CLASSICAL TEST THEORY OR ITEM RESPONSE THEORY: THE SWEDISH EXPERIENCE

Christina Stage

EM No 42, 2003

ISSN 1100-696X
ISRN UM-PED-EM--42--SE

Paper written for, and published in Spanish by Centro de Estudios Públicos, Santiago, Chile. Available at www.cepchile.cl

This paper is a description of the efforts made to investigate whether item response theory (IRT) would be applicable to the Swedish Scholastic Aptitude Test (SweSAT). The aim has been to examine whether a switch from classical test theory (CTT) to IRT, in the process of item development, test design, scoring or equating, would improve the quality of the test. The paper consists of three parts. Part I, "The applicability of IRT models to the SweSAT sub-tests", is a summary of five reports (Stage, 1996, 1997a, b, c and d) describing the different steps taken, in investigating if an IRT model could be fitted to the five SweSAT sub-tests separately. Part II, "Comparison between item analysis based on IRT and CTT", is a summary of three earlier reports (Stage, 1998a, b and 1999) in which comparisons were made between CTT difficulty and discrimination indices and IRT difficulty and discrimination parameters on the three sub-tests ERC, READ and WORD. In part III, "Applicability of IRT to SweSAT: the total test", a description is given of an attempt to fit an IRT model to the total SweSAT. The conclusion was that since the model data fit was somewhat dubious, especially for the total test, there was nothing to be gained by switching from CTT to IRT.

Introduction

The Swedish Scholastic Aptitude Test (SweSAT) is a norm-referenced test, which is used for selection to higher education in Sweden. The test is administered twice a year, once in spring and once in autumn. After each administration that particular test is made public and therefore a new version has to be developed for each administration. As test results are valid for five years it is important that results from different administrations are comparable.

Since 1996 the test consists of 122 multiple-choice items, divided into five sub-tests:

1. **DS** a data sufficiency sub-test measuring mathematical, reasoning ability by 22 items.
2. **DTM** a sub-test measuring the ability to interpret diagrams, tables and maps by 20 items.
3. **ERC** an English reading comprehension sub-test, consisting of 20 items.
4. **READ** a Swedish reading comprehension sub-test, consisting of 20 items.
5. **WORD** a vocabulary sub-test consisting of 40 items.

Ever since the first version of SweSAT was taken into use in 1977, the development and assembly of the test as well as the equating of forms from one administration to the next has been based on the classical test theory (CTT).

In the classical test theory (CTT), which began to evolve with the Binet test almost a hundred years ago, the test score is viewed as made up of two components, a "true score" and an error. The true score and error are regarded as completely independent. The true score is viewed as unchanging from one form of a test to a parallel alternate form, and from one occasion to another. The error is considered to be unique to the specific measurement, and to be entirely

independent of the error that might be expected to appear on another measurement of the same construct. The true score can never be directly observed. It can only be inferred from consistency of performance from one test score to another.

CTT has been a productive model that led to the formulation of a number of useful relationships:

- the relation between test length and test precision (reliability)
- estimates of the precision of difference scores and change scores
- the estimation of properties of composites of two or more measures
- the estimation of the degree to which indices of relationship between different measurements are attenuated by the error of measurement in each.

Although CTT's major focus is on test-level information, item statistics (i.e. item difficulty and item discrimination) are also important. At the item level CTT is relatively simple, since there are no complex theoretical models to relate an examinee's ability or success on a particular item. The proportion of a well-defined group of examinees, that answers an item correctly (empirically examined) - the *p*-value - is used as the index for the item difficulty (actually it is an inverse indicator of difficulty, since higher values indicate easier items). The ability of an item to discriminate between high ability examinees and low ability examinees is expressed statistically as the correlation coefficient between the scores on the item and the scores on the total test.

CTT models are often referred to as "weak" models, because the assumptions of these models are easily met by test data.

There are, however, some shortcomings with CTT. One shortcoming is that item difficulty and item discrimination indices are group dependent; the values of these indices depend on the group of examinees in which they have been obtained. Another shortcoming is that observed and true test scores are dependent. Observed and true scores rise and fall with changes in test difficulty. Another shortcoming has to do with the assumption of equal errors of measurement for all examinees. The ability estimates are in fact less precise both for low and for high ability students than for students of average ability.

During the last decades a new measurement system, item response theory (IRT), has been developed and it has become an important complement to classical test theory in the design, construction and evaluation of tests. Within the framework of IRT it is possible to obtain item characteristics which are *not* group dependent; ability scores, which are *not* test dependent; and a measure of precision for each ability level.

According to Hambleton et al. (1991):

> *IRT rests on two basic postulates: a) the performance of an examinee on a test item can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; and b) the relationship between examinee's item performance and the set of traits underlying item performance can be described by a monotonically*

> *increasing function called an item characteristic function or item characteristic curve. This function specifies that as the level of the trait increases, the probability of a correct response increases. (p. 7)*

There are several different IRT models, but they all have in common the use of a mathematical function to specify the relationship between observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on the test. In any practical application of latent trait models one must specify the mathematical form of the item characteristic curves and obtain estimates of the item parameters needed to describe the curves. In the three-parameter model these parameters are:

1. item difficulty "b"
2. item discrimination "a"
3. a pseudo guessing parameter "c"

In the two-parameter model no guessing is assumed to exist, and in the one-parameter model item discrimination is assumed to be the same for all items.

IRT models are referred to as "strong" models, since the assumptions may be difficult to meet with test data. One important assumption included in the most common IRT models is the assumption of unidimensionality, which means that only one ability is measured by the items that make the test. What is required for the unidimensionality assumption to be met adequately is the presence of one dominant factor that influences test performance. Another and related assumption is that of local independence. Local independence means that when the abilities influencing test performance are held constant, examinee's responses to any pair of items are statistically independent.

Once a latent trait model is specified, the precision with which it estimates examinee ability can be determined for different ability levels. The information varies with ability level, which makes it possible to determine the standard error of estimate for different ability levels. The item information function gives information of the usefulness of the item in measuring ability at a particular ability level.

Presently IRT is receiving increasing attention from test agencies in test-design, test-item selection, in addressing item-bias and equating and reporting test scores. The potential of IRT for solving these kind of problems is substantial. It is essential, however, in order to achieve the possible advantages from an IRT model, that there is fit between the model and the test data of interest. A poorly fitting IRT model will not yield invariant parameters.

> *In many IRT applications reported in the literature, model-data fit and the consequences of misfit have not been investigated adequately. As a result, less is known about the appropriateness of particular IRT models for various applications than might be assumed from the voluminous IRT literature. (Hambleton et al., 1991. p.53)*

Hambleton et al. (1991) further warn against placing too much confidence in statistical tests, since these tests have a serious flaw: their sensitivity to examinee sample size. Instead the

authors recommend that judgements of fit of the model to test data be based on three types of evidence:

1. Validity of the assumptions of the model for the test data
2. Extent to which the expected properties of the model (e.g. invariance of item and ability parameters) are obtained.
3. Accuracy of model predictions using real and, if appropriate, simulated test data.

In the following parts of this paper results from several types of analyses are reported. The aim of these investigations have been to find different kind of evidence for or against fit of an IRT model to SweSAT test data.

I. The applicability of IRT models to the SweSAT sub-tests

IRT has a great potential for solving many problems in testing and measurement. The success of specific IRT applications is not assured, however, simply by processing test data through one of the computer programs…. The advantages of item response models can be obtained only when the fit between the model and the test data of interest is satisfactory. (Hambleton et al., 1991, p. 53)

For the investigation of whether an IRT model could be successfully fitted to each one of the five sub-tests DS, DTM, ERC, READ, and WORD a random sample of three percent of the 82,506 examinees, who took part in the SweSAT in spring 1996 was used. The sample consisted of 2,461 test-takers: 1,349 females and 1,112 males. The results of these examinees on the separate sub-tests were the data, which have been analyzed in different ways.

The first step was to perform a standard classical item analysis, the outcome of which is presented below.

Classical item analysis

The CTT item analysis of the DS sub-test gave a range of p-values from .40 to .81, and a range of biserial correlations from .25 to .70. The reliability coefficient, alpha, was r =.82.

The CTT item analysis of the DTM sub-test gave a range of p-values from .28 to .82 and a range of biserial correlations[1] from .19 to .56. The reliability coefficient, alpha, was r =.72.

The CTT item analysis of the ERC sub-test gave a range of p-values from .28 to .82 and a range of biserial correlations[1] from .19 to .56. The reliability coefficient, alpha, was r =.72.

The CTT item analysis of the READ sub-test gave a range of p-values from .34 to .84, and a range of biserial correlations[1] from .21 to .45. The reliability coefficient, alpha, was r =.68.

The CTT item analysis of the WORD sub-test, finally, gave a range of p-values from .16 to .96, and a range of biserial correlations[1] from .02[2] to .64. The reliability coefficient, alpha was r = .85.

---

[1] The biserial correlations are calculated as the correlation between the item and the total score without that item.

[2] There was one deviating item, which did not work properly; hence the very low biserial correlation.

The ranges of biserial correlations indicate that there is a substantial variation in the discrimination power of the items in all five sub-tests. Sometimes, though, the range may be deceptive because of a couple of ”outliers”. Moreover high biserial correlations are sometimes associated with very easy items. Such discrimination indices do not really reveal effective items, and therefore the p-values were plotted against the biserial correlations for all the items in each sub-test. In figure 1 the plot for the WORD sub-test is shown as an example.

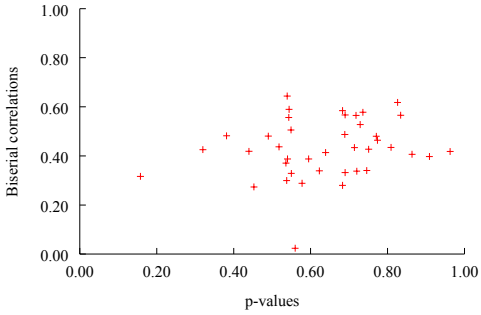

Figure 1. Biserial correlations plotted against p-values of the 40 items in the WORD sub-test.

The plots, which were similar for all five sub-tests, gave support to the assumption that there actually was variation in the discriminating power of the items in all sub-tests. There did not seem to be any connection between easy items and high biserial correlations. The conclusion was that there seems to be a need for an item discrimination parameter, and therefore a one-parameter IRT model seemed unsuitable for the results of all sub-tests.

To examine whether guessing had taken place in the tests, the test-takers with the lowest results were studied. All test-takers below the first percentile on each sub-test were selected, and difficult items were defined as items with p-values lower than .50.

The results of these poor examinees on the most difficult items of each sub-test were studied, and the result was that on the DS sub-test the p-values for these poor examinees on the eight most difficult items were:

p = .11, .30, .08, .14, .20, .13, .11, and .17

on the DTM sub-test:

p = .26, .14, .11, .06, .17, .18, .11, and .20

on the ERC sub-test:

p = .21, .24, .12, .22, .19, .15, and .35

on the READ sub-test:

p = .17, .16, .12, and .15

on the WORD sub-test:

p = .13, .01, .14, .01, .11, and .22

These results indicated that guessing can hardly be excluded on any of the sub-tests, and therefore a two-parameter model also appeared to be unsuitable to fit the data.

Factor analysis

An assumption common to all IRT models is that the set of test items is unidimensional. A crude measure of unidimensionality is the reliability coefficient, alpha, as this coefficient is a measure of the internal consistency of the items in a test. The coefficient alpha varied between .68 and .85 for the sub-tests. The coefficient r = .68 indicates that the sub-test is not very homogenous, but this is only a rude measure. A more appropriate method for assessing the unidimensionality of a test is factor analysis (Hambleton & Rovinelli, 1986). If the factor analysis shows only one dominant factor, this is support for unidimensionality. The results of the factor analyses were:

For the DS sub-test the analyses resulted in three factors with eigenvalues: 4.77, 1.21, and 1.09 respectively. The variance explained by the first factor was 21.7 percent, and all items had substantial loadings on the first factor (between .24 and .64).

For the DTM sub-test the result was four factors with eigenvalues: 3.3, 1.2, 1.1, and 1.0 respectively. The variance, explained by the first factor, was 16.4 percent.

For the ERC sub-test the result was also four factors with eigenvalues: 3.8, 1.1, 1.0, and 1.0 respectively. The variance, explained by the first factor, was 19.4 percent.

For the READ sub-test the result was five factors with eigenvalues: 2.9, 1.1, 1.0, 1.0, and 1.0 respectively. The variance, explained by the first factor, was 14.5 percent.

For the WORD sub-test, finally, the unrotated factor analysis resulted in nine factors with eigenvalues: 6.1, 1.4, 1.2, 1.2, 1.1, 1.1, 1.0, 1.0 and 1.0 respectively. The variance explained by the first factor was 15.4 percent.

All eigenvalues were plotted, and the plots for the two sub-tests with the smallest first eigenvalues, i.e. the READ and the WORD sub-tests are shown in Figure 2.

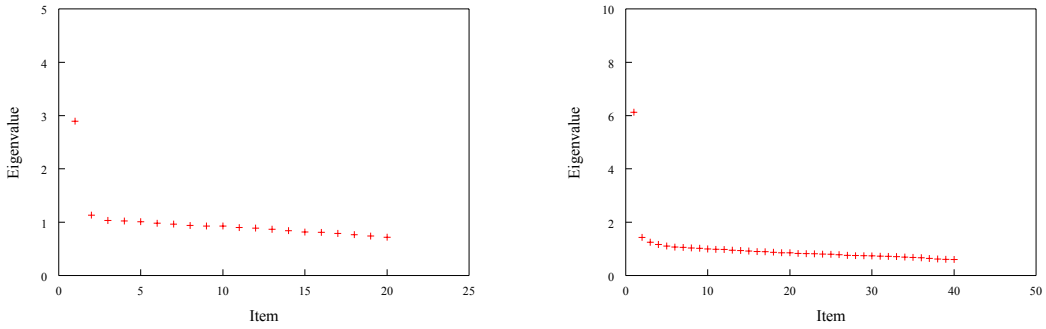

Figure 2. Plot of eigenvalues for the READ (to the left) and the WORD (to the right) sub-tests.

In Figure 2 it is shown that, after all, there seems to be one dominant first factor in both sub-tests, since according to Hambleton and Rovinelli (1986):

> *The number of "significant" factors is determined by looking for the "elbow" in the plot. The number of eigenvalues to the left of the elbow is normally taken to be the number of significant factors underlying test performance. (p. 289)*

Even though it would have been better if the amount of variance explained by the first factor generally had been greater it is not implausible or unreasonable to assume a single factor with the test data for any of the sub-tests.

The three-parameter logistic IRT model

An attempt was made to fit the results of each sub-test to the three-parameter logistic IRT model by means of the program BILOGW (Mislevey & Bock, 1990). When the number of items is 20 or greater, approximate chi-square statistics for the goodness of fit of each item are included as output of the program. For this purpose, the cases in the calibration sample are sorted into successive intervals of the latent continuum according to the estimates of their ability rescaled to mean = 0 and standard deviation = 1. This gives a reasonable test of fit if the number of items is large enough to make an assignment of cases accurate, and if the sample size is large enough to retain three or more intervals. In these studies the least number of items was 20 and the sample of examinees was large. The number of intervals used was 10 for most of the items. The outcome of the goodness of fit analyses were:

For the DS sub-test the outcome was that for 11 items there was model data misfit at $\alpha = .05$ level, and for four of these items the misfit was significant at $\alpha = .01$ level. The reliability index was $r = .84$, which may be compared with coefficient alpha in the classical analysis, which was $r = .82$.

For the DTM sub-test the outcome was that for ten items there was significant model data misfit, and for eight of these items the misfit was significant at α = .01 level. The reliability index was r = .74, as compared with coefficient alpha, which was r = .72.

For the ERC sub-test the outcome was 14 items with significant model data misfit, and for seven of these items the misfit was significant at α = .01 level. The reliability index was r = .80, while coefficient alpha was r = .76.

For the READ sub-test the outcome was nine items with significant model data misfit, and for seven of these items the misfit was significant at α = .01 level. The reliability index reported was r = .72, and coefficient alpha was r = .68.

And finally the outcome of the goodness of fit analysis for the WORD sub-test was that for eight of the items there was a model data misfit which was significant at α = .05 level and for one of these eight items the misfit was significant at α = .01 level. The reliability index reported was r = .87, while the reliability coefficient, alpha was r = .85.

Residual analyses

Another type of goodness of fit analyses were made by means of the program RESID (Rogers, 1994). In carrying out these analyses, examinees are first sorted into ability categories. The number of ability levels was specified to eight and the observed proportions of examinees in each ability category, answering the item correctly, were calculated. Expected proportions correct for each ability interval were obtained by computing the probability of success on the item on each ability level. Residual values (observed - expected) and standardized residuals were then computed. The program also contains chi-square fit statistics for each item as output.

The outcome of the RESID analyses was:

For the DS sub-test the differences between observed and expected values were statistically significant for two items, both at the .05 level.

For the DTM sub-test the differences were statistically significant for six items, five of which on .01 level.

For the ERC sub-test the differences were significant for seven items, two of which at .01 level.

For the READ sub-test the differences were statistically significant for five items, two of which at .01 level.

Finally for the WORD sub-test the differences were statistically significant for six items, one of which at the .01 level.

Residuals make it possible to compare predicted and actual performance. Raw residuals are the differences between expected and observed performance on an item at a specified performance level. Standardized residuals (SRs) take into account the sampling error associated with each performance level as well as the number of examinees at that particular level of performance. When the model fits the data the SRs might be expected to be small and

randomly distributed about 0. Within the framework of regression theory it is common to assume that the distribution of SRs is approximately normal. In Table 1 a summary of the SRs from the goodness of fit analyses is given.

Table 1. Absolute values of the standardized residuals (per cent) for the five sub-tests

SRs	DS	DTM	ERC	READ	WORD
| 0-1 |	72.16	65.63	62.50	68.75	70.31
| 1-2 |	22.16	25.00	31.25	24.38	26.25
| 2-3 |	5.11	7.50	5.00	6.88	3.13
| >3 |	.57	1.88	1.25	0.00	.31

The results in Table 1 show that even though the distributions for the DTM and ERC sub-tests are a bit too flat, all the distributions of SRs are fairly close to the normal distribution which is supposed to be strong support for model data fit.

Hambleton et al. (1991) have given the following recommendations regarding assessment of model data fit:

> *In assessing model-data fit, the best approach involves a) designing and conducting a variety of analyses designed to detect expected types of misfit, b) considering the full set of results carefully, and c) making a judgment about the suitability of the model for the intended application. Analyses should include investigations of model assumptions, of the extent to which desired model features are obtained, and of differences between model predictions and actual data. Statistical tests may be carried out, but care must be taken in interpreting the statistical information. The number of investigations that may be conducted is almost limitless. (p. 74)*

For the DS sub-test a comparison was made between item parameters estimated on two different samples of test-takers. The correlation between b-values was r = .95, and the correlation between a-values was r = .72.

For the DTM sub-test one comparison was made between parameters estimated within IRT and indices calculated within CTT. The correlation between the estimated b-values and the calculated p-values was r = -.94, and the correlation between estimated a-values and calculated biserial correlations (r_{bis}) was r = .82. The correlation between test-scores and estimated ability parameters was r = .95. Another comparison was made between parameters estimated on male and female test-takers, and the outcome of this comparison was that the correlation between b-values estimated on male and female examinees was r = .89, and between a-values r = .91.

The same comparisons were made for the ERC sub-test. The correlation between estimated b-values and calculated p-values on this sub-test was r = -.88, and between estimated a-values and calculated r_{bis} the correlation was r = .73. The correlation between b-values estimated on male examinees and b-values estimated female examinees was r = .93, and the correlation between a-values was r = .83.

The same comparisons were also made for the READ and WORD sub-tests. For the READ sub-test the correlation between p- and b-values was r = -.98, and for the WORD sub-test this correlation was r = -.74. For the READ sub-test the correlation between r_{bis} and a-values was r = .84, and for the WORD sub-test it was r = .82. The correlation between b- values estimated on males and b-values estimated on females on the READ sub-test was r =.92, and on the WORD sub-test this correlation was r = .85. The correlation between a-values estimated on male examinees and a-values estimated on female examinees was r = .76 on the READ sub-test and r = .77 on the WORD sub-test. The results of these group comparisons are given in Table 2.

Table 2. Relations between item parameters estimated within IRT and item indices calculated within CTT, and between IRT item parameters estimated on female and male examinees.

correlation	DTM	ERC	READ	WORD
p- and b-values	-.94	-.88	-.98	-.74
r_{bis} and a-values	.82	.73	.84	.82
F and M est. b	.89	.93	.92	.85
F and M est. a	.91	.83	.76	.77

Discussion

The results from these attempts to fit a three-parameter logistic IRT model separately to each one of the five sub-tests in SweSAT, are somewhat mixed. The results from the initial CTT analyses gave support for the need of a three-parameter model for all the five sub-tests. The factor-analyses gave support for unidimensionality in the DS sub-test, for which the first factor could explain 21.7 per cent of the test variance. For the other sub-tests the support from the factor analyses for unidimensionality was weaker: 19.4 percent explained variance by the first factor for the sub-test ERC, 16.4 percent for the sub-test DTM, 15.4 percent for the sub-test WORD, and 14.5 percent for the sub-test READ. When the explained variance is less than 20 percent, it is uncertain whether the test can be claimed to be unidimensional.

The assumption of local independence is also problematic at least for three of the SweSAT sub-tests. The READ sub-test consists of four texts with five questions in relation to each text. Even though these five items are independent of each other they belong to the same text. The DTM sub-test consists of ten figures, tables and maps with two questions to each graph. The ERC sub-test consists of a varying number of texts with two to five questions to each text. It is doubtful whether the items actually are locally independent in tests of this format.

Another problem is the different results from the two statistical tests. As a rule more items were found to have statistically significant model data misfit by the BILOG-program than by the RESID-program, and this could at least partly be explained by the fact that BILOG divides into more ability groups than RESID. However, some of the items, which were found to have significant model data misfit by RESID, were not found so by BILOG. This was true for two items in the DS sub-test, two in the DTM sub-test, one in the ERC sub-test, two in the READ sub-test, and three in WORD sub-test.

So far the analyses performed have neither fully confirmed nor completely rejected model data fit of the three-parameter logistic IRT model to SweSAT data.

II. Comparison between item analysis based on Item Response Theory and on Classical Test Theory

As for all high-stake tests, the pilot- or pre-testing of items for SweSAT is a crucial part of the test development process. The pre-testing of items has several purposes (Henrysson, 1971) of which the most important for SweSAT are:

- To determine the difficulty of each item so that selection may be made, that will give a difficulty level of the sub-test, which is parallel to earlier versions of the same sub-test.
- To identify weak or defective items with non-functioning distractors.
- To determine for each item its power to discriminate between good and poor examinees in the achievement variable measured.
- To identify (gender) biased items.

Ever since SweSAT first was taken into use in spring 1977, the development and assembly of the test, as well as the equating of different forms from one administration to the next, has been based on classical test Theory (CTT). On the basis of the data obtained in the pre-test the items are rejected or selected for the final test, and the statistics which are used in the item analysis are:

p-values of the items

p-values of the distractors

biserial correlations (r_{bis})

p-values of males and females

the item test regression

The major limitation of CTT in this regard is that the person statistic (i.e. the test score) is dependent on the sample of items (i.e. the test), and the item statistics are dependent on the sample of examinees. The primary argument for the use of IRT models over CTT procedures is that IRT should result in sample free measurements. With IRT a person should theoretically receive the same estimate of ability, regardless of the test given, and item statistics should remain stable across different groups of individuals. Hence the great advantage of IRT is the item parameter invariance. One drawback of IRT is that big sample sizes are needed for the estimation of parameters.

IRT has been vigorously researched by psychometricians, and numerous books and articles have been published (Fan, 1998). The empirical studies available, however, have primarily focused on various applications of IRT, and very few studies have actually compared CTT and IRT for item analysis and test design. Fan (1998) continues:

> *It is somewhat surprising that empirical studies examining and/or comparing the invariance characteristics of item statistics from the two measurement frameworks are so scare. It appears that the superiority of IRT over CTT in this regard has been taken for granted in the measurement community, and no empirical scrutiny has been deemed necessary. The empirical silence on this issue seems to be an anomaly. (p. 361)*

Since spring 1996 pre-testing of items for SweSAT has been performed in connection with the regular test administration, which means that the examinee sample, on which the pre-testing is performed, is a sample from the true examinee population and it contains 1000 examinees as a minimum. This new procedure for pre-testing would make possible the use of IRT for item analysis and compilation of new test versions.

Aim

The purpose of this study was to compare the item statistics from the CTT framework with the item parameters from the IRT framework, and to examine the stability from pre-testing to regular testing of the two sets of item characteristics. Specifically the following questions were addressed:

1. How do item difficulty indices from CTT compare to item difficulty parameters estimated by IRT?

a) for pre-test data?

b) for regular test data?

2. How do item discrimination indices from CTT compare to item discrimination parameters estimated by IRT?

a) for pre-test data?

b) for regular test data?

3. How stable are the CTT item indices from pre-test data to regular test data?

4. How stable are the IRT item parameters from pre-test data to regular test data?

Method

Classical test theory

In the regular test in spring 1997, there were 20 WORD, 16 READ, and 14 ERC items, which had been pre-tested in spring 1996. For these items p-values and biserial correlations were calculated. The same indices were calculated for the corresponding items in the pre-test data, and the values were compared.

Item response theory

The five WORD pre-test combinations, on which the above-mentioned 20 WORD items had been dispersed, were run in BILOGW together with the regular WORD test from spring 1996, and the a-, b- and c-parameters were estimated. The WORD sub-test from spring 1997 was also run in BILOGW, and the item parameters were estimated. Finally the parameters estimated for the 20 common items were compared.

The eight READ pre-test versions from spring 1996 were run in BILOGW together with the regular READ sub-test from spring 1996, and the a-, b-, and c-parameters were estimated. The READ sub-test from spring 1997 was run in BILOGW and the item parameters were estimated. The parameter estimates for the 16 common items were noted and compared.

The four ERC pre-test versions spring 1996 were run in BILOGW together with the regular ERC sub-test from spring 1996, and the a-, b- and c-parameters were estimated. The ERC sub-test from spring 1997 was run in BILOGW and the item parameters were estimated. The item parameter estimates for the 14 common items were noted and compared.

Results

One problem when analyzing the stability of the item parameters is that pre-testing has two purposes. One purpose is to get information about the difficulty level and discrimination power of the items in order to be able to compile tests of equal difficulty. The other purpose is to make sure that all items function in a satisfactory way. If an item is not working well enough it will be changed or excluded. If there are major changes the item will be pre-tested once more before it is taken into use, but if the changes are minor the item will be used in the regular test. Such changes, however, mean that the items are not exactly the same in the pre-test version as in the regular test. Another problem is that items may be placed in different order in the pre-test booklet and the regular test booklet, and items tend to become more difficult in the end of the booklet. In Tables 3 to 5 the placement of the item in the pre-testing booklet as well as in the regular test are given, and items for which minor changes had been made between the two occasions are marked with *.

The WORD sub-test

In Table 3 the difficulty and discrimination indices calculated within the CTT framework are shown for 20 WORD-items from the pre-testing as well as from the regular test. In the same Table the difficulty and discrimination parameters estimated within the IRT framework are presented.

Table 3. Item characteristics for 20 WORD items calculated within the framework of CTT and estimated within the framework of IRT.

Item No		CTT difficulty		IRT difficulty		CTT discrimin		IRT discrimin	
pre-test	reg. test	pre-test	reg. test	pre-test	reg. test	pre-test	reg. test	pre-test	reg. test
8	1*	.73	.71	-.43	-.43	.60	.58	1.29	1.14
20	4*	.79	.72	-.96	-.64	.46	.41	.73	.62
39	5*	.78	.74	-1.94	-1.19	.25	.33	.32	.43
18	9	.68	.71	-.25	-.59	.44	.43	.71	.63
36	10*	.75	.72	-.92	-.81	.50	.53	.72	.78
27	11*	.80	.82	-1.49	-1.79	.35	.35	.48	.46
36	15*	.71	.58	-.55	.11	.40	.37	.59	.55
14	16*	.65	.70	-.02	-.65	.44	.48	.81	.72
5	19	.46	.42	.46	.71	.42	.37	.55	.50
16	23	.65	.62	.11	.08	.35	.40	.58	.72
38	24	.58	.56	.08	.16	.47	.30	.75	.40
12	25	.51	.59	.17	-.02	.58	.58	.97	1.13
24	27	.69	.66	.37	.33	.36	.35	1.07	.83
4	28*	.53	.44	.35	.51	.56	.52	1.55	1.04
4	29	.42	.42	1.16	1.08	.33	.26	.71	.61
5	35*	.31	.38	1.59	.89	.32	.46	.45	.95
37	36*	.71	.62	-.37	-.07	.43	.44	.70	.70
6	38*	.41	.46	1.25	.49	.31	.37	.70	.48
28	39*	.27	.40	1.61	1.22	.28	.32	1.13	1.18
39	40*	.31	.31	2.27	2.45	.23	.21	.42	.45

The correlation between p-values calculated on pre-test data and p-values calculated on regular test data was r = .93.

The correlation between b-values estimated on pre-test data and b-values estimated on regular test data was r = .92.

The correlation between p- and b-values was r = - .93, for the pre-test data as well as for the regular test data.

The correlation between r_{bis} calculated on pre-test data and r_{bis} calculated on regular test data was r = .81

The correlation between a-values estimated on pre-test data and a-values estimated on regular test data was r = .74.

The correlation between the item discrimination r_{bis} and the item discrimination parameter a was r = .65 for the pre-test data and r = .64 for the regular test data.

The Read sub-test.

Table 4. Item characteristics for 16 READ items calculated within the framework of CTT and estimated within the framework of IRT.

ItemNo		CTT difficulty		IRT difficulty		CTT discrim.		IRT discrimin.	
pre-test	reg. test	pre-test	reg. test	pre-test	reg. test	pre-test	reg-test	pre-test	reg-test
5	5	.74	.74	-.61	-.69	.30	.32	.50	.59
7	6*	.30	.68	2.16	-.52	.23	.36	.46	.64
6	7	.78	.71	-1.16	-.45	.37	.38	.54	.73
8	8	.80	.81	-1.31	-1.09	.40	.35	.59	.63
17	9	.64	.81	.37	-.88	.37	.43	.85	.90
20	10	.52	.69	1.32	.27	.25	.22	.64	.42
17	11	.61	.75	1.42	-1.20	.18	.11	.52	.37
20	12	.45	.52	1.92	.71	.17	.33	.72	.86
14	13	.59	.66	.48	-.14	.35	.36	.83	.69
15	14*	.36	.50	1.85	.74	.24	.30	.53	.72
14	15*	.24	.30	1.67	1.08	.34	.43	.87	1.23
16	16*	.28	.57	2.54	.33	.17	.28	.59	.54
17	17*	.35	.53	1.33	.51	.35	.36	.72	.69
19	18	.63	.76	.82	-.77	.29	.32	.76	.60
19	19	.56	.65	.84	.04	.29	.31	.69	.70
20	20	.57	.60	.38	.22	.34	.34	.57	.91

The correlation between p-values of the items in the pre-test versions and p-values of the corresponding items in the regular test was r = .78.

The correlation between b-values estimated on pre-test data and b-values estimated on regular data was r = .55.

The correlation between p-values calculated on pre-test data and b-values estimated on pre-test data was r = -.90.

The correlation between CTT and IRT regarding difficulty indices on regular test data was r = -.92.

The correlation between r_{bis} of the items in the pre-test versions and r_{bis} of the corresponding items in the regular test was r = .66.

The correlation between a-values estimated on pre-test data and a-values estimated on regular test data was r = .54.

The correlation between CTT and IRT discrimination indices on pre-test data was r =.35. The correlation between CTT and IRT discrimination indices on regular test data was r = .78.

The ERC sub-test

Table 5. Item characteristics for 14 ERC items calculated within the framework of CTT and estimated within the framework of IRT.

Item No		CTT difficulty		IRT difficulty		CTT discrimin.		IRT discrimin.	
pre-test	reg. test	pre-test	reg. test	pre-test	reg. test	pre-test	reg. test	pre-test	reg.test
1	1	.33	.38	1.65	1.31	.30	.33	.59	.79
2	2	.72	.66	.20	.34	.34	.33	.83	.76
3	3	.35	.29	1.70	1.74	.28	.29	-.72	.71
4	4	.41	.47	.83	.54	.45	.47	.76	.83
5	5*	.62	.73	.78	.35	.16	.54	.27	1.05
1	6	.77	.78	-.90	-.62	.62	.54	1.00	1.10
2	7	.54	.50	.09	.49	.48	.46	.67	.81
3	8	.53	.53	.77	.63	.37	.42	.90	.99
11	9*	.41	.60	.81	-.07	.41	.38	.56	.55
5	10	.67	.58	-.52	.19	.57	.51	.84	1.02
14	12	.60	.51	-.13	-.10	.51	.46	.71	.73
13	13	.68	.62	-.27	-.03	.56	.56	.96	1.10
14	14	.65	.65	-.40	-.19	.57	.52	.85	.91
10	15	.77	.74	-.85	-.52	.52	.52	.80	.95

The mean of p-values was .58 for the pre-test data as well as for the regular test. The correlation between p-values of items in the pre-test versions and p-values from the corresponding items in the regular test was r = .86.

The mean of b-values was .27 for the pre-test items and .29 for the regular test items. The correlation between b-values estimated on pre-test data and b-values estimated on regular test data was r = .88.

The correlation between r_{bis} of the items in the pre-test versions and r_{bis} of the corresponding test items in the regular test was r = .57.

The correlation between a-values estimated on pre-test data and a-values estimated on regular test data was r = .34.

For the 12 items, which had not been changed between pre-test and regular test, the correlation was r = .95 for the p-values, and r = .96 and ρ = .94 for $r_{bis.}$

For the 12 unchanged items the correlation between b-values from pre-test and regular test was r = .96, and between a-values r = .82.

The correlation between CTT difficulties and IRT difficulties was r = -.90 for pre-test data as well as for regular test data. The correlation between discriminations calculated within CTT and estimated within IRT was r = .74 for pre-test data and r = .76 for regular test data.

The results for the three sub-tests are summarized in Table 6.

Table 6. Stability of CTT item indices, and IRT item parameters from pre-test to regular test versions. Relations between CTT item indices and IRT item parameters on pre-test and regular test.

sub-test	pre- and reg. test		pre- and reg. test		p- and b-values		r_{bis} and a-values	
	p	b	r_{bis}	a	pre	reg.	pre	reg
ERC	.86 (.95)	.88 (.92)	.57	.74	-.90	-.90	.74	.76
READ	.78	.55	.66	.54	-.90	-.92	.35	.68
WORD	.93	.92	.81	.74	-.93	-.93	.64	.65

Discussion

For the WORD and READ sub-tests (the three-parameter logistic model) none of the pre-test items was identified as significantly misfitting to the three-parameter logistic model. For three pre-test items in the ERC, however, there was misfit, which was significant at $\alpha = .01$ level. In the regular sub-tests there was one item in the WORD sub-test, one in the READ sub-test, and two in ERC sub-test, for which there was a model data misfit, which was significant at $\alpha = .01$ level. These items were No 10 in the regular WORD sub-test, and No 11 in the regular READ sub-test and items No 9 and 14 in the ERC sub-test.

The overall conclusion of the studies is that the agreement between results from the item-analyses within the two different frameworks CTT and IRT was reasonably good. The correlation between item difficulties for the regular test versions was r = -.93 for the WORD sub-test, r = -.92 for the READ sub-test, and r = -.90 for the ERC sub-test.

Regarding the stability from pre-test to regular test data there were no great differences between the two theories. For the WORD sub-test agreement between difficulties in pre-test and regular test was r = .93 within CTT and r = .92 within IRT. For the READ sub-test the correlation between CTT difficulties was r = .78 and between IRT difficulties the correlation was r = .55. For the ERC sub-test the correlation within CTT was r = .87 and within IRT r = .88. On the whole the correlations between item difficulties in pre-test and regular test data were actually higher for the CTT indices than for the IRT parameters.

> *Because IRT differs considerably from CTT in theory, and commands some crucial theoretical advantages over CTT, it is reasonable to expect that there would be appreciable differences between IRT- and CTT-based item and person statistics. Theoretically such relationships are not entirely clear, except that the two types of statistics should be monotonically related under certain conditions (Crocker & Algina, 1986, Lord, 1980) but such relationships have rarely been empirically investigated, and, as a result they are largely unknown. (Fan, 1998, p. 360)*

The overall conclusion from these comparisons is that the prediction from pre-test data to regular test data is acceptable, but that is true for CTT as well as for IRT. Actually the predictions made within the CTT framework were generally more correct than the predictions made within the IRT framework. The IRT item parameters were not completely invariant. Since the groups on which the pre-testing had been performed were large and representative

samples from the actual examinee population this outcome may be expected. The dilemma is, however, that in order to be able to estimate the item parameters within the IRT framework, large samples are a necessity, and when the samples are large enough the item indices within the CTT framework are very stable as well.

What is usually regarded as the main shortcoming of CTT is that item statistics, such as item difficulty and item discrimination depend on the particular examinee sample in which they are obtained. The invariance of the corresponding IRT item statistics across samples is usually considered as the main theoretically superiority. *The invariance of item parameters across groups is one of the most important characteristics of item response theory (Lord, 1980, p. 35).* In the studies reported here the item parameters estimated within the IRT framework were not superior to the statistics derived within the CTT framework regarding invariance across groups. The problem may be that in order to achieve this invariance of IRT parameters, there must be perfect model-data fit. Unfortunately there are no objective criteria on model data fit, but according to Hambleton et al. (1991) *...invariance and model data fit are equivalent concepts (p. 24).*

III. Applicability of IRT to the SweSAT: the total test

SweSAT is scored in accordance with classical test theory (CTT), the raw score for each examinee is the number of items answered correctly. All items are of multiple choice format and scored dichotomously, i.e. "1" for a right answer and "0" for a wrong one. It has been found that on this kind of items examinees differ in their tendency to guess, or to omit items, for which they do not know the correct answer, and this may cause irrelevant variance. Several methods have been invented to correct for the effect of guessing, but empirical studies have not supported the use of correction methods (Crocker & Algina, 1986, p.403). The SweSAT test-takers are encouraged to mark every item. The total test score is equated and transformed to a normed score, and this normed score is used in selection to higher education.

SweSAT consists of five sub-tests measuring, word knowledge (WORD), logical thinking (DS) Swedish reading comprehension (READ), English reading comprehension (ERC), and the ability to interpret diagrams, tables and maps (DTM). In earlier studies the applicability, of a three-parameter logistic IRT-model, was investigated for each sub-test (Stage, 1996, 1997a, 1997b, 1997c, 1997d). The outcome of these studies was neither a confirmation nor a rejection of model data fit of the three-parameter logistic IRT model to the SweSAT data.

Since it is the total test score that constitutes the result for the SweSAT examinees, the applicability of an appropriate IRT-model to the total test is of importance. In this study the aim was to investigate whether a three-parameter logistic IRT-model could be successfully applied to the total test.

The three-parameter model was chosen, since it had turned out to be the most appropriate model for each of the sub-tests. There was no reason to believe that the total test should be different from its parts regarding the need of a discrimination parameter as well as a pseudo guessing parameter.

Unidimensionality

An un-rotated factor analysis of the total test scores resulted in a first factor of 12.1, a second factor of 3.7, a third factor of 2.0, and 34 factors above 1.0. The first factor explained 9.9 percent of the variance, the second factor 3.0 per cent and the remaining factors from 1.7 to 0.8 per cent each. In Figure 3 a plot of the eigenvalues are shown.

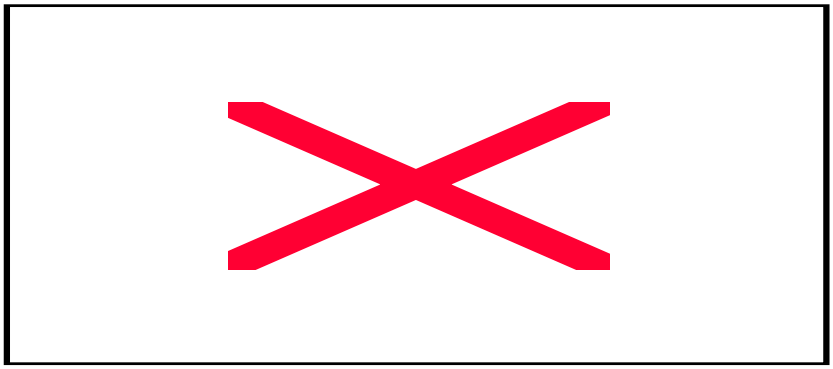

Figure 3. Plot of eigenvalues for the total SweSAT test

In Figure 3 it may be seen that even though there is a kind of "elbow" in the graph, and the first factor is dominant the assumption of unidimensionality is uncertain, since the second factor is somewhat too strong.

A factor analysis was also performed on sub-test level. This analysis resulted in a first factor with eigenvalue 3.05, a second factor with eigenvalues 0.83, a third with 0.41, a fourth with 0.38, and a fifth factor with eigenvalue 0.35. The first factor explained 61 per cent of the variance. This analysis gave more support for unidimensionality.

The three-parameter logistic model

The program BILOGW was used to fit the test results of the SweSAT to the three-parameter logistic IRT model. The outcome of the goodness of fit analysis of the items was that for 67 items there was a model data misfit which was significant at the $\alpha = .05$ level, and for 44 of these items the misfit was significant at the $\alpha = .01$ level. The program was run on 28 505 test-results, and since statistical tests of model fit are very sensitive to sample size, such results were to expected. This is what Hays (1969) calls the fallacy of evaluating a result in terms of statistical significance alone:

> *Virtually any study can be made to show significant results if one uses enough subjects, regardless of how nonsensical the content may be. (p. 326)*

The program was also run on two different random samples of 1000 test-results. For the first sample the number of items for which there was significant misfit had decreased to seven, and for four of these items the misfit was significant at the $\alpha = .01$ level. For the second sample the number of significantly misfitting items was only six, none of which was significant at $\alpha = .01$ level.

A residual analysis was made on the whole population with the program RESID, and this analysis resulted in only 10 items with significant model data misfit, three of which were at $\alpha = 0$ .01 level. Six of these items did not, however, have significant model data misfit according to BILOGW.

The standardized residuals (see p. 8 for explanation) were distributed:

SRs	**percent**
0 – 1	73.91
1 – 2	23.91
2 – 3	1.88
> 3	0.31

Since the distribution of SRs is very close to the normal distribution there is support for model data fit, by the residual analysis.

Comparison between item statistics

The correlation between CTT p-values and IRT b-values was r = -.84, and between CTT r_{bis} and IRT a-values the correlation was r = .63. The correlation between CTT test scores and IRT ability estimates was r = .96. In Figure 4 the b-values are plotted against the p-values.

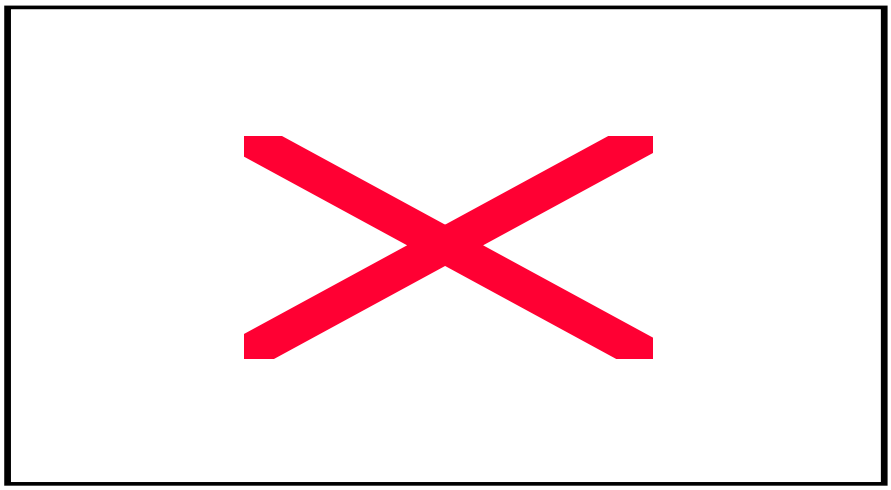

Figure 4. Scatter-plot of IRT b-values against CTT p-values.

In Figure 4 it looks like there is a curvilinear relationship between b- and p-values. Since p-values actually expresses item difficulty on an ordinal scale, a transformation was made to an interval scale by normalization (see e. g. Aiken, 1991, p. 96). This normalization only raised the correlation between b- and p-values to r = .85, however.

In order to further investigate the parameter invariance two different random samples, of 1000 individuals each, were taken from the population of test-takers. For each sample the CTT item statistics were calculated and the IRT item parameters were estimated. Comparisons were made between p-values and biserial correlations from the two samples and between b-values and a-values. Comparisons were also made with the same values from the population. The outcome of these comparisons are shown in Table 7.

Table 7. Stability of item statistics from the two measurement frameworks. Correlations between CTT and IRT based item difficulties and item discriminations

correlation between	population–sample A	population–sample B	sample A – sample B
p-values	.995	.996	.989
b-values	.940	.946	.960
r_{bis}	.931	.957	.882
a-values	.764	.775	.683

Discussion

These attempts to fit a three-parameter logistic IRT model to the SweSAT test, given in fall 2002, were not very successful. It is difficult to get an unequivocal answer regarding model data fit, since there are no objective criteria. However, the factor analysis on item level was discouraging, and so was the BILOGW chi-square test on the total population of test-takers. On the other hand, the RESID test was encouraging, and so was the BILOGW, chi-square tests on two samples of 1000 test-takers each. The item-parameters, however, were not invariant, and according to Hambleton et al. (1991)…*invariance and model data fit are equivalent concepts* (p. 24). Therefore the conclusion must be that the three-parameter logistic model did not fit the SweSAT data.

On the other hand the results from CTT analyses were quite encouraging. The item difficulty indices as well as the item discrimination indices were very stable between the two random samples, as well as between the population and the two samples.

Concluding remarks

Since SweSAT has been developed within the framework of CTT it has seemed reasonable to compare the outcomes of the IRT analyses with the corresponding outcomes within the CTT framework. Empirical comparisons between results from the two theories are not very common. Lawson (1991) compared the one-parameter model and CTT on three data sets and found *...remarkable similarities between the results obtained through classical measurement methods and results obtained through one-parameter latent*[3] *trait methods (p. 163).* Fan (1998) examined items and person statistics derived from IRT and CTT on a large-scale statewide assessment database. His findings indicated that the person and item statistics derived from the two measurements frameworks were quite comparable. He also found that the invariance of item statistics across samples, which is usually considered to be the superiority of IRT models, appeared to be similar for the two measurement frameworks. Both Lawson (1991) and Fan (1998) conclude their articles by quoting from Thorndike's opening speech, in 1982, at an Australian Conference, focused on IRT models:

> *For the large bulk of testing, both with locally developed and with standardized tests, I doubt that there will be a great deal of change. The items that we will select for a test will not be much different from those we would have selected with earlier procedures, and the resulting test will continue to have much the same properties. (p. 12)*

In the studies reported in this paper, the CTT item indices were not only comparable to the IRT item parameters, they were generally more invariant between different samples of test-takers. One possible explanation for these results is that the IRT model did not fit the test data. But even if the results are due to poor model data fit, the only reasonable conclusion is that for SweSAT data, CTT seems to work better than IRT.

SweSAT is well accepted by Swedish test-takers as well as by universities. The most important demand on the test is that it should rank test-takers as fairly as possible with regard to their expected study success. Other requirements on the test are that:

- 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 fast, cheaply and objectively
- It should not be possible for an individual to improve her/his 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 unbiased recruitment should be observed. No group should be discriminated against because of, e.g., gender or social class.

---

[3] IRT models are sometimes called latent trait models

All changes, which could improve the test in any of these aspects, are worth striving for. That would, however, be a matter of changes, which would improve the validity of the test. The main goal of any change must surely be this, to improve the validity of the test, rather than to make it fit a specific test theory.

IRT consists of a family of models that have been argued to be useful in the design, construction and evaluation of educational tests. As further research is carried out, the remaining technical problems associated with applying the models will hopefully be resolved. In addition it is expected that newer and more applicable models will be developed in the coming years, enabling IRT to provide even better solutions to important measurement problems (Hambleton et al., 1991). As mentioned earlier one important assumption of IRT models is unidimensionality, which means that the items in a test measure one single ability. There are models in which it is assumed that more than one ability is necessary to account for the performance on a test. These so-called multidimensional models are, however, more complex, and have not been as well developed as the unidimensional models.

Although at present IRT does not seem to be applicable for the construction and design of SweSAT, work is underway to investigate whether IRT can be used for the equating of different test versions. Emons (1998) made a thorough study, "*Nonequivalent groups IRT observed score equating. Its applicability and appropriateness for the Swedish Scholastic Aptitude Test",* in which he used items, which had been pre-tested in spring 1996 as anchor items for equating the spring 1997 test. In this study, however, the number of anchor items may have been too few to constitute a proper link for the equating. For some sub-tests the link consisted of only 2 or 3 items. Similar studies are continuously carried out, and at present the outcomes of the traditional "equivalent groups equipercentile equating" method are continuously compared to IRT equating with links of different sizes.

Another area of studies on SweSAT where IRT models are presently used is for studying differential item functioning (DIF). The item characteristics curves, in a very good way, illustrate the DIF problem, since the curves show the probability to give a correct answer to an item, given a certain ability level. The comparison of the curves for different groups of test-takers (for SweSAT mainly males and females) exactly corresponds to the most accepted definition of DIF. For these kinds of studies the sub-test score can be used, and on sub-test level the unidimensionality seems to be an acceptable assumption.

Finally, in the case of computerized adaptive testing (CAT) or tailored tests, IRT is the only proper theoretical framework. In CAT, which items a test-taker gets depends on his/her performance on the earlier items in the test. Only the items that are most informative about the test-taker are administered. High-ability test-takers do not need to get the very easy items, and low ability test-takers do not need to get the very difficult items. In this way the test can be shortened considerably, and still give the same information and the same measurement precision as a longer conventional test. If, in the future, SweSAT, or some version of SweSAT, should become transformed to CAT, then the application of IRT will be a necessity.

REFERENCES

Aiken, L. R. (1991). *Psychological Testing and Assessment*. Massachusetts: Allyn & Bacon.

Crocker, L. L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart & Winston.

Emons, W., H., M. (1998). Nonequivalent Groups IRT Observed Score equating. Its Applicability and Appropriatenss for the Swedish Scholastic Aptitude Test. (Eduactional Measurement No 32). Umeå: Umeå University, Department of Educational Measurement.

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educational and Psychological Measurement, Vol. 58 No 3, 357-381*.

Hambleton, R. K. & Rovinelli, R. J. (1986) Assessing the Dimensionality of a Set of Test Items. *Applied Psychological Measurement, 10, 287-302*.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991) *Fundamentals of Item Response Theory* (Newbury, Sage).

Hambleton, R. K. & Jones, R. W. (1993). Comparison of Classical Test Theory and item Response Theory and their Applications to Test Development. *Educational Measurement: Issues and practice, 12 (3), 535-556.*

Hays, W.L. (1969) *Statistics* (London, Holt, Rinehart and Winston).

Henrysson, S. (1971). Gathering, Analyzing, and Using Data on Test Items. In R. L. Thorndike (Ed.) *Educational Measurement, 2nd Edition, 130-159.* Washington DC: American Council on Education.

Lawson, S. (1991). *One-Parameter Latent Trait Measurement: Do the Results justify the Effort?* The Annual Series of the Southwest Educational Research Association, Vol. 1, 159-168.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems,* Hillsdale NJ: Lawrence Erlbaum.

Mislevey, R. L. & Bock R. D. (1990). BILOG-W. Manual for BILOG-W. Chicago: International Scientific Software International. Item Analysis and test Scoring with Binary Logistic Models.

Rogers, J. (1994). RESID. Assessment of Fit for Unidimensional IRT Models. Program developed at the University of Massachusetts School of Education.

Stage, C. (1996). *An Attempt to Fit IRT Models to the DS Sub-test in The SweSAT* (Educational Measurement No. 19). Umeå: Umeå University, Department of Educational Measurement.

Stage, C. (1997a). *The Applicability of Item Response Models to the SweSAT. A Study of the DTM Sub-test* (Educational Measurement No. 21). Umeå: Umeå University, Department of Educational Measurement.

Stage, C. (1997b). *The Applicability of Item Response Models to the SweSAT. A Study of the ERC Sub-test* (Educational Measurement No. 24). Umeå: Umeå University, Department of Educational Measurement.

Stage, C. (1997c). *The Applicability of Item Response Models to the SweSAT. A Study of the READ Sub-test* (Educational Measurement No. 25). Umeå: Umeå University, Department of Educational Measurement.

Stage, C. (1997d). *The Applicability of Item Response Models to the SweSAT. A study of the WORD Sub-test.* (Educational Measurement No. 26). Umeå: Umeå University, Department of Educational Measurement.

Stage, C. (1998a). *A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Sub-test WORD*. (Educational Measurement No. 29). Umeå: Umeå University, Department of Educational Measurement.

Stage, C. (1998b). *A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Sub-test ERC*. (Educational Measurement No. 30). Umeå: Umeå University, Department of Educational Measurement.

Stage, C. (1999). *A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Sub-test READ*. (Educational Measurement No. 33). Umeå: Umeå University, Department of Educational Measurement.

Thorndike, R. L. (1982). Educational Measurement – Theory and Practice. In D. Spearritt (Ed.), *The improvement of measurement in education and psychology: Contributions of latent trait theory* (p.3-13). Princeton, NJ: ERIC Document Reproduction Service No. ED 222 545.