, and more flexible testing time; more information, since you could test more at the same time; immediate feed back. It is also popular. The main disadvantages are the costs; the enormous amount of items needed; and the big demand on testing facilities. It would only be worthwhile if new item types are introduced. It is more suited for low-stake than for high-stake tests. What would be achieved by computerizing the test?

A writing test would add nothing to the validity, and the reliability would drop. On the other hand writing tests have an important signal value and increase the face validity. It would possibly be worthwhile if e-rater could be applied to reduce costs by allowing for only one human rater.

Response time

Wim van der Linden

It has long been known that response times on test items are an important source of information on the person´s behaviour, but we had to wait for the advent of computer-based testing to make the recording of response times a routine part of test administration. Now that testing is widely computerized, the question of how to model response times has become urgent.

Two versions of a normal and lognormal model for response times of examinees were investigated. The models had a parameter structure analogous to the 2PL response models in IRT, with a parameter for the speed of each person as parameters for the time intensity and discrimination power of each item. It was shown how thw parameters can be estimated by a Markov chain Monte Carlo (MCMC) method (Gibbs sampler). The method was used to analyse response times for the adaptive version of a test from the Armed Services Vocational Aptitude Battery (ASVAB). We tested the validity of the models using posterior predictive checks on the response times. The lognormal model showed an excellent fit to the data, whereas the normal model appeared unable to allow for a characteristic skew-ness of the response time distributions. The addition of an equally constraint on the item discrimination parameters did not lead to appreciable loss of fit.

The potential advantages for the practice of testing of having good estimates of these parameters are numerous. In principle, any of current applications of IRT can be improved by using response times as an extra source of information on the persons and items. Examples of such applications are: (1) the use of response times as covariates in IRT calibration, (2) improved item selection in CAT using response times on the previous items in the test, (3) empirical determination of test speeded-ness, (4) empirical study of speed-accuracy trade off in testing, (5) person fit analyses using response times, for example, to detect cheating or pre-knowledge about the items, and (6) more accurate estimates of the extra time to be offered to students who need accommodation.

What is the opinion on SweSAT in Sweden?

I The views of some groups of admitted students

Ewa Andersson

Question:
SweSAT is intended to measure general ability to study at university and will therefore serve as a selection test for all university programmes. Do you, from that perspective, consider the different subtests to be relevant as selection test?

WORD	52 %	Social Work, seniors (68 %)
		Technical. Engin., seniors (33%)
DS	62 %	Medical Education, seniors (89 %)
		Social Work, seniors (44%)
READ	82 %	Business Adm., seniors (91%)
		Technical. Engin., freshmen (62%)
DTM	55 %	Business Adm., seniors (77%)
		Social Work, seniors (44%)
ERC	81 %	Business Adm., seniors (96%)
		Technical Engin. freshmen (62%)

Comments:

- The results indicate that all subtest are considered to be relevant by at least 50 % of the students.
- READ & ERC = Most relevant subtests
- WORD & DTM > Least relevant subtests
- Students at different education programmes view the subtests in different ways

Question:

Which subtest do you consider to be the *least* relevant in a selection test like the one described in the last question?

WORD	(46%)
DTM	(29%)

DS (17%)
READ (4%)
ERC (3%)

Social Work, freshmen	DS (33%), DTM (33%), READ (31%)
Social Work, seniors	DS (40%), DTM (32%), WORD (28%)
Medical Education, freshmen	WORD (57%), DTM (31%)
Medical Education, seniors	WORD (50%), DTM (31%)
Business Adm., freshmen	WORD (54%), DTM (27%)
Business Adm., seniors	WORD (46%), DTM (27%), READ (14%)
Technical. Engin., freshmen	WORD (54%), READ (15%), DS (15%)
Technical. Engin., seniors	WORD (39%), DTM (28%), DS (22%)

Comments:

- The results indicate that nearly half (46%) of the students considered WORD to be the least relevant subtest.
- Similarities and differences between study programs:
- Social Work, Technical Engineering, & Business Administration have more of a diversified view
- Medical Education has more of a concentrated view
- All of the students – except students at Social Work – seem to view WORD as the least relevant subtest

Question:

Which subtest do you consider to be the *most* relevant in a selection test like the one described in the first 25?

READ (68%)
DS (17%)
ERC (7%)
WORD (6%)
DTM (4%)

Social Work, freshmen	READ (82%)
Social Work, seniors	READ (78%), DS (13%)
Medical Education, freshmen	READ (74%), ERC (9%)
Medical Education, seniors	READ (69%), DS (17%)
Business Adm., freshmen	READ (66%), DS (21%)
Business Adm., seniors	READ (41%), DS (36%)
Technical Engin., freshmen	READ (38%), DS (31%), ERC (23%)
Technical Engin., seniors	READ (50%), DS (28%)

Comments:

- The results indicate that two out of three (68 %) of the students considered READ to be the <u>most</u> relevant subtest.
- Similarities and differences between study programs
- Business Administration, & Technical Engineering have more of a diversified view
- Social Work, & Medical Education have more of a concentrated view
- All of the students seem to view READ as the most relevant subtest

Overall conclusions (all questions)

Most relevant subtest:: READ

Least relevant subtest:: WORD

Females:

Most relevant: READ

Least relevant: WORD (DTM)

Males:

Most relevant: READ (DS)

Least relevant: WORD (DTM)

No difference between social groups, regarding views of which subtest they consider to be the most or least relevant, but….

students with parents where the highest level of education is compulsory school or lower secondary school (grundskola/realskola), seem to be more positive view toward all the subtests in SweSAT.

Biggest difference between social groups: WORD

II What are the test-takers opinions?

Stig Eriksson

Up until spring 1993 there were regular questionnaires, but much has happened since spring 1993:

- The subtests GI (General Information) and STECH (Study Technique) have been abolished
- The sub-test ERC (English Reading Comprehension) has been introduced
- The day of testing is divided into five sections, each 50 minutes
- One of the sections is a pre-test section
- New ways of information: SweSAT guide, and SweSAT homepage

The purpose of this pilot study:

- Lay ground for a new version of the questionnaire
- Renew and revise the questionnaire, relate it to the new circumstances
- Make it shorter, with a better focus
- Eliminate open answers
- Prepare for computerized scoring

BUT also to get information about: What are the opinions on the SweSAT of the test-takers of today?

Regular test day spring 2002. Group: Test-takers in Umeå. Questionnaire : 38 questions; 13 open answers. Returned: Males: 111, Females: 207

Some results:

Do You think the subtests of SweSAT are relevant?

	Yes			No	Doubtful
	M	F	Tot		
WORD	64.2	58.5	60.5	11.5	28.0
DS	70.6	58.8	62.9	10.9	26.2
READ	78.0	84.9	82.5	5.1	12.4
DTM	69.7	65.7	67.1	5.8	27.2
ERC	79.1	82.8	81.5	5.4	13.1

Which SweSAT subtest is the least relevant?

	Males	Females	Total
WORD	38.5	33.0	35.0
DS	18.3	32.0	27.2
READ	13.8	5.0	8.1
DTM	14.7	24.0	20.7
ERC	13.8	5.0	8.1

Which SweSAT subtest is the most relevant?

	Males	Females	Total
WORD	7.4	8.0	7.8
DS	15.7	7.5	10.4
READ	53.7	62.0	59.1
DTM	6.5	6.0	6.2
ERC	15.7	15.5	15.6

Are there types of subtests missing in the SweSAT?

YES: Males: 23,4 Females: 22,9

What do you think about the pre-test section?

	Males	Females
No objection	36,9	33,8
It takes too much time and energy	55,0	60,4
I would prefer another alternative	7,2	5,3

Alternative pre-test: Skip the section and put pre-test items in the subtest-booklets (+ more time) Would such an alternative suit you better?

	Males	Females
YES	41,8	50,2
NO	25,5	20,3
Don´t know	26,4	24,2
Other alternative.	6,4	3,9

III Opinions of some university study program directors on the subtest DTM.

Per-Erik Lyrén

Occurrence of graphical representations in student literature:

Extent/Type of graphical representation	None	Small	Rather large	Large
Tables	-	1	3	3
Bar charts, histograms	-	2	3	2
Pie charts	3	1	2	1
Curves	1	1	1	4
Maps	4	2	-	1
Other (e.g. flow charts	-	3	3	1

Relevance of the DTM subtest:

How relevant is the DTM subtest for giving a general prognosis of study success?

	not at all relevant	to a small extent	to a large extent	very relevant
within your field of education?	-	1	3	3
in general?	-	1	5	1

How important is it that the tasks and the graphical representations in DTM are relevant?

	not at all	less important	important	very important
to your study program	-	4	2	1
to the test-takers	-	1	3	2

Cognitive requirements:

How important is it that your students can use graphical information to

	not at all important	less important	important	very important
Identify information?	-	1	1	5
Do simple arithmetic based on given information?	-	1	1	5
Do rather complex arithmetic based on given information?	1	1	2	3
Compare and analyse given information?	-	1	1	5

Tentative changes in the DTM sub-test

Gunilla Ögren

The subtest DTM (i.e. Diagrams, Tables, and Maps) is an ability test intended to measure the ability to locate and process information found in various kinds of statistical and graphic material. The test consists of ten sets of figures with 20 multiple-choice questions. The format of DTM is fixed: each set of figures is followed by two items with five options each. However, the authentic material used in the subtest does not always fit this rather strict format. Although highly relevant, some sets of figures may not allow more than one item, while others could well be used for more than two items. Moreover, the character of the source material often makes it difficult to construct four distractors that are both plausible and incorrect. These observations have motivated three different studies with the following aims:

Study I: To study whether the character of the subtest changes if each set of figures is followed by three items instead of two, i.e. if the test taker has any advantage of having solved two items already when solving the third.

Study II: To study whether the subtest changes with regard to difficulty and measurement quality if the number of options in each item is reduced from five to four.

Study III: To validate the results of the two earlier studies by implementing both changes in the same version of the subtest.

Method: The three studies were made among students doing their third year on theoretical programs in upper secondary school.

In all of the three studies, special test versions were put together. These were based upon two try-out versions of the subtest DTM, each consisting of ten sets of figures followed by two items each (20 items in total). All ten sets of figures and ten out of 20 items were the same in both try-out versions, while the remaining ten items were unique to each version. The ten items found in both versions were so-called anchor items. These two try-out versions of DTM formed the basis for the test versions used in the three studies, which were put together as follows:

In **study I,** one of the try-out versions was used with five items from the other try-out version added to it. Thus, the test version used in study I consisted of 25 items in total, i.e. five sets of figures were followed by two items each and the other five sets by three items each.

Study II made use of the same try-out version as study I, but here one option was eliminated from each item. The test version used in this study consisted of 20 items with four options each.

Study III was intended to be based upon exactly the same testing material as study I and II, but as the items following one of the sets of figures had not functioned satisfactorily, this set was replaced in study III. The test version used in study III consisted of ten sets of figures and a total of 25 items. The number of items following each set varied from one to four. All items had four options each.

Main results

The results can be summarized in the following way:

- For some sets of figures it would be possible to increase the number of items from two to three without any interdependence between the items.
- If the number of options in the items is reduced from five to four, the average p-value becomes higher.
- If the number of items in the test is increased by five, the average p-value becomes lower.
- If the number of options in the items is reduced from five to four and the number of items in the test is increased from 20 to 25, the average p-value becomes about the same as compared to a test version consisting of 20 items with five options each.
- An increase in the number of items would compensate for a reduction in the number of options, provided that the number of sets of figures remains unchanged.
- An increase in the number of items does not seem to have any negative effects on the result differences between females and males.

After these studies our intention is to recommend that the test is changed to 25 items, 4 options and unchanged test-time.

The test would be more realistic if you allowed a more flexible use of the figures, it would also be both practically and financially advantageous if the material was used in a more flexible way.

As reliability generally increases with the number of items in a test, a DTM test containing 25 items instead of 20 would gain also in measurement quality. The results of this study do not indicate that a larger number of items lead to greater differences in p-values between males and females.

Models for predicting item difficulty from pre-test data on the WORD and DS sub-tests.

Anders Lexelius

Pre-testing history:

Before 1996:	Since 1996:
Number of test-takers:	Number of test-takers:
100 – 150 per/item	1000 – 2000 per/item
No motivation	Motivation
Pre-test in upper secondary school	Pre-test within the regular test

Since the pre-test booklets are not randomly, but geographically distributed, there are some pre-test groups, which perform above, and other groups, which perform below the overall average results. In this study three correction models are compared on the WORD and DS subtest. These subtests were chosen since in both there is independence between the items, and anchor items have been used in the pre-testing.

Three correction models:

- Constant model. $C=(\text{Mean}_{\text{total group}} - \text{Mean}_{\text{pre-test group}})/n$
- Linear regression model.
- Structural equation model

Table 1. The sub-test DS: pre-test values and predicted averages on seven test occasions.

Occasion	Pre-test	Constant	Linear	SEM
1	12.49	11.83	11.87	11,94
2	12.29	12.06	12.03	12.08
3	11.97	11.98	11.98	11.95
4	12.29	11.87	11.81	11.90
5	13.11	12.86	12.82	12.88
6	12.96	11.67	11.62	11.78
7	11.36	11.62	11.53	11.62

Table 2. The subtest WORD: pre-test values and predicted averages on 15 anchor items, used in nine test occasions.

Occasion	Pre-test	Constant	Linear	SEM
1	8.56	8.71	8.65	8.63
2	8.78	8.78	8.80	8.81
3	8.49	8.94	8.88	8.82
4	9.81	9.51	9.35	9.53
5	9.68	9.23	9.15	9.28
6	8.37	8.67	8.61	8.56
7	9.68	9.08	9.08	9.14
8	8.84	8.99	8.98	8.98
9	8.13	8.58	8.57	8.53

The effects of differential scoring on different groups

Mats Hamrén, Christina Jonsson

There have been complaints about the weight of the subtest Word on the total test score. 40 items, answered in 15 minutes, constitute 33 per cent of the total test score. In this small study, the effects of two different models for weighting the subtests were compared with the un-weighted total score for gender, age groups, groups with different educational background and, groups with different social backgrounds.

Models:

I ½ Word + DS + READ + DTM + ERC

II WORD + 2(READ + ERC) + 3(DS + DTM)

Results:

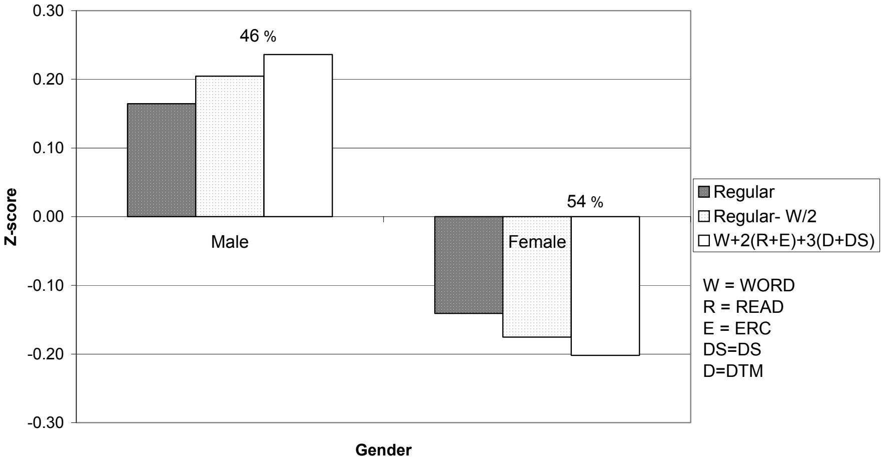


Figure 1. Results for males and females of the two weighting models in comparison with the regular test score.

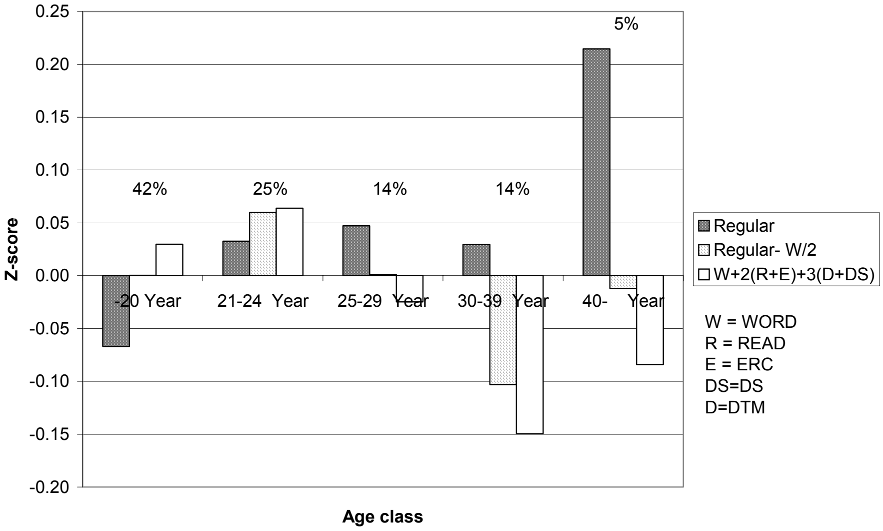


Figure 2. Results for different age groups.

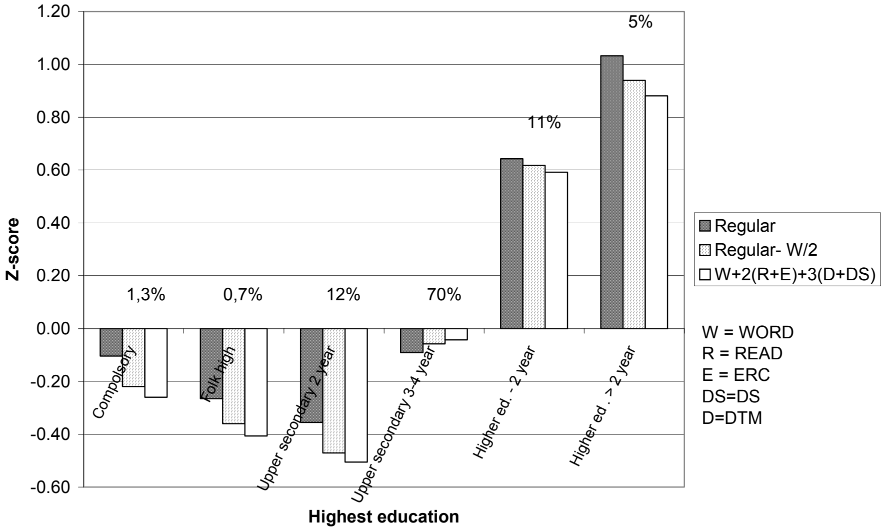


Figure 3; Results for groups with different educational backgrounds.

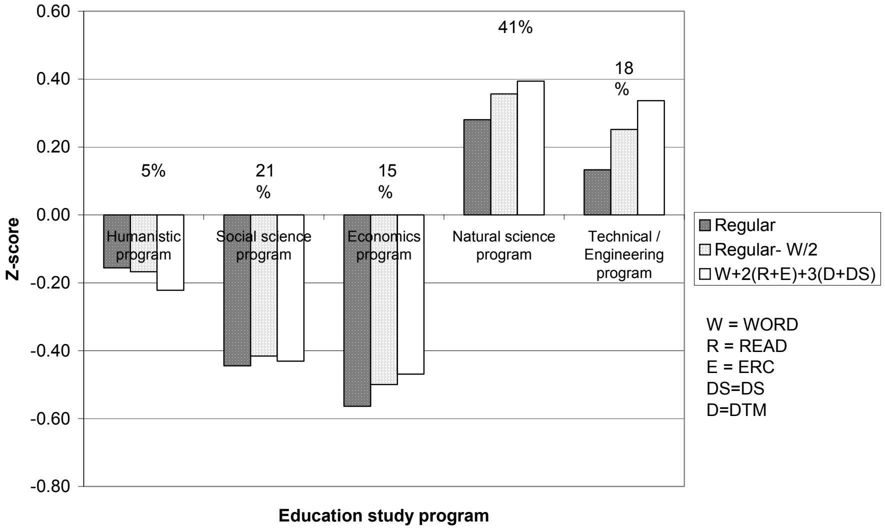


Figure 4. Results for groups from different study programs in upper secondary school.

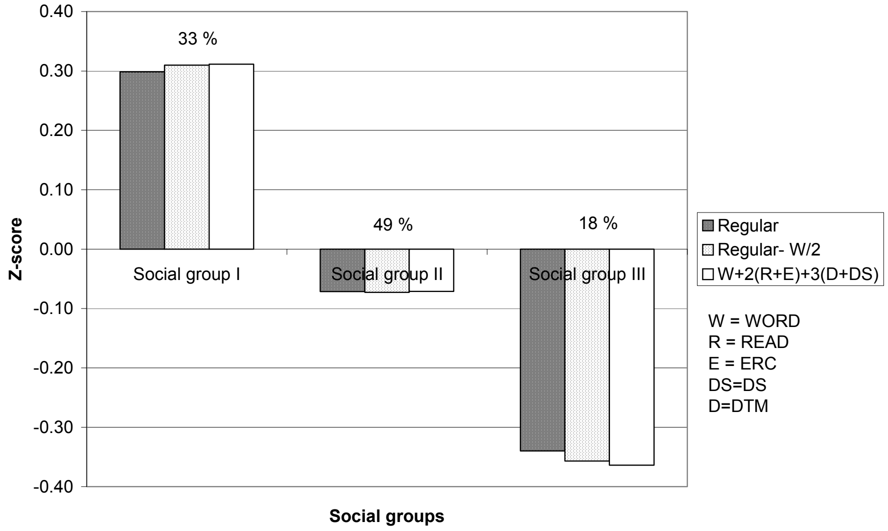


Figure 5. Results for different social groups.

SweSAT divided into one verbal and one quantitative part – the effects on admission

Nils Olsson

The aim of this study was to investigate whether a division of SweSAT into one verbal and one quantitative part would have any effect on the admission to two different study programs. The three sub-tests WORD, READ and ERC made up the verbal part, and the two sub-test Ds and DTM made up the quantitative part. The study programs chosen were the law program and the engineer program. For the applicants to the civil engineer program, only the quantitative sub-tests were used, and for the applicants to the law program only the verbal sub-tests were used. The outcome regarding admittance on this divided SweSAT score was compared with the regular admittance.

The results were that when only the verbal part was used for admittance to the law program there were very small changes in the admitted group. As for the engineer program, where the quantitative part only was used, there was a change in the distribution between females and males. !6 per cent less females were accepted on their test scores. On the other hand, if the total admitted group was regarded, i.e. those admitted on basis of grades as well, the difference became much less. In the total group there were only two per cent more males admitted by using only the quantitative sub-tests in SweSAT.

SweSAT results for different social groups

I Differences on item level

Christina Stage

The main aim was to study differences between social groups on item level. Since results from gender studies have had some impact on the item construction a second aim was to compare the results of social classes with the results of males and females. The tests spring 1992 and spring 1997 were analyzed by means of the Mantel-Haenszel method on social groups and on males/females.

Table 1. Group differences in grades and test results.

	males/females	effect size	social groups	effect size $_{\text{I-III}}$
Sec. school	F > M	≈ .40	I > II > III	≈ .80
Up.sec. sch.	F > M	≈ .35	I > II > III	≈ .65
SweSAT	M > F	≈ .36	I > II > III	≈ .60
WORD	M ≈ F	≈ .0	I > II > III	≈ .60
DS	M > F	≈ .50	I > II > III	≈ .60
READ	M ≈ F	≈ .0	I > II > III	≈ .60
DTM	M > F	≈ .60	I > II > III	≈ .60
ERC	M > F	≈.40	I > II > III	≈ .60

Table 2. Number of test-takers, distribution on different groups and average test results in spring 1992 and spring 1997.

Group	Tot	S I	S II	S III	Females	Males
1992	N= 16 354					
%	100	37	49	14	55	45
Test sc.	.97	1.07	1.01	.86	.87	1.11
2002	N= 9 723					
%	100	30	46	20	54	46
Test sc.	.88	1.05	.88	.71	.81	.97

For both test versions there were very few items, which showed DIF for social groups. The only sub-test where B-items were found for social groups was the WORD sub-test. The Word sub-test was also where the most C-items for females and males were found. In Table 3 the number of items that showed significant DIF for females or males are shown divided into the categories C (serious DIF) B (moderate DIF, A (negligible DIF), and non significant.

Table 3. Outcome of Mantel-Haenszel analyses of the WORD sub-tests spring 1992 and spring 2002, for males and females, and social groups I and III.

Favour females 1992	social groups	Favour males	social groups
6 C-items:	1 B_I 2 A_I 1 A_{III} 2 not sign	2 C-items:	1 A_{III} 1 not sign
2 B-items:	2 A_{III}	1 B-item:	1 not sign
9 A-items:	2 A_{III} 7 not sign	3 A-items:	1 A_I 2 not sign
2002			
4 C-items:	1 B_I 1 A_I 2 not sign.		
6 B-items:	2 A_{III} 4 not sign.	1 B-item:	1 not sign
10 A-items:	10 not sign 10 not sign	7 A-items:	4 A_{III} 3 not sign

No real tendency was found in the results except for the two items, which were B-items in favor, of social group I, and A-items in favor of females. Both these items were fancy dishes (aioli and parfait).

II Differences in repeated test taking

Birgitta Törnkvist

The aim of this study was to investigate the effects of the social group of the test-taker and its relation to repeated test taking. On the one hand whether there is a difference concerning the willingness to repeat SweSAT, and, on the other hand, whether changes in observed score are related to social group.

Table 1. Social group and sex distribution of the 20 415 test-takers born between 1972 and 1983, who took the SweSAT in fall 2000 as their first, second, third or fourth test.

Number of tests	Social group I	Social group II	Sovial group III	N	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	13	422	40
N	6 522	9 868	4 025	20 415	10 850
%	32	48	20	100	53

From earlier studies it is known that test takers with high scores repeat the test more often than test takers with low scores. In this study the mean score on the first test for the test takers, who had taken the test four times, was 0.97, and the mean score on the fourth test was 1.19. The mean score of the test-takers in this study, who took the test for the first time, was 0.79.

Table 2. Mean normed score (M), and number of test-takers at the first test for test-takers in different social groups who took SweSAT as their first (1) and fourth (4) test respectively, in fall 2000.

Social group	Sex	1 test M	N	4 tests M	N
I	Male	0.94	2 126	1.07	119
	Female	0.87	2 318	1.01	61
	Total	0.90	4 444	1.05	180
II	Male	0.82	3 288	0.96	108
	Female	0.74	3 948	0.93	81
	Total	0.78	7 236	0.95	189
III	Male	0.73	1 250	0.94	28
	Female	0.63	1 850	0.60	25
	Total	0.67	3 100	0.78	53
Total	Male	0.84	6 664	1.01	255
	Female	0.75	8 116	0.91	167
	Total	0.79	14 780	0.97	422

The analyses of the effect of repeated test taking on the mean scores are based on multivariate models with the variables sex, age, and social group controlled for.

As may be seen in Figure 1 the strongest effect of repeated test taking was achieved from the first to the second test for all social groups and for both sexes. The changes in mean scores are similar in all compared groups except at the last test administration for social group III

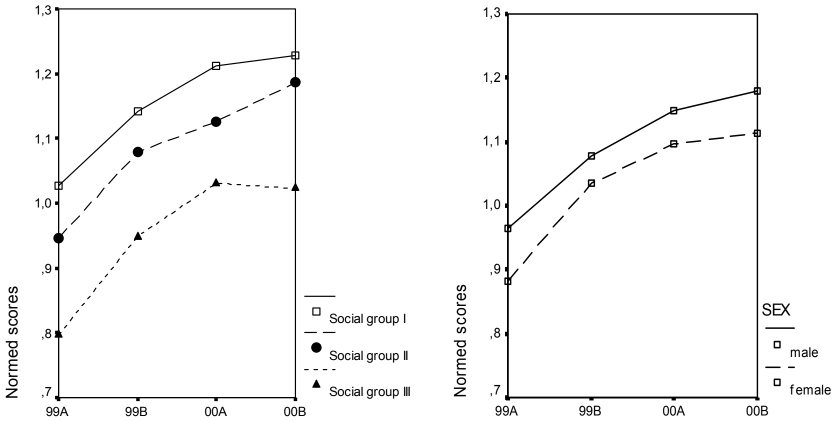


Figure 1. Mean normed SweSAT scores for different social groups, when controlled for sex, and for males and females respectively, when controlled for social background. Evaluated at age 20 years.

On the sub-test level the differences between the mean scores for the different social groups and sexes are about the same for the sub-tests DS and DTM, but differ for the sub-tests READ and WORD in comparison with the mean normed scores.

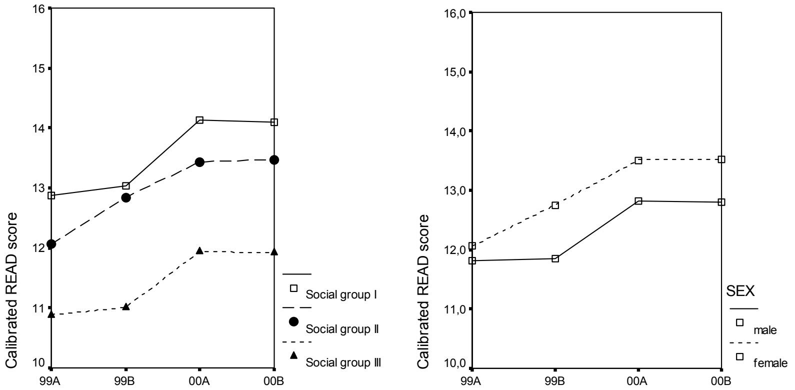
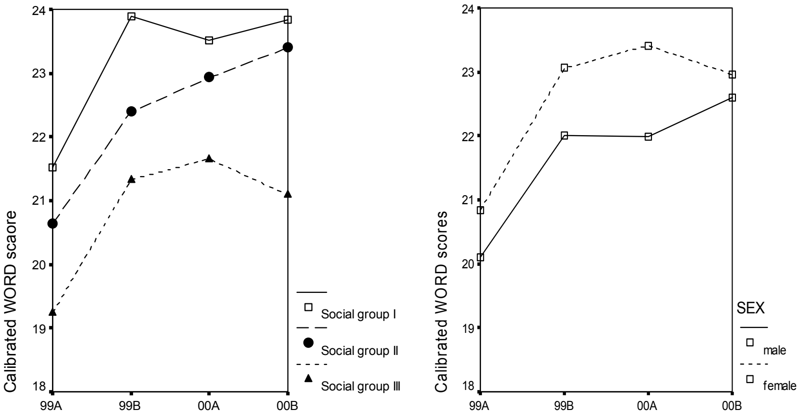


Figure 2. Mean calibrated score for subtest READ for different social groups, when controlled for sex, and for males and females, respectively, when controlled for social group. Evaluated at age 20 years old.



Figur3. Mean calibrated score of subtest WORD for different social groups, when controlled for se, and for males and females respectively, when controlled for social group. Evaluated at age 20 years old.

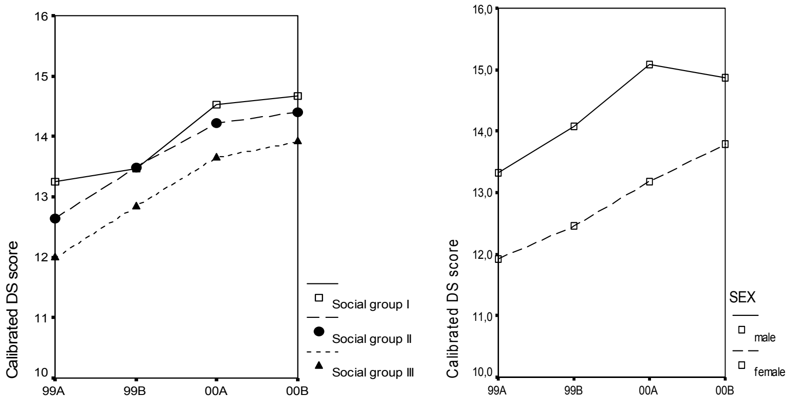


Figure 4. Mean calibrated score for the sub test DS for different social groups, when controlled for sex, and for males and females, respectively, when controlled for social group. Evaluated at age 20 years old.

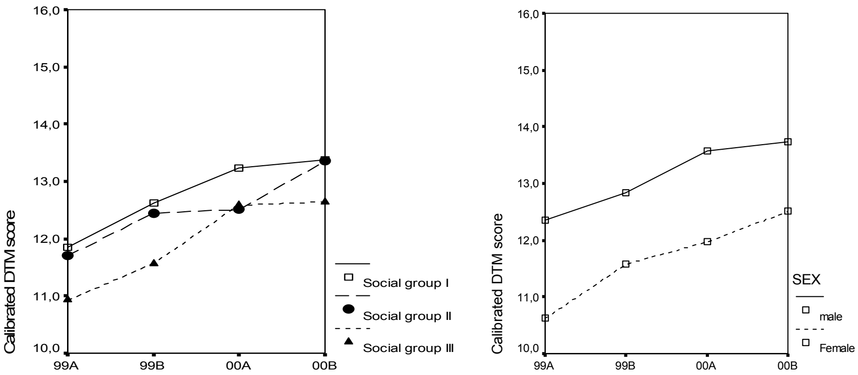


Figure 5. Mean calibrated scores for the sub test DTM for different social groups, when controlled for sex, and for males and females, respectively, when controlled for social group. Evaluated at age 20 years old.

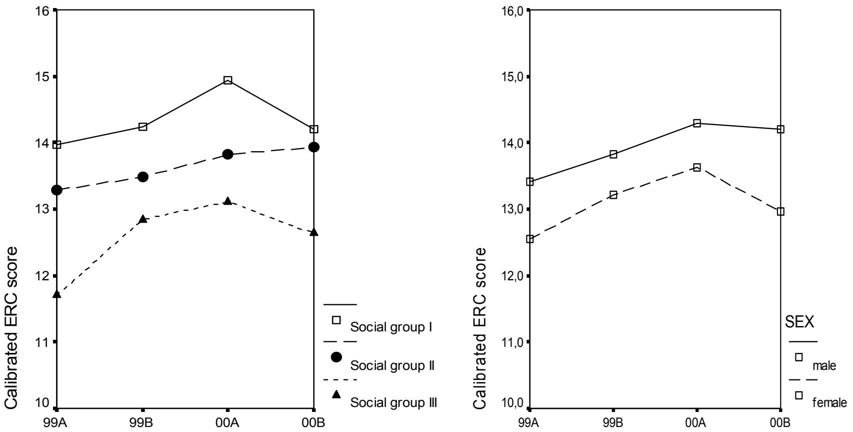


Figure 6. Mean calibrated scores for the sub test ERC for different social groups, when controlled for sex, and for males and females, respectively, when controlled for social group. Evaluated at age 20 years old.

Conclusions: Self-selection is operating; test-takers with high SweSAT scores repeat the test more often than those with low scores. Males repeat more often than females. Social group I repeat more often than social groups II and III.

Prediction studies

I *Study progress in the Civil Engineer Program*

Allan Svensson

The aim was to study if there are any differences in academic performance among students admitted on the basis of their:

- grade point average from upper secondary school (GPA)
- SweSAT-scores (SweSAT)
- SweSAT-scores with additional scores for work experience (SweSAT + WE)

Academic performance is defined by the number of credit points obtained after each academic year. (Sweden has a system of credit points, where one term of successful full-time studies with a work-load of 40 hours a week can yield 20 credit points, and one year yields 40 credit points).

Table 1. Students included in the study.

Enrolment	Year of birth	Age limit at enrolment	Follow up time
1993	72 – 76	21 years	7 academic years
1994	72 – 77	22 "	6 " "
1995	72 – 78	23 "	5 " "
1996	72 – 79	24 "	4 " "
1997	72 – 80	25 "	3 " "
1998	72 – 81	26 "	2 " "
1999	72 – 82	27 "	1 " "
2000	72 – 83	28 "	0 " "

Table 2. The number of admitted students within different selection groups according to enrolment year.

Enrolment	GPA	SweSAT	SweSAT+WE	Total
1993	2 449	874		3 323
1994	2 591	1 021	1	3 613
1995	2 975	1 181		4 156
1996	3 168	1 171	10	4 349
1997	3 299	1 237	20	4 556
1998	3 571	1 289	41	4 901
1999	3 685	1 448	60	5 193
2000	3 929	1 607	82	5 618
Total	25 667	9 828	214	35 709

The social background of the students has been classified on the basis of information about the occupation 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
Group IV.	No information available

Table 3. The proportion of students from different social groups. Per cent

Group	GPA	SweSAT	SweSAT+WE	Total
I	45	49	29	46
II	43	42	54	43
III	9	8	13	9
0	3	1	4	3
Total	100	100	100	100

Table 4. Credits obtained during the first academic year for different selection groups. Data based on students enrolled from 1993 to 1999.

	GPA		SweSAT		SweSAT+	WE*
	M	sd	M	sd	M	sd
Women	28.55	10.98	25.25	12.22	19.53	12.44
Men	29.44	11.60	24.93	12.77	24.11	13.21
Soc gr.I	29.76	11.17	25.45	12.67	22.12	12.19
Soc.gr.II	29.32	11.34	24.85	12.53	22.66	13.74
Soc.gr.III	27.19	12.09	23.79	12.89	25.31	12.81
Total	29.19	11.43	24.99	12.66	23.00	13.13

- This group is very restricted. It contains 32 women and 100 men; 42 students from gr.I, 64 from gr. II, and 21 from gr. III.

Table 5. Credit points obtained after the first to the fifth academic year among students who started 1993, 1994, and 1995.

	Selectiongroup		
Academic year	GPA	SweSAT	Diff
One	28.91	24.64	4.27
Two	33.22	29.46	3.76
Three	31.55	29.24	2.31
Four	34.04	31.66	2.38
Five	28.81	27.32	1.49

II *Study Progress in the Economics and Social Study Programs*

Kent Löfgren

The aims and designs of these studies were the same as in Allan´s study.

Table 1. Enrolment year and group of admittance for the social study program.

Year	GPA	SweSAT	SweSAT+WE	Total	N
1993/94	79.1	20.9	0.0	100	320
1994/95	75.5	24.5	0.0	100	355
1995/96	75.4	24.6	0.0	100	346
1996/97	71.8	27.9	0.0	100	383
1997/98	74.8	23.1	0.3	100	432
1998/99	69.6	25.0	2.1	100	541
199/00	70.5	25.4	4.1	100	579
Total	73.2	24.6	2.2	100	2 956

Table 2. Enrolment year and group of admittance for the economics study program.

Year	GPA	SweSAT	SweSAT+WE	Total	N
1993/94	76.0	23.9	0.1	100	1 538
1994/95	72.1	27.9	0.1	100	2 451
1995/96	86.6	31.3	0.1	100	2 518
1996/97	67.4	31.5	1.0	100	2 784
1997/98	66.9	31.5	1.7	100	3 098
1998/99	63.8	31.3	3.4	100	3 385
199/00	64.4	32.8	4.0	100	3 619
Total	67.6	31.6	1.8	100	19 393

Table 3. The proportion of students from different social groups; social study program

Group	GPA	SweSAT	SweSAT+WE	Total	N
I	26.1	36.0	11.1	28.2	833
II	49.1	47.8	61.9	49.1	1 451
III	22.8	14.6	27.0	20.8	616
0	2.0	1.6	0.0	1.9	56
Total	100	100	100	100	1 956

Table 4. The proportions of students from different social groups; economics study program

Group	GPA	SweSAT	SweSAT+WE	Total	N
I	37.2	42.6	27.4	38.6	7 494
II	47.2	45.5	55.8	46.8	9 079
III	13.1	10.6	16.0	12.4	2 404
0	2.5	1.4	0.9	2.1	416
Total	100	100	100	100	19 393

Table 5. Credit points obtained during the first academic year for different selection groups, and separately for students enrolled in Spring and in Autumn. The social study program

Enrolment	Autumn		Spring	
year	GPA	SweSAT	GPA	SweSAT
1993/94	34.5	35.8	18.6	17.6
1994/95	35.8	37.8	18.7	18.0
1995/96	37.4	37.7	18.8	17.4
1996/97	37.9	36.5	17.6	16.8
1997/98	37.6	35.8	19.2	18.7
1998/99	38.2	37.9	18.6	18.2
1999/00	36.7	36.2	18.5	17.7
Total	36.7	36.2	18.5	17.7
N	1 127	388	1 004	333

Table 6. Credit points obtained during the first academic year for different selection groups, and separately for student enrolled in Sprin and in Autumn. The economics study program.

Enrol	Autumn			Spring		
Year	GPA	SweSAT	Swe+We	GPA	SweSAT	Swe+We
1993/94	32.8	29.8	-	16.0	15.0	-
1994/95	32.5	30.1	26.8	16.0	14.6	-
1995/96	31.4	30.1	25.3	15.8	15.6	-
1996/97	30.6	28.5	32.1	15.6	15.4	15.9
1997/98	31.2	27.1	31.2	23.3	23.9	17.2
1998/99	30.7	28.6	30.4	16.6	14.3	15.9
1999/00	30.5	27.1	29.7	16.7	15.0	17.8
Total	31.3	28.5	30.2	17.9	17.1	16.4
N	10 936	4 996	289	2 026	873	59

Grades and test results

Anders Lexelius

The database

- 73 773 students with examination certificates from 1997
- 37 088 females 36 685 males
- 63 423 born 1978 7 880 born 1977
- Courses-level-credits-marks
- SweSAT 97A, total test and sub test scores for 19596 students

Math courses, level and credits*

- Mathematics, A – 110 (Core subject)
- Mathematics, B – 40
- Mathematics, C – 50
- Mathematics, D – 40
- Mathematics, E – 60

English courses, level and credits*

- English, A – 110 (Core subject)
- English, B – 40
- English, C – 30

Swedish courses, level and credits*

- Swedish, A – 80 (Core subject)
- Swedish, B – 120 (Core subject)
  Language B1
  Literature B2
- Swedish, C – 50
  Written and oral communication

* One credit is approximately one hour of teaching

Table 1. Correlations between English and SwAT 97A

	WORD	DS	READ	DTM	ERC	Tot
Eng A N=19535	0.58	0.33	0.48	0.36	0.61	0.60
Eng B N=18867	0.61	0.31	0.50	0.36	0.65	0.62
Eng C N= 6146	0.44	0.25	0.38	0.27	0.49	0.47

Table 2. Correlations between mathematics and SweSAT 97A

	WORD	DS	READ	DTM	ERC	Tot
Math A N=19583	0.33	0.58	0.40	0.57	0.35	0.54
Math B N=18810	0.27	0.49	0.35	0.48	0.30	0.46
Math C N=16791	0.29	0.50	0.36	0.48	0.30	0.47
Math D N=9221	0.27	0.40	0.30	0.40	0.26	0.40
Math E N=7778	0.27	0.41	0.31	0.40	0.27	0.41

Table 3. Correlations between Swedish and SweSAT 97A

	WORD	DS	READ	DTM	ERC	Tot
Swe A N=19576	0.41	0.24	0.37	0.28	0.36	0.42
Swe B1 N=19566	0.44	0.26	0.40	0.29	0.39	0.45
Swe B2 N=19566	0.38	0.21	0.35	0.23	0.33	0.38
Swe C N=6580	0.34	0.25	0.33	0.26	0.32	0.38

The validity of the GPA

Christina Wikström

An empirical study of the effects of programme enrollment on GPA in Swedish upper secondary schools.

This study investigates the effect of programme enrolment on the grade point average, with focus on its function as selection instrument to higher education. Two questions are raised. First, are students graded differently, depending on which programme he or she is enrolled in? Second, how does the course composition in different upper secondary programmes affect GPA and hence the student ranking? The empirical analyses are based on data from Swedish elementary and upper secondary schools. Grades from compulsory courses in core subjects as well as overall GPA are investigated. The results show that theoretically oriented programmes are harder graded than programmes with a vocational orientation. It also shows that the course composition in different programmes does affect the students GPA and hence their rank positions. Students in theoretically oriented programmes are affected negatively, while students in practical- and vocationally oriented programmes are affected positively. The conclusion is that students in theoretically oriented programmes are “punished” twice, which affects their competitive strength when competing for attractive study positions at university level. First, by being harder graded than students in other programmes, and second by having to include a larger proportion of grades from harder and more hard grading courses in their GPA.

Appendix 1

Program for the 10th SweSAT Meeting

Tuesday June 1st Faculty Club, University Campus

09.00 Welcome and opening address

The SweSAT program during the last years

Coffee

Government Commission on admission to higher education (Ewa)

Tests adapted to persons with special needs (Margaretha)

Comments on the commission proposals (Nils, Michal)

Discussion

12.00 Lunch: Hotel Björken

13.00 The WORD subtest. What does it measure? (Sandra)

The experience of writing/essay tests at ETS (Michal)

Coffee

14.30 Demonstration of the OSN system (Michal)

16.00 Bus via Uman Home Hotel to the Elk Farm

19.00 Dinner at the Elk Farm

Wednesday June 2nd The Comfort Uman Home Hotel

09.00 Test score reporting (Ron)

Coffee

10.15 Domain specific tests (Ingemar) Discussion

12.15 Lunch: Comfort Home Uman

13.15 Response time (Wim)

What is the opinion on SweSAT in Sweden? (Ewa, Per-Erik, Stig)

Tentative changes in the DTM subtest (Gunilla)

Coffee

Models for predicting item difficulty from pre-test data (Anders)

The effects of differential scoring on different groups (Mats, Christina J)

SweSAT divided into one verbal and one quantitative part – the effects on admission (Nils)

18.30 Dinner at Sjöbris

Thursday June 3rd Faculty Club, University Campus

09.00 SweSAT results for different social groups (Christina S, Birgitta)

Prediction studies (Allan, Kent)

Coffee

Grades and test results (Anders)

The validity of the GPA (Christina W)

13.00 Lunch: Sävargården

Appendix 2

Participants

Advisory board:

Michal Beller, USA (Israel)
Ronald K. Hambleton, USA
Wim van der Linden, The Netherlands
Allan Svensson, Gothenburg
Widar Henriksson, Umeå
Christina Stage, Umeå

The National Agency for Higher Education:

Håkan Forsberg, (Wednesday, Thursday)
Margaretha Hallgren
Nils Olsson,
Ingemar Wedman, The Council on Access to Higher Education (Tuesday pm., Wednesday am.)

The SweSAT program, Umeå:
Stig Eriksson
Mats Hamrén
Christina Jonsson
Ingegerd Jonsson
Jenny Lindberg
Per-Erik Lyrén
Sandra Scott
Marit Sigurdsson
Gunilla Ögren

The VALUTA-project, Umeå:
Ewa Andersson
Kent Löfgren
Christina Wikström

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 UmeD, 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