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

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 item 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 poower 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. The statistics which are used from 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)

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 detailled 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 of IRT is that big sample sizes are 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 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 ERC subtest between item analysis based on CTT and item anlysis besed on IRT. In an earlier study (Stage, 1998) the same comparison was made for the WORD subtest and the conclusion from that study was that the results from the two analyses were very similar in spite of the differences between the theoretical frameworks.

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

In the SweSAT given in spring 1997 the subtest ERC contained 14 items which had been pretested on four different samples from the examinee population in spring 1996. The aim of this study is to compare, for these 14 items, the stability of item parameters estimated by IRT (BILOGW) with item statistics obtained by CTT.

In an earlier study (Stage, 1997b) of the applicability of IRT on the subtest ERC, the unidimensionality was assessed by factor analysis and the first three eigenvalues were 3.8, 1.1 and 1.0. An analysis of the standardized residuals between observed and model predicted performance gave as a result that 1.25 % of the standardized residuals had an absolute value higher than three, 5 % had an absolute value between two and three, 31.25 % between one and two and 62.5 % 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 seven items 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 14 ERC-items in the regular test spring 1997, which had been 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 four ERC pretest combinations spring 1996 were run in BILOGW together with the regular ERC subtest from spring 1996 and the a-, b- and c-parameters were estimated. The ERC subtest from spring 1997 was run in BILOGW and the item parameters were estimated. The parameter estimates for the corresponding 14 items were noted and compared.The ICCs for the corresponding items were also compared (Figure 5 to 18).

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. Such changes had been made on two of the items in the ERC subtest, namely items No 5 and No 9. 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 presentation in the pretest booklets may differ from the order in the regular test. Even though the ERC subtest is not speeded changes in the order of presentation may change the item in some unknown way.

Results

Classical test theory

In Table 1 the p-values and the r_{bis} obtained from the four pre-test versions and from the spring 1997 test are presented for the the 14 common items. The order of presentation in the pretest versions and the regular test version is also given.

Table 1. *CTT-based item indices: p-values and r_{bis}, and order of presentation for 14 items.*

Item No		Pretest		Regular test	
pre	**reg**	**p**	**r_{bis}**	**p**	**r_{bis}**
1	1	.33	.30	.38	.33
2	2	.72	.34	.66	.33
3	3	.35	.28	.29	.29
4	4	.41	.45	.47	.47
5	5	.62	.16	.73	.54
1	6	.77	.62	.78	.56
2	7	.54	.48	.50	.46
3	8	.53	.37	.53	.42
11	9	.41	.41	.60	.38
5	10	.67	.57	.58	.51
14	12	.60	.51	.62	.46
13	13	.68	.56	.62	.56
14	14	.65	.57	.65	.52
10	15	.77	.52	.74	.52

The mean of the p-values was .58 for the pretest as well as for the regular test, the standard deviations were .15 and .14, and the sums were 8.05 and 8.15 respectively. This may be interpreted as equivalent achievements of the two groups of examinees and/or as equivalence of difficulty level of the two tests.

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 = .86 and ρ = .87. A plot of the p-values is shown in Figure 1.

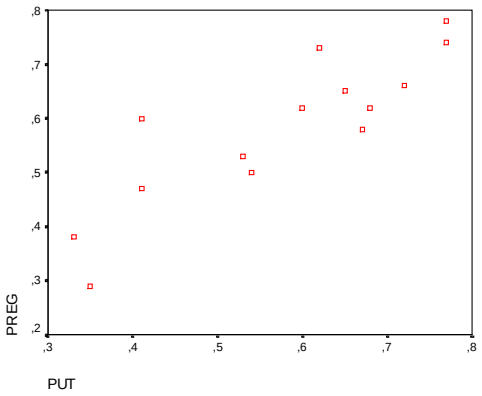


Figure 1. *Plot of p-values calculated on regular test data against p-values calculated on pretest data.*

The items deviating most from the linear regression line were No 5 and No 9 (above the line) and No 10 (below the line). When the items which had been changed between pretest and regular test i.e. No 5 and No 9 were removed, the correlation between p-values for the remaining 12 items was r = .95 and ρ = .94.

The correlation between r_{bis} of the items in the pretest versions and the corresponding items in the regular test version was r = .57 and ρ = .64; the plot of r_{bis} is shown in Figure 2.

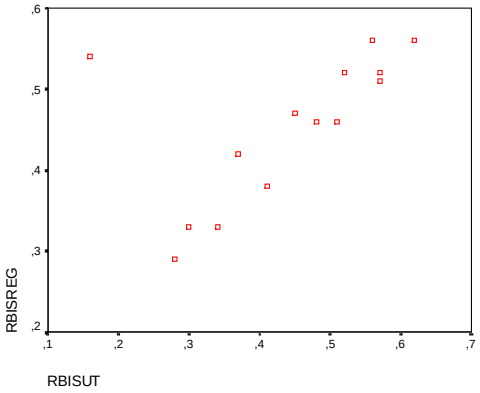


Figure 2. *Plot of r_{bis} calculated on regular test data against r_{bis} calculated on pretest data.*

The item which was most deviating from a linear regression line was No 5 (above the line). When the changed items, No 5 and No 9, were removed the correlation between the two values of r_{bis} increased to r = .96 and ρ = .94.

Item response theory

Table 2. *IRT-based item statistics*: *estimated b-, a- and c-parameters for 14 items, and order of presentation.*

Item No		**Pretest**			**Regular test**		
pre	**reg**	**b**	**a**	**c**	**b**	**a**	**c**
1	1	1.65	.59	.17	1.31	.79	.21
2	2	.20	.83	.50	.34	.76	.41
3	3	1.70	.72	.22	1.74	.71	.16
4	4	.83	.76	.15	.54	.83	.17
5	5	.78	.27	.34	.35	1.05	.31
14	6	-.90	1.00	.14	-.62	1.10	.30
2	7	.09	.67	.12	.49	.81	.20
3	8	.77	.90	.33	.63	.99	.30
4	9	.81	.56	.11	-.07	.55	.18
5	10	-.52	.84	.12	.19	1.02	.24
14	12	-.13	.71	.15	-.10	.73	.21
13	13	-.27	.96	.24	-.03	1.10	.22
14	14	-.40	.85	.12	-.19	.91	.21

10	15	-.85	.80	.24	-.52	.95	.28

The mean of b-values was .27 for the pretest items and .29 for the regular test items. This can be interpreted as equivalence in difficulty level of the two item sets.

The correlation between b-values estimated on pretest data and b-values estimated on regular test data was r = .88 and ρ = .83 The plot of b-values is shown in Figure 3.

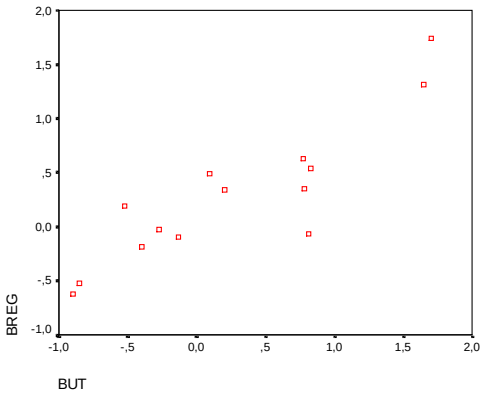


Figure 3. *Plot of b-values estimated on regular test data against b-values estimated on pretest data.*

The items which were deviating most from the linear regression line were No 9 (below the line) and No 10 (above the line). When the changed items, No 5 and No 9 were removed, the correlation between the two b-values increased to r = .96 and ρ = .94.

The correlation between a-values estimated on pretest data and on regular data was r = .34 and ρ = .58. The plot of a-values is shown in Figure 4.

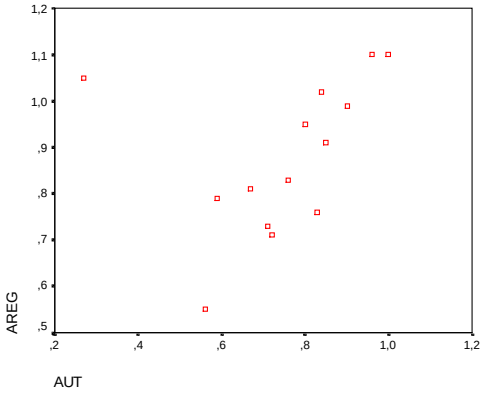


Figure 4. *Plot of a-values estimated on regular test data against a-values estimated on pretest data.*

The items which were deviating most from the regression line were No 5 (above the line) and No 9 (below the line). When these items (which were also the changed items) were removed the correlation increased to r = .82 and ρ = .80

The correlation between c-values estimated on pretest data and c-values estimated on regular test data was r = .80 and ρ = .58.

Item characteristic curves of 14 items.

In Figures 5 to 18 the ICCs of each item from the pretest as well as from the regular test are shown.

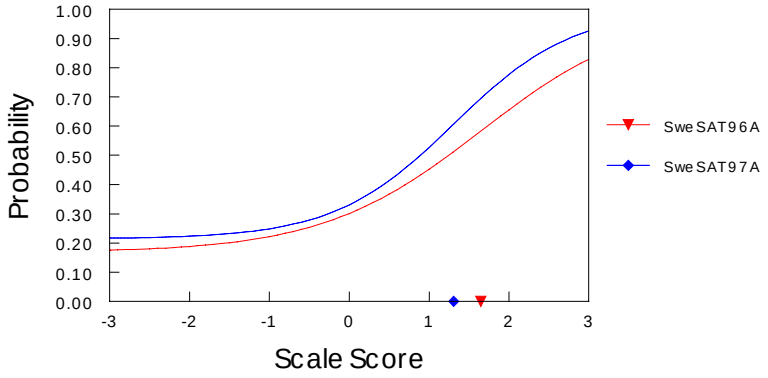


Figure 5. *ICCs of item No 1*.

For item No 1 the estimated b-value had decreased from 1.65 to 1.31 from pretest to regular test. The estimated a-value had increased from .59 to .79. Hence the item was slightly easier and better discriminating in the regular test.

For the same item the p-value had increased from .33 to .38 and the r_{bis} had increased from .30 to .33.

The results were the same from the two analyses: item No 1 was slightly easier and better discriminating in the regular test than in the pretest.

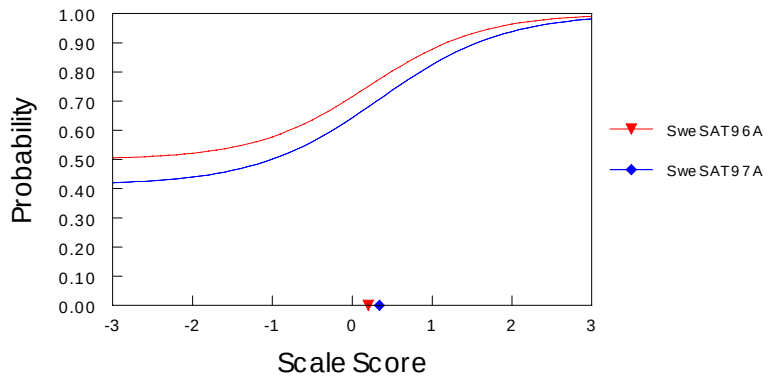


Figure 6. *ICCs of item No 2.*

For item No 2 the estimated b-value had increased from .20 to .34 while the estimated a-value had decreased from .83 to .76. For this item the estimated value of the pseudo guessing parameter was unusually high for the pretest as well as for the regular test.

For the same item the p-value had decreased from 72 to .66 and the r_{bis} from .34 to .33.

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

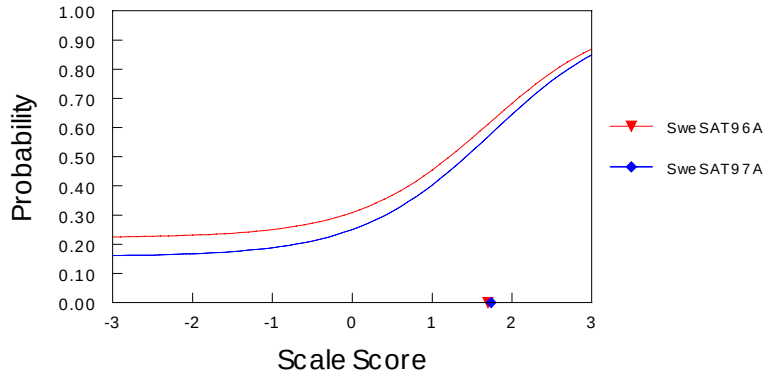


Figure 7. *ICCs of item No 3.*

For item No 3 the estimated b-value had increased from 1.70 to 1.74 and the estimated a-value had decreased from .72 to .71.

For the same item the p-value had decreased from .35 to .29 and the r_{bis} had changed from .28 to .29.

According to both analyses the item was unnoticeably more difficult but had about the same discrimination power in the regular test as in the pretest.

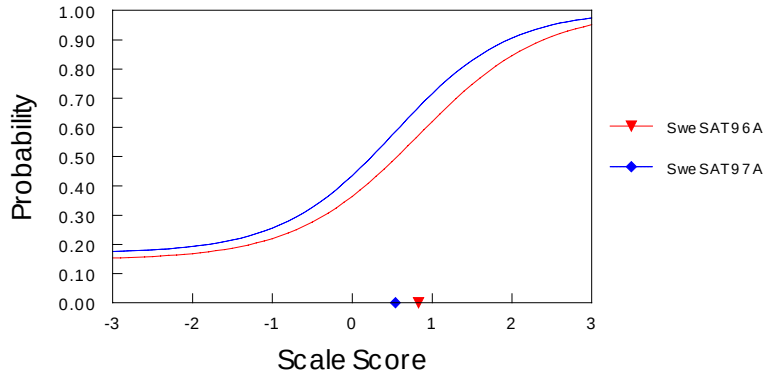


Figure 8. *ICCs of item No 4.*

For item No 4 the estimated b-value had decreased from .83 to .54 and the estimated a-value had increased from .76 to .83.

For the same item the p-value had increased from .41 to .47 and the r_{bis} from .45 to .47.

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

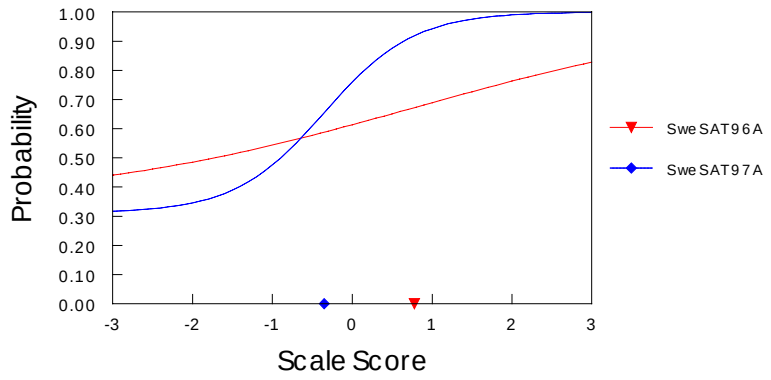


Figure 9. *ICCs of item No 5.*

For item No 5 the b-value had decreased from .78 to .35 from the pretest to the regular test while the a-value had increased from .27 to 1.05.

For the same item the p-value had increased from .62 to .73 and the r_{bis} from .16 to .54.

According to both analyses the item was easier and better discriminating in the regular test than in the pretest. This was also one of the items which had been changed between pretest and regular test.

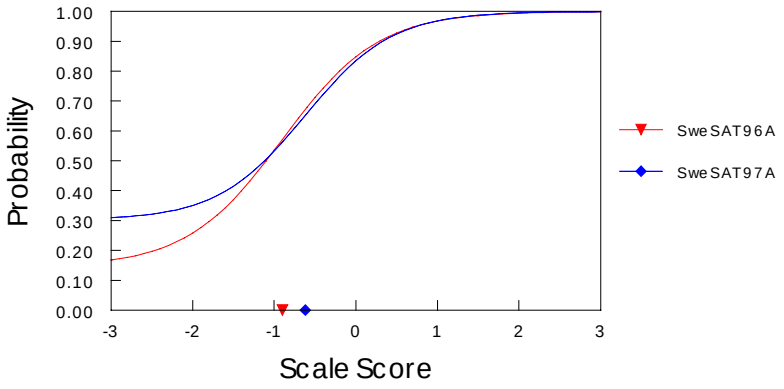


Figure 10. *ICCs of item No 6.*

For item No 6 the b-value had increased from -.90 to -.62 and the a-value had increased from 1.0 to 1.1 from pretest to regular test. As may be seen in Figure 10 for this item also the c-value had increased considerably.

For the same item the p-value was almost the same (.77 and .78 respectively) but the r_{bis} had decreased from .62 to .56.

For this item the two analyses differed regarding conclusions about the changes between pretest and regular test. According to IRT the item was slightly more difficult but had about the same discrimination power in the regular test as in the pretest, but according to CTT the difficulty was the same while the discrimination was poorer in the regular test than in the pretest.

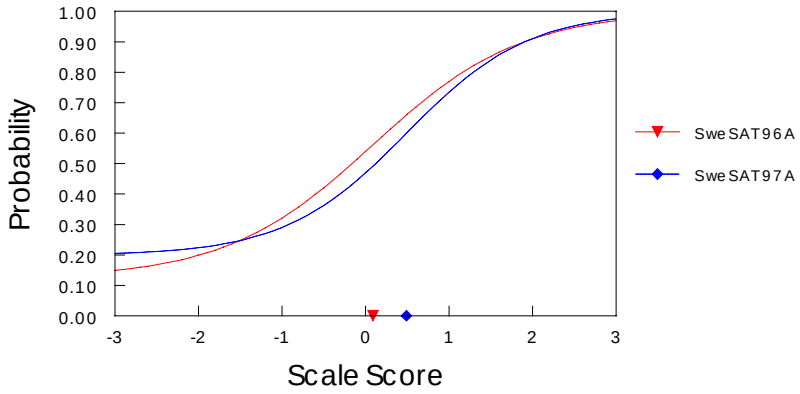


Figure 11. *ICCs of item No 7.*

For item No 7 the b-value had increased from .09 to .49 and the a-value had increased from .67 to .81 from pretest to regular test.

For the same item the p-value had decreased from .54 to .50 and the r_{bis} from .48 to .46.

According to both analyses the item was somewhat more difficult in the regular test than in the pretest but according to IRT it was also better discriminating in the regular test while accordning to CTT the discrimination was slightly better in the pretest.

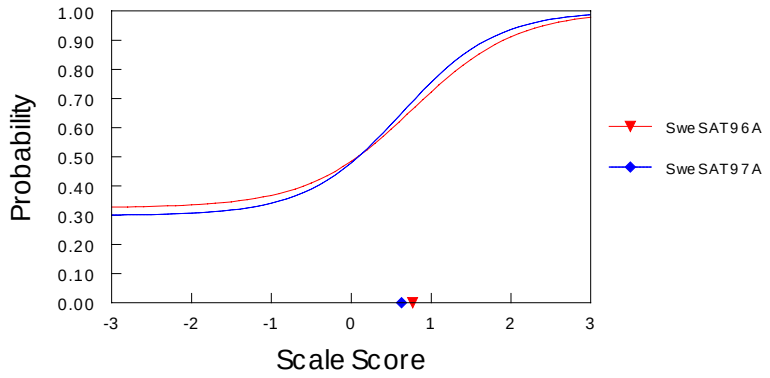


Figure 12. *ICCs of item No 8.*

For item No 8 the b-value had decreased from .77 to .63 while the a-value had increased from .90 to .99 from pretest to regular test.

For the same item the p-value was the same (.53) in the pretest as in the regular test while the r_{bis} had increased from .37 to .42.

According to IRT this item was slightly easier in the regular test than in the pretest while according to CTT the difficulty level was the same. According to both analyses the discrimination was somewhat better in the regular test than in the pretest.

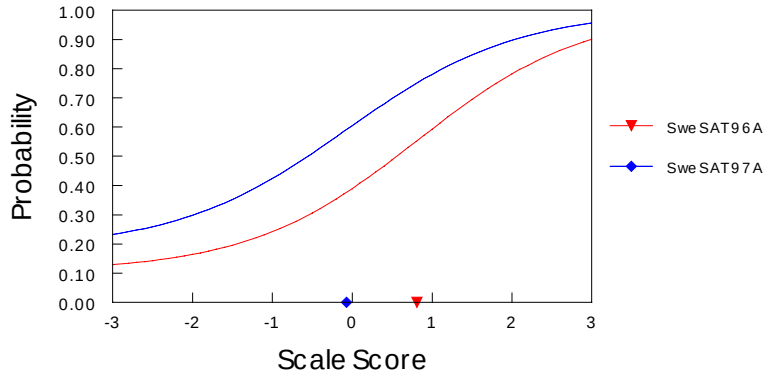


Figure 13. *ICCs of item No 9.*

For item No 9 the b-value had decreased from .81 to -.07 while the a-value was about the same (.56 and .55 respectively) in the pretest and the regular test.

For the same item the p-value had increased from .41 to .60 while the r_{bis} had decreased from .41 to .38 from pretest to regular test.

Hence according to both analyses the item was easier and slightly less discriminating in the regular test than in the pretest. This item had been changed between pretest and regular test and, not suprisingly, the item deviated considerably from the predictions regarding p-, b- and a-value.

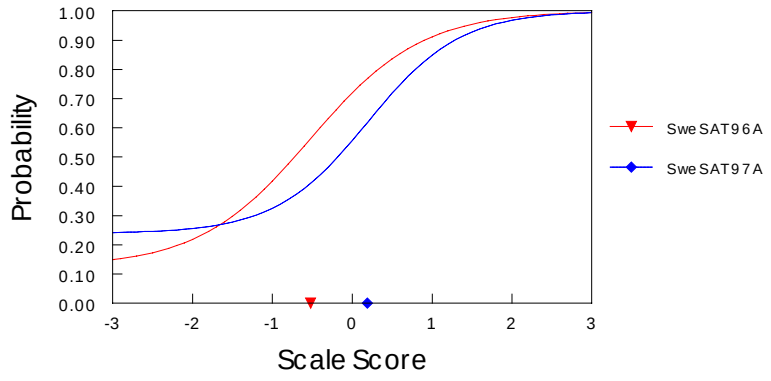


Figure 14. *ICCs of item No 10.*

For item No 10 the b-value had increased from -.52 to .19 and the a-value from .84 to 1.02 from pretest to regular test.

For the same item the p-value had decreased from .67 to .58 and the r_{bis} from .57 to .51.

According to both analyses this item was more difficult in the regular test than in the pretest but according to IRT it was also more discriminating while according to CTT is was less discriminating in the regular test than in the pretest.

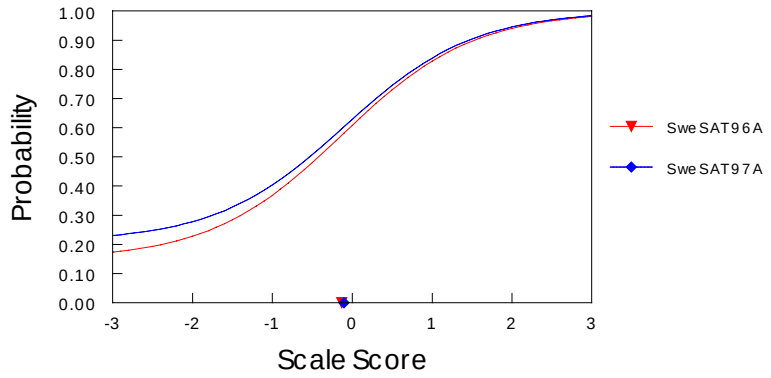


Figure 15. *ICCs of item No 12.*

For item No 12 the b-values were almost the same (-.13 and -.10 respectively) in the pretest as in the regular test and so were the a-values (.71 and .73 respectively).

For the same item the p-values too were very close (.60 and .62 respectively) while the r_{bis} had decreased from .51 to .46 from pretest to regular test.

According to both analyses the difficulty level of the item was about the same in the pretest and the regular test but according to IRT the discrimination power was slightly better in the regular test and according to CTT it was slightly poorer.

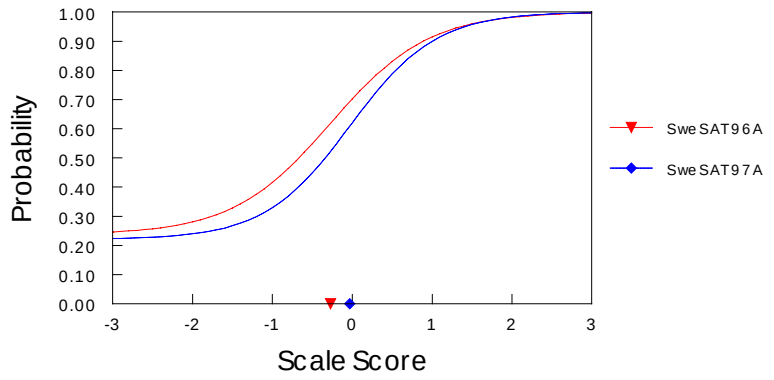


Figure 16. *ICCs of item No 13.*

For item No 13 the b-value had increased from -.27 to -.03 and the a-value from .96 to 1.10 from pretest to regular test.

For the same item the p-value had decreased from .68 to .62 while the r_{bis} was the same (.56) in the pretest and the regular test.

According to both theories this item was slightly more difficult in the regular test than in the pretest and the discrimination was slightly better according to IRT while it was exactly the same according to CTT.

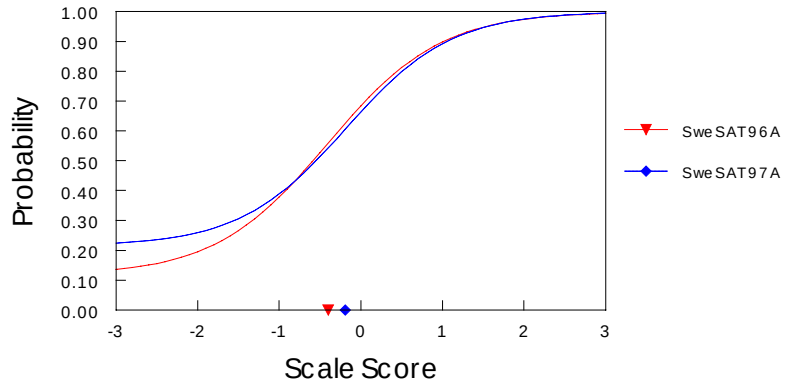


Figure 17. *ICCs of item No 14.*

For item No 14 the b-value had increased from -.40 to -.19 and the a-value from .85 to .91 between pretest and regular test.

For the same item the p-value was the same (.65) while the r_{bis} had decreased slightly (from .57 to .52).

According to IRT this item had become slightly more difficult and better discriminating in the regular test than in the pretest. According to CTT the difficulty level was the same while the discrimination was slighly poorer in the regular test than in the pretest.

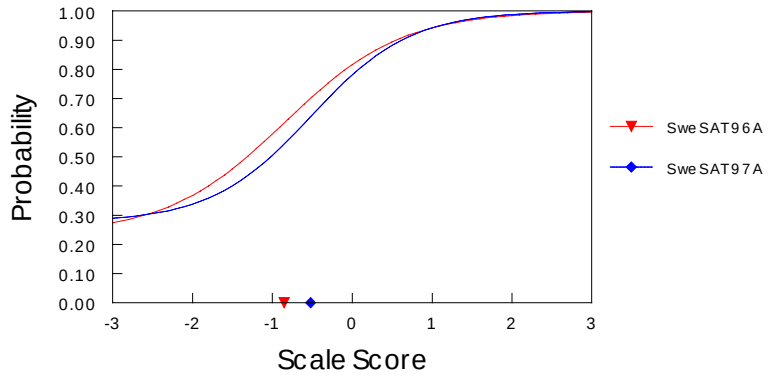


Figure 18. *ICCs of item No 15.*

For item No 15 the b-value had increased from -.85 to -.52 and the a-value from .80 to .95 from pretest to regular test.

For the same item the p-value had decreased from .77 to .74 while the r_{bis} was the same (.52) in the regular test and the pretest.

According to both analyses item No 15 was a little more difficult in the regular test than in the pretest but according to IRT the discrimination power had increased while it was the same according to CTT.

Discussion

The agreement between results from the item-analyses performed within the two different theoretical frameworks IRT and CTT was very good. For six items the decisions about changes between pretest and regular test were exactly the same, for five items the decisions about difficulty changes were the same, while the conclusions about discrimination differed slightly and for the remaining three items the deviations were very small as well.

The correlation between the IRT estimated b-values and the CTT calculated p-values was r = -.90 for the pretest items as well as for the regular test items (ρ = -.88 and -.82 respectively). A plot of the difficulty statistics from the two theories, for the regular test items, is shown in Figure 19.

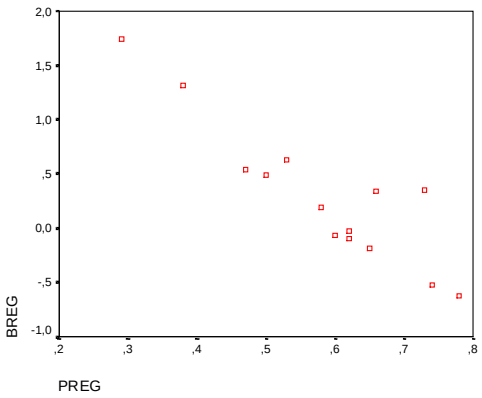


Figure 19. *Estimated b-values of 14 regular test items plotted against p-values of the same items.*

The two most deviating items (see Figure 19) were No 2 and No 5; for these two items the standardized residuals were larger than one standard deviation.The same items were the most deviating in the pretest data.

The correlation between r_{bis} and estimated a-values was r = .74 for pretest items and r = .76 for regular test items (ρ = .66 and .85 respectively). A plot of item discrimination statistics for the regular test items is shown in Figure 20.

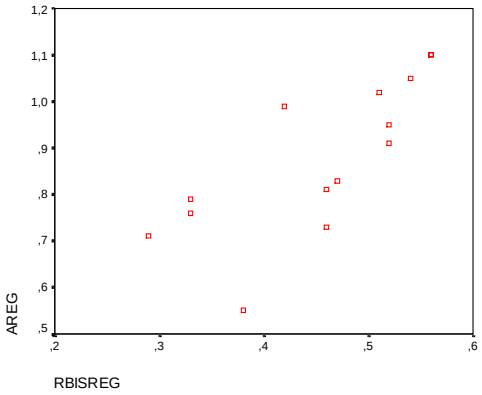
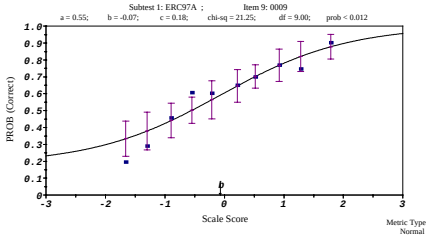


Figure 20. *Estimated a-values of 14 regular test items plotted against* r_{bis} *for the same items.*

The most deviating items were Nos 8, 9 and 12; for these three items the standardized residuals were larger than one standard deviation.

Regarding the agreement between pretest and regular test data there were no great differences between the two theories regarding difficulty. For CTT the correlation between pretest and regular test item difficulty was r = .87 (ρ = .86) for IRT it was r = .88 (ρ = .83). For the discrimination statistics the agreement was somewhat poorer; the correlation between CTT indices was r = .57 (ρ = .64) and for IRT the correlation was r = .34 (ρ = .58).

The assessment of model data fit for IRT showed that for two items in the regular test, Nos 9 and 14 there was a model data misfit which was significant at ∝ = .01 level. For the pretest items there was model data misfit for three items at ∝ = .01 level, these items became Nos 5, 12 and 14 in the regular test. In Figure 21 the model data fit for items No 9 and No 14 in the regular test is shown.



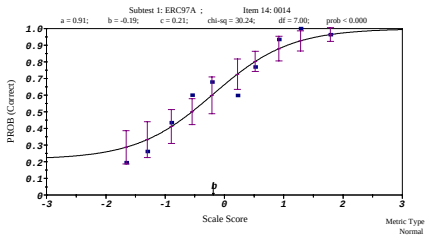


Figure 21. *The model data fit for items No 9 (left) and 14 (right).*

As may be seen in Figure 21 the model data misfit does not seem to be very serious. Item No 9 seems to be problematic, since it has been deviating in most analyses and certainly not behaved as expected. Item No 14 on the other hand has not turned out as problematic in the earlier analyses, even though the decisions about the item based on CTT and IRT diferred slighly.

The items for which the differences in ICCs were greatest between pretest and regular test were No 5 and No 9 (Figures 9 and 13). Since these two items had been changed between pretesting and regular test this is a very reasonable outcome. These items also turned out to be deviating in most regression analyses between pretest and regular test statistics (all, except for the b-values). Interesting, however, is the fact that while for item No 5 the change seemed to have improved the item, since the r_{bis} as well as the a-value of this item had increased considerably, the change in item No 9 only made the item easier, but not better functioning, since the r_{bis} as well as the a-value had decreased.

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

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. (p. 360)

What is usually mentioned 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 (see i.e. Hambleton & Swaminathan, 1985), while this is not the case for 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 examinee groups used in this study it is difficult to find any obvious advantage or greater invariance in the IRT based item statistics.

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. (1998). A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory. A Study of the SweSAT Subtest WORD. (Educational Measurement). Umeå University, Department of Educational Measurement