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

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 multiple-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 which will give a subtest with the same level of difficulty as 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.

On the basis of the data obtained in the pretest the items are improved and selected for the final test.
The statistics which are used in the item analysis are:
p-values of the items
p-values of the distractors

p-values of males and females
biserial correlations (r_{bis})
(the item test regression)

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).

There are some shortcomings with CTT, however, one of which is that the item statistics are sample dependent. This may cause problems, especially if the sample on which the pretesting was made differs in some unknown way from the examinee population. Another limitation which may be important 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 that there is 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 el, 1991, p. 7). The item statistics are in order of interest: 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 pseudo 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 invaraiance 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, however, 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 anomal (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 makes it possible to use IRT for item analysis and compilation of new test versions.

The present study has been performed within a research project[1] with the general aim of examining whether the use of IRT would improve the quality of SweSAT. In earlier studies the applicability of IRT models to SweSAT was examined (Stage, 1996, 1997a, b, c, d) and the conclusion was that a three parameter logistic IRT model fitted the data reasonably well. In this study a comparison is made on the READ subtest between item analysis based on CTT and item anlysis based on IRT. In earlier studies (Stage, 1998a, b) the same comparisons were made for the WORD and ERC subtests and the conclusion from both studies was that the results were very similar in spite of the differences between the theoretical frameworks.

In the SweSAT given in spring 1997, the subtest READ contained 16 items, which had been pretested on eight different samples from the examinee population in spring 1996. The aim of this study is to compare, for these 16 items, the stability of item parameters estimated by IRT (BILOGW) with 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, 1997c) of the applicability of IRT on the subtest READ, the unidimensionality was assessed by factor analysis and the first eigenvalues were 2.9, 1.1 and 1.0. An analysis of the standardized residuals between observed and model predicted performance gave as a result that none of the standardized residuals had an absolute value higher than three, 6.9 % had an absolute value between two and three, 24.4 % between one and two and 68.8 % of the residuals had an absolute value lower than one. The test of individual item misfit which is included in the BILOGW program resulted in seven items misfitting at the α =.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 with item difficulty parameters estimated by IRT?

a) for pretest data?

b) for regular test data?

2. How do item diccrimination indices from CTT compare with 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 16 READ-items in the regular test in spring 1997, which were pretested in spring 1996, the p-values and the biserial correlations

(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 eight READ pretest versions in spring 1996 were run in BILOGW together with the regular READ subtest from spring 1996, and the a-, b- and c-parameters were estimated. The READ subtest from spring 1997 was run in BILOGW and the item parameters were estimated. The parameter estimates for the corresponding 16 items were noted and compared. The ICCs for the corresponding items were also compared (Figure 5 to 20).

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 tests of equal difficulty. 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. The changes mean that these items are not exactly the same in the pretest version as in the regular test. Another problem is that the order of presentetation in the pretest booklets may differ from the order in the regular test. Changes in the order of presentation may also change the item in some way. The order of presentation in the pretest and the regular test is given in Tables 1 and 2.

Results

Classical test theory

In Table 1 the p-values and the r_{bis} obtained from the eight pre-test versions and from the spring 1997 test are presented for the the 16 common items.

Table 1. *CTT-based item indices: p-values and r_{bis} for 16 items.*

Item No		**Pretest**		**Regular test**	
pre	**reg**	**p**	$\mathbf{r_{bis}}$	**P**	$\mathbf{r_{bis}}$
5	5	.74	.30	.74	.32
7	6	.30	.23	.68	.36
6	7	.78	.37	.71	.38
8	8	.80	.40	.81	.35
17	9	.64	.37	.81	.43
20	10	.52	.25	.69	.22
17	11	.61	.18	.75	.11
20	12	.45	.17	.52	.33
14	13	.59	.35	.66	.36
15	14	.36	.24	.50	.30
14	15	.24	.34	.30	.43
16	16	.28	.17	.57	.28
17	17	.35	.35	.53	.36
19	18	.63	.29	.76	.32
19	19	.56	.29	.65	.31
20	20	.57	.34	.60	.34

The correlation between p-values of the items in the pretest versions and p-values of the corresponding items in the regular test version was r = .78 and Spearman rank ρ = .82. A plot of the p-values is shown in Figure 1.

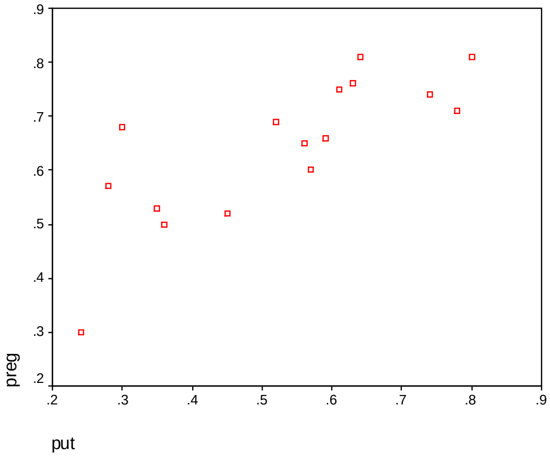

Figure 1. *Plot of p-values calculated on regular test data against p-values calculated on pretest data.*

The correlation between r_{bis} of the items in the pretest versions and the corresponding items in the regular test version was r = .66 and and ρ = .69. The plot of r_{bis} is shown in Figure 2.

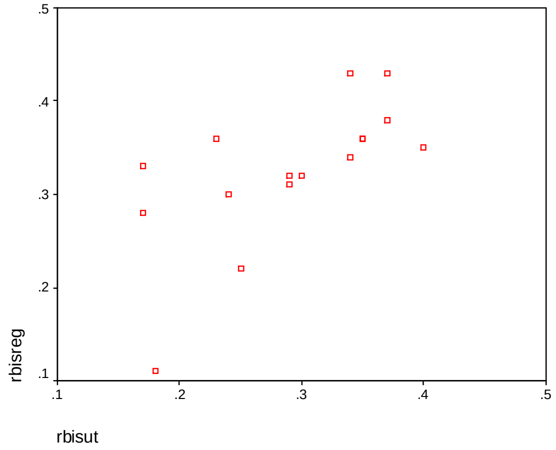

Figure 2. *Plot of r_{bis} calculated on regular test data against r_{bis} calculated on pretest data.*

Item response theory

Table 2. *IRT-based item statistics*: *estimated b-, a-, and c-parameters*

Item No		Pretest			Regular test		
pre	**reg**	**b**	**a**	**c**	**b**	**a**	**c**
5	5	-.61	.50	.33	-.69	.59	.23
7	6	2.16	.46	.16	-.52	.64	.22
6	7	-1.16	.54	.23	-.45	.73	.24
8	8	-1.31	.59	.21	-1.09	.63	.26
17	9	.37	.85	.39	-.88	.90	.27
20	10	1.32	.64	.37	.27	.42	.41
17	11	1.42	.52	.48	-1.20	.37	.28
20	12	1.92	72	.36	.71	.86	.30
14	13	.48	.83	.35	-.14	.69	.29
15	14	1.85	.53	.21	.74	.72	.25
14	15	1.67	.87	.13	1.08	1.23	.12
16	16	2.54	.59	.21	.33	.54	.25
17	17	1.33	.72	.17	.51	.69	.23
19	18	.82	.76	.46	-.77	.60	.31
19	19	.84	.69	.36	.04	.70	.30
20	20	.38	.57	.25	.22	.91	.34

The correlation between b-values estimated on pretest data and b-values estimated on regular test data was r = .55 and ρ = .56. The plot of b-values is shown in Figure 3.

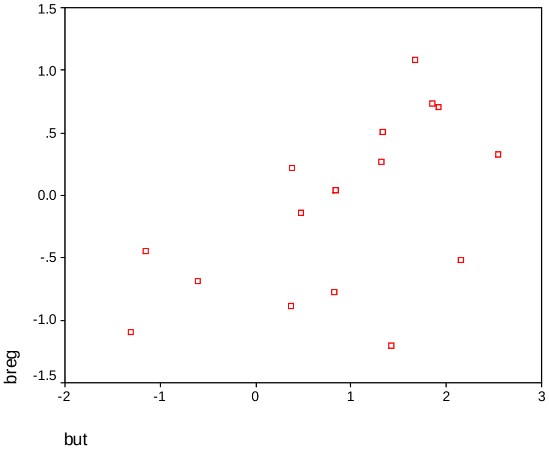

Figure 3. *Plot of b-values estimated on regular test data against b-values estiamted on pretest data.*

The correlation between a-values estimated on pretest data and on regular data was r = .54 and ρ = .51. The plot of a-values is shown in Figure 4.

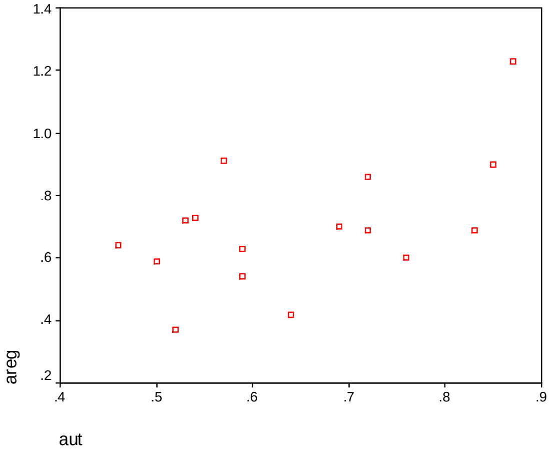

Figure 4. *Plot of a-values estimated on regular test data against a-values estimated on pretest data.*

The correlation between c-values estimated on pretest data and c-values estimated on regular data was r = .61 and ρ = .75.

Item Characteristic Curves of 16 items.

In Figures 5 to 20 the ICCs of each item are shown for the pretest and the regular test.

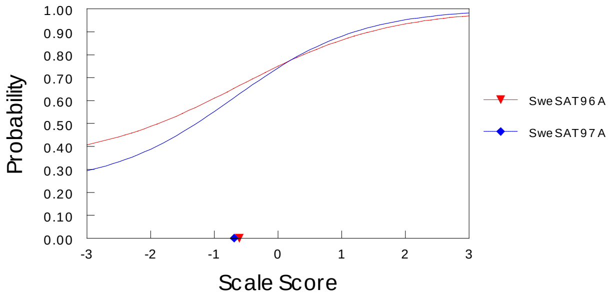

Figure 5. *ICCs of item No 5 in the regular test and No 5 in the pretest.*

For item No 5 the estimated b-value had decreased from -.61 to -.69 from pretest to regular test. The estimated a-value had increasd from .50 to .59. Hence the item was slightly easier and better discrimiating in the regular test.

For the same item the p-value was the same (.74) in the pretest and regular test while the r_{bis} had increased slightly from .30 to .32.

The results were almost the same from the two analyses: item No 5 had about the same difficulty and was slightly better discriminating in the regular test than in the pretest.

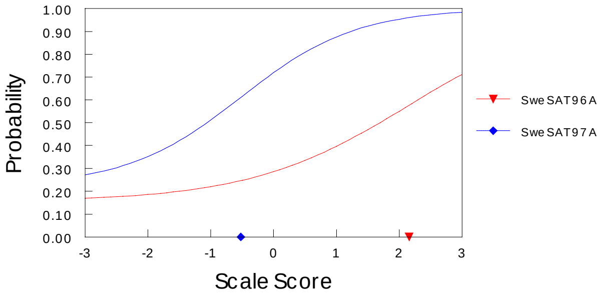

Figure 6. *ICCs of item No 6 in the regular test and No 7 in the pretest.*

For item No 6 the estimated b-value had decreased from 2.16 to -.52 and the estimated a-value had increased from .46 to .64.

For the same item the p-value had increased from .30 to .68 and the r_{bis} from .23 to .36.

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

This item had been extensively changed between pretest and regular test.

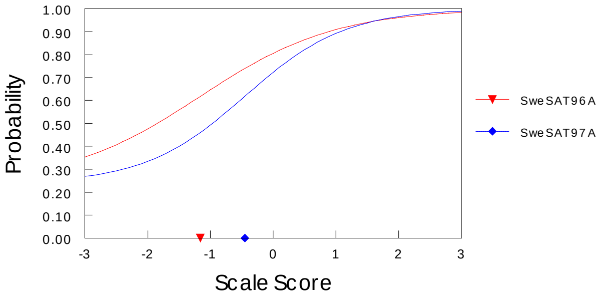

Figure 7. *ICCs of item No 7 in the regular test and No 6 in the pretest.*

For item No 7 the estimated b-value had increased from –1.16 to-.45 and the estimated a-value had increased from .54 to .73.

For the same item the p-value had decreaed from .78 to .71 ad the r_{bis} had changed from .37 to .38.

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

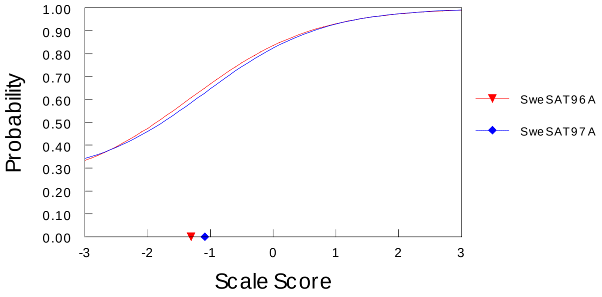

Figure 8. *ICCs of item No 8 in the regular test and No 8 in the pretest.*

For item No 8 the estimated b-value had increased from -1.31 to –1.09 and the estimated a-value had increased from .59 to .63 from pretest to regular test.

For the same item the p-value had increased from .80 to .81 and the r_{bis} had decreased from .40 to .35.

According to both analyses the item was imperceptibly easier in the regular test but according to IRT the item was imperceptibly more discriminating and according to CTT slightly less discriminating in the regular test than in the pretest.

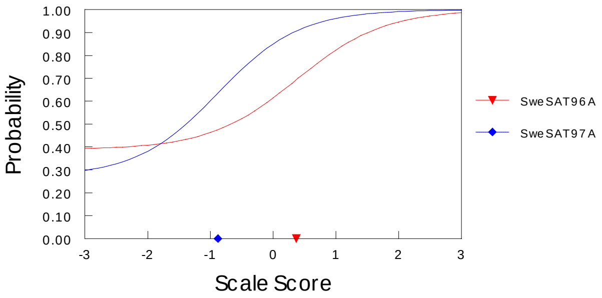

Figure 9. *ICCs of item No 9 in the regular test and No 17 in the pretest.*

For item No 9 the b-value had decreased from .37 to -.88 from the pretest to the regular test while the a-value had increased from .85 to .90.

For the same item the p-value had increased from .64 to .81 and the r_{bis} from .37 to .43.

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

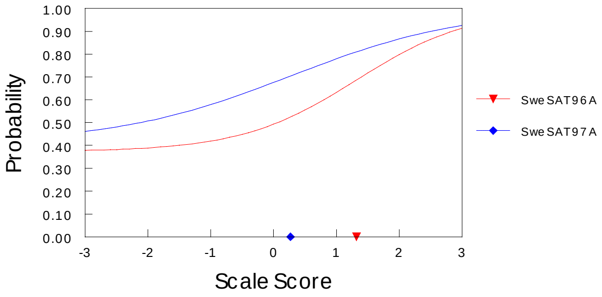

Figure 10. *ICCs of item No10 in the regular test and No 20 in the pretest.*

For item No 10 the b-value had decreased from 1.32 to .27 and the a-value had decreased from .64 to .42 from pretest to regular test.

For the same item the p-value had increased from .52 to .69 and the r_{bis} had decreased from .25 to .22.

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

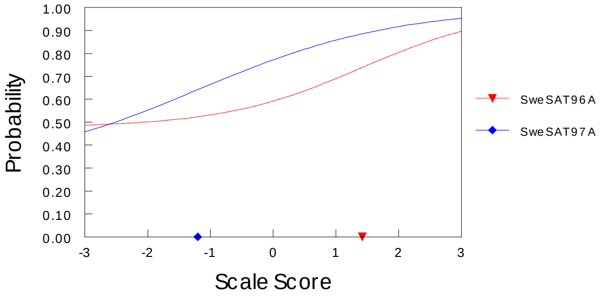

Figure 11. *ICCs of item No 11 in the regular test and No 17 in the pretest.*

For this item the b-value had dereased, from 1.42 to -1.20 and the a-value had decreased from .52 to .37 from pretest to regular test.

For the same item the p-value had increased from .61 to .75 and the r_{bis} had decreased from .18 to .11.

According to both analyses this item was much easier and slightly less discriminating in the regular test than in the pretest.

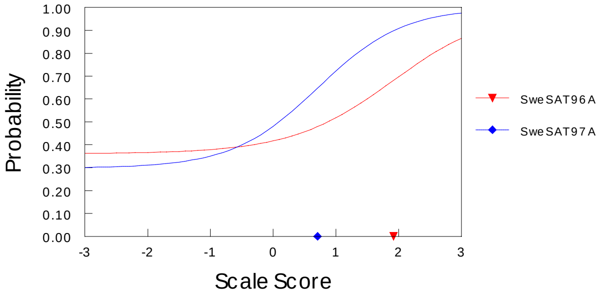

Figure 12. *ICCs of item No 12 in the regular test and No 20 in the pretest.*

For item No 12 the b-value had decreased from 1.92 to .71 while the a-value had increased from .72 to .86 from pretest to regular test.

For the same item the p-value had increased from .45 to .52 and the r_{bis} had increased from .17 to .33.

According to both analyses this item was easier in the regular test than in the pretest and also according to both analyses the discrimination was better in the regular test than in the pretest.

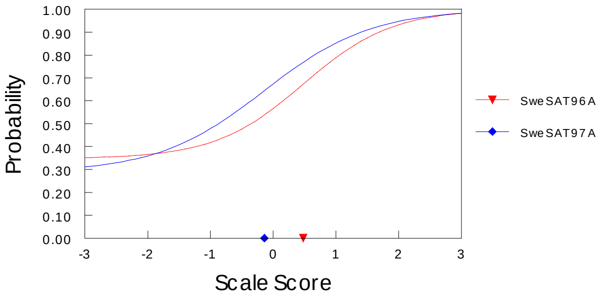

Figure 13. *ICCs of item No 13 in the regular test and No 14 in the pretest.*

For item No 13 the b-value had decreased from .48 to -.14 and the a-value had decreased from .83 to .69 from the pretest to the regular test.

For the same item the p-value had increased from .59 to .66 while the r_{bis} had increased slightly from .35 to .36 from pretest to regular test.

Hence according to both analyses the item was easier in the regular test than in the pretest, but according to IRT the discrimination was worse in the regular test and according to CTT it was about the same.

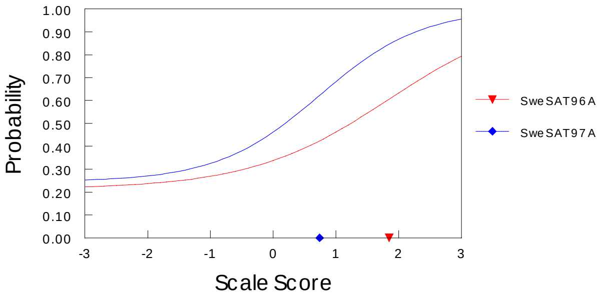

Figure 14. *ICCs of item No 14 in the regular test and No 15 in the pretest.*

For item No 14 the b-value had decreased from 1.85 to .74 and the a-value had increased from .53 to .72 from pretest to regular test.
For the same item the p-value had increased from .36 to .50 and the rbis from .24 to .30.

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

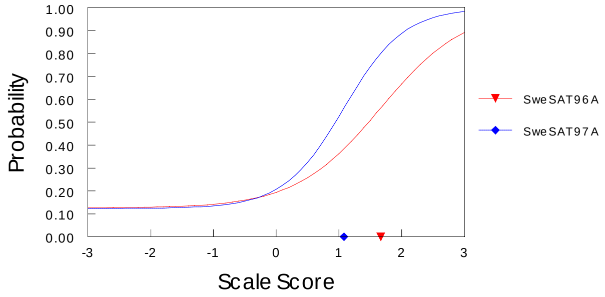

Figure 15. *ICCs of item No 15 in the regular test and No 14 in the pretest.*

For item No 15 the b-value had decreased from 1.67 to 1.08 from the pretest to the regular test and the a-value had increased from .87 to 1.23.

For the same item the p-value had increased from .24 to .30 and the r_{bis} had increased from .34 to .43 from pretest to regular test.

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

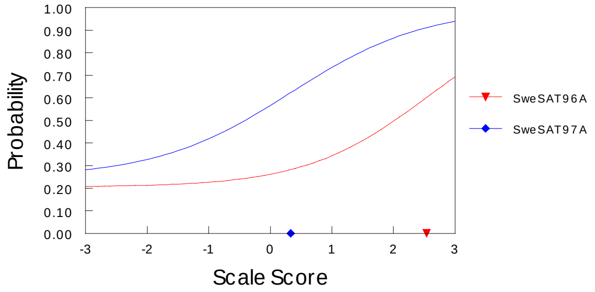

Figure 16. *ICCs of item No 16 in the regular test and No 16 in the pretest.*

For item No 16 the b-value had decreased from 2.54 to .33 and the a-value from .59 to .54 from pretest to regular test.

For the same item the p-value had increased from .28 to .57 and the r_{bis} from .17 to .28.from the pretest to the regular test.

According to both theories this item was easier in the regular test than in the pretest but according to IRT the discrimination was somewhat lower and according to CTT somewhat higher in the regular test than in the pretest.

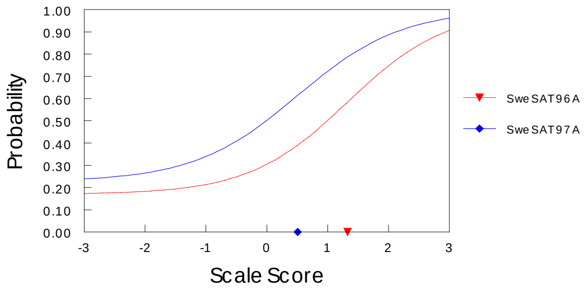

Figure 17. *ICCs of item No 17 in the regular test and No 17 in the pretest*.

For item No 17 the b-value had decreased from 1.33 to .51 and the a-value from .72 to .69 between pretest and regular test.

For the same item the p-value had increased from .35 to .53 and the r_{bis} from .35 to .36 from pretest to regular test..

According to IRT as well as CTT this item had become easier in the regular test than in the pretest. According to IRT the discrimination was imperceptibly lower, while according to CTT it was imperceptibly higher in the regular test than in the pretest.

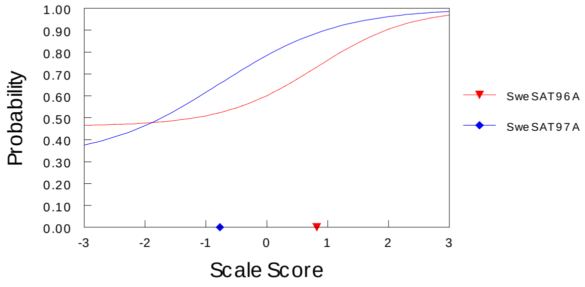

Figure 18. *ICCs of item No 18 in the regular test and No 19 in the pretest.*

For item No 18 the b-value had decreased from .82 to -.77 and the a-value from .76 to .60 from pretest to regular test.

For the same item the p-value had increased from .63 to .76 and the r_{bis} from .29 to .32 between pretest and regular test.

According to both analyses the item was easier in the regular test than in the pretest but according to IRT the discrimination power had decreased while it had increased according to CTT.

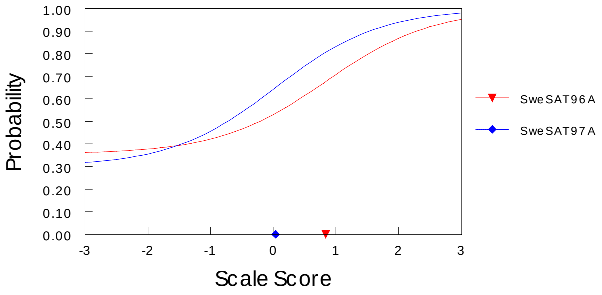

Figure 19. *ICCs of item No 19 in the regular test and No 19 in the pretest.*

For item No 19 the b-value had decreased from .84 in the pretest to .04 in the regular test while the a-value remained almost the same (.69/.70).

For the same item the p-value had increased from .56 to .65 and the r_{bis} from .29 to .31 from pretest to regular test.

Hence, according to both analyses, this item had become easier and somewhat better discriminating in the regular test than in the pretest.

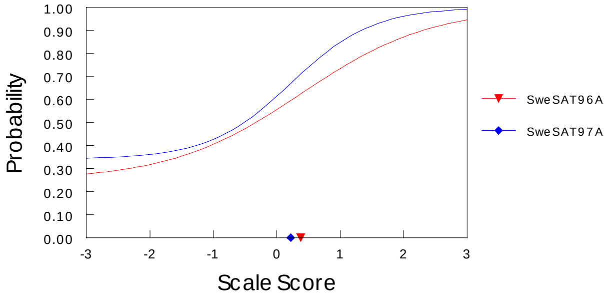

Figure 20. *ICCs of item No 20 in the regular test and No 20 in the pretest.*

For item No 20 the b-value had decreased from .38 to .22 and the a-value had increased from .57 to .91 from pretest to regular test.

For the same item the p-value had increased from .57 to .60, while the r_{bis} remained the same (.34) between pretest and regular test.

According to both analyses item No 20 was easier in the regular test, but according to IRT the discrimination power had increased and according to CTT it was the same between pretest and regular test.

Comparison between estimated IRT parameters and CTT statistics

The correlation between the estimated b-values and the p-values was $r = -.90$ and $\rho = -.88$ for the pretest and $r = -.92$ and $\rho = -.95$ for the regular test items. A plot of the difficulty statistics for the regular test items is shown in Figure 21.

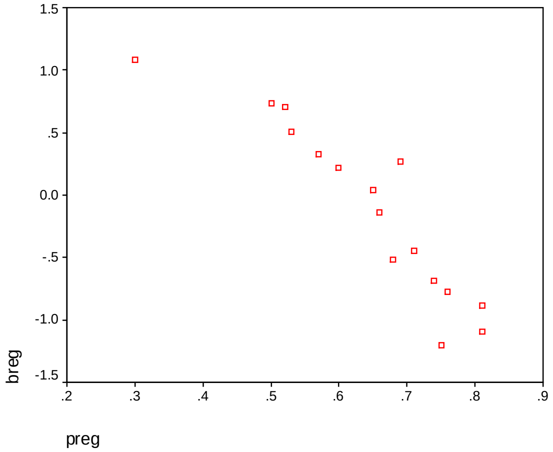

Figure 21. Estimated *b-values of 16 regular test items plotted against p-values of the same items*.

The most deviating items (see Figure 21) were No 6 and No 7.

The correlation between r_{bis} and a-values was $r = .35$ and $\rho = .32$ for the pretest items and $r = .78$ and $\rho = .68$ for the regular test items. The plot for regular test items is shown in Figure 22.

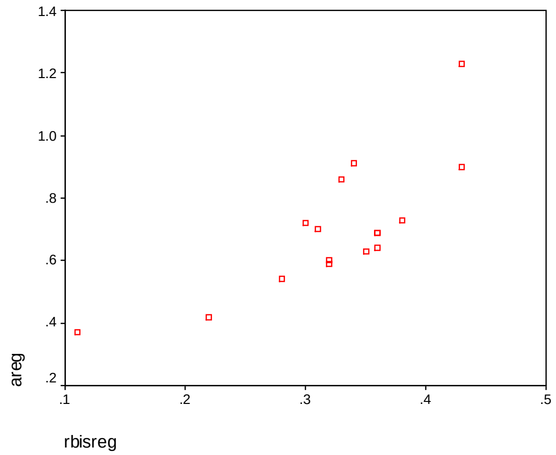

Figure 22. *Estimated a-values of 16 regular test items plotted against r_{bis} for the same items.*

The most deviating items were No 8, 11 and 16, for these items the standardized residuals were larger than one standard deviation.

Regarding item No 11 there seems to be a discrepancy as to difficulty as well as discrimination. As may be seen in Figure 11 item No 11 was a bad item according to IRT and the same is true according to CTT, since the r_{bis} of the item was as low as .11. This item should never have been included in the test.

The assessment of the IRT model data fit, also showed that for one item in the regular test, No 11, there was a model data misfit which was significant at $\alpha = .01$ level.

Discussion

The agreement between results from item-analyses performed within the two different frameworks IRT and CTT was very good. For 15 out of 16 items the result regarding difficulty changes of the items was the same and for 11 the changes of discrimination power was the same; for the remaining items as well the discrepancies were very small.

For CTT the correlation between pretest and regular item difficulties was r = .78 and ρ = .82; for IRT the correlation was r = .55 and ρ = .56. For the CTT item discrimination indices the correlation was r = .66 and ρ = . 69; for IRT the corresponding correlation was r = .54 and ρ =.51. Hence the pretest item statistics seem to be more consistent within the CTT framework. On the other hand quite a few items had been changed between pretest and regular test and the IRT indices might be more sensitive to such changes.

The correspondance between IRT and CTT regarding difficulty indices for the regular test items was very good, the correlation was r = -.92 and ρ = -.95.

The overall conclusion is that the predictions made from pretest data to regular test data are satisfactory, but that is true for CTT as well as for IRT. Since the groups on which pretesting had been performed were large and representative samples from the real examinee population this outcome may be expected. However, as expressed by Fan (1998):

Because IRT differs 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 (p.360).

What is usually mentioned as a main shortcoming of CTT is that the item statistics: p-values and r_{bis} , are sample dependent (see i.e. Hambleton et.al., 1991, p. 3), while this is not the case with IRT. *The invariance of item parameters across groups is one of the most important characteristics of item response theory* (Lord, 1980, p.35). For the authentic and very large examinee sample which has been used in this study, it is difficult to find greater invariance or any other obvious advantages in the IRT based item indices.

References

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 Mea surement.

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.

Stage, C. (1998a). *A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Subtest WORD*. (Educational Measurement No 29). 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 Subtest ERC*. (Educational Measurement No 30). Umeå University, Department of Educational Measurement.