Notes from the Twelfth International SweSAT Conference

Umeå, June 15–17, 2008

Christina Stage

EM No 63, 2008
ISSN 1103-2685

Preface

The Swedish Scholastic Assessment Test (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 twelfth meeting of the scientific advisory board. A list of 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 SweSAT Program during the last two years

Christina Stage

This board has by now been in existence for 16 years, and this is the twelfth meeting, and all five guests, Michal, Ron, Wim, Jan-Eric, and Allan have been with us since the beginning. We are very fortunate to have such distinguished advisors. **You are most welcome to this twelfth meeting.**

Of the original members of this board three have by now sadly passed away. Professor emeritus Sten Henrysson died in 1998, Professor Sven-Eric Reuterberg died in 2003, and as you all know in January this year Professor Ingemar Wedman died only 62 years old. I hope you will join me in a minute of silence in the memory of them.

Today SweSAT has been in existence for 31 years, and more than 2 million tests in 64 different versions have been administrated to more than a million unique test-takers. These are big numbers in a small country like Sweden.

When SweSAT had been in existence for 25 years The National Agency for Higher Education arranged an international evaluation, which was carried out by John Fremer, David Lohman, and Werner Wittman. As you may remember the outcome of the evaluation was rather positive. (See *Notes from the Tenth International SweSAT Conference*, Em 52:2004)

The year before SweSAT had been in existence for 30 years the National Agency arranged a procurement of two of the sub-tests Data Sufficiency (DS) and Diagrams, Tables, and Maps (DTM). They wanted to find someone willing to construct the sub-tests to a better price, given some specified, quality demands. This became a very long and cumbersome affair, which took plenty of time and effort for us, as well as for the National Agency. There turned out to be one competitor to us, a private agency. In November the evaluation of the answers was finished, and the outcome was that we had higher quality scores on both tests, while the private agency had lower prices on both tests. After having done a weighting of quality and price the decision was that we got the cheaper test (DS), while the private agency got the more expensive test (DTM).

Since we found the evaluation very biased, we appealed against this decision in the County Court (in Stockholm), In February 2007 the judgement was given that we had won. According to court the private company had not fulfilled the qualitative requirements, and should be excluded from the evaluation.. The National Agency in turn appealed against this decision in the next court, the Administrative County court. Already in March the verdict came that the decision of the county court would not be changed. The

National Agency appealed again to the next court, Supreme Administrative Court. In July the final decision came, and we got both the sub-tests. It should be noted, however, that the people at the National Agency were, never of one opinion regarding this procurement.

In August Gunilla and I were invited to National Agency to give a seminar about SweSAT. A lot of people attended the seminar, and SweSAT was described as the flagship of National Agency.

As may be seen in Figure 1, the number of test-takers has continued to decrease. But this spring the number was only about one percent lower than last spring. The positive thing, however, is that in Stockholm the number had increased, and was 10 percent higher than last spring, and Stockholm is usually ahead of the rest of the country. So it looks somewhat more hopeful for the future

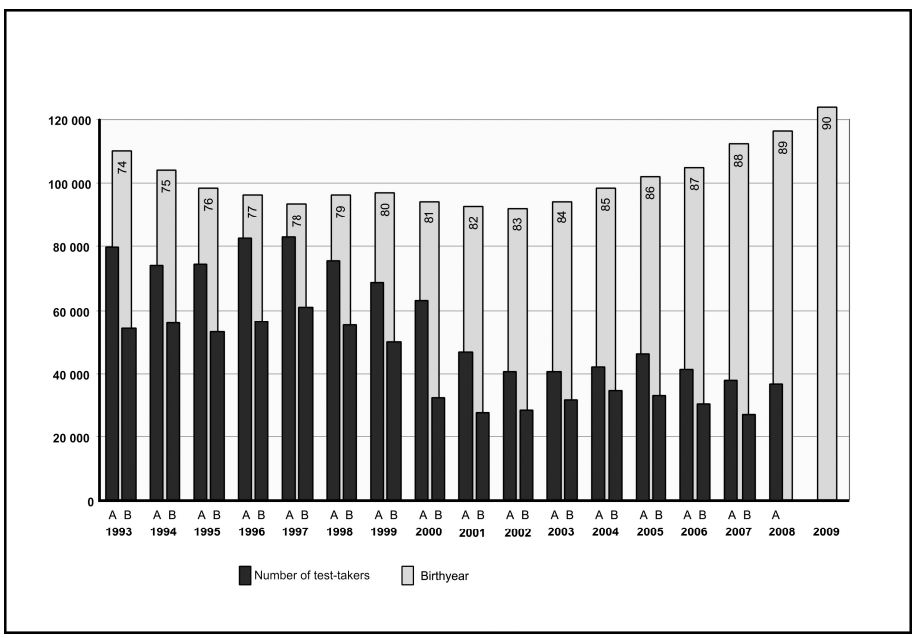

Figure 1.The number of test-takers

We have got an assignment by the National Agency to construct and field-test three types of new items (new for SweSAT), and to give recommendations for a changed test. This is going to be the main subject of this meeting, and we are looking forward to get your opinions and your advice tomorrow.

Now Widar, who has been asked by the dean of faculty to return as head of our department, will tell you something about what has happened and eventually will happen at the department.

News from the Department of Educational Measurement

Widar Henriksson

Some quick facts about Uneå University:

There are 28 724 students of which 62 per cent are females, and of which 1 359 are doctoral students (53 % females)

The median age of the undergraduate students is 29.9 years.

The number of employees is 4 143 of which 52 per cent are females.

There are 262 professors of which 20 per cent are females.

There are five faculties and 50 departments/units and 21 research centres.

An evaluation of the faculty of Education was ordered by the vice chancellor, and performed by Sigrid Bömeke, Ulf P. Lundgren and Roger Säljö, three very respected educational researchers in Sweden. The Report *Research in Educational Sciences at Umeå University – An Evaluation* was presented in December 2007. The main conclusion was that the scientific quality was poor, with a few exceptions (among them educational measurement).

The consequence of this evaluation was that the faculty of Teacher Education was dissolved, and the concept School of Education (SoE) was established. The departments, units, centres, and teachers at the former faculty of teacher education will be distributed among the remaining four faculties (which will need a considerable work). The SoE will define courses and the departments within the university can offer plans for teaching the courses, and motivate why they are the most suitable. A SoE committee will review the proposals and decide which department should get it.

Research grants will be distributed in accordance with the evaluated level of scientific quality A, B or C. Four areas are defined as level A, while the B and C levels are still not defined.

There are clear indications that the number of departments, units etc at the faculty of social science must be reduced, by merging of the present 14 departments, and different centres, and units into larger departments.

News from the Advisory Council on Access to Higher Education

Jan-Eric Gustafsson

The tasks of this council is to give advice regarding:

1. SweSAT
2. Development of domain tests
3. Rules for special eligibility
4. Rules for exceptions of standard rules of eligibility
5. Use of alternative selection methods
6. Principles of validation of competence
7. Current admission issues

Earlier grades and SweSAT were the only selection instruments, but the interest for alternative selection has increased, and in September there will be a conference on this topic.

The present members of the council are:

- Jan-Eric Gustafsson (chairman, Gothenburg Iniversity)
- Lars-Ove Farnebo (the Karolinska institute)
- Mohammad Fazlhashemi (Umeå University)
- Anna Petersson (National Association of Student Unions in Sweden)
- Sophie de Goude (National Association of Student Unions in Sweden)
- Sara Thyberg (the Royal Institue of Technology)
- Elisabeth Svensk (the National Agency for Higher Education)
- Jan Sydhoff (the National Agency for Education)
- Jörgen Tholin (University College of Borås)

New rules for eligibility and selection from 2010

- Domain eligibility: set of requirements in term of courses in similar educations along with identification of courses which give “merit points”.

- Three main grounds for selection: Grade (minimum 1/3), SweSAT (minimum 1/3), and locally determined selection grounds (30%)

- Students who have supplemented grades will be placed in a special quota group, with few places.

What does a Chief Research Scientist at CTB/McGraw Hill do?

Wim van der Linden

CTB – California Testing Bureau was bought by McGraw-Hill some 20 years ago, and the company has in all 20 000 employees. It is engaged in education and publishing, information and media, school education, and test programs. It produces 20 million test forms a year.

The work is very much the same as for a university professor, only no staff and no students: scrutinizing soft-wear to find flaws, producing new soft-wear to launch. Help out in psychometric emergencies.

Two main projects are

- To develop CAT which, hitherto, never has been produced at CTB.
- Improve automatic test assembly (including developing software), for which there is a huge request.

Own research on law school admission and response time.

RAMA – The National Authority for Measurement and Evaluation in Education

Michal Beller

RAMA is a task force for the advancement of education. “If you want to change the world, you must change education”

RAMA´s activities are: research, formative assessment, tests, questionnaires, surveys, reporting and providing data banks, provide guidance and recommendations.

Three times there have been proposals in Knesset to abolish PET (Psychometric Entrance Test), and the main reason is that coaching schools are supposed to make the results unfair.

RAMA:

- Conduct periodic evaluations of education system and evaluation in schools
- Assessment for learning
- Activities:
    - Large scale tests

    - Formative assessment in service of learning
    - Reporting to the public and providing data banks

Meitzav – state assessment

- a set of school indicators
- administered every second year
- reports include information on
    - Pedagogical environment
    - School climate
    - Student achievement

Negative effects are that there are: between subjects reallocation of time, teaching to the test, and gaming the system. RAMA wants to implement "**assessment for learning**". They want to strengthen both the internal and external examinations.

Through New Horizon (a new reform):

The teacher salaries should be raised, there should be more quality time between teachers and students, the class size should be reduced, and there should be differential investment according to social economic status.

Score Reporting

Ron Hambleton

Considerable investments of time and money have been made to address technical problems in large scale assessment – test specifications, IRT-modeling and equating, reliability assessment, DIF-analyses, and validation studies.

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. This is unfortunate because:

- Reporting scales are confusing for many persons (e.g., percent, percentile, IQ, SAT vs. ACT, NAEP etc.)
- Quantitative literacy is not high among policy makers and the public (half of the population in the US can not read bys schedules)

Research from several NAEP-related score reporting studies:

Hambleton & Slater (1995)
Hambleton & Smith (2001)
Wainer, Hambleton & Meara (1999)
Hambleton & Meara (2000)
Hambleton, Allalouf & Slater (2001)

Some ideas for important score reports:

- Bench-marking (item-mapping)
    - Capitalizes in IRT

New SAT skills report:

- Originally intended for curriculum people
- Descriptors for different score categories and content strands; skills they can do and what the next steps are.

There is much to do, and much needs to be done.

Development of the SweSAT

Christina Stage

One of the problems with the test as it is at by now, is the great dominance of verbal items. As you can see the number of Swedish verbal items (WORD and READ) is 60, and then we have ELF, which is verbal as well. This makes 80 out of 122 items or 65%. Table 1 will remind you of how the test is composed at present.

Table 1. SweSAT since 1996.

Sub-test	**In short**	**Items**	**Time**
Data Sufficiency	DS	22	50 min
Diagrams, Tables and Maps	DTM	20	50 min
English Reading Comprehension + Vocabulary	ERC + WORD	20 40	35 + 15 min
Swedish Reading Comprehension	READ	20	50 min
One of the above (pre-testing)			50 min
Total test		122	4 h 10 min

The verbal dominance in SweSAT is the main reason why the technical educations are not very interested in the test. On the other hand they have also declined the test that was specially designed for them. Maybe Nils will talk

about that tomorrow. They would prefer the SweSAT, if it would be supplemented with a mathematical sub-test.

As I said yesterday the National Agency for Higher Education wants changes of the test, primarily in order to increase the predictive validity. Results from the VALUTA-project showed that grades are better predictors of success in higher education than are test results. Which I think is a common phenomenon for selection tests everywhere.

We have got an assignment to construct and field-test three types of items, which eventually will be included in a changed SweSAT, and to give recommendations how a future SweSAT should be composed. That is what this day is going to be all about, and the main topic for this meeting. As a background Christina will begin by presenting a review, she has made of selection tests internationally.

A Review of Selection Tests Internationally

Christina Wikström

Internationally we can find many tests similar to SweSAT:

USA: SAT, ACT, LSAT, GRE, GMAT, MCAT etc.

Israel: PET

Australia & New Zealand: ENTER, UMAT

Canada: MCAT, DAT etc.

UK: Law, Biomedicin

India: CAT, GATE etc.

…..and many more…

Psychometric Entrance Test (PET)

"a tool for predicting academic performance"

Quantitative reasoning

- questions and problems
- graph and table comprehension
- quantitative comparisons
- number series

Verbal reasoning

- analogies
- sentence completion
- logic
- reading comprehension

English

- sentence completion
- restatement
- reading comprehension

A College Test (ACT)

"Measures the knowledge, understanding and skills that you have acquired up to now"

English test:
Standard written English and rhetorical skills (5 passages, 75 MC)

Reading test:
Passages from social studies, natural sciences, prose fiction, and the humanities (40 MC)

Maths test:
Algebra, geometry/trigonometry on different levels (60 MC)

Science test:
Presents data (graphs) research summaries or conflicting points of view, to measure interpretation, reasoning, analysis, and problem solving skills

SAT

Reading section:

Critical reading and sentence level reading (MC)

Writing section:

Develop a point of view, support a point of view, follow conventions of standard written English. Improving sentences, identifying sentence errors, improving paragraphs (Essay, MC)

Maths section:

Numbers and operations, algebra and functions, geometry and measurement, data analysis, statistics and probability (student produced response, MC)

Common arguments for test revisions have been:

- making group differences smaller (increasing verbal weight, revising use of language, and types of texts)
- making it more relevant to specific programs (splitting verbal - quantitative, more focus on eligibility demands)
- making it look more of an achievement test (curriculum based, high school content)
- making it appear more relevant for students/test-takers (curriculum based, including typical high school related content)
- Sending signals to schools regarding what is important (reading, writing, mathematics)

Table 2. Comparison between PET, SweSAT and SAT:

PET	SweSAT	SAT
Quantitative reasoning	**Quantitative/analytical Section**	**Quantitative section**
Questions and problems	Data sufficiency	Mathematics MC+grid
Graph or table comprehension	Diagrams, Tables and Maps	
Quantitative comparisons	*Quantitative comparisons*	
Number series	*Analytic reasoning*	
Verbal reasoning	**Verbal section**	**Verbal section**
Logic	Vocabulary	
Reading comprehension	Reading comprehension	Critical reading
Sentence completion	*Sentence completion*	Writing (incl essay)
Analogies	*Analogies*	
English reasoning		
Sentence completion		
Restatements		
Reading comprehension	English reading comprehension	

Development of the SweSAT *(continued)*

In the present SweSAT the WORD sub-test with 40 items is a problem in itself. The sub-test is a good one, since it is fairly easy to construct, it measures something essential, it is time economic, and it is possible to weigh male items against female items and collate a gender neutral test. One problem is that it favours older test takers in front of young, another is that it takes only 15 minutes out of the total testing-time, which is 4 h 10 minutes, which means that it altogether has too great impact on the final score. That is 6% of the testing time and 33% of the total number of points. The sub-test has also been criticized for being unfair against immigrants, since they may

have special problems with knowing the meaning of words presented without a context.

As may be seen in Table 2, the verbal sub-tests also have fairly high inter-correlations.

Table 3. Inter-correlations between the verbal sub-tests; (averages for the last three test administrations). Reliabilities in bold, correlations corrected for attenuation in italic.

Sub-test	WORD	READ	ERC
WORD	**.86**	*.77*	*.76*
READ	.61	**.72**	*.73*
ERC	.61	.61	**.75**

One way to increase the predictive validity could be to have two distinct parts in the test: one part measuring verbal ability and the other part measuring analytical and quantitative ability. Then different educations could choose to give different weights to, or to use only one of the two parts depending of what is most important for that education. This implies that the two parts can be equated separately. For the SweSAT that would make it necessary to expand the quantitative part of the test. The present 42 quantitative items in total are not enough for equating into 21 steps. At a minimum 60 items are needed to make an equating process meaningful. The result then would be three different equated scores for the test-takers: one verbal, one quantitative/analytical, and one total.

As already mentioned we have got an assignment by the National Agency to construct, and field-test three types of items: “sentence completion” (SEC), “analogies” (ANA) and a sub-test measuring quantitative/analytical ability (QC). The outcome of the field-testing, of verbal items (78 SEC) and (30 ANA) will be presented by Maria, Ragnar and Sandra, and the outcome of the field-testing of QC (75) will be reported by Anders. Finally Gunilla and I will present the recommendations we intend to give, and on which we are very eager to hear your comments.

Sandra has been responsible for the WORD subtest since 2001, but will finish this summer. She will describe the verbal sub-tests. Maria, who has just been engaged in the project as successor of Sandra, has constructed most of the field-tested items, and will tell you about her experiences of that.

Ragnar who is mainly working with the READ sub-test, will tell you about the outcome of the field-testing, and of the enquiry, which the test-takers answered.

Field-Testing of some Verbal Sub-tests

Ragnar Haake, Maria Johansson, Sandra Scott

Two types of verbal items have been constructed and field-tested Sentence Completion (SEC) and analogies.

Example of a SEC-item in Swedish

Ett svart hål är egentligen inte ett hål I /---/att det inte finns någonting där. Det är precis tvärtom; ett svart hål är ett objekt i rymden som har otroligt stor massa, så stor att inte ens ljuset kan komma därifrån. Det beror på att /--/ i ett svart hål är så stark att ingen materia kan lämna det.

A betydelsen – dimensionen

B definitionen - -magnetismen

C gebitet – densiteten

D bemärkelsen - gravitationen

The same example translated to English

A black hole is in fact not a whole in /---/ that there is nothing there. It is just the other way round; a black hole is an object in space that has an immense mass, so big that not even light can off. This is due to the fact that /---/ in a black hole is so strong that no substance can leave it.

A the meaning - the dimension

B the definition - the magnetism

C the domain - the density

D the sense - the gravitation

Example of two analogy items in Swedish, and the same examples translated to English

	kaputt - hel	**ruined - whole**
A	fatal - olycklig	fatal - unhappy
B	modest - gammal	modest - old
C	ringa - stor	trifling - big
D	trivial - grov	trivial - coarse

	duva - red	**dove - peace**
A	ring - symbol	ring - symbol
B	triangle - fara	triangle - danger
C	orm - bett	snake - bite
D	kors - tak	cross- ceiling

The experience is that SEC-items are easy to construct, while ANA-items are difficult to construct. Another advantage with the SEC-items is that it is difficult to improve the score by mechanical practice.

Results from the try-outs of Sentence Completion and Analogies:

SEC 102 items in six different booklets:

- 15 items p >.80
- 13 items p < .30
- 4 items with r_{bis} > .30
- 16 items with p-diff > .15
- 54 items OK

ANA 75 items in six different booklets:

- 4 items with p > .80
- 15 items with p < .30
- 2 items with r_{bis} < .30
- 16 items with p-diff > .15
- 45 items OK

Table 4. Inter-correlations: averages for three groups, within brackets the number of items, and the smallest and highest value. In italics correlations corrected for attenuation.

	WORD (10)		SEC (20)		ANA (10/15)	
WORD	**.69**	(.66-.76)	*.95*		*.87*	
SEC	.65	(.59-.73)	**.71**	(.66-.76)	*.89*	
ANA	.56	(.31-.75)	.65	(.42-.77)	**.58**	(.35-.75)

Table 5. Results from the surveys; all numbers are percentage.

Question	SEC	ANA	WORD
Most difficult	20	57	23
Easiest	45	15	40
Best measure of language	66	19	15
Best measure of vocabulary	12	37	49
Most meaningful	59	19	21
Least meaningful	10	58	32
Most fun	33	28	39
Most boring	28	40	31

Field testing of a Quantitative Sub-test

Anders Lexelius

The Quantitative Comparison (QC) items test the ability to reason quickly and accurately about the relative sizes of two quantities. The quantities are arithmetic, algebra, geometry, functions, and statistics.

An accepted item should have a p-value between 0.2 and 0.8, a biserial correlation higher than 0.3, and the difference between males and females should be less than 0 .2.

Five different QC sub-test containing 20 items each were tried out together with 10 DS-items on five different groups of students from the third grade in upper secondary school, the natural science program. The results are shown in Table 1.

Table 6. Results for males and females on 20 different QC-items in five booklets, and 10 DS-items, the same in all five booklets. .

	QC			DS		
Booklet	Male	female	alpha	male	female	N
1	10.32	8.03	.63	5.81	4.31	155
2	10.37	9.19	.68	5.26	5.88	111
3	13.29	11.40	.75	7.76	7.47	91
4	10.77	10.35	.65	5.84	4.88	86
5	12.59	11.52	.53	6.54	6.22	64

Table 7. Inter-correlations.

Booklet	QC/DS	QC/math-grade	DS/math-grade
1	.56	.36	.52
2	.70	.54	.49
3	.59	.52	.32
4	.43	.04	.00
5	.65	.34	.23

In a questionnaire, which was answered right after the test was finished, 72 percent of the test-takers regarded the QC sub-test as having right difficulty level, while 19 percent found it too hard, and 8 percent too easy. In comparison with DS 47 percent found DS more difficult, 29 percent found the two

subtests equally difficult, and 25 per cent found QC more difficult. And finally 53 per cent found QC more meaningful, while 45 percent found DS more meaningful.

A New Model for the SweSAT

Gunilla Ögren, Christina Stage

Unfortunately there are a lot of restraints when you want to build a complete test. The general requirements on the test, which are included in the contract with the National Agency, are:

* The test should be in line with the aims and content of higher education
* The test must not have negative effects on the education in upper secondary school.
* It should be possible to score the test quickly, cheaply, and objectively.
* It should NOT be possible for an individual to improve his/her result by means of mechanical exercises or by learning special principles for problem-solving.
* The test-takers should experience the test as meaningful and suitable.
* The demand for un-biased recruitment should be observed. No group should be discriminated against because of gender or social class.

Besides there are several administrative restraints: the whole test should not take longer than today i.e. 4h and 10 min. Pre-test items should be included in a way, which makes it impossible for the test-takers to identify them, and that makes the scoring as convenient as possible. There should be as equable distribution as possible between verbal and analytical items regarding number of items as well as time.

A specific problem is the English reading comprehension test. It does not fit naturally in neither the verbal nor the analytical part. The arguments when it was introduced in 1992 were: 1) the students should be given a signal about the importance of English, and 2) it should contribute to decrease the gender differences on the test. One argument against the sub-test, in 1992 and which is still valid, is that English ability is measured by the grades (and is a mandatory subject for eligibility for higher education). Other arguments against the sub-test are that it consistently contributes to increase the gender differences, it disfavours people with dyslexia. There are also consistent ceiling effects in the results. Furthermore young people today generally have a much better mastery of English, than they had in 1992. The conclusion is that we would like to abolish that sub-test.

A preliminary suggestion what a new SweSAT could look like is presented in Table 8:

Table 8. Suggestion for a new SweSAT.

Verbal part:

Sub-test	In short	Items	Score
Vocabulary	WORD	20	20
Reading Comprehension	READ	20	40
Sentence Completion	SEC	20	20
Analogies	ANA	8	8
Total		68	88

Quantitative/Analytical part:

Subtest:	In short	Items	Score
Analytical Reasoning	AR	8	8
Data Sufficiency	DS	12	12
Quantitative Comparisons	QC	18	18
Diagrams, Tables and Maps	DTM	25	50
Total		63	88

"Authentic" items are recommended. At present we have two sub-tests which can be regarded as authentic, the READ sub-test and the DTM sub-test, the problem with both these sub-tests is that each item is time-consuming (2.5 and 2.3 min). We want to keep these two sub-tests, since they are generally regarded as meaningful by the test-takers. However, we would like to give two points for each correct item in these sub-tests. The reason would be to motivate the test-takers to work on these time-consuming items, and not ignore them. It would not affect the reliabilities, and hardly the ranking, but it would give a signal that these items are important, and worth working on.

Summary of the changes

Verbal section:

- Reduce the number of items in the WORD sub-test by half to 20 items.
- Introduce two new types of sub-tests Sentence Completion (SEC) with 20 items, and Analogies (ANA) with 8 items.
- Keep the Swedish Reading Comprehension (READ) sub-test with 20 items but give two points for each correct answer.

- Arrange two booklets, which contain all four types of items, and let the test-takers use the allotted time at their own disposal.

Quantitative/analytical section:

- Expand the Diagrams, Tables and Maps (DTM) sub-test to 25 items, with the same number of figures as today (i.e. 10), and give two points for each correct answer.
- Reduce the number of Data Sufficiency (DS) items to 12, and add 8 Analytical reasoning (AR) items
- Introduce a quantitative comparison (QC) sub-test with 18 items.
- Arrange two booklets, which contain all four types of items, and let the test-takers use the time at their own disposal.

The results will be presented by three norm scores: one for the verbal section, one for the quantitative/analytical section, and one for the total test.

Comments

There was agreement between the participants that it is only confusing with a quantitative/analytical section. There should be two pure sections, one verbal and one quantitative, but both sections should include some measure of analytical ability. Hence the AR sub-test would rather belong to the verbal section since it is a verbal sub-test.

There was also agreement that the ANA sub-test should be abolished. The items are difficult to construct. There exist a lot of bad experiences with the item-type. The items are often culture-dependent. Anyhow 8 items are two few to give any contribution but rather confusion.

Everybody also agreed about the necessity to do more piloting, both of the suggested sub-tests and other alternative sub-tests. And that it is important to find out the inter-correlations between the sub-tests.

A majority of the participants (but not all) agreed that the ERC sub-test did not really fit in anywhere, and was not really necessary, and also too short to stand on its own. Furthermore English is a school subject which is providing eligibility for higher education, and if there is a general wish for better knowledge of English, the qualification requirements could easily be raised. Against abolishing ERC was the argument that the correlations with WORD, and READ are high, and that it is the only sub-test on which test-takers with another native language than Swedish succeed as well as Swedes.

The sub-test READ was criticized for being too un-efficient. It is necessary to get more information out of a test which takes 50 minutes. It should either have more items of the same kind to each text or maybe complemented with some cloze items. According to Lord: *speed is not a bad predictor*.

Finally the invited guests were asked to write down, and give us a list of suggestions for changes, and their most important recommendations. These suggestions and recommendations are reproduced in their entirety in appendices.(Appendix 1: Micahl Beller, Ron Hambleton and Wim van der Linden, Appendix 2: Jan-Eric Gustafsson, and Appendix 3: Allan Svensson)

A new model for pre-testing

Gunilla Ögren, Christina Stage

At present a test-day consists of five sections of 50 minutes each. Four sections consist of regular items and one consists of pre-test items. The order between the sub-tests is not fixed, but varies between test-occasions. The sub-tests DS, DTM, and READ each make up one section, while the subtests ERC and WORD together form one section, where 35 minutes are allotted to ERC and 15 minutes to WORD. This pre-testing model means that all test-takers have to take two versions of one of the sub-test sets, but they do not know which one is the pre-test.

The problems with the present pre-testing model with all pre-test items in one block are:

1. The tests are not exactly parallel from one test occasion to another since the sub-tests have to be presented in different order. (Otherwise it would be too easy for the test-takers to find out which block contains pre-test items, and the data would be very unreliable.)

2. The tests are not exactly the same for all test-takers, since they are doing different subtests and different number of items.

3. Many test-takers are upset by having to spend 50 minutes on items, which are of no use to them.

We would like to change the pre-testing model so that the pre-test items would be hidden in the regular sub-tests. Such a model would be fairer to the test-takers both at each regular test, and between test occasions, since the order of the sub-tests can be fixed, and all test-takers do the same number of pre-test items and sub-tests.

The advantages with the new model would be:

The sub-tests will be in the same order at different test administrations.

All test-takers get the same number and types of items.

Even if the gain in total time for pre-test items is small (about 10 minutes) it will probably not be experienced as oppressive by the test-takers, as doing a separate block of 50 minutes.

It is not as big a disaster as today if a test-booklet would disappear during the administration.

Comment

The risk with such a pre-test model is that if any of the pre-test items does not function properly, this fact might destroy the score fore test-takers who get stuck on it.

The present model for equating

Christina Stage

As you know, results of SweSAT are valid for five years, which makes equating between different versions of the test a serious undertaking. The conversion of raw scores from to norm scores should make it possible to compare scores from one test occasion to another, i.e. it should be as easy or difficult to obtain a certain norms score on one test as on another.

At present the norm score has a range from 0.0 to 2.0, the latter being the top result. Each correct answer is given one point and the total number of correct answers represents the raw score. In order to ensure that scores on different test administrations are comparable, the raw scores are converted to norm scores. The strategy applied to define scale limits for the norm scores is based on a combination of comparisons.

Pre-equating

The test-developers aim at assembling parallel versions of each sub-test. Parallel according to a) subject areas, content b) cognitive level c) difficulty.

The equating procedure (equi-percentile-equating)

The total group of test-takers is examined and compared to earlier populations regarding sex, age, and background education.

A reference population I is selected through proportional stratified selection from the total group. They are selected to give the same proportion of gender, ages, and background education.

A reference population II consists of those among the test-takers who are 18 or 19 years old, and are still registered on the third year in upper secondary school.

It is assumed that these two sub-populations have equivalent ability distributions over the years.

1. Study of the equivalence between the groups on this test with the corresponding groups on earlier tests. To check the assumption the results of these three groups are studied simultaneously at sub-test level and for the whole

test. As you know the crucial task is to find out whether eventual raw score differences from earlier test versions are caused by the test being easier or more difficult or by the test-takers being better or worse. If this test is easier or more difficult does it depend on the test or on the test-takers?

By using the cumulative frequency distributions of results the proportions of test-takers in each group which corresponds best to the proportions from earlier years on each norm score level are found. In that way three equating functions are defined, and the final equating function is a weighted mean of the three equating functions.

Since 1997 we have also done IRT-equating as a complement. We started with two reference tests, but in 2002 we changed to only one reference test, since it was too difficult two interpret two (the results could be very different). The problem is the limited number of common items. The test-developers are instructed not to use more than four items, which have been pre-tested together. As most we have had 26 common items, and at that occasion the test-developers had been instructed to use as many items as possible, which had been pre-tested at one special test-occasion. The normal number of common items is between 18 and 22.

We use the results from the IRT-equating as a fourth group of the test-takers, and make a weighing of the four groups to find the appropriate limits for each score. Table 9 (which is from the test this spring) will illustrate how it is done. It is not a weighing with equal weights, if one of the groups deviates very much, it gets a lower weight.

Table 9. Equating Table

Norm score	Total group	Refpop I	Refpop II	IRT	New score
0	0-33	0-34	0-33	0-33	0-33
0.1	34-39	35-40	34-42	35-41	34-40
0.2	40-43	41-44	43-46	42-43	41-44
//					
1.0	71-73	72-75	69-72	67-69	68-71
//					
1.9	105-108	105-107	106-108	104-108	105-108
2.0	109-122	108-122	109-122	109-122	109-122

Systematic Error in the SweSAT Equating Procedure

Per-Erik Lyrén

An assumption in the equating procedure is that the reference groups have the same ability level across administrations. The school system, policy, and society changes:

- organizational changes
- Syllabi makeovers (e.g. math courses)
- Students are poorly prepared for higher education (DN Debatt, 2002)
- Fewer men go to higher education (Tha National Agency for Higher Education, 2008)

Conclusion: Things are happening that are likely to influence:

- potential test-takers´ decisions of whether or not to take the SweSAT
- The ability level of potential test-takers

The purpose of the study was to examine whether the equal ability assumption is violated or not.

Data:

Potential linking items are available (WORD and DS items), which have been administered as try-out items (N typically 1000 – 2000). For WORD 15 items were imbedded in the try-out form in the administrations from fall 1997 to spring 2004 (14 administrations). For DS 22 items were administered as a complete try-out form from fall 1999 and onwards.

Main results - WORD

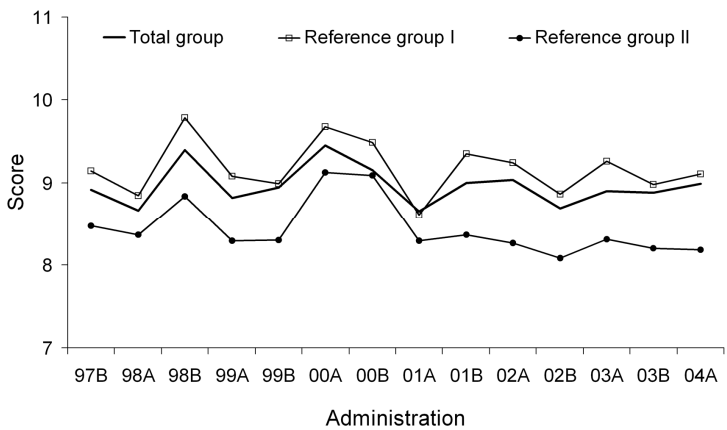

Figure 2. Estimated mean scores on the WORD anchor items.

Main results - DS

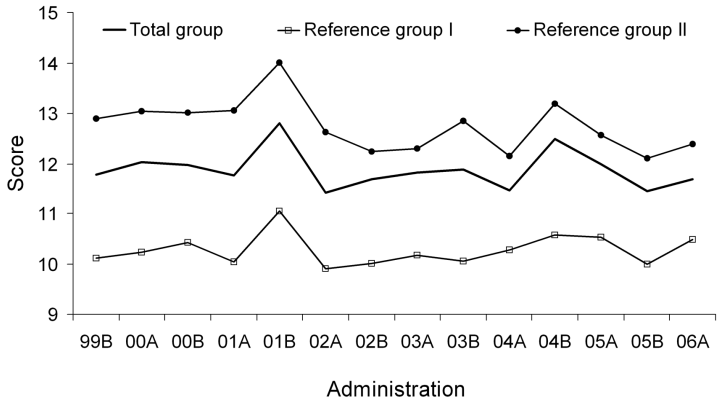

Figure 3. Estimated mean scores on the DS anchor items.

The Experience of Domain Specific Tests

Nils Olsson

The domain specific test, which has been developed by the National Agency for Higher Education for the technical educations, will not be used. At a meeting with representatives from the technical educations it was made clear that, since selection is hardly a problem at these educations, they are not interested in the test. They would, however, be interested in a SweSAT with more mathematical content.

The domain test for medical and paramedical training had this spring been tried out on 227 applicants to the college of veterinary medicine in Uppsala. The test consists of three sub-tests: one personality test, one test of communication skills, and one test of general knowledge. The test is computerized.

The technicalities had functioned satisfactorily. The correlation with SweSAT results was high. The general knowledge test had a low reliability, and many of the items did not have acceptable characteristics.

Comments

Personality tests are very risky in a high stake test. Faking is unavoidable. Most test-takers realize what they should answer irrespectively of what they really think. To have simulated situations is a better way of measuring personality traits.

Appendix 1

The Second Generation of the SweSAT: Some Suggestions for Consideration

Michal Beller, Ronald K. Hambleton, Wim van der Linden[1]
SweSAT Technical Advisory Committee Members

In this brief report we want to try and respond to the main question from Christina Stage, Director of the SweSAT Program in the Department of Educational Measurement at Umea University: **What is most important to think about as the SweSAT is being revised for the next generation of candidates?** Let us say first that we are very pleased to see the Department of Educational Measurement and the National Agency moving forward with a major study of the next generation of the SweSAT. It is timely to do so and we are pleased with the general direction of the proposed revisions.

Background

We would like to see some information collected from university professors about their thoughts about the knowledge and skills that might be assessed on the SweSAT. We would ask them what they think are the necessary skills for successful students in university—we expect they might say reading for comprehension, reading quickly, having study skills, having a basic aptitude for handling quantitative information, writing clearly, and so on. We doubt there will be any big surprises but it would be useful to confirm what they think the SweSAT should measure (separate from the national exams). We might ask them not to confuse the national exams (which measure school curricula) with developed abilities or aptitudes which the SweSAT can measure. We might make the professor survey with both structured (e.g., What is your opinion about the current version of the SAT? Excellent, very good, etc.) and open-ended questions (e.g., What suggestions do you have for the revision of the SweSAT?), and we might even give the professors a list of skills and have them choose or rank in terms of importance. We might even share some test items with them to see how suitable they think the items are. This information could be collected from surveys, or even surveys and focus groups. We agree that we have a pretty good idea of what professors are going to say, but we see value in completing the exercise.

[1] All three members are equally responsible for this report, and so the names appear in alphabetical order.

We would supplement what we might learn from professors with information from candidates, university students, the public, and the national agency, about the current SweSAT and what the SweSAT could/should look like in the future. The information collected would almost certainly support the intent to organize the knowledge and skills around critical reading and quantitative skills, and possibly writing.

Finally, we would continue the efforts to monitor parallel efforts in the US (with the SAT and the ACT), and Israel (with the PET), and perhaps other countries too. Changes in these other programs might be considered for implementation in Sweden too. Of particular interest are the relatively recent changes in SAT I (see Appendix A).

More Test Efficiency, Reliability, and Validity

We want to address the number of score points per minute of testing on the SweSAT. It seems quite low. We think with the SAT in the US, the SAT is getting at least 60 or more points per 70 minutes of testing or .86 points a minute. We think many testing programs are getting between .67 and close to a point a minute. With the exception of the Word subtest, the SweSAT is getting relatively low points per minute, often less than .50. We would like to see the SweSAT push the candidates a bit more by adding questions without lengthening the available time. We think you will be surprised that adding items is quite possible and will enhance score reliability and validity, and a small element of speededness is likely to increase validity too. That the test may become slightly speeded is not such a bad thing, and may even enhance predictive validity, since an ability to work quickly is a useful skill to have in university. The idea of counting the reading comprehension component more in the total score is definitely worth considering, though we recommend that more test items be included in this section first, before considering the weighting option.

We think too that you should try to find ways to get more measurement opportunities from the reading passages. Apparently they are long and complex. That's fine to "model" reading passages the candidates will see in university, but if candidates spend several minutes reading the passages, then you need to get, perhaps, five to seven or eight questions worth of information. Passages that take multiple minutes or more to read cannot be justified with as few as four questions unless they are being polytomously-scored (and that is not the case with the SweSAT). Take a look at how the US handle multiple questions and score points from their passages—check NAEP, the SAT, the GMAT, Massachusetts tests (see, www.doe.mass.edu), etc.

English Language Subtest

We think the English Language Sub-Test might be best removed from the new SweSAT. It does not fit easily into the new scheme, and is not long enough to serve as a stand-alone score for reporting. English proficiency is addressed in the national exams, and for those students who need to document their English proficiency, the SweSAT subtest on English proficiency will not be reliable or valid enough to be useful (unless it is substantially lengthened). Universities can require TOEFL scores to assess writing, speaking, listening, and reading in English if they want the English proficiency information. The TOEFL (or the British equivalent) is computer-administered, and with excellent psychometric properties.

Organization of Subtests into Verbal and Quantitative Sections

Let us begin by saying that we support the idea of dividing the test into Q and V sections/scales (both including reasoning items). It will be beneficial to explore the Q and V constructs as reported by ETS for SAT and GRE:
http://www.ets.org/research/researcher/RR-93-22.html
http://www.ets.org/research/researcher/RM-03-01.html
http://www.ets.org/research/researcher/RM-02-01.html

Regarding the V section, we would like to say that we are not strong advocates for including the Analogy questions in the SweSAT. Not only are eight items too few to report a subtest score, but we know the SAT program in the US dropped analogies because of several criticisms: They are not linked in any way to the high school curricula or university programs, they are probably more coachable than other item formats, sometimes they are culturally biased, and on occasion it may be difficult to defend the correct answers.

We don't know how the proposed eight subtests in the new SweSAT actually are correlated. Our inclination would be to drop analogies, and broaden reading comprehension to incorporate the sentence completion item format (but there is not consensus on this point among the three of us). Our verbal section would emphasize vocabulary (but not straight vocabulary questions but rather determining meaning in context) and reading comprehension. Multiple formats would be fine if it can be demonstrated that these non-traditional item formats (e.g., sentence completion, or modified Cloze format) increase test score validity.

As for the quantitative section, we agree with the suggestion that perhaps Analytic Reasoning may belong in the Verbal Section. The remaining subtests might be reconsidered too and their placement in the test scales. Also, we wondered if a quantitative section might benefit from algebra word problems, and perhaps some geometric problems. We were encouraged by

the results for the Quantitative Comparisons, but we encourage you to continue to do research on new item formats (please note that they, along with the Analogies, were dropped from the New SAT). Check out the validity of this type of item on the GRE in the US. We expect there are a number of studies that have been done especially at ETS with the GRE exam.

Also, it might be beneficial to explore the structured constructed response item type (student produced responses) included in the SAT, where the undesired guessing factor is eliminated. Our understanding is that these "gridding problems" have been very successful and appreciated by the students, high schools and universities.

We want to reinforce a discussion at the meeting concerning data collection designs (e.g., some extra testing of high school and college students) that would permit the collection of data for studying the "structure" of the proposed new SweSAT. For example, a plan has been put forward for organizing a number of subtests into Verbal and Quantitative sections. Do the intercorrelations among the subtests support this proposed organization? We think the ultimate organization of subtests into two sections ought to be driven by empirical evidence in additional to any logical evidence that can be mounted. With the right data-collection design, and sufficient numbers of students, structural equation modeling can be done (with lots of missing data) to determine how the battery of sub-tests actually are related and might best be organized and weighted.

Pretesting and Equating

We remember the days when pretesting was not being done as well as it is today, and equating was not being done at all. It is exciting to see the progress that has been made on these two topics. In the next generation, we think you want to determine if the "pretest block" can be used for both pretesting and equating, and with a newly considered design—a pretest block would be one solution, embedding items in the test would be another, and of course there are other designs possible too. We are not able to advance a best solution now, but we do feel strongly that improvements need to be made in the collection of pretest data, and we want to see equating of scales over time being done with a defensible and valid equating design, one that does not depend on an assumption of equivalent groups of students over time.

We want to endorse too the idea of using IRT models to calibrate the test items, build the scales, identify item biases, select the test items (perhaps using automated test assembly), equate the test scores, report the scores, and much more. Many test agencies are finding that an IRT based examination system can help with many technical problems.

Computer-Based Testing of the SweSAT

We completely endorse the decision to investigate administering the SweSAT on a computer. Of course many challenges need to be overcome for the computer-based (CB)-based SweSAT notably practical ones (will there be sufficient computers for candidates and can the technological problems be resolved?), security issues, and technical (e.g., choice of a computer-based test design, the choice of item formats that capitalize on the power of the computer for assessing new skills, etc.). But the transition will be challenging, and there will be much to learn and problems to overcome. Also, your National Agency for Higher Education should be willing to invest in computerization of the SweSAT since the developmental costs will be high. At the same time, the movement to computer-based testing is inevitable. In anticipation of this transition, the national agency and the DEM will want to begin soon to investigate computer-based testing with exams such as the GRE, GMAT, MCAT, TOEFL, etc.

Communications with Candidates, and the High Schools, Universities, and National Agency

Regardless of the ultimate decisions about score reporting (1, 2, or 3 scales, or may be more), we would like to see serious thought given to expanding information about the scores given to candidates, the schools, and even the universities and national agency. Expanded information to candidates about their strengths and weaknesses would be helpful (see, for example, some of the new types of reports the College Board is working on), at the same time, reports to schools, counties, and the national agency would be valuable too in assessing curricula, and instructional strategies.

Final Comments

The National Agency for Higher Education wants to increase the validity of the SweSAT scores. This is an important and realistic goal. While we strongly support the current initiatives, we want to repeat a comment made by one of our members—any changes are likely to last for quite some time, and so it would be best to move carefully and slowly in making modifications. It will be difficult to implement change, and so you want to take the time to research the proposed changes, and to be able to defend the changes that are being made. In addition to all of the suggestions offered earlier, we would also suggest that any recommendations that are ultimately made to the National Agency should be supported by either logical or empirical evidence. Don't hesitate to involve members of the advisory committee in the next steps. Occasional reports will keep us up-to-date and give us an opportunity to provide input as our time permits. Best of success as you move forward!

Appendix A

New SAT I Structure (from Wikipedia: http://en.wikipedia.org/wiki/SAT#Questions):

2005 Changes

In 2005, the test was changed again, largely in response to criticisms by the University of California system. Because of issues concerning ambiguous questions, especially analogies, certain types of questions were eliminated (the analogies from the verbal and quantitative comparisons from the Math section). The test was made marginally harder, as a corrective to the rising number of perfect scores. A new writing section, with an essay, based on the former SAT II Writing Subject Test, was added, in part to increase the chances of closing the opening gap between the highest and mid-range scores. Other factors included the desire to test the writing ability of each student in a personal manner; hence the essay. The New SAT (known as the SAT Reasoning Test) was first offered on March 12, 2005, after the last administration of the "old" SAT in January of 2005. The Mathematics section was expanded to cover three years of high school mathematics. The Verbal section's name was changed to the Critical reading section.

Section	Average Score	Time (Minutes)	Content
Writing	497	60	Grammar, usage, and word choice
Mathematics	518	70	Number and operations; algebra and functions; geometry; statistics, probability, and data analysis
Critical Reading	503	70	Critical reading and sentence-level reading

Version: Prepared on June 23, 2008.

Comments on the document
Development of the SweSAT
Jan-Eric Gustafsson 2008-06-25

The document "Development of the SweSAT" presented at the Umeå SweSAT conference by Christina Stage proposes several changes of the SweSAT. These proposals are discussed below.

A main change that is suggested is to split the subtests of the SweSAT into two parts to allow separate Verbal and Quantitative/Analytic scores. In order to make this possible it is proposed that at least 60 items are needed in each part. The following structure of the test is suggested:

Verbal

Subtest	*Label*	*Items*	*Score*
Vocabulary	WORD	20	20
Reading Comprehension	READ	20	40
Sentence Completion	SEC	20	20
Analogies	ANA	8	8
Total		*68*	*88*

Quantitative/Analytical

Analytical Reasoning	AR	8	8
Data Sufficiency	DS	12	12
Quantitative Comparisons	QC	18	18
Diagrams, Table and Maps	DTM	25	50
Total		*63*	*88*

General comments
Let me first of all make some comments on the distinction between the two parts. It is a well established finding that the predictive validity of the SweSAT is poor, particularly for educational programs with a mathematical/science content (e. g., Cliffordson, 2004a, 2008; Svensson, Gustafsson & Reuterberg, 2001). The most reasonable explanation for this is the dominance of verbal items in the current version of the test (80 items out of 122). Furthermore, 40 out of these 80 items derive from the WORD subtest, which is less analytical than are the other verbal subtests (READ and ERC).

Thus, to increase the predictive validity of the SweSAT it is necessary to increase the number of quantitative items. It probably also would be beneficial to increase the proportion of analytical items in the verbal part of the test. One reason for this is that it is likely to improve predictive validity, and another reason is that it would reduce the highly undesirable improvement

on the verbal test as a function of age, which is primarily due to the strong relation between age and performance on the WORD test (see, e. g., Cliffordson, 2004b; Gustafsson, Andersson & Hansen, 2000).

Thus, what we would need is one Verbal/Analytical and one Quantitative/Analytical part. However, the proposed structure of subtests does not fulfill this requirement particularly well. It is true that the proposed Verbal part is more analytical than are the current verbal subtests in the SweSAT, but it is not presented as aiming to measure Verbal/Analytical ability. One consequence of this is that the newly developed AR test is placed in the Quantitative/Analytical part rather than in the Verbal part. However, this test has been developed to fill the great need to have an analytical test with verbal content to suit, among others, law students, and it is a good example of a verbal/analytic test.

It could, of course, be argued that the AR test fits quite nicely with the DTM and DS subtests in the Quantitative/Analytical part. However, this is due to the fact that the DTM and DS subtests pose heavy verbal demands on the test takers, in the form of reading and vocabulary skills (Carlstedt & Gustafsson, 2005). Thus, these tests are not good measures of Quantitative/Analytic ability for the simple reason that they are too heavily verbally loaded.

In summary, this analysis suggests that the distinction between the two parts is inadequate to meet the needs of the new SweSAT, for the reasons that the Verbal part should be made more clearly analytical and the Quantitative/Analytical part should be made less verbal and more quantitative.

Below I make some more specific comments on the subtests proposed for the two parts.

The Verbal part

It is argued that one problem with the WORD test is that it is so efficient that it only takes 6 % of the testing time, while it contributes no less than 33 % of the score points. While this dominance in the total score points is unfortunate it does not seem to be a good solution to decrease the number of items, because they contribute to the reliability of the test, which will become even more of a concern if separate part scores are to be constructed. Another possibility would be to keep this subtest in its current form with 40 items, but reduce its impact on the total score by assigning 0.5 points for each correct answer.

The newly developed SEC test seems to be an interesting and useful test, even though more empirical data on the measurement characteristics of the

test are needed. The current response format for the items with two or three gaps should also be revised to allow independent choice of response option for each gap.

It is argued that the READ subtest should be kept unchanged, because it is a so called authentic test. However, this is also an extremely inefficient test when it comes to the yield of score points per minute (2.5 minutes/item). There is also the problem that there is a dependence among the items belonging to the same passage, which causes the reliability to become lower than is shown by standard formulas such as Cronbach's alpha. This problem will be further aggravated if the READ score is multiplied by two when computing the total score. When redesigning the SweSAT there are, therefore, strong reasons to abolish this test, or to redesign it into a more efficient test format. One possibility would be to use a cloze format similar to the newly proposed SEC test, but with more sentences, and with more gaps.

It is also suggested that a short analogies (ANA) test with 8 items should be included in the Verbal part, and that the items should be constructed in such a way that they primarily reflect vocabulary skill. While it has been shown that analogies items can be designed to measure analytical skills and vocabulary to a different extent (Ullstadius, Carlstedt, & Gustafsson, 2008), it seems that such items primarily measure analytical skills, and that it is quite difficult to construct unequivocal analogy items which primarily reflect vocabulary. It is, therefore, recommended that the ANA subtest is dropped.

In the document it is also argued that the English Reading Comprehension (ERC) subtest should be abolished. One reason for this is that English is taught as a subject matter and that it is part of the eligibility requirements for entrance to higher education. Another reason mentioned is that the ERC test does not fit naturally into neither the Verbal, nor the Quantitative/Analytical parts. Further arguments against the ERC test which are brought forward are that the test shows an unfavorable gender difference in favor of males, and that persons with dyslexia perform poorly on the test. It is also argued that there are ceiling effects on the ERC test, and that young people have better mastery of English than they had at the time when the test was introduced.

However, these arguments against the ERC test certainly are not sufficient to abolish the test. It is true that English is a subject matter taught at upper secondary school, but this is true for mathematics and Swedish as well, so following the same logic most other subtests should be dropped as well. It is quite obvious that the ERC test fits perfectly with the other verbal tests, and particularly so with the READ and SEC tests. The reason why it is argued that ERC does not fit with either part is probably the unclear nature of this distinction, as has already been pointed. While it is true that males tend to

perform slightly better on this test than females, this probably reflects the fact that the males taking the test is a somewhat more select group than the females (see Reuterberg, 1998) rather than any bias in the ERC test per se. The fact that dyslexics perform poorly on this test rather attests to its validity, than showing that the test is biased. The ceiling effect on the ERC test should be regarded as a desirable feature, given that there is a group of test-takers who are native speakers of English, and who should not be given an undue advantage of this in their total SweSAT score. The fact that young people have good mastery of English should also be seen as an advantage, given that the verbal score tends to be positively correlated with age.

While the arguments against abolishing the ERC test thus are weak or non-existent, there are very strong reasons for keeping this subtest. The most important reason is that this is the only subtest on the current SweSAT which does not show any bias against test-takers who do not have Swedish as their mother tongue (see Reuterberg, 2003; Reuterberg & Hansen, 2001). Given that the SweSAT is already severely biased against this group of test-takers, it would be highly unfortunate to make the situation even worse by dropping the ERC subtest. Another good reason for keeping it on the SweSAT is that a high-stakes selection test efficiently signals what are important things to learn.

The Quantitative/Analytical part

It is proposed that the newly developed AR subtest with 8 items is included in the Quantitative/Analytical part. However, as has already been discussed this subtest does not fit within this category, because it is a verbal/analytic test. It probably would fit better in the Verbal/Analytic part, but more empirical information is needed to understand the properties of this subtest.

The DS subtest is also suggested to be included in the Quantitative/Analytical part, where it fits quite well. However, the number of items is suggested to be reduced from 20 to 12, which is not a good idea, given that this subtest is the one which has the strongest quantitative component of all the existing SweSAT subtests (Carlstedt & Gustafsson, 2005; Åberg-Bengtsson, 2005). Thus, if the test is to be included it should have 20 items rather than 12 items.

The QC test is a newly developed test with 18 items, which holds great promise as a quantitative/analytical test. One particularly attractive feature of QC is that this subtest is almost completely non-verbal, which would make it less biased than most other subtests against those who do not have Swedish as mother-tongue. However, more empirical work is required to understand the measurement properties of this subtest.

The DTM test is also suggested to be included in the Quantitative/Analytical part, and it is proposed that that the number of items should be increased to 25. It is, furthermore, proposed that the DTM items should be given double weight when computing the total score. This implies that the DTM subtest would account for 50 out of 88 score points according to the suggested structure for the Quantitative/Analytical part. This is a highly questionable recommendation for two reasons. One is that the fact that the DTM items are relatively time consuming is not a good reason for giving them such an emphasis in this part, particularly since the DS and QC items are also relatively time consuming, and particularly so in comparison with the WORD items. The other reason is that it is doubtful if the DTM items at all belong in the Quantitative/Analytical part, given that there is very little quantitative content in this test (Åberg-Bengtsson, 2005). Truly enough this subtest poses substantial analytical demands, but it also poses demands for reading skills, general knowledge and vocabulary. Thus, given that the purpose is to make a clear distinction between a verbal and a quantitative/analytical part, it may be that another subtest, which has a more obvious mathematical content, should be found to replace the DTM subtest in this part. One possibility may be an algebra or numerical subtest of the kind used in the Math section of the SAT. However, it can also be argued that the DTM test has very good properties as a measure of analytic ability, which makes it essential that it is included in the SweSAT.

As it currently stands, the proposal for a Quantitative/Analytical part is highly problematic, particularly in that the quantitative content is very limited indeed. Only the 12 DS items and the 18 QC items could be claimed to be clearly numerical, while the rest of the items are not. On top of that the DS items pose relatively high demands on reading ability.

Recommendations

As has already been made clear there are numerous problems with the proposed new structure of the SweSAT. The distinction between the two main parts is not conceptually clear, because it mixes the two dimensions verbal/quantitative and analytical/non-analytical. One possible way to solve this problem is to construct one Verbal/Analytical part, and one Quantitative/Analytical part. Another possibility would be to construct three parts: one Verbal, with emphasis on vocabulary and reading; one Quantitative, with emphasis on quantitative manipulations, such as in QC; and one Analytical, with emphasis on problem solving, such as in AR, DTM, and DS. I here only sketch upon the alternative with two parts, and suggest the following structure:

Verbal/Analytical

Subtest	*Label*	*Items*	*Score*
Vocabulary	WORD	40	20
Cloze Test	CLOZE	20	20
Sentence Completion	SEC	20	20
English Reading Comprehension	ERC	20	20
Total		*100*	*80*

Quantitative/Analytical

Data Sufficiency	DS	20	20
Quantitative Comparisons	QC	20	20
Diagrams, Table and Maps	DTM	20	20
Algebra/Numerical operations	ALG	20	20
Total		*80*	*80*

The Verbal/Analytical part is suggested to be composed of the current WORD-test, but where each correct score in only awarded half a point. The current READ-test is replaced with a completion test of the cloze type with altogether 20 items. The newly developed SEC test, with the previously suggested changes, is also included. The SEC test could possibly be combined with the cloze test into one long test with 40 items. Finally, the current ERC test is suggested to be included in this part.

The Quantitative/Analytical part is suggested to be composed of the DS and DTM tests, as they currently exist. In addition the QC subtest would be included with 20 items, along with a newly developed Algebra/Numerical operations test with 20 items.

While the Verbal/Analytical part should be possible to administer within the same time frame as the current verbal tests, somewhat more time may be needed for the Quantitative/Analytical tests.

References

Carlstedt, B.., & Gustafsson, J-E. (2005). Construct Validation of the Swedish Scholastic Aptitude Test by means of the Swedish Enlistment Battery. *Scandinavian Journal of Psychology, 46*(1), 31-42.

Cliffordson, C. (2004a). De målrelaterade gymnasiebetygens prognosförmåga. *Pedagogisk Forskning i Sverige, 9*(2), 129-140.

Cliffordson, C. (2004b). Effects of Practice and Intellectual Growth on Performance on the Swedish Scholastic Aptitude Test (SweSAT). *European Journal of Psychological Assessment, 20*(3), 192-204.

Cliffordson, C. (2008). Differential Prediction of Study Success Across Academic Programs in the Swedish Context: The Validity of Grades and Tests as Selection Instruments for Higher Education. *Educational Assessment, 13*(1), 56-75.

Gustafsson, J.-E., Andersson, A., & Hansen, M. (2000). Prestationer och prestationsskillnader i 1990-talets skola. I: *Välfärd och skola,* SOU 2000:39, pp. 135-211. Stockholm: Statens offentliga utredningar.

Reuterberg, S-E. (1998). On Differential Selection in the Swedish Scholastic Aptitude Test. *Scandinavian Journal of Educational Research, 42*(1), 81-97.

Reuterberg, S-E. (2003). Vilken betydelse har utländsk bakgrund för resultatet på högskoleprove Del II. Högskoleverkets rapportserie 2003:23R. Stockholm: Högskoleverket.

Reuterberg, S.-E., & Hansen, M. (2001). Vilken betydelse har utländsk bakgrund för resultat på högskoleprovet?. Högskoleverkets rapportserie 2001:3 R. Stockholm: Högskoleverket.

Svensson, A., Gustafsson, J.-E., & Reuterberg, S.-E. (2001). Högskoleprovets prognosvärde. Sambandet mellan provresultat och framgång första studieåret vid civilingenjörs-, jurist- och grundskollärarutbildningarna. Högskoleverkets rapportserie 2001:19 R. Stockholm: Högskoleverket.

Ullstadius, E., Carlstedt, B., & Gustafsson, J.-E. (2008). The multidimensionality of verbal analogy items. *International Journal of Testing, 8*(2), 166-179.

Åberg-Bengtsson, L. (2005). Separating the quantitative and analytic dimensions of the Swedish Scholastic Aptitude Test (SweSAT). *Scandinavian Journal of Educational Research, 49*, 359-383.

Appendix 3

Allan Svensson
2008 06 22

Comments on changes of the SweSAT

To start with, I will point out that I am very positive to a division of the test two parts – not least considering that high-schools now are allowed to accept around one third of the applicants in a free quota group. Since many colleges probably will have great difficulties in developing relevant selection instruments of their own, it is important that they will have a greater freedom in choosing how to use the different parts in a new SweSAT.

I am, however, a bit sceptical to the proposals presented on the meeting – partly because two of the sub-tests will have such great impact (READ in the verbal part and DTM in the quantitative part), and partly because some of the new sub-tests seem to be hard to construct, and also would contain few items. Furthermore I am of the opinion that ERC should be kept, since English is getting a more important role within higher education. Besides, this sub-test, as far as I know, has worked well.

Regarding the difficulties in construction, the reliability, and the acceptability of the test, I would like ti give the following general recommendations:

Verbal part: ANA is abolished. Read is substituted by SEC and modified in a way that makes it possible to become the new sub-test on reading comprehension. WORD and ERC are kept. Each sub-test should have 30 items, which will give a total score of 90 on this part.

Quantitative part: AR is abolished. DS and DTM are kept. QC is added. The best would be if each of these sub-tests could have 30 items. If that is not possible some type of weighting should be applicable to make the maximum score the same in the two parts. One idea might be that each sub-test ha 15 items, each of them with the weight 2.

Appendix 4

Preliminary Program for the 12th SweSAT Meeting, June 15 - 17
Comfort Home Hotel/Uman, Storgatan 52 Umeå tel 127220.

Sunday June 15th

12.00 Lunch

13.15 Welcome and opening address (Chistina Stage)
News from the Department of Educational Measurement (Widar Henriksson)
News from the Advisory Council on Access to Higher Education (Jan-Eric Gustafsson)
What does a Chief Research Scientist at CTB/McGraw-Hill do? (Wim van der Linden.)

14.45 Coffee

15.15 RAMA - The National Authority for Measurement and Evaluation in Education.
(Michal Beller)
Score Reporting (Ron Hambleton)

Monday June 16th

8.30 Development of the SweSAT (Christina Stage)
A review of selection tests internationally (Christina Wikström)

9.45 Coffee

Field-testing of some verbal subtests (Ragnar Haake, Maria Johansson, Sandra Scott)
Field-testing of a quantitative subtest (Anders Lexelius)

12.00 Lunch

13.00 A new model for the SweSAT (Gunilla Ögren, Christina Stage)
A new model for pre-testing SweSAT (Gunilla Ögren, Christina Stage)

15.30 Coffee

16.00 Bus leaving for Norrbyskär

Tuesday June 17th

8.30 The present model for equating (Christina Stage)
Systematic equating error (Per-Erik Lyrén)

9.45 Coffee

10.15 The experience of domain specific tests (Nils Olsson)
Concluding remarks

12.00 Lunch

Appendix 5

Participants

Advisory board:

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

The National Agency for Higher Education:

Margaretha Hallgren
Nils Olsson,

The SweSAT program, Umeå:
Ragnar Haake
Mats Hamrén
Christina Jonsson
Maria Johansson
Anders Lexelius
Jenny Lindberg
Per-Erik Lyrén
Sandra Scott
Gunilla Ögren

The Department of Educational Measurement, Umeå:

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 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

Em No 58. NOTES FROM THE ELEVENTH INTERNATIONAL SweSAT CONFERENCE. Umeå, June 12–14, 2006. Christina Stage

2007

Em No 59. PROCEEDINGS FROM THE CONFERENCE: THE GDE-MODEL AS A GUIDE IN DRIVER TRAINING AND TESTING. Umeå, May 7–8, 2007. Widar Henriksson, Tova Stenlund, Anna Sundström, Marie Wiberg

Em No 60. MEASURING AND DETECTING DIFFERENTIAL ITEM FUNCTIONING IN CRITERION-REFERENCED LICENSING TEST. A theoretic comparison of methods. Marie Wiberg

2008

Em No 61. SYSTEMATIC EQUATING ERROR WITH THE RANDOMLY-EQUIVALENT GROUPS DESIGN: AN EXAMINATION OF THE EQUAL ABILITY DISTRIBUTION ASSUMPTION. Per-Erik Lyrèn

Em No 62. PSYCHOMETRIC EVALUATION OF A SELF-EFFICACY SCALE FOR DRIVER COMPETENCE USING THE RATING SCALE MODEL. Anna Sundström

The Swedish Scholastic Assessment Test (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 twelfth meeting of the scientific advisory board. A list of 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.

Department of Educational Measurement
Umeå University 901 87 Umeå
http://www.edmeas.umu.se
Tel: 090-786 50 00 Fax: 090-786 66 86