# NOTES FROM THE FIFTH INTERNATIONAL SweSAT CONFERENCE Umeå, May 31 - June 2, 1997

**Christina Stage**

Em No 28, 1997


ISSN 1103-2685
ISRN UM-PED-EM--28--SE

The SweSAT has by now been in existence for 20 years and has become an integrated and generally accepted part of the Swedish educational system. The International Scientific Advisory Board, was set up in 1992, and met for the first time in Umeå in May 1993[1]. Since then the board has met once a year[2] and the fifth meeting was held in Umeå in May/June 1997.

This report is a condensed summary of the fifth meeting. The main topics for the meeting were: Test Development, Item Response Theory (IRT) and Validity. The program for the conference as well as a list of participants are enclosed.

## *The SweSAT program since April 1996*

***Christina Stage***

The most important event in the SweSAT program during the past year is the administration of the ”new” SweSat one week after the fourth meeting of this board. As you all know we had received permission, on a two years experimental basis, to pretest new items in the regular test administration. On the very first meeting in 1993 members of this board recommended us to pretest items on the regular test-takers in stead of students in upper secondary school, which was the case earlier. This necessitated some changes and restructuring of the test and also a lot of negotiations with the central authorities. In No-

---

[1] Wedman, I. & Stage, C. (1994) *Notes from the first International SweSAT Conference May 23-25, 1993* (Educational Measurement No 9). Umeå: Umeå university, Department of Educational Measurement.

[2] Henriksson, W., Henrysson, S., Stage, C., Wedman, I. & Wester, A. (1994). *Notes from the Second International SweSAT Conference. New Orleans, April 2, 1994.* (Educational Measurement No 10). Umeå: Umeå university, Department of Educational Measurement.

Stage, C. & Henriksson, W. (1995). *Notes from the Third International SweSAT Conference. Umeå, May 27-30, 1995*. (Educational Measurement No 16). Umeå: Umeå university, Department of EducationalMeasurement.

Stage, C. (1996). *Notes from the Fourth International SweSAT Conference. New York, April 7, 1996*. (Educational Measurement No 20). Umeå: Umeå univeristy, Department of Educational Measurement.

vember 1995 a decision was taken by the National Agency for Higher Education that, starting spring 1996 and on a trial basis, the test should be restructured and that it should contain one part with pre-test items.

In spring 1996 the number of testtakers was larger than ever, perhaps because a lot had been written in the newspapers about the ”new” test. Everything went well and the testtakers were astonishingly agreeable; of course there were some complaints, but twenty to thirty complaints from 82 500 testtakers is not overwhelming, considering the largest change since the test was introduced. The general information subtest had been excluded, the testtakers were no longer allowed to keep the test booklets, the order of the subtests was not known in advance and the testtakers had to work for 50 minutes on pretest items.

In the fall nothing serious happened with the test either. There are always fewer testtakers in fall than in spring but the number was larger than any fall before, 56 000.

This spring the number of testtakers increased still further to 82 973. Fortunately everything went well again and later on we will present some preliminary results from the new pretesting.

This spring the National Agency for Higher Education has established an Advisory Board for the SweSAT and Other Special Selection Tests. The members of this Board have been personally chosen by the director-general of the Agency. So far the Board has only met once and among the mebers are Jan-Eric Gustafsson and Christina Stage.

### *SweSAT in 20 years from now - A prophecy*

***Ingemar Wedman***

Background - then, now and the future
SweSAT - an intelligence test
The marking system
A more encompassing concept of validity
Research in connection with SweSAT

Reporting
Psychometrics
CRT-information out of NRT-information
SAT/ACT in Sweden (incl. computerized testing)

There have never been any big problems with SweSAT maybe because it has developed to become more of an intelligence test than a performance test which was the original intention. The exclusion of the subtests on study techniques and general information has in a way destroyed the test. As it is now we could as well use translations of SAT or ACT as develop a test of our own.

The intensive debate of the marking system in upper secondary school is probably the main reason why SweSAT has not been much discussed. If/when that discussion has calmed down there will probably be more focus on the SweSAT.

There should be a more encompassing concept of validity and more adjustment to the concepts of Messick and Cronbach.

The reporting of test results to testtakers and massmedia should be improved; it should be possible to get CRT-information out of NRT-information.

The research in connection to the test must be intensified; the answers should be there when the questions are asked.

*”The fact that it is statistically difficult to evaluate the predicitive ability in an admission instrument does not mean, and must not mean, that the selection of an instrument is arbitrarily made, on the opposite. At the lack of conclusive statistical data great effort has to be made to judge and contentwise evaluate the instruments. A choice of a certain instrument based on the fact that this instrument reflects an important content is fully acceptable even though conventional prediction studies later on show low correlation coefficients between the predictor and the criterion. This attitude is in fact more inviolable than an attitude to uncritically choose an instrument solely based on statistically matters. (Henriksson, Henrysson, Stage & Wedman, 1985)*

*"The conclusion that can be drawn, even if the sample is small, is that the predictive ability of a selection instrument may not be captured by one correlation coefficient only. The reality of life of students is much more complicated than this, which is not a new observation. Hackman & Taber in a study back in 1979 already noted that success and failure in academic studies may be a result of various interactive factors." (Wester, 1995)*

### *News from Germany, Israel, the Netherlands and the US.*

*Michal Beller* gave a short briefing on what had happened in Israel during the last year. There had been no controversies about applications. The activities of coaching schools are still at the same level (everybody seems to take coaching courses). NITE is using CAT for the test of English as a foreign language. NITE has also got National Assessment on contract with the Ministry of Education. The pendulum against MC questions is now going back again. The pressure on Matriculation exams will probably ease; which topics will be given as exams will be decided by lottery.

*Ron Hambleton* reported that CAT had been receiving increased attention from testing agencies in the US. Credential agencies seem especially interested, though major admissions testing programs such as GRE and GMAT are using the CAT paradigm. But there remain major questions about CAT and one of them is test security. The current lawsuit between ETS and is a good example of the problem of test security.

Another problem of CAT is the size of the item bank. It is well known that item banks will need to be very large to address the problem of test security. At the same time, the number of items in an item bank may be misleading because some current research suggests that many items in a bank are never used. For example, perhaps there are 2000 items in a bank but about 300 items may be the ones being selected for tests. Current interest is centered on forcing some of the less attractive items into the CATs.

Related to the problem of item bank size is the costs associated with adding calibrated and validated items to an item bank. It was reported recently, that College Board might spend 50 million dollars to create a new item bank for the SAT. This is a huge expenditure.

Professor Hambleton also mentioned that performance assessment remains strong in the US but there is less demand to replace 100% of the selected response items with constructed response items in tests. Even the most ardent supporters of constructed response items seem willing to recognize that in the future testing programs may be built around both selected response and constructed response items.

Finally professor Hambleton noted the important publication of Nancy Cole and WarrenWillingham on gender differences, and their findings about the significance reduction in gender differences in mathematics ans science over the last 30 years. The book which is titled *Gender and Fair Assessment* is the culmination of four years of work by several researchers using data from more than 400 different tests.

Among other things *Wim van der Linden* observed that in the Netherlands admittance to higher education had been discussed during the last year. A GPA-based lottery system has been used and one case which has attracted a great deal of attention during the year was a girl with a GPA of 9.5 (the average is 6-7) who was not admitted. A committee was appointed which after six months suggested that 40 percent of the places should be allocated directly on GPA, and the remaining places should be allocated by lottery. The decision of the Ministry was that 10 percent of the places will be allocated on GPA and the rest by lottery. CITO is at present working with a new item-banking system.

*Günter Trost* reported that the TMS has been abolished in Germany, mainly for economical reasons (in Germany there have been no fees for the testtakers). Another reason was that the number of applicants has decreased to a ratio of 2:1. In Zwitzerland, however, the German speaking cantons except for Zürich will use the TMS for selection to Medical studies beginning next year. Belgium as well is introducing a new selection system for Medical studies; a very comprehensive program in which three parts of the TMS are included.

## Test Development

### *Experiences from the new pretesting model*

***Kerstin Andersson & Gunilla Ögren***

Starting spring 1996 the SweSAT consists of five subtests with a total of 122 items and with four hours and 10 minutes effective testing time. Each subtest constitutes a 50-minute section with one exception: ERC and WORD together form one 50-minute section where 35 minutes are taken up by ERC and 15 minutes by WORD.

**Tabell 1** *The SweSAT since 1996*

| **Block** | **Subtests** | **Items** | **Time** |
|---|---|---|---|
| I | DS | 22 | 50 min |
| II | DTM | 20 | 50 min |
| III | ERC + WORD | 20 + 40 | 35 + 15 min |
| Total test | | 122 | 4 h 10 min |

The order between the subtests is no longer fixed; in Table 1 they are simply listed in alphabetical order.

The test is given on the same day all over the country and since Sweden is rather longish it is necessary to decentralise the administration. The test is administered to various places in the country by the 23 centres of higher education. The number of testtakers varies a lot between these centres: Stockholm, Lund and Gothenburg are the three biggest centres. Together they account for about 40 per cent of the testtakers, while the three northenmost centres together account for only about 11 per cent.

As already pointed out the pretesting is carried out in a hidden section of the test. Each time there are roughly 20 different booklets with new material. These booklets are distributed over the different centres of higher education so that each centre gets to do one pretesting section which is the same for all testtakers at that particular centre. The cen-

tres with small numbers of testtakers are combined with other centres so that the minimum number for each pretest booklet will be approximately 2000. So far this pretesting procedure has been used three times and some of these items have now been used in regular tests.

Not only do the number of testtakers vary between the centres of higher education, but so do the scores obtained at these centres. For the whole test given on spring 1996 the mean score was 76.22. The minimum score obtained was 72.05 and the maximum score was 81.87. This means that the difference between the extreme centres was 9.82.

When the subtests are assembled there are above all three item indices that are decisive. First of all the biserial correlations must be acceptable, which usually means that they must be at least. 30 for an item to be chosen; secondly the level of difficulty should be varied in a special way for each subtest and finally the gender differences must be acceptable. Hence these are the crucial results from the pretesting.

In the last test - spring 1997 - quite a number of items had been pretested with the new model and so the consistency at item level may be studied.

So far the DS subtest has only used one item from the new pretests and the result seems very promising as the level of difficulty as well as the gender difference are roughly the same from pretest to regular test.

As for the two items used in the DTM subtest one item has become more difficult and one has become easier.

In the ERC subtest where 14 items have been pretested in the new system some items have become easier, and some items have become more difficult. In sum, however, the difference is not alarming, only .04 more difficult. The sum of the gender differences on these 14 items is exactly the same for the pretest version and the regular test.

The READ subtest as a whole has become almost two points easier. One reason for this may be that text number two used to be the last text in the pretest, meaning that it was probably more difficult at the

pretest since the testtakers were probably short for time. In the regular test the text was number two and all items have become easier. More must be found out about the significance of the position of a text for the level of difficulty.

In the WORD subtest as many as 31 items out of the total 40 had been pretested with the new model. Some of the items have become easier, and some have become more difficult. As a whole the subtest has become .37 easier, which is not too bad a result, considering the large number of items.

One great problem is that items are often revised after the pretesting. These revisions are necessary but may also be one of the explanations for the differences in results between the pretest and the regular test.

From this overview it may be concluded that the results of the new pretesting procedure are not so consistent as we had hoped for, but they are far better than the results of the old pretesting procedure.

The general problems with pretesting items for the SweSAT are:

- The obvious problem of secrecy
- The Swedish principle of public access to official records
- The administration, which makes it impossible to randomise the tests on testtakers instead of centres of higher education
- The economy, which prohibits the possibility of items hidden in each test booklet
- The differences in results between the centres of higher education
- The time for the pretest - 50 minutes is actually too much for the pretest
-

## *Estimating Item Statistics with Judgmental Data and Small Examinee Samples*

***Ronald K. Hambleton***

### Literature Review:

Research spans 70 years and has been rather continous. We located over 50 published articles in our review. Mislevy and Sheehan´s work is especially encouraging (but is very labor intensive). Tatsuoka too, has some promising results, but this work, too, is labor intensive.

Often raters are capable of getting the order of difficulty correct, but the absolute levels of difficulty are elusive. Panelists are often unaware of the ability levels of the candidate pool.

Some of our recent work on standard-setting and score reporting suggest new ways to anchor the difficulty estimates and potentially improve the estimates of item difficulty.

### Examples of Factors Affecting Item Difficulty:

1. Negations: the greater the number, the more difficult the item
2. Referential: the greater the number, the more difficult the item
3. Vocabulary: the more multisyllabic words used, the more difficult the item
4. Sentence and paragraph lenghts affect item diffficulty
5. Abstraction of text
6. Location of relevant text: apparently, when the relevant material is in the middle of a passage, the item is harder for the candidates
7. The levels and numbers of cognitive skills needed to solve the problem affect the difficulty

8. The novelty of the item format

9. The placement of the item in the test

**Table 2** *Anchor-Based Method*

| Proportion Correct | | | | |
|---|---|---|---|---|
| .00 | .25 | .50 | .75 | 1.00 |
| | Desc. (.25) | Desc. (.50) | Desc. (.75) | |

1. Review the three anchor points descriptions (chosen to be .25, .50 and .75 on the item difficulty scale for this field-test) in terms of content, cognitive skills, item format, and a sample item or two, etc.
2. Read each item or set of items (associated with a common stimulus such as a passage or problem statement) and then decide whether individual exam items are harder or easier than those with known item difficulty levels. Estimate item difficulty for these new items and place this item difficulty estimate in the column provided for the rating form. (Round 1)
3. Recieve feedback on panelists´ item difficulty estimates, and discuss this information, and ultimately revise your item difficulty if you feel revisions are in order. (Round 2)

Item Mapping Method Steps

**Table 3** *Proportion Correct*

| .00 | .20 | .40 | .60 | .80 | 1.00 |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | |

Items

(Actually, item p-values will be used)

1. Review the six exam items mapped onto the item difficulty scale for this field-test. Try to determine what makes some items more difficult or easier than others.

2. Read each set of items (associated with a common stimulus such as a passage or problem statement) and then sort into categories:

   category 1 : .00 to .24

   category 2: .25 to .49

   category 3: .50 to .74 and

   category 4: .75 to 1.00

   Record your ratings on the round 1 Rating Form. Estimate item difficulty and place the estimate in the column provided on the Rating Form. (Round 1)

3. Receive feedback on panelists´ placement of items and item difficulty estimates, and discuss this information, and ultimately revise the category placement and item difficulty estimates. (Round 2)

4. Receive feedback on the difficulties of several exam items and then reconsider item difficulty estimates. (Round 3)

5. Evaluate the advantages and disadvantages of prior collateral information in the estimation process compared to candidate sample sizes.

## Item Response Theory

### *Optimal Constrained Adaptive Testing*
***Wim van der Linden***

The basic idea underlying CAT was to increase the precision of ability estimates. The idea is to start at a certain point, choose the best item, reestimate the ability and choose the next item which is optimal at the current ability estimate etc.. Because the ability estimate is updated after each new response, the procedure is adaptive and item selection converges to optimality. Advantages with CAT is shorter test length at the same precision and flexibility of test administration.

The problems with unconstrained CAT are a) face validity, since the tests do not have equal composition among examinees. b) the risk to overexpose some items (security problem) and c) that sets of items may be linked to a common stimulus or items may contain clues to other items.

These problems can be solved by constrained CAT: the content and distribution of items in the adaptive test must meet the specifications of the original test. Wainer & Kisley for example do not select items but small testlets. This is a clever way to build in a content factor specification. Sets of items may be handled as a constraint - prestructuring. Item exposure as well may be handled as one of the constraints.

## *IRT and the SweSAT*

***Christina Stage & Kristian Ramstedt***

As you all know, hitherto SweSAT has been constructed and assembled in accordance with the classical test theory. This year, however, we have started to examine the possibilities to use IRT. In two aspects especially we expect to to be able to improve the test by using IRT and that is in the test design or choice of test items and in equating of different versions of the test.

So far some preliminary studies only have been performed on the possibility to use IRT in the development and evaluation of SweSAT. Attempts have been made to adjust IRT models to the test results from spring and fall 1996. One conclusion of these preliminary analyses is that a three parameter model is needed for the SweSAT data. There is quite a difference in the discriminating power of the items and guessing is definitely present. We have been working with the three parameter logistic model and mainly the BILOG program.

Initially the model was adjusted to the five subtests separately but according to the statistical tests of model data fit there were quite a few items in each subtest for which there was significant misfit (see Table 1). One reason for the large number of misfitting items on subtest level was thought to be the small number of items in each subtests. The small number of items could result in bad estimates of the ability parameter. This assumption is supported by the fact that the number of missfitting items was the least in the WORD subtest, where the number of items is the highest.

Gustafsson (1996) has shown that even though there are two main dimensions measured by test there is also a considerable relation between these two dimensions which indicates a general factor. In Table 1 the intercorrelations between the five subtests are shown.

**Table 1** *Correlations between the subtests. (Within brackets reliabilities - $KR_{20}$)*

| | DS | DTM | ERC | READ | WORD |
|---|---|---|---|---|---|
| DS | (.82) | | | | |
| DTM | .68 | (.72) | | | |
| ERC | .43 | .43 | (.76) | | |
| READ | .44 | .44 | .58 | (.68) | |
| WORD | .31 | .35 | .57 | .61 | (.85) |
| TOT | .72 | .72 | .77 | .78 | .80 |

Since there are fairly high intercorrelations between the subtests (see Table 5) and it seems reasonable to assume a general factor which is present in all the subtests the next step was to adjust the three parameter logistic model to the test as a whole. The number of statistically significantly missfitting items decreased substantially when the whole test of 122 items was used. In Table 6 the items in each subtest for which there were significant model data misfit are presented.

**Table 2** *The significantly misfitting items in each subtest, when the subtests are run separately and when the total test is used. (* significant at .01 level*

| | |
|---|---|
| WORD - WORD | 6*, 9, 14, 18, 19, 21 |
| WORD - total test | 39* |
| DS - DS | 1, 3, 4*, 5*, 6, 7*, 8, 9, 15*, 16, 20 |
| DS - total test | 14* |
| READ - READ | 2*, 5*, 6*, 9*, 10, 13*, 15*, 16*, 19* |
| READ - total test | 2 |
| DTM - DTM | 1*, 2*, 6*, 7*, 8*, 13*, 14*, 18*, 19 |
| DTM - total test | 2*, 12 |
| ERC - ERC | 4*, 7*, 9*, 11, 12, 13, 16*, 17, 18, 20* |
| ERC - total test | - |

As may be seen in Table 2 there are only two items for which there is significant misfit on subtest as well as the whole test level; these items

are number 2 in the READ subtest and number 2 in the DTM subtest. The item fit ot these items are shown i Figure 1 for the READ item and Figure 2 for the DTM item.

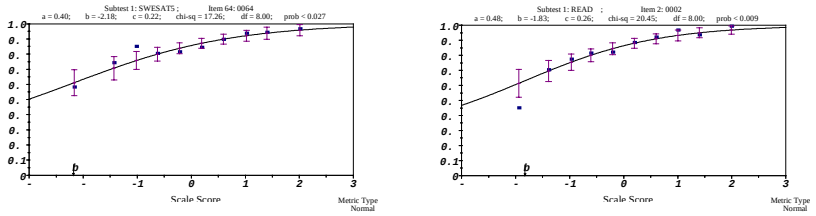


**Figure 1** *Model data fit for item 2 in the READ subtest parameters estimated on the total test results (left) and parameters estimated on subtest results (right).*

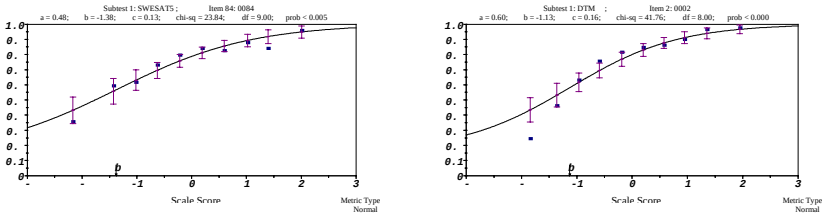


**Figure 2** *Model data fit for item two in the DTM subtest.*

As may be seen in Figure 1 and Figure 2 the misfit does not seem to be very serious for neither the READ nor the DTM item.

Another comparison was of the item parameters estimated on subtest level and total test level. In Figure 3 the b-parameters estimated on subtest-level are plotted against the b-parameters estimated on total test level.

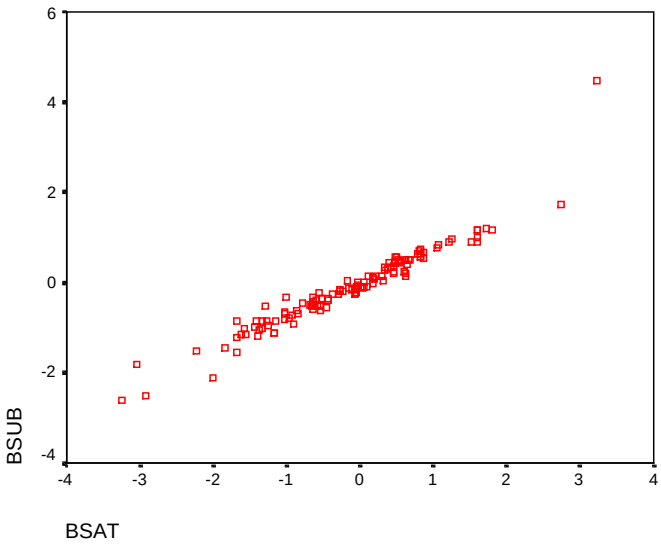


**Figure 3** *b-parameters estimated on subtest level plotted against b-values estimated on total test.*

As may be seen in Figure 3 the estimates correspond fairly well, even though the b-values estimated on separate sub-tests generally seem to be lower than those estimated on the total test. The correlation between b-parameter estimates was .96. The correlations for the different sub-tests were: DS r = .97, DTM r = .99, ERC r = .99, READ r = .99 and WORD r = .95.The most deviating item was item seven in the WORD sub-test which, however, had not turned out as significantly misfitting in any statistical test. The ICC of item seven is shown in Figure 4.

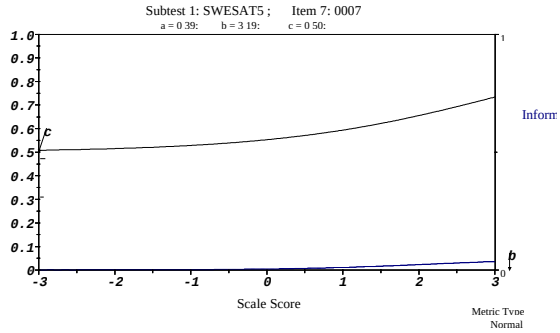


**Figure 4** *ICC of item seven in the WORD subtest.*

As may be seen in Figure 4 item seven in the WORD subtest was a very poor item with very low discrimination and also very low information. This item should have been removed from the test had the analysis been made by IRT. When item seven is removed the correlation between the b-parameters for the sub-test WORD is r = .98.

In Figure 5 the discrimination parameters - a - estimated on separate sub-tests are plotted against the a-parameters estimated on the total test.

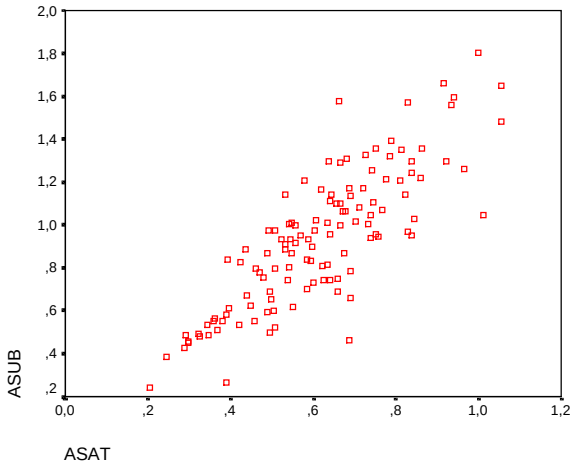


**Figure 5** *a-parameters estimated on subtest level plotted against a-parameters estimated on total test.*

As may be seen in Figure 5 the correspondence between a-parameters estimated in different ways is fairly good even though the parameters estimated on sub-tests are generally higher than the parameters estimated on the total test. The correlation between the a-parameters was r = . 83. For the separate sub-tests the correlations were: DS r = .88, DTM r = .91, ERC r = .93, READ r = . 91 and WORD r = .74.

In Figure 6 the pseudo guessing parameters - c - estimated on subtest level are plotted against c-values estimated on total test level are shown.

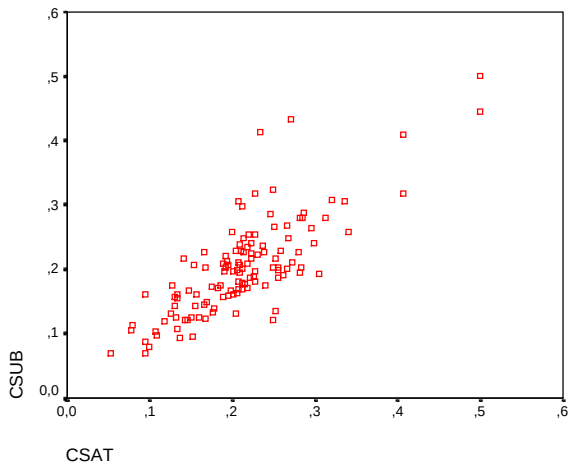


**Figure 6** *c-parameters estimated on subtest level plotted against c-parameters estimated on total test.*

The correlation between the c-parameters estimated on subtests and the total test was r =.80. The correlation for the DS subtest separately was r = .69, for the DTM subtest it was r = .84, the ERC subtest r = . 85, the READ subtest r =.74 and the correlation for the Word subtest was r = .83.

Even though the results so far seem to be promising they are still very preliminary and a lot of work remains to do and several questions need to be answered before IRT can be taken into regular use. Some crucial questions are 1) the importance of violation against the assumption of local independence. (In three of the subtests two or more questions are

to common passages.) 2) the assumption of unidimensionality. The unidimensionality is less when the total test score is used still the model data fit seem to be better on total test level than on subtest level. 3) the different computer programs (BILOG and XCALIBRE) give different parameter estimates etc.

### *Equating of the SweSAT*

***Christina Stage***

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 to normed scores should make it possible to compare scores from one test administration to another, i.e. it should be as easy or difficult to obtain a certain normed score on one test as on another.

The normed 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 represent the raw score. In order to ensure that scores on different test administrations are comparable the raw scores are converted to normed scores. The strategy applied to define scale limits for the normed scores is based on a combination of comparisons.

#### Preequating:

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

#### Equating:

a) The total group of testtakers is examined and compared to earlier populations regarding sex, age and background education

b) Reference population I is selected through proportional stratified selection from the total group in order to provide the same distribu-

tion of sex, age and education. Reference population I is approximately 10 percent of the total number of testtakers.

c) Reference population II consists of those testtakers who are 18 years old and are still registered in a 3-year upper secondary theoretical study program. Reference population II usually is around 20 percent of the testtakers.

The results of these three groups are studied simultaneously at subtest level as well as for the whole test. The work starts from the top, i.e. normed scores 2.0 and 1.9 by matching the results of this particular test as closely as possible against the results of previous test administrations. The raw scores are then distributed over the score intervals of the normed scores. Each normed score interval represents three or four raw scores and the aim is, of course, to find as optimal boundaries as possible compared with previous tests.

### *Research Study: Validity of IRT Equating of the SweSAT*

***Wilco Emons***

A problem related to equating two tests which have been administered to different populations, is the possible existence of differences in both the difficulty of the examinations and the ability-distribution of the examinees. The equating procedure has to account for both fluctuations. The problem sketched above can be addressed by item response theory (IRT) equating, because item and examinee characterstics are modeled by separate sets of parameters.

In February 1997, a study has been started to the applicability and appropriateness of IRT based equating for the SweSAT. The purpose of the study is to explore if IRT equating can be used to improve and simplify equating of the SweSAT. Therefore, an IRT equating set-up will be developed, implemented and evaluated.

The equating method considered is IRT Observed Number of Correct Score Equating. In IRT observed score equating, an appropriate IRT model is used for generating estimated observed score distributions.

Using these distributions, the scores are then equated by a conventional equipercentile equating method. IRT-observed score equating entails choice of a test administration design, choice of an IRT model, accurate parameter estimation, evaluation of modelfit, generating observed score distributions and equipercentile equating. A more detailed description of these elements will be given below.

## Test-administration design

The test administration design for a single subtest is depicted inFigure 1. This design was introduced in 1996. From Figure 1, it can be seen that each test consists of the actual examination accompanied by a booklet with try-out items. A number of the try-out items are used in the next examination. Only common items which have not been altered between try-out administration and actual administration can support equating.

**Figure 1.** *The SweSAT test-administration design.*

An overview of the number of common items is given in Table 1. The numbers of common items which are changed are also given. It should be noted that the number of unaltered common items is relatively small. The expectation is that this number will increase the next years.

The number of common items needed for successful equating will be a point of study in further research.

**Table 1.** *Number of common items in the 1996:A and B and the 1997:A administration.*

Data Sufficiency

| | | |
|---|---|---|
| 96A | 96B | none |
| 96A/B | 97A | 1 item |

Diagrams, tables and maps

| | | |
|---|---|---|
| 96A | 96B | 2 items, both changed |
| 96A/B | 97A | none |

English reading comprehension

| | | |
|---|---|---|
| 96A | 96B | none |
| 96A/B | 97A | 14 items, two changed |

Swedish reading comprehension

| | | |
|---|---|---|
| 96A | 96B | none |
| 96A/B | 97A | 16 items, 6 changed |

Word

| | | |
|---|---|---|
| 96A | 96B | 9 items, 5 changed |
| 96A/B | 97A | 31 items, 22 changed |

## Choice of an IRT model

An important step in IRT-equating is the choice of an appropriate IRT-model, that is a model that fits the data satisfactorily. In the present situation, only IRT models for dichotomously scored items have to be

considered, because all items are scored right or wrong. A number of questions need to be considered for determination of an appropriate IRT model: Do items discriminate differently? Is a common factor underlying the responses? Do items differ in level of difficulty? Is guessing involved? Item characteristic curves (icc´s) are studied to get answers to these questions. From these icc´s, it was found that items do indeed discriminate differently and do have different difficulties. It was also found that one factor is underlying the responses and that guessing occurs. The conclusion can be drawn that the three parameter logistic model seems to be the most appropriate model. This finding corresponds to the conclusion found in a study on modelfit of an IRT model for the Data Sufficiency subtest and the Diagram,Tables and Maps subtest[3].

## Estimation of parameters and fixing the scale

The parameters will be concurrently estimated and therefore placed on a common scale. Advanced computer software, such as Bilog and BilogMg, is available for estimating all model parameters in incomplete designs. A related problem is the indeterminacy of the scale, which will be solved by fixing the mean and variance of the ability distribution of the reference population to zero and one, respectively.

## Evaluating modelfit

Evaluating modelfit has always been emphasized as a very important aspect when applying IRT. However, moderate efforts will be invested in trying to find an IRT model which fits perfectly on item level, because the purpose is reliable equating, rather than finding the best fitting IRT model. So in the present framework, modelfit relates to the question whether applying more complicated models will result in significantly different equating functions.

## Generating observed score distributions

For each population, expected observed score distributions are generated for both the old and the new examination. The new examination

[3] Stage, C. (1996) Em No 19 and (1997) Em No 21.

is not actually made by the population presented the old examination, so the expected new examination score distribution must be viewed as an estimate of the performance of the old populationconfronted with the new examination. Then for each population, the expected raw scores are equated by conventional linear equating, that is scores on the same percentile rank are assumed t be equivalent. To illustrate the principle, consider the population from autumn 1996, depicted in Figure 1. For this population, two distributions will be generated. Firstly, the expected observed score distribution of the SweSAT 1996:B, that is, the test they have actually made. Secondly, the expected observed score distribution for for 1997:A, which is an estimate of the score distribution in case this population had been administered the SweSAT of 1997:A. Because, the ability distribution of the population is kept fixed, differences between the expected score distributions should be explained by differences in difficulty between the two tests.

### Evaluation and further research

Confidence intervals for the estimated observed score frequencies will be estimated to evaluate the reliability of equating. Further research will focus on the number of common items needed. Also the question whether to equate on subtest level or total test level will be considered.

## Validity

### *Meta-analytic studies of validity*[4]

***Michal Beller***

A meta-analysis was conducted across 1,888 studies which investigated the predicitve validity of the means of selection to universities in Israel. The criterion was the grade point average at the end of the first

[4] The complete study is printed in Kennet-Cohen, T., Bronner, S. & Oren, C. (1995). *A Meta-Analysis of the Predictive Validity of the Selection Process to Universities in Israel*. 202, NITE, Jerusalem.

year of university studies, The predictors were the components of the admissions procedure: the Psychometric Entrance Test (PET), the average of the grades on the high school matriculation certificate (Bagrut) and a composite score consisting of PET and Bagrut. In addition the validity of each of the PET´s three subtests as individual predictors, was examined. The validity studies were conducted at the departmental level, for all areas of the study, at six universities, for the academic years 1984-1992.

A meta-analysis of the correlations using artifact distributions was conducted. The meta-analysis corrected for three artifacts: sampling error, restriction of range of the independent variable, and error of measurement in the dependent variable.

The findings regarding the actual (i.e. corrected) correlations of the six predictors with the criterion were as follows: the predictive validity of the composite score (0.66) was higher than the individual validities of each of its two components (0.54 and 0.47 for PET and Bagrut respectively). Similarly, the validity of PET was higher than the individual validities of each of its subtests (0.45, 0.38 and 0.30 for Quantitative Reasoning, Verbal Reasoning and English respectively).

### *Item-bias on the SweSAT*[5]

***Christina Stage***

As study was conducted with the main purpose to investigate whether the total test score had the same interpretation for males and females. The examination was performed by comparing the results on subtest and item level of groups of males and females who had the same normed score on the test. The normed score 1.3 was chosen as it represents a result clearly above average, which is competitive for many educations, and it is also a level where the numbers of males and females are roughly the same.

[5] This study is described in Em No 23, Department of Educational Measurement, Umeå University

The comparison of males and females with a normed score of 1.3 demonstrated that the composition of scores was very different for the two groups. The effect sizes of differences on the subtests varied from .03 to .65: on the DS and DTM subtests the effect sizes were .55 in favour of males, on the READ and WORD subtests the effect sizes were in favour of females and .65 and .40 respectively, only on the ERC subtest was the effect size negligible (.03 in favour of females).

On test item level there were significant differences in p-values between males and females on 95 items out of the total number of 122 items and on 30 of these items the differences were larger than .10. This further supports that males and females with the same overall result (i.e. the same normed score) have achieved this score in different ways.

In order to examine whether there were items in the test which should be judged as gender biased two MH analyses were performed. One analysis was made on the males and females who had a normed score of 1.3 and with the subtest scores as matching variable. As a comparison an analysis was also performed on a random sample of testtakers and with the total normed score as matching variable.

In the MH analysis on results of the testtakers with a normed score of 1.3 items from the subtest WORD only were flagged as large DIF[6], six of the flagged items were favouring females and five items were favouring males.

In the MH-analysis on random groups four DS items, three DTM items and three WORD items were flagged as favouring males while one READ item and five WORD items were flagged as favouring females.

[6] For a description of negligible, intermediate and large DIF see Dorans & Holland (1993). DIF Detection and Description: Mantel-Haenszel and Standardization. In Holland & Wainer (Eds.) *Differential Item Functioning*. Hillsdale, New Jersey, Lawrence Erlbaum Associates.

For illustrative purposes the ICC:s of the items flagged as large DIF were provided.

## *SweSAT and Grades - A comparison*

***Jan-Eric Gustafsson***

The presentation was structured into two different parts, one dealing with effects of local dependence among test items on estimates of reliability, and the other with taking selectivity into account when investigating relations between aptitude tests and school performance.

### Local dependence and reliability

It was first demonstrated, within a framework of confirmatory factor analysis, that local dependence among groups of items affects reliability adversely whenever the local dependence is caused by factors which may be regarded as random sources of variation. As an example, it was argued that reading tests composed of a limited number of texts to which a larger number of items is related are afflicted by a random form of local dependence (or random multidimensionality), because choice of texts is arbitrary and non-replicable. It was also demonstrated that the effect on estimates of reliability is a function of the number of sources of local dependence, the number of items affected by each source, and of the size of the effect.

Using data from the reading subtest of the SweSAT 92A it was then demonstrated that a unidimensional confirmatory factor model fits the 24 items reasonably well. Some improvement of fit was, however, obtained when additional latent variables representing local dependence due to the four texts were introduced. The estimate of reliability based on the unidimensional model was .687. The estimate of reliability based on the multidimensional model was .694 assuming that the text-factors contribute systematic (”true”) variance, while it only was .642 assuming that the text-factors contribute non-systematic (”error”) variance.

It is concluded that standard techniques for estimating reliability of tests within which there is local dependence among items grossly overestimate reliability.

### Estimating relations between aptitude and achievement in selected groups

It is a considerable challenge to determine the exact nature of the patterns of relations between school achievement and aptitude variables. One reason for this is that school achievement is multidimesional, and another reason is that all school systems sooner or later involve some kind of differentiation, with different groups of students taking courses with a different orientation at different levels of difficulty. In Sweden, for example, the nine-year comprehensive school is more or less undifferentiated, all students receiving grades in 17 subject matter areas. Data on the SweSAT is, however, only available for a subset of the students who typically are in academic programs. Because students elect to take the test on the basis of interest and previous achievement there are strong effects of self-selection.

In the presentation a latent-variable modeling approach for incomplete data is used to solve these problems. A starting point is taken in grades assigned at the end of the comprehensive school, with structurally missing data for those not having taken the SweSAT. In this way a latent-variable model comprising all grades and all students is fitted.

The study uses data about grades from official registers for all students born in 1972 who took the SweSAT 92A (some 15 000 students). For grades from comprehensive school a model with four latent variables is fitted: one general school achievement factor, a language factor, a math-science factor, and a non-verbal achievement factors. For the SweSAT a two-factor model with one general (reasoning) and one verbal factor was fitted. According to the model there was a high relation between the general achievement factor and the general SweSAT factor (.68), as well as between the language factor and the SweSAT verbal factor.

Comparisons also were made with the results obtained when analyzing only cases with complete data. It is concluded that this methods tends to yield lower and less consistent estimates of the amount of relation between individual differences in school achievement and performance on the SweSAT.

## *Fairness of TMS towards Female and Male Applicants*[7]

***Günter Trost***

Admission to all medical schools in Germany except one is primarily based on two selection criteria:
a) the average mark in the secondary school leaving certificate
b) the total score in the ”Test for Medical Studies” (TMS) a scholastic aptitude test

In the context of a large-scale longitudinal study the fairness of the admission system towards relevant sub-groups of applicants was analysed. The sub-groups were defined by:

- gender
- the type of secondary school they had attended
- their socio-econimic status; and
- practical experience in the medical field

One of the central questions of the study was: Did the introduction of the test into the selection system enhance or impair the fairness of the selection process?

In order to answer the question, the fairness model proposed by Anne Cleary was applied.

With respect to the fairness of the admission system towards male and female applicants, the results can be summarised as follows:

Male applicants are disadvantaged regardless of whether admission is based on school marks or on the test score. The degree of unfairness,

[7] From a study titled *Fairness in Admission to Medical Schools in Germany* by Günter Trost & Mathias Meyer, Institute for Test Development and Talent research, Bonn, Germany.

however, is higher if the selection is based on school marks alone, and it is lower if the test score is taken into account.

## *SOCIOECONOMIC AND GENDER DIFFERENCES ON THE SWESAT*

***Sven-Eric Reuterberg***

It is a well known fact that there are substantial gender differences in SweSAT scores in favour of male test takers. The socioeconomic differences are not equally well known since information on the test takers socioeconomic background has not been collected regularly. Within the frame of the ETF project (Evaluation Through Follow-up), however, this information is available for those persons who were born in 1972 and who took the SweSAT in the spring of 1991. For this group ETF has also collected scores on three intelligence tests (Verbal, Spatial and Numerical-logical factors). Furthermore, the leaving certificates from compulsory school are known.

The number of test takers with this information is 1030 and they have been categorized into two socioeconomic groups: Upper middle class and Lower middle class.

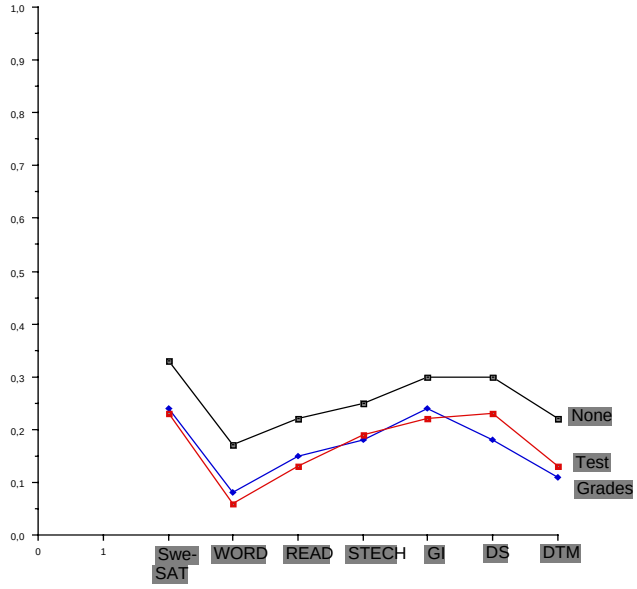


**Figure 1.** *Socioeconomic differences in SweSAT scores: actual and with test or grades as control variable.*

Figure 1 shows the mean differences in the total SweSAT score and in each subtest between the two socioeconomic groups. These differences are expressed as effect sizes. When no account is taken of differences in intelligence or grades from compulsory school, the total SweSAT score is about 1/3 of a standard deviation higher for test takers from upper middle class and when we control for differences in intelligence or grades these differences are decreased by about 0.1 unit of a standard deviation. Among the subtests, WORD shows the smallest differences and GI (General information) together with DS (Data sufficiency) show the highest.

As shown by Figure 2 the gender differences are substantially higher, and on the total SweSAT score males outperform females by 0.7 standard deviation units which corresponds to about 10 raw score points. Taking gender differences in intelligence into account reduces the SweSAT differences to 0.5 units but controlling for grades leads to an

increase of the SweSAT differences up to 0.9 units, and this is due to the fact that the female test takers have the highest grades. We can also see that the gender differences are particularly strong in the quantitative subtests DS and DTM.

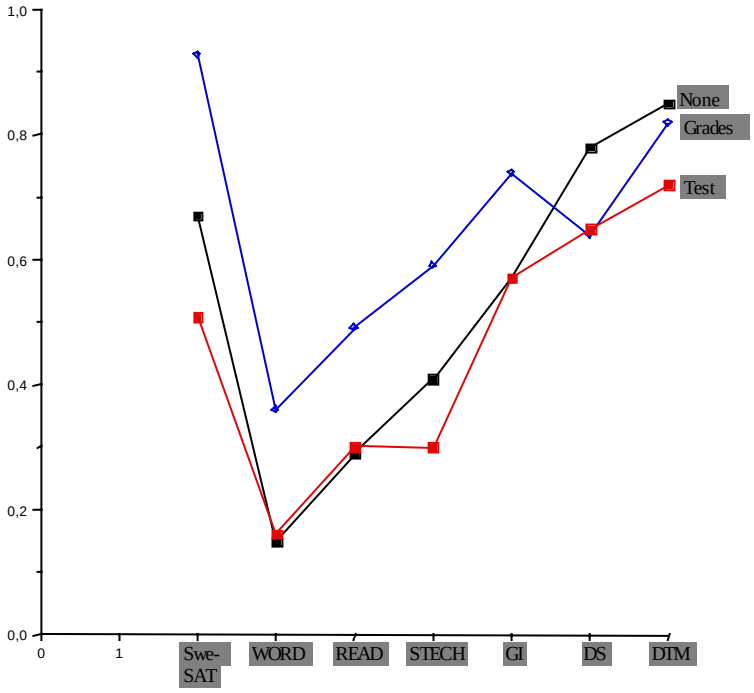


**Figure 2.** *Gender differences in SweSAT scores: actual and with test or grades as control variable. Effect sizes.*

In summary, there are both socioeconomic and gender differences in SweSAT scores and the gender differences cannot, by far, be explained by differences in intelligence or grades. The socioeconomic differences are much smaller, but there is a difference in favour of test takers from upper middle class - a difference which cannot fully be explained by differences in intelligence or grades.

The gender differences have been analyzed one step further in order to find out whether the differences are caused by differences in the latent factors captured by all or some of the SweSAT subtests or by specific factors connected to each subtest.

As shown by Gustafsson, Wedman and Westerlund (1992) the SweSAT measures two latent variables: a general analytic factor (A) with relations to each subtest and a verbal-knowledge factor (K‘) with relations to the verbal subtests. Besides these broader latent variables there are reasons to expect subtest specific factors. Table 1 shows for each subtest the significant gender differences in A, K‘ and the specific factors, respectively.

**Table 1.** *Gender differences in latent variables and subtest specific factors***.**

| Subtest | A | K´ | Spec |
|---|---|---|---|
| **WORD** | **+** | **-** | |
| **READ** | **+** | **-** | |
| **STECH** | **+** | **-** | |
| **GI** | **+** | **-** | **+** |
| **DS** | **+** | | **+** |
| **DTM** | **+** | | |

**+: Males outperform feamels.**
**-: Females outperform males.**

As shown by Table 1 the male test takers have the highest scores on the A-factor while females have higher scores on the K‘-factor. This means that the gender differences in the verbal subtests caused by A are reduced by the female superiority on the K‘-factor, but this is not the case for the quantitative subtests DS and DTM. This is why these two subtests show the greatest gender differences. The GI and DS subtests also are influenced by subtest specific factors favouring the males, and consequently, these specific factors contribute to the gender differences.

### *Conference participants*

Michal Beller, Israel
Ronald, K. Hambleton, USA
Wim van der Linden, The Netherlands
Günter Trost, Germany

Jan-Eric Gustafsson
Widar Henriksson
Sven-Eric Reuterberg
Christina Stage
Allan Svensson
Ingemar Wedman
Anita Wester

Inger Rydén Bergendahl

Kerstin Andersson
Wilco Emons, The Netherlands
Kristian Ramstedt
Simon Wolming