A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Subtest WORD.

Christina Stage

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 the 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 mutiple-choice items, divided into five subtests:

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

As for all high-stake tests the pretesting of items for SweSAT is a crucial part of the test development The pretesting of items has several purposes (see Henrysson, 1972) of which the most important for SweSAT are:

- to determine the difficulty of each item so that a selection may be made that will give a difficulty level of the subtest which is parallel to earlier versions of the same subtest.
- to identify weak or defective items with nonfunctioning 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 was first taken into use in spring 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 classical test theory

(CTT). On the basis of the data obtained in the pretest the items are improved and 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)

There are some shortcomings with CTT, however, one of which is that the item statistics are sample dependent; this may especially cause problems if the sample on which the pretesting was made differs in some unknown way from the examinee population. Another limitation which may be of importance in item analysis is that CTT is test oriented rather than item oriented.

During the last decades a new measurement system, item response theory (IRT) has been developed and has become an important complement to CTT in the design and evaluation of tests. The potential of IRT for solving different kinds of testing problems is substantial provided fit between the model and the test data of interest.

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 examinees´ 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 (ICC). (Hambleton et al., 1991, p. 7) The item statistics of interest are b, a, and c (for the three parameter model) plus corresponding item information functions. The b-parameter is an item difficulty parameter, a is an item discrimination parameter and c is a psuedo guessing parameter. (for more detailed descriptions of IRT see i.e. Lord, 1980, Hambleton & Swaminathan, 1985, Hambleton, Swaminathan & Rogers, 1991).

One great advantage of IRT is the item parameter invariance. The property of invariance of ability and item parameters is the cornerstone of IRT. It is the major distinction between IRT and classical test theory. (Hambleton, 1994, p. 540). The property of item parameter

invariance is also the property which would be of most value in the design of SweSAT. One drawback of IRT is that a big sample size is necessary for the estimation of parameters.

IRT has been vigorously researched by psychometricians and numerous books and articles have been published. The empirical studies available, however, have primarily focused on the application in test equating and very few studies have compared CTT and IRT for item analysis and test design. *It is somewhat surprising that empirical studies examining and/or comparing the invariance characteristics of item statistics from the two measurement frameworks are so scarce. 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. (Fan, 1998 p.361)*

Since spring 1996 pretesting of items for SweSAT has been performed in connection with the regular test administration, which means that the examinee sample on which pretesting is performed is a sample from the true examinee population and it contains 1500 examinees as a minimum. This new procedure for pretesting would make possible the use of IRT for item analysis and compilation of new test versions.

The present study has been performed within a project[1] with the general aim to examine whether the use of IRT would improve the quality of SweSAT. In earlier studies the applicability of IRT models to the SweSAT subtests was examined (Stage, 1996, 1997a, b, c, d) and the conclusion was that the three parameter logistic IRT model fitted the data reasonably well. In this study a comparison is made on the WORD subtest between item analysis based on CTT and item anlysis based on IRT. In studies to come the same comparisons will be made for other SweSATsubtests.

In the SweSAT given in spring 1997 the subtest WORD contained 20 items which had been pretested on five different samples from the examinee population in spring 1996. The aim of this study is to compare, for these 20 items, the stability of the item parameters estimated by IRT (BILOGW) with the item statistics obtained by CTT.

[1] The project is financed by The Swedish Council for Research in the Humanities and Social Sciences (HSFR).

In an earlier study (Stage, 1997d) of the applicability of IRT on the subtest WORD, the unidimensionality was assessed by factor analysis and the first three eigenvalues were 6.1, 1.4 and 1.2. An analysis of the standardized residuals between observed and model predicted performance gave as a result that 0.31 % of the standardized residuals had an absolute value higher than three, 3.13 % had an absolute value between two and three, 26.25 % between one and two and 70.31 % of the residuals had an abslute value lower than one. The test of individual item misfit which is included in the BILOGW program resulted in one item misfitting at the $\propto$=.01 level.

Aim

The purpose of the present study was to compare the item statistics from the CTT framework with those from the IRT framework and to examine the stability from pretest to regular test of the two sets of item statistics. Specifically the study addresses the following questions:

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

a) for pretest data?

b) for regular test data?

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

a) for pretest data?

b) for regular test data?

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

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

Method

Classical test theory

For the 20 WORD-items in the regular test spring 1997, which had been pretested in spring 1996, the p-values and the biserial correla-

tions (r_{bis}) were calculated. The same indices were calculated on the corresponding items in the pretest data and the values were compared.

Item response theory

The five WORD pretest combinations spring 1996 were run in BILOGW together with the regular WORD subtest from spring 1996 and the a-, b- and c-parameters were estimated. The WORD subtest from spring 1997 was run in BILOGW and the item parameters were estimated. The parameter estimates for the corresponding 20 items were noted and compared. The ICCs for the corresponding items were also compared (Figure 5 to 24).

One problem when analysing the stability of the item parameters is that pretesting has two purposes. One aim is to get information about the difficulty level and the discrimination power of the items in order to be able to compile parallel tests.The other purpose is to make sure that all the items function in a satisfactory way, and if an item is not working well enough one or more distractors may be changed. These changes mean that the corresponding items are not always exactly the same in the pretest version as in the regular test. Another preblem is that items are usually presented in different order in the pretest booklets and the regular test booklet.Even though the WORD subtest is not speeded, items may be more difficult when presented in the end of the booklet than when they are presented in the beginning In connection with the ICCs (pp. 11-30) the changes made between pretest and regular test will be described.

Results

Classical test theory

In Table 1 the p-values and the r_{bis} obtained from the five pretest versions spring 1996 and from the regular test spring 1997 are presented for the the 20 common items.

Table 1 *CTT-based item statistics.*

Item No		Pretest		Regular test	
pre	**reg**	**p-value**	**r_{bis}**	**p-value**	**r_{bis}**
8	1	.73	.60	.71	.58
20	4	.79	.46	.72	.41
39	5	.78	.25	.74	.33
18	9	.68	.44	.71	.43
36	10	.75	.50	.72	.53
27	11	.80	.35	.82	.35
36	15	.71	.40	.58	.37
14	16	.65	.44	.70	.48
5	19	.46	.42	.42	.37
16	23	.65	.35	.62	.40
38	24	.58	.47	.56	.30
12	25	.51	.58	.59	.58
24	27	.69	.36	.66	.35
4	28	.53	.56	.44	.52
4	29	.42	.33	.42	.26
5	35	.31	.32	.38	.46
37	36	.71	.43	.62	.44
6	38	.41	.31	.46	.37
28	39	.27	.28	.40	.32
39	40	.31	.23	.31	.21

The Spearman rho between p-values from pretest and regular test was ρ =.92 and for the same p-values transformed to delta, the correlation was r =.93.

In Figure 1 the p-values from the regular test spring 1997 have been plotted against the p-values from the pretest versions.

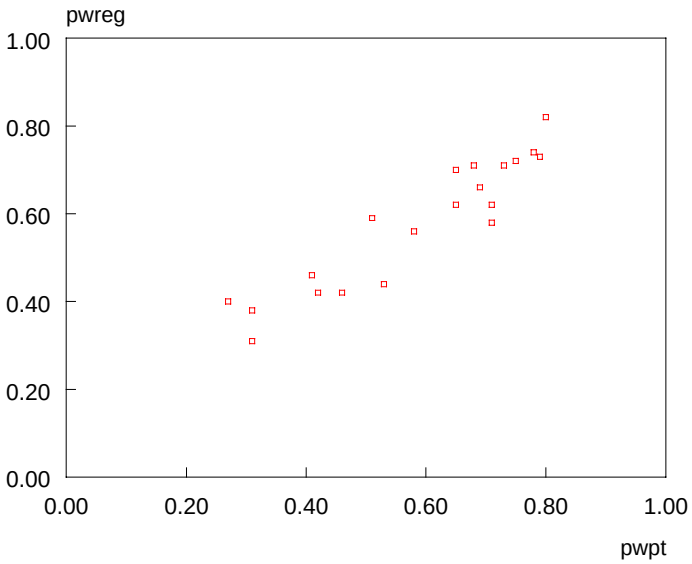

Figure 1 *The p-values from the regular test plotted against the p-values from the pretest*

In Figure 2 the r_{bis} of the items in the regular test have been plotted against the r_{bis} of the same items in the pretest

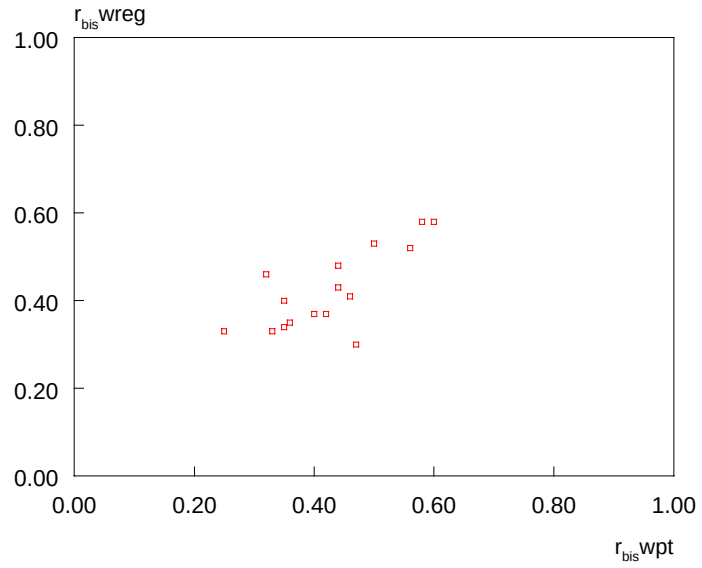

Figure 2 *The r_{bis} from the regular test plotted against the r_{bis} from the pretest.*

For the r_{bis} the correlation between pretest and regular test was r = .81.

Item response theory

In Table 2 the a-, b- and c-parameters are presented from the pretest versions spring 1996 and the regular test spring 1997.

Table 2 *IRT-based item statistics.*

Item No		**Pretest**			**Regular test**		
pre	**reg**	**b**	**a**	**c**	**b**	**a**	**c**
8	1	-.43	1.29	.27	-.43	1.14	.23
20	4	-.96	.73	.25	-.64	.62	.23
39	5	-1.94	.32	.20	-1.19	.43	.17
18	9	-.25	.71	.27	-.59	.63	.23
36	10	-.92	.72	.15	-.81	.78	.11
27	11	-1.49	.48	.20	-1.79	.46	.19
36	15	-.55	.59	.26	.11	.55	.20
14	16	-.02	.81	.29	-.65	.72	.16
5	19	.46	.55	.08	.71	.50	.09
16	23	.11	.58	.32	.08	.72	.28
38	24	.08	.75	.18	.16	.40	.16
12	25	.17	.97	.12	-.02	1.13	.17
24	27	.37	1.07	.48	.33	.83	.41
4	28	.35	1.55	.25	.51	1.04	.14
4	29	1.16	.71	.23	1.08	.61	.19
5	35	1.59	.45	.09	.89	.95	.16
37	36	-.37	.70	.29	-.07	.70	.21
6	38	1.25	.70	.23	.49	.48	.09
28	39	1.61	1.13	.19	1.22	1.18	.27
39	40	2.27	.42	.15	2.45	.45	.18

The correlation between the b-values estimated on pretest and regular test data was r = 92. A plot of the b-values is shown in Figure 3.

The correlation between a-values estimated on pretest and regular test data was r = .74 and the plot is shown in Figure 4.

The correlation between c-values was r = .74

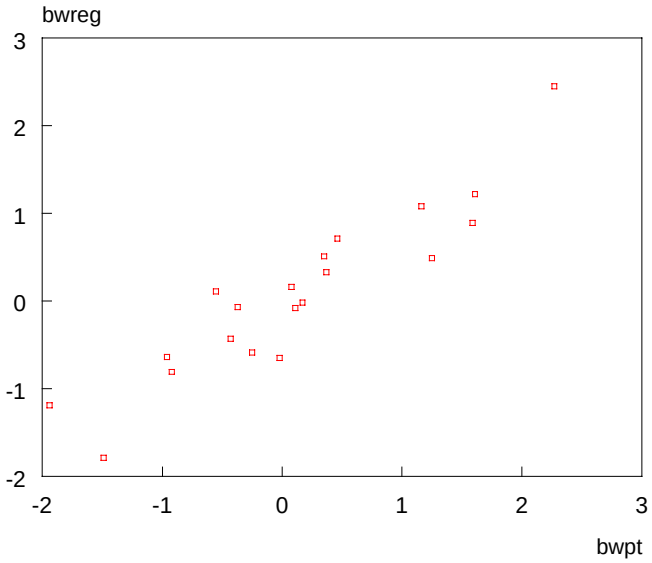

Figure 3 *Plot of b-values from the regular test against b-values from the pretest.*

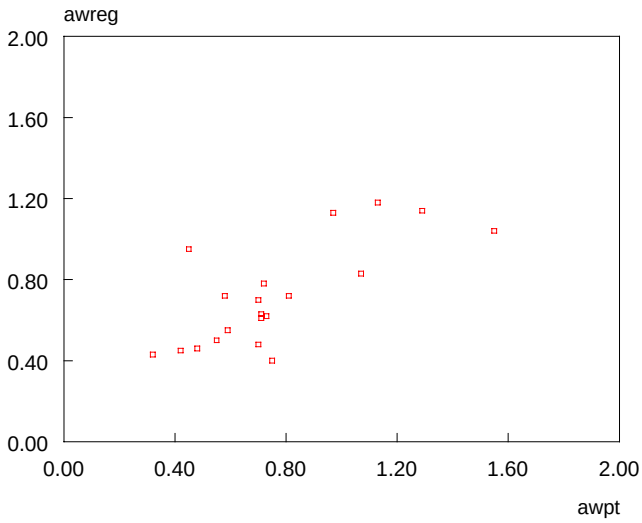

Figure 4 *Plot of a-values from the regular test against a-values from the pretest.*

Item Characteristic Curves of the 20 word items 1997 which were pretested in spring 1996

In Figures 5 to 24 the ICCs of each item from the pretest as well as the ICCs of the corresponding items from the regular test are shown.

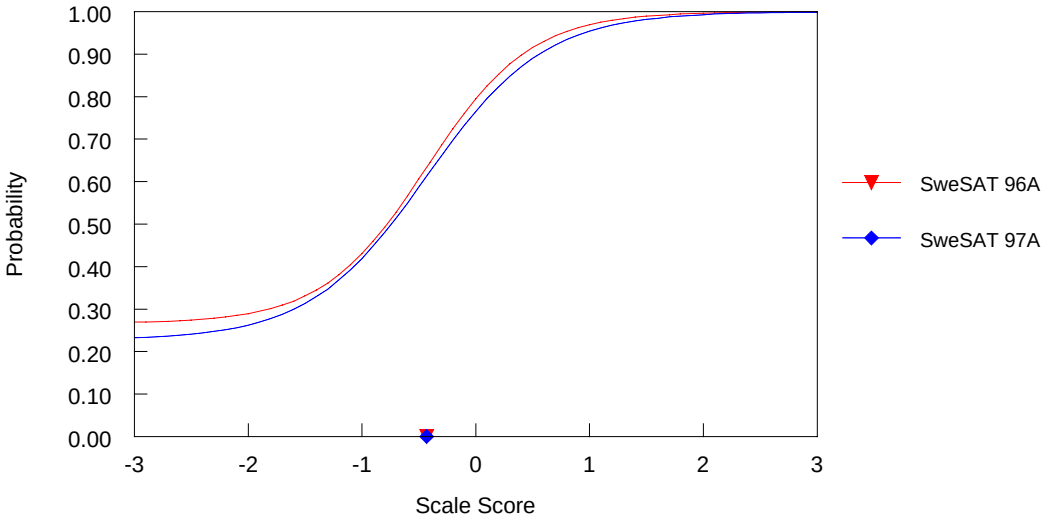

Figure 5 *ICCs for item No 1 in the Spring 1997 WORD subtest.*

In this item two of the distractors had been changed after the pretest and the item was number 8 in the pretest booklet.

As may be seen in Figure 6 the two ICCs correspond very well, the b-values were exactly the same (-.43) in the pretest as in the regular test, while the a-value in the pretest was 1.29 and in the regular test 1.14, i.e. there was a very small decrease in discrimination power from pretest to regular test.

The p-value for this item in the pretest was .73 and in the regular test the p-value was .71; i.e. a very small decrease; the r_{bis} in the pretest was .60 and in the regular test it was .58; a small decrease as well.

On the whole the results from CTT and IRT correspond very well and according to both analyses this item seems to work in the same way in the pretest as in the regular test.

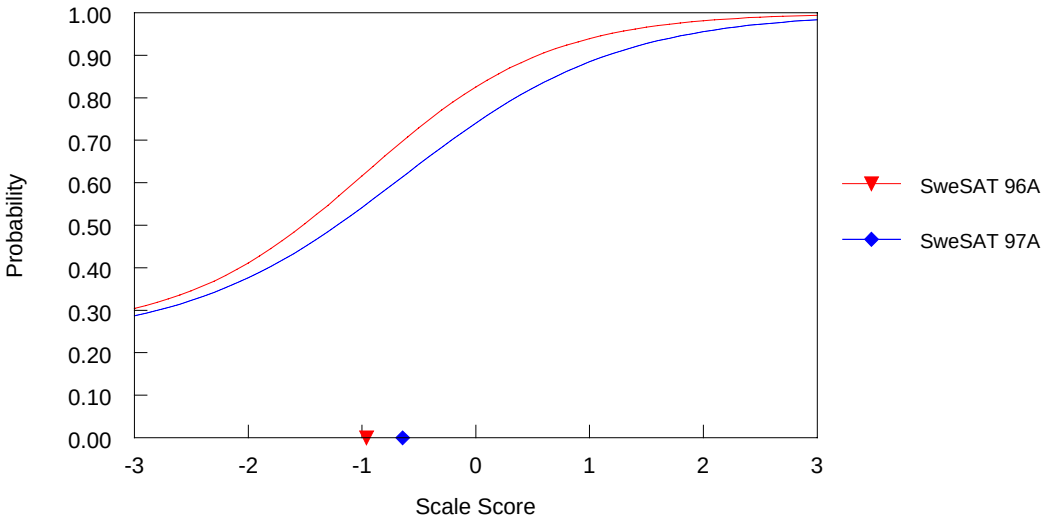

Figure 6 *ICCs of item No 4.*

In item No 4 one distractor had been changed after the pretest and the position in the pretest booklet was No 20.

For this item the b-value had increased from -.96 in the pretest to -.64 in the regular test, i.e. the item was a bit more difficult in the regular test; the a-value had decreased from .73 to .62, hence the discrimination is somewhat lower in the regular test than in the pretest.

From the CTT the conclusions are the same: the p-value has decreased from .79 to .72 and the r_{bis} has decreased from .46 to .41; i.e. the item had become somewhat more difficult but less discriminating in the regular test than it was in the pretest.

For item No 4 as well the conclusions from the two analyses are the same.

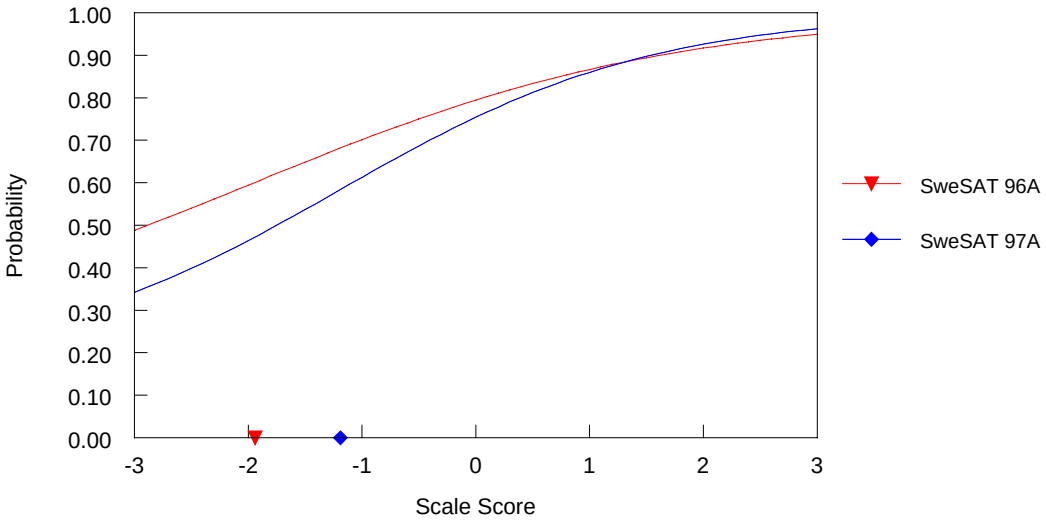

Figure 7 *ICCs of item No 5.*

In item No 5 two distractors and the correct answer had been changed slightly; the position in the pretest booklet was No 39

For this item the b-value as well as the a-value had increased from pretest to regular test (from -1.94 to -1.19 and from .32 to .43) which means that the item is more difficult and also a bit more discriminating in the regular test than it was in the pretest.

The same conclusions are drawn from CTTas the p-value had decreased from .78 to .74 and the r_{bis} had increased from .25 to .33.

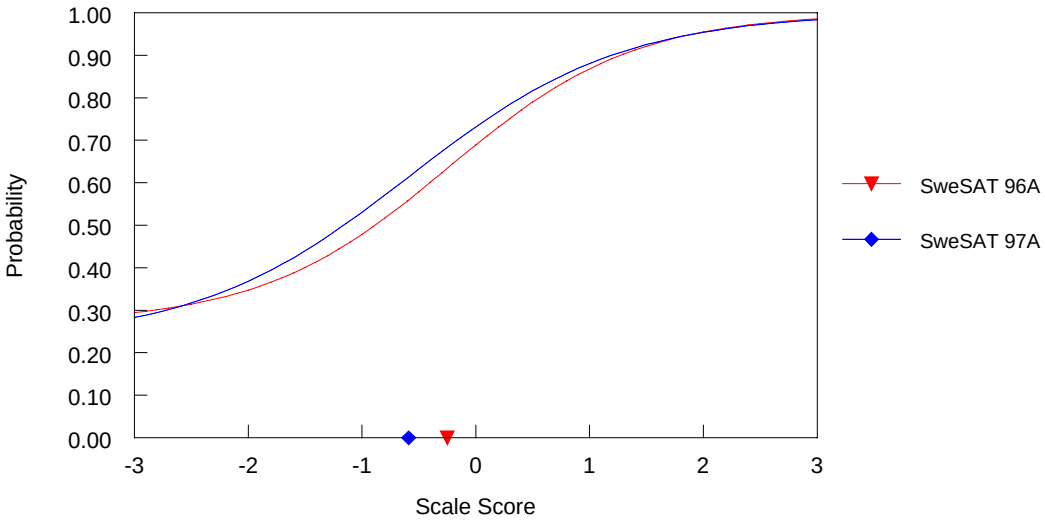

Figure 8 *ICCs of item No 9.*

In item No 9 only the order of distractors had been changed but the position in the pretest booklet was No 18.

For this item the b-value had decreased from -.25 to -.59 and the a-value from .71 to .63. from the pretest to the regular test. This means that the item was a bit easier and less discriminating in the regular test than in the pretest.

The p-value had increased from .68 to .71 while the r_{bis} is almost the same (.44/.43).

Hence the conclusions from the two theories are the same: the item was somewhat easier and slightly less discriminating in the regular test than in the pretest.

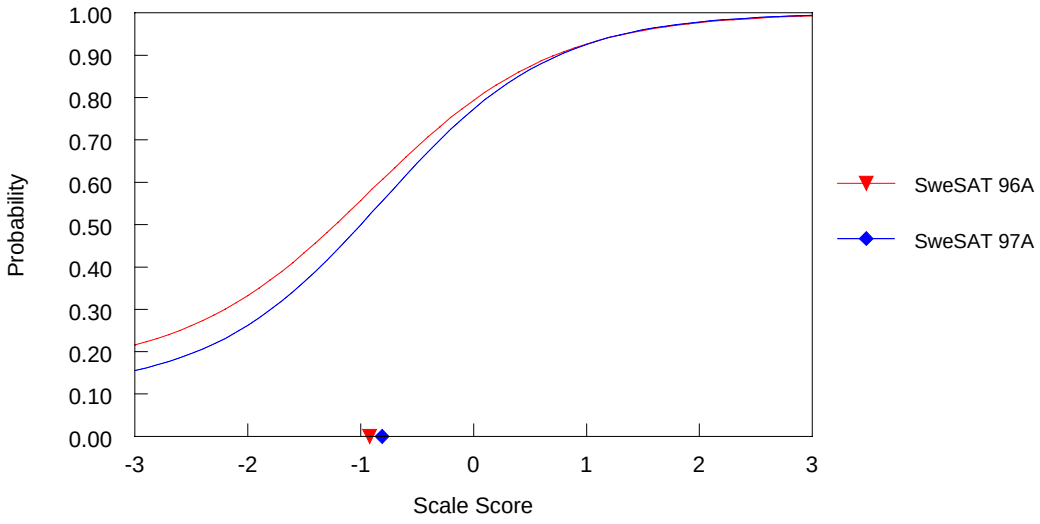

Figure 9 *ICCs of item No 10.*

In item No 10 one distractor had been changed and the position in the pretest booklet was No 36.

For this item the b-value had increased very little (from -.92 to -.81) from the pretest to the regular test and so had the a-value (from .72 to .78). This means that the item is a little more difficult and better discriminating in the regular test than it was in the pretest.

The same conclusion is drawn from the classical test theory since the p-value had decreased from .75 to .72 and the r_{bis} had increased from .50 to .53.

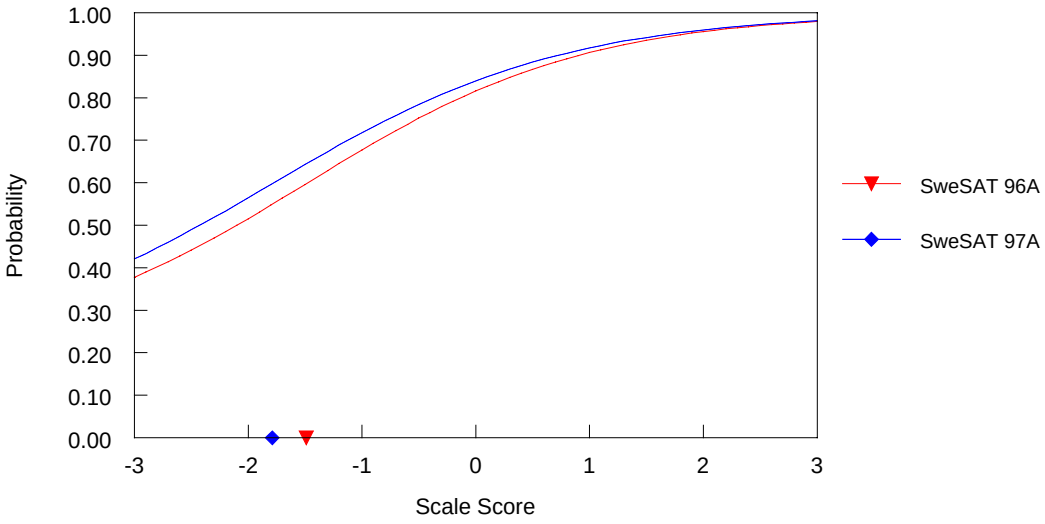

Figure 10 *ICCs of item No 11*

In item No 11 one distractor had been changed and the position in the pretest booklet was No 27.

For this item the b-value had decreased (from -1.49 to -1.79), while the a-value was very much the same (.48/.46), i.e. the item was a little bit easier in the regular test than in the pretest, while the discrimination is about the same.

The p-value had increased from .80 to .82, while r_{bis} was the the same .35 in the regular test as in the pretest.

Again the conclusions are the same from CTT and IRT.

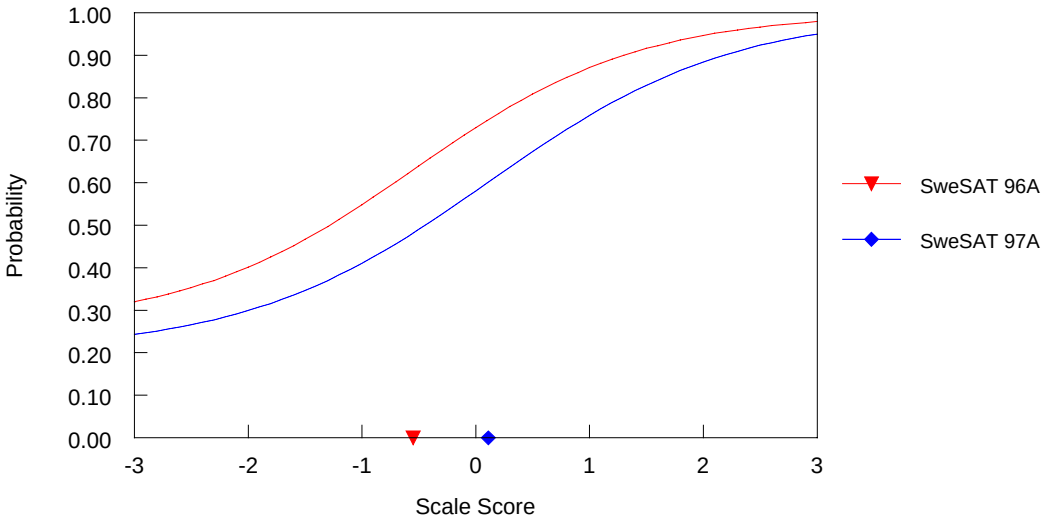

Figure 11 *ICCs of item No 15.*

In item No 15 one distractor had been changed; and the position in the pretest booklet was No 36.

For item 15 the b-value had increased from -.55 to .11, while the a-value had decreased to a very small extent (from .59 to .56). The p-value had decreased from .71 to .52 and the r_{bis} from .40 to .37.

Again the conclusions are the same, the item was more difficult and somewhat less discriminating in the regular test than in the pretest.

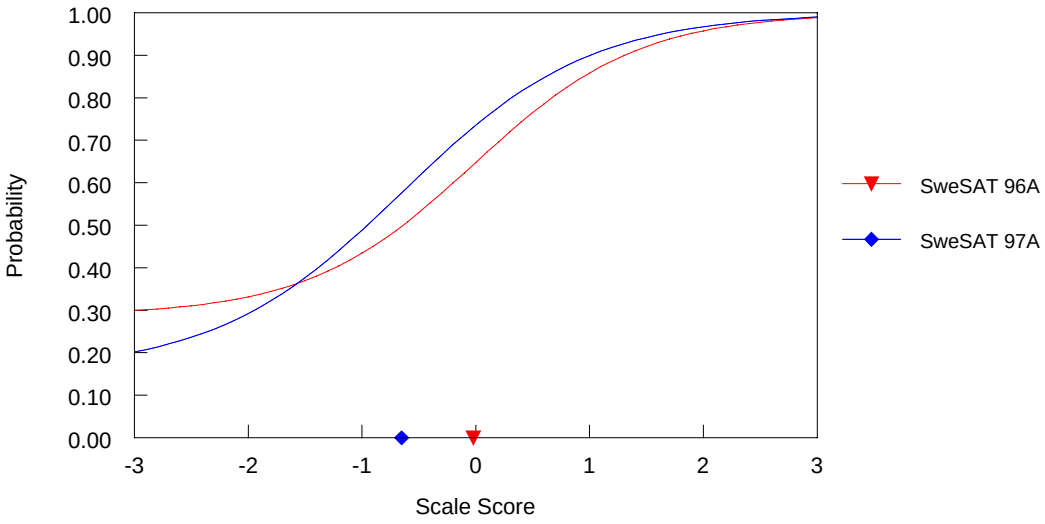

Figure 12 *ICCs of item No 16.*

In item No 16 as well one distractor had been changed; the position in the pretest booklet was No 14.

For item No 16 the b-value had decreased from -.02 to -.65 and the a-value had decreased from .81 to .72, which means that the item was easier and somewhat less discriminating in the regular test than in the pretest.

The p-value had increased from .65 to .70 and the r_{bis} had increased from .44 to .48.

For this item the conclusion from both item-analyses is that the item was easier in the regular test than in the pretest but according to IRT it was less discriminating and according to CTT it was more discriminating in the regular test than in the pretest.

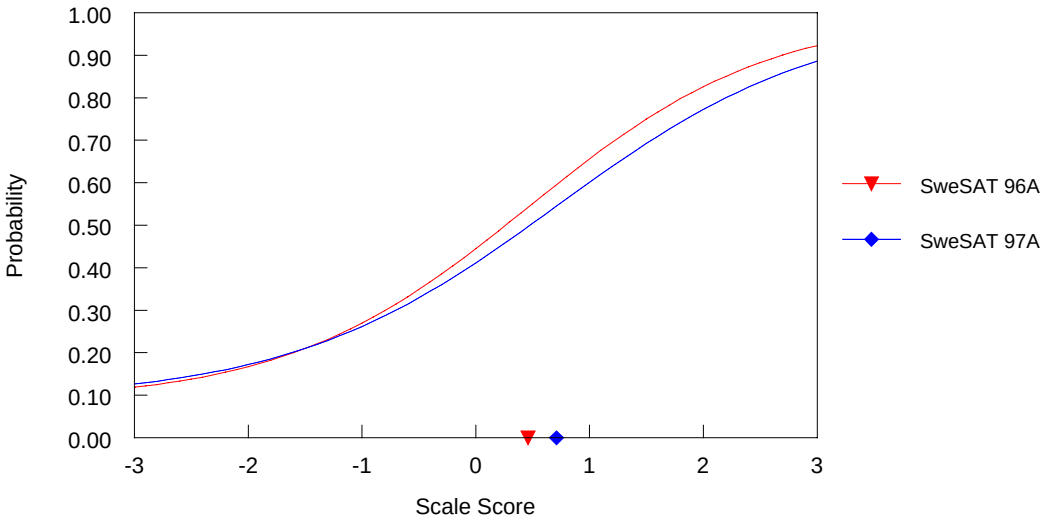

Figure 13 *ICCs of item No 19.*

In item No 19 nothing had been changed, but the position in the pre-test booklet was No 5.

For item No 19 the b-value had increased from .46 to .71, while the a-value had decreased from .55 to .50.

The p-value had decreased from .46 to .42 and the r_{bis} from .42 to .37.

According to both analyses the item was a bit more difficult and less discriminating in the regular test than in the pretest.

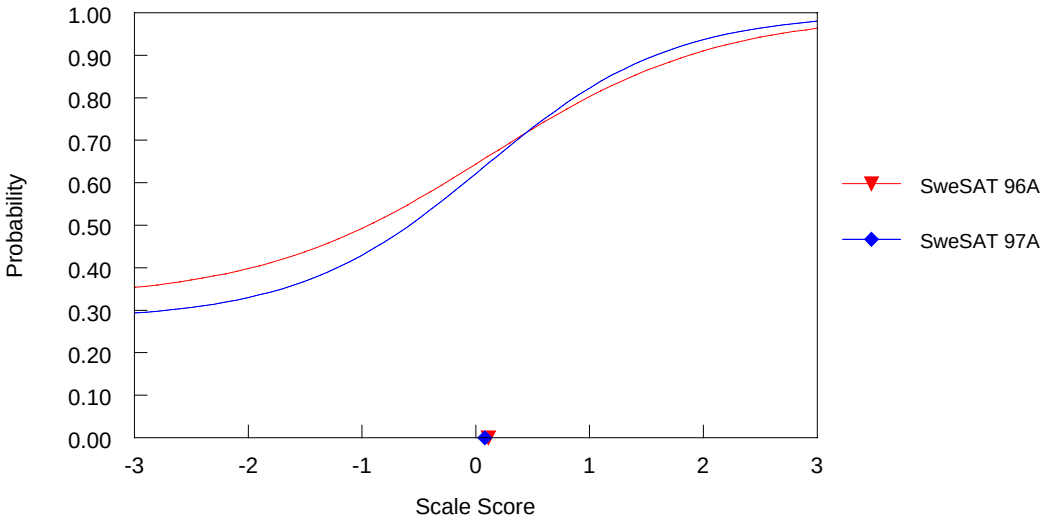

Figure 14 *ICCs of item No 23.*

In item No 23 no changes had been made and the position in the pretest booklet was No 16.

For item No 23 the b-value was almost the same (.11/.08) in the regular test as in the pretest but the a-value had increased from .58 to .72.

The p-value had decreased from .65 to .62 and the r_{bis} had increased from .35 to .40.

According to IRT this item was unnoticeably easier but had better discrimination power in the regular test than in the pretest. According to CTT the item was a bit more difficult and but also better discriminating in the regular test than in the pretest.

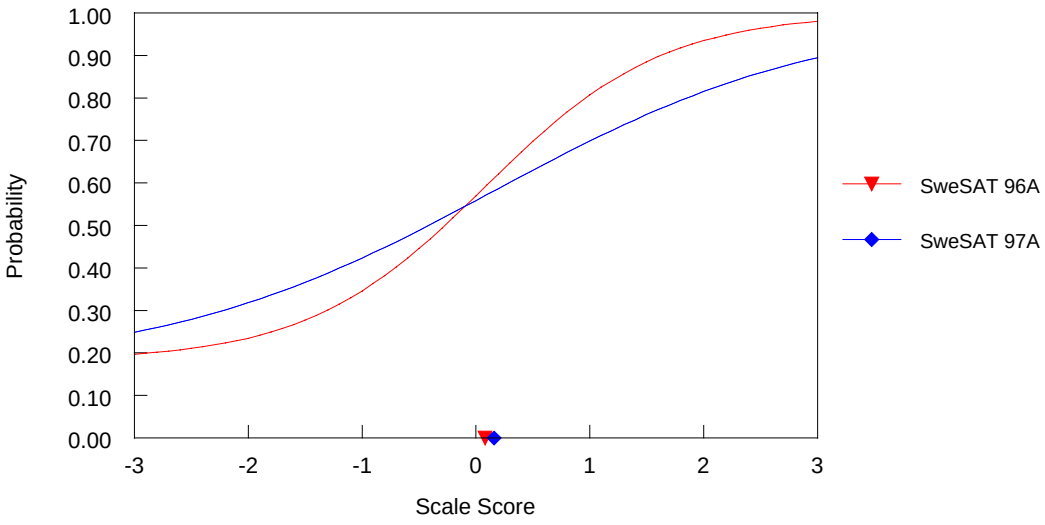

Figure 15 *ICCs of item No 24.*

In item No 24 no changes had been made but the position in the pre-test booklet was No 38.

For item No 24 the b-value had increased slightly from .08 to .16, while the a-value had decreased from .75 to .40 from pretest to regular test.

The p-value had decreased from .58 to .56 and the r_{bis} had decreased from .47 to .30.

According to both analyses the item had become a bit more difficult but less discriminating in the regular test than in the pretest.

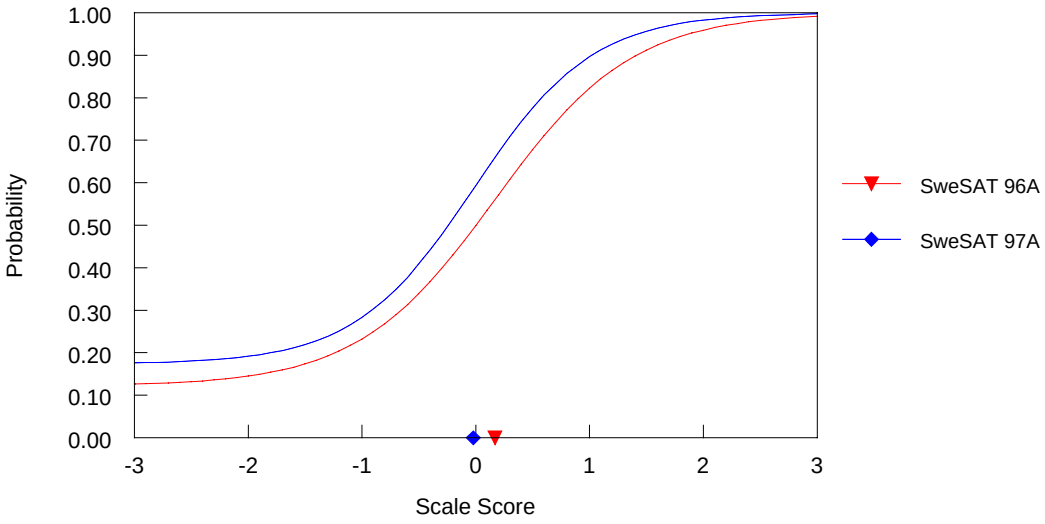

Figure 16 *ICCs of item No 25.*

In item No 25 no changes had been made but the position in the pretest booklet was No 12.

For item No 25 the b-value had decreased from .17 to .12, while the a-value had increased from .97 to 1.13 from the pretest to the regular test.

The p-value had increased from .51 to .59 and the r_{bis} was unchanged (.58).

Hence according to both analyses the item was a bit easier in the regular test but according to IRT it was also slightly more discriminating in the regular test than in the pretest.

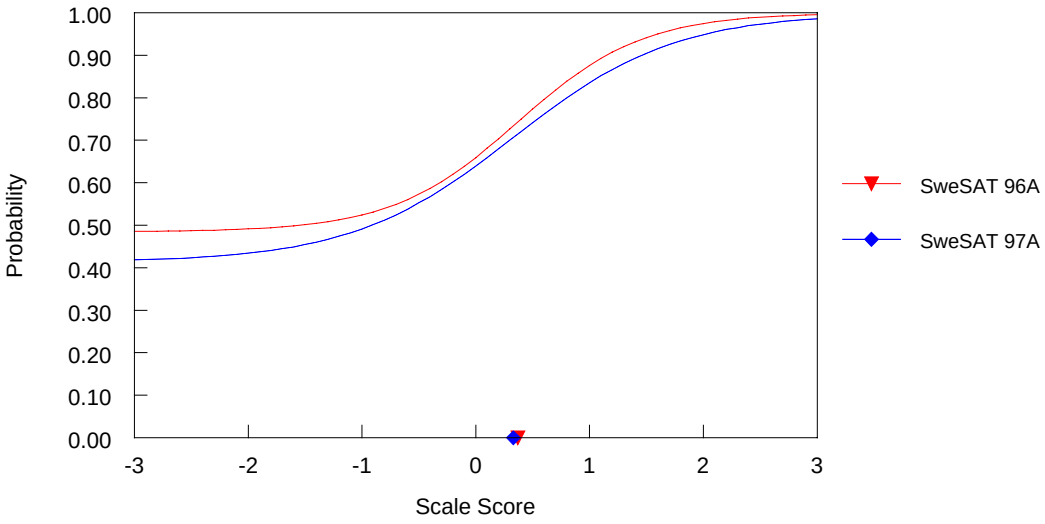

Figure 17 *ICCs of item No 27.*

In item No 27 no changes had been made and the position in the pretest booklet was No 24.

For item No 27 the b-value had decreased from .37 to .33 and the a-value had decreased from 1.07 to .83 from pretest to regular test.

The p-value had decreased from .69 to .66 and the r_{bis} had decreased slightly (from .36 to .35) from pretest to regular test.

Hence according to IRT this item was somewhat easier in the regular test than in the pretest, while according to CTT the item was somewhat more difficult in the regular test. According to both analyses the discrimination power had decreased to a small extent from pretest to regular test.

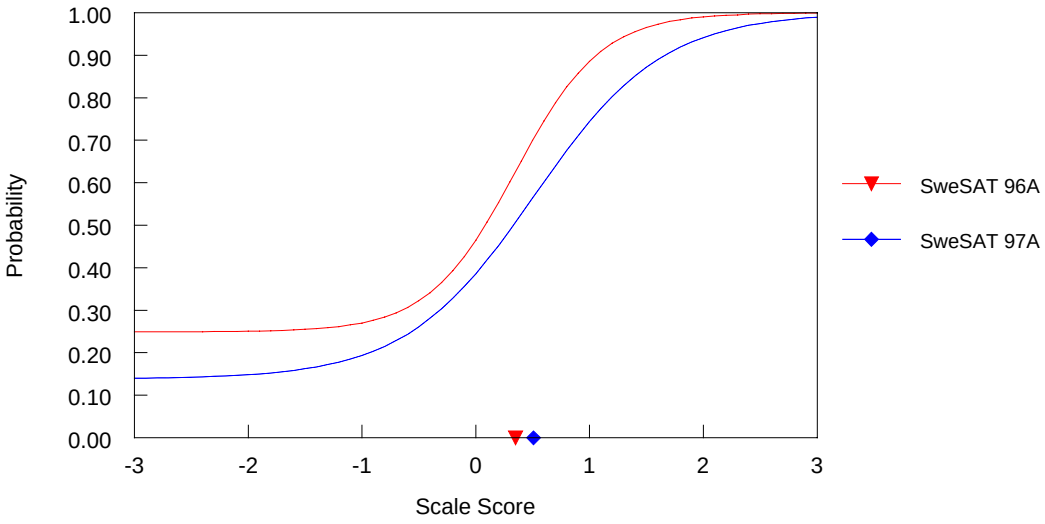

Figure 18 *ICCs of item No 28.*

In item No 28 one distractor and the correct answer had been changed; the position in the pretest booklet was No 4.

For item No 28 the b-value had increased from .35 to .51, while the a-value had decreased from 1.55 to 1.04 from pretest to regular test.

The p-value had decreased from .53 to .44 and the r_{bis} had decreased from .56 to .52.

According to both theories the item was more difficult but less discriminating in the regular test than in the pretest.

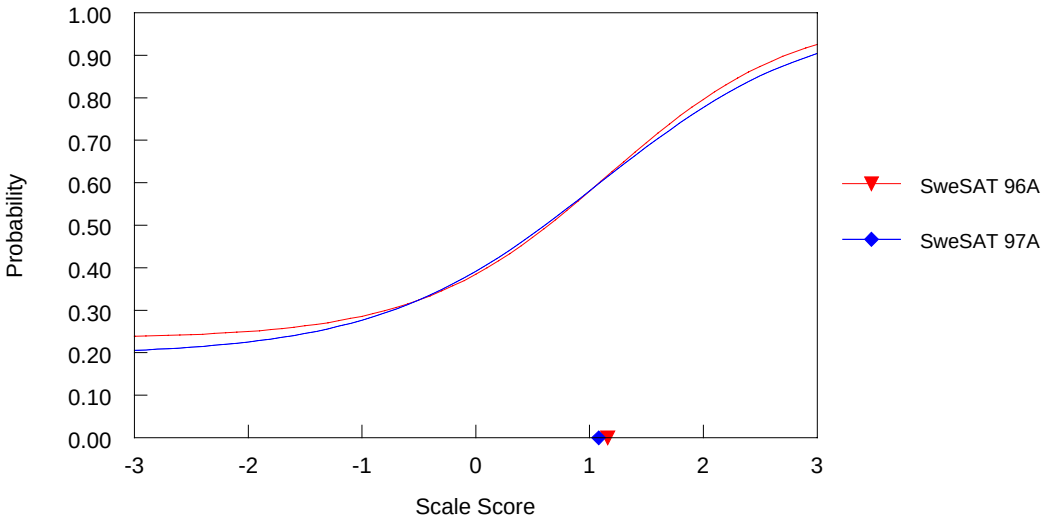

Figure 19 *ICCs of item No 29.*

In item No 29 no changes had been made but the position in the pre-test booklet was No 4.

For item No 29 the b-value had decreased slightly from 1.16 to 1.08 and the a-value had decreased from .71 to .61 from pretest to regular test.

The p-value was the same (.42) in the pretest as in the regular test and so was the r_{bis} (.33).

Hence according to IRT the item was somewhat easier and poorer discriminating in the regular test than in the pretest. According to CTT the difficulty level was exactly the same and so was the discrimination power.

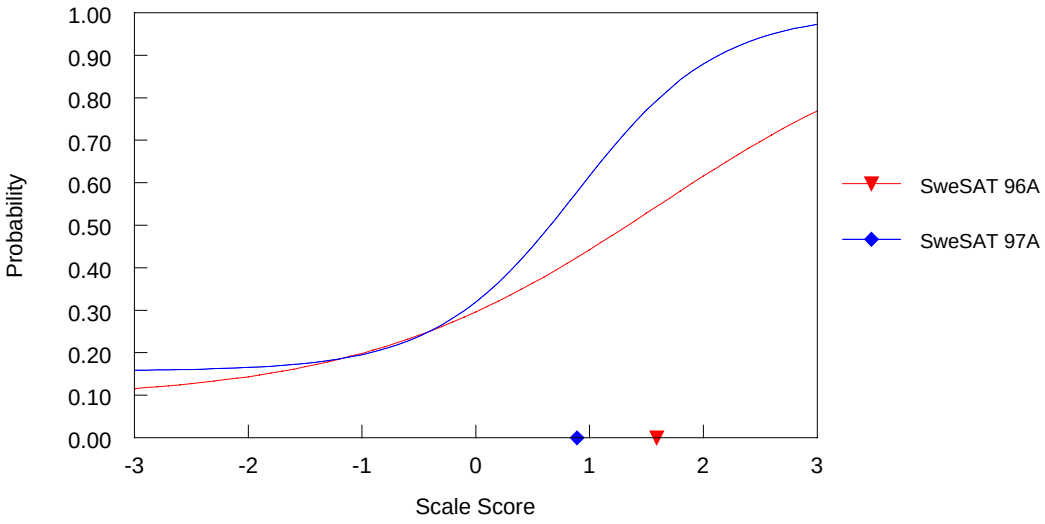

Figure 20 *ICCs of item No 35.*

In item No 35 one distractor had been changed and the position in the pretest booklet was No 5.

For item No 35 the b-value had decreased substantially (from 1.59 to .89) but the a-value had increased (from .45 to .95) between pretest and regular test.

The p-value had increased from .31 to .38 and the r_{bis} had increased from .32 to .46 from pretest to regular test.

Hence according to both analyses this item was easier but more dicriminating in the regular test than in the pretest.

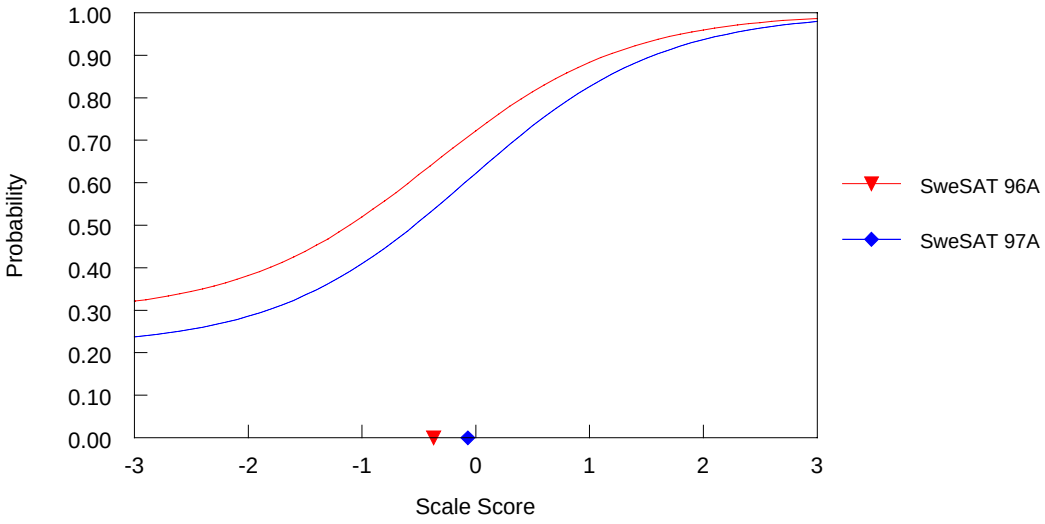

Figure 21 *ICCs of item No 36.*

In item No 36 one distractor had been changed and the position in the pretest booklet was No 37.

For item No 36 the b-value had increased from -.37 to -.07 from pretest to regular test, while the a-value remained the same (.70).

The p-value had decreased from .71 to .62 and the r_{bis} had increased slightly (from .43 to .44) from pretest to regular test.

According to both analyses the item was more difficult in the regular test than in the pretest, while the discrimination remained approximately the same.

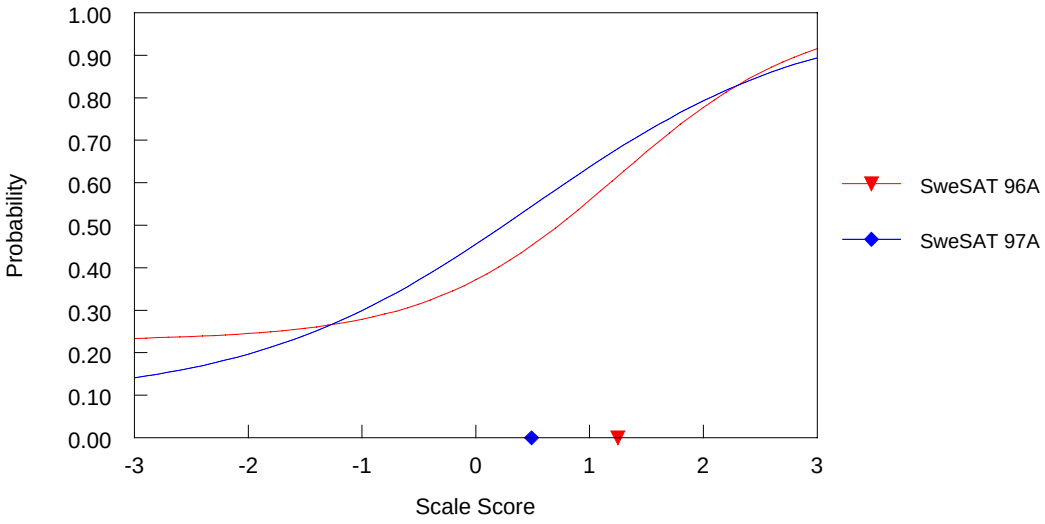

Figure 22 *ICCs of item No 38.*

In item No 38 as well one distractor had been changed and the position in the pretest booklet was No 6.

For item No 38 the b-value had decreased from 1.25 to .49 and the a-value had decreased from .70 to .48 from the pretest to the regular test.

The p-value had increased from .41 to .46 and the r_{bis} had increased from .31 to .37.

According to both analyses the item was easier in the regular test than in the pretest, but according to IRT the discrimination was relatively better in the pretest version and according to classical test theory it was better in the regular test.

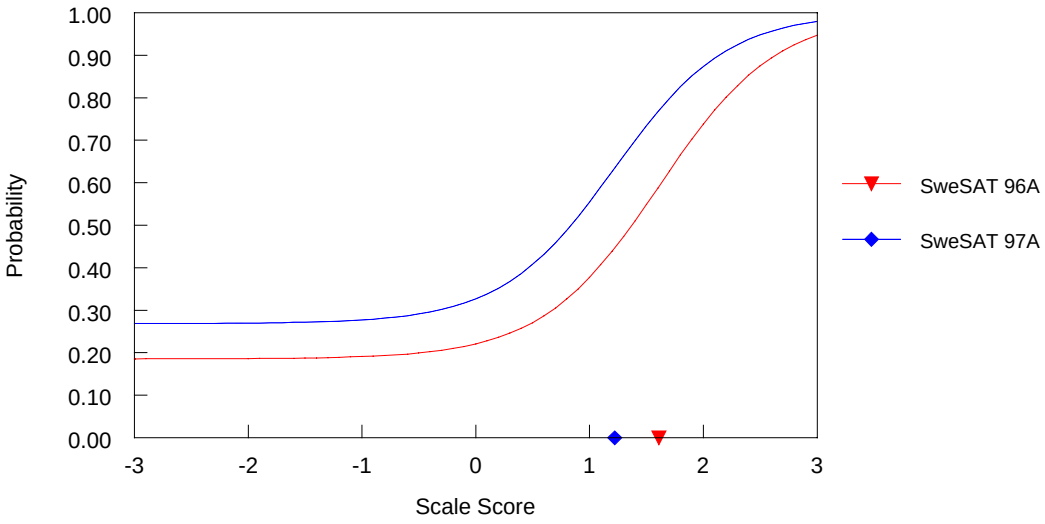

Figure 23 *ICCs of item No 39.*

In item No 39 three distractors had been changed and the position in the pretest booklet was No 28.

For item No 39 the b-value had decreased from 1.61 to 1.22, while the a-value had increased slightly from 1.13 to 1.18 from the pretest to the regular test.

The p-value had increased from .27 to .40 and the r_{bis} had increased from .28 to .32.

According to both analyses the item was easier and more discriminating in the regular test than in the pretest.

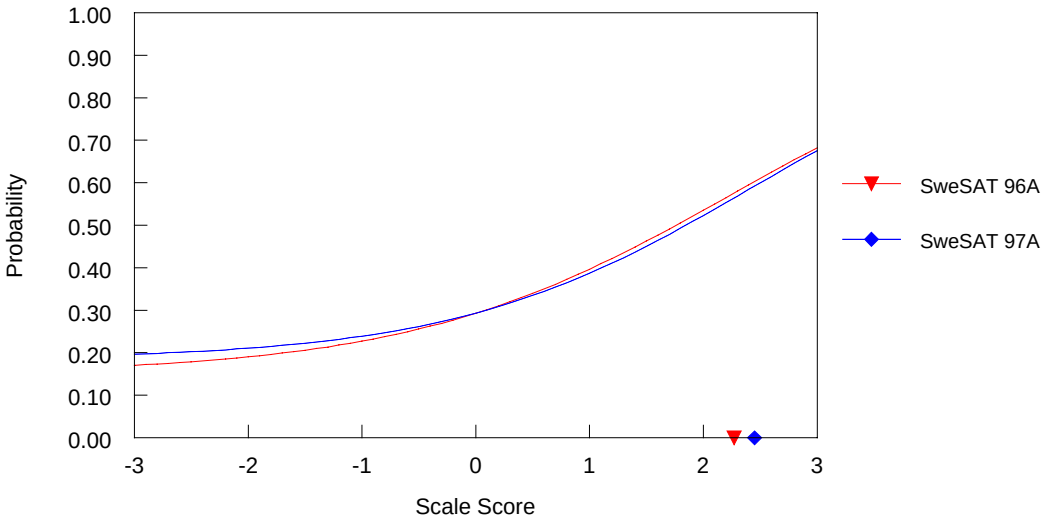

Figure 24 *ICCs of item No 40.*

In item No 40 two distractors had been changed and the position in the pretest booklet was No 39.

For item No 40 the b-value had increased from 2.27 to 2.45 and the a-value had increased from .42 to .45 from pretest to regular test.

The p-value was the same (.31) in both versions but the r_{bis} had decreased slightly from .23 to .21 from pretest to regular test.

According to both analyses this item was very difficult and poorly discriminating, but according to IRT the item was somewhat more difficult and slightly better discriminating in the regular test than in the pretest, while according to CTT the difficulty level was the same for the two versions and the discrimination power was slightly poorer in the regular test.

Discussion

The agreement between results from item-analysis performed by IRT and CTT is very good. For 13 of the 20 items analysed the conclusions about the change between pretest and regular test regarding difficulty level as well as discrimination were exactly the same. For only one item (No 40) the conclusion about both difficulty and discrimination differred, but according to both theories this item was very difficult and had low discrimination. For two items only (No 25 and 29) there were differences regarding the change of difficulty level and for three items there were differences regarding the change of discrimination. All differences, however, were minor.

As for model data fit if the IRT model used (three parameter logistic model) none of the 20 items was identified as misfitting at the $\alpha = .01$ level in the pretest versions. In the regular test version one item (No 10) was identified as missfitting at $\alpha = .01$ level[2]. For item No 10 the results from the two theories were the same, however.

There are at least two complications for the prediction of the regular data, when using actual data as in this study. One complication is that the items are not always exactly the same in the pretest as in the regular test. Distractors which did not work in the pretest were changed before the item was included in the regular test; and the effect of such changes is not always possible to foresee. The other complication is that items are presented in different order in the pretest and the regular test. Even though the test is not actually speeded, items seem to be more difficult when placed in the end of the booklet than when they are placed in the beginning. For the two items which had changed most in difficulty level according to CTT (No 15 and 39) it is impossible to tell the reason; for item No 15 one distractor had been changed but the item was number 36 in the pretest booklet; for item No 39 three distractors had been changed and also the item was number 28 in the pretest booklet. The items which had changed most in difficulty level according to IRT were No 5 and No 35 and these changes were difficult to interpret as well; in item No 5 two distractors had been changed and the order in the pretest booklet was 39, but the item had become more difficult; for item No 35 one distractor had been changed and the position in the pretest booklet was 5, but the item had be-

[2] There was a discrepancy between model predicted and observed response pattern.

come easier. For the items where no distractors had been changed (Nos 9, 23, 24, 25, 27 and 29) the ICCs are very similar and all changes in p-values were very small as well.

The overall conclusion from this study is that the prediction from pretest to regular test data is satisfactory and the major part of the discrepancy in the prediction can be explained by changes of the items. This conclusion, however, is true for both analyses regardless of theoretical framwork..

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)

In this empirical study the correspondence between results from the item analyses performed within the two different theoretical frameworks was very good. The comparability of IRT- and CTT-based item statistics was examined by correlating IRT and CTT item statistics obtained from the same sample of participants. The correlation between the item difficulty parameter ”b” from the IRT model with the CTT item difficulty value ”p” was r = .93 for pretest results as well as for regular test results. The correlation between the IRT item discrimination parameter ”a” and CTT item discrimination index ”r_{bis}” was r = .65 for the pretest results and r= .64 for the regular test results. And also for the individual items the accuracy of the predictions made from pretest results to regular test results were very similar.

For the CTT item indices the correlation between pretest and regular test were, for p-values r = .93 and for r_{bis} r = .81. For the IRT parameters the correlation between pretest and regular test item parameters were for b-values r = .92 and for a-values r = .74.

What is important when compiling a test like SweSAT, however, is to be able to predict the difficulty level of the regular test from the pretest data. As for the discrimination power of the items it is enough to

know that every item is discriminating satisfactorily, you do not need to predict the exact level of discrimination.

In this study where the pretesting had been performed on large and representative samples it does not seem to be of any importance for the test design whether the item analysis has been performed within the IRT framework or within the CTT framework.

References

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

Fan, X. (1998). Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educational and Psychological Measurement, 58 (3),* 357-381.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Boston: Kluwer.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: 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),* 38-47.

Hambleton, R. K. (1994). Item Response Theory: A Broad Psychometric Framework for Measurement Advances. *Psicothema, 6 (3),* 535-556.

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

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

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

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

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

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

Stage, C. (1997d). *The Applicability of Item Response Models to the SweSAT. A Study of the WORD Subtest*. (Educational Measurement No 26). Umeå University, Department of Educational Measurement.