Predicting Gender Differences in WORD Items. A Comparison of item Response Theory and Classical Test Theory.

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 of items 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.

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 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)

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. al., 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 psuedo guessing parameter (for more detailed descriptions of IRT see i.e. Lord, 1980, Hambleton & Swaminathan, 1985, Hambleton et. al., 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, 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 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 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 earlier studies (Stage, 1998a, b) comparisons were made, for the WORD, READ and ERC subtests, between item analysis based on CTT and item anlysis based on IRT. The conclusion of those studies was that the results from the two analyses were very similar despite the differences between the theoretical frameworks. In those studies, however, only the whole group of examinees was studied. Since one important part of the item analysis performed for SweSAT is to make sure that the tests are not gender biased it is of great

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

interest to compare the outcome of analyses made for males and females separately.

Test fairness is an important issue in all testing programmes and for SweSAT there have been a lot of discussion of whether the test is gender biased. Items for which the pretest results show great differences between males and females are usually abandoned (or, sometimes, counterbalanced) to make sure that the test is not biased. Delta-plot-, χ^2 –, and Mantel Haenszel-analyses have been performed (Stage, 1985, 1972, 1997) but routinely only the p-values of males and females from the pretesting are compared.

To investigate bias empirical evidence is needed concerning the performance on test items of different groups of interest. Empirical evidence of differential performance is a necessary, but not sufficient, indication of item bias. The mean difference concept is the most uniformly rejected of all criteria of test bias by psychometricians (Reynolds, 1982). To distinguish the empirical evidence from the actual conclusion, the term differential item functioning (DIF) is used.

According to Hambleton et al. (1991) a definition of DIF which is generally accepted by psychometricians is:

"An item shows DIF if individuals having the same ability, but from different groups, do not have the same probability of getting the item right." (p. 110)

The importance of the item bias issue has led to considerable attention being devoted to the development and evaluation of methods for detecting DIF. The Mantel-Haenszel (MH) method has emerged as one of the most popular procedures. The MH method compares the probabilities of a correct response for groups of examinees of the same ability. Details of the MH method may be found in papers by Holland & Thayer (1986, 1988).

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 gender differences estimated by IRT (BILOGW) with gender differences obtained by CTT. For comparison a MH analysis was also performed on the same items.

In an earlier study (Stage, 1997b) of the applicability of IRT on the subtest WORD, 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 the 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 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 for the two groups, males and females. Specifically the study addresses the following questions:

1. How do item difficulty indices for males and females from CTT compare with item difficulty parameters for males and females estimated by IRT?

a) for pretest data?

b) for regular test data?

2. How do item discrimination indices for males and females from CTT compare with item discrimination parameters for males and females estimated by IRT?

a) for pretest data?

b) for regular test data?

3. How stable are the CTT based gender differences from pretest data to regular test data?

4. How stable are the IRT based gender differences from pretest data to regular test data?

5. How do the results compare with the outcome of MH analysis?

Method

Classical test theory

For the 20 WORD-items used in the regular test in spring 1997, which were pretested in spring 1996, the p-values and the biserial correlations (r_{bis}) were calculated for males and females separately. The same indices were calculated for the corresponding items in the pretest data and the values were compared.

Item response theory

The five WORD pretest combinations in spring 1996 were run in BILOGW together with the regular WORD subtest from spring 1996, and the a-, b- and c-parameters were estimated for males and females separately. The WORD subtest from spring 1997 was run in BILOGW and the same 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.

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. Such 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 WORD subtest is not speeded, changes in the order of presentation may still change the item in some way. All changes made on the items are presented in connection with the ICCs (Figures 1 to 20). The order of presentation in the pretest and the regular test is also given in Tables 1 and 2.

Mantel-Haenszel analysis

In order to examine whether there was DIF in any of the 20 examined items a MH analysis was performed in which the subtest score of the 40 WORD items in the regular test from spring 1997 was used to match males and females.

Statistical tests are too sensitive to sample size, a well known fact which Hays (1969) has expressed as follows:

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

At Educational Testing Service in Princeton where the MH method is used as a standard procedure, the items are classified into three categories: A – negligible DIF, B – intermediate DIF, and C - large DIF. The category into which an item is placed depends on two factors: the absolute value of the difference (transformed to the Δ-scale) and whether the difference is statistically significant or not. An item is placed in category **A:** if either MH Δ-diff is not significantly different from zero or the absolute value is less than one unit. **C:** if MH Δ-diff exceeds 1.5 in absolute value and is statistically significant, and in **B:** for all cases in between. (see Dorans & Holland, 1993).

These categories are used in the examination of the 20 WORD items in this study.

Results

Classical test theory

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

Table 1. *CTT-based item indices, p-values and biserial correlations, for males and females.*

Item No			**pretest**					**regular test**			
Pre	**reg**	**male**		**female**			**male**		**female**		
		p	**r_{bis}**	**p**	**r_{bis}**	**p-diff**	**p**	**r_{bis}**	**p**	**r_{bis}**	**p-diff**
8	1	.76	.61	.71	.60	+.05	.72	.54	.68	.55	+.04
20	4	.72	.50	.84	.49	-.12	..65	.37	.76	.45	-.11
39	5	.76	.22	.82	.23	-.06	.67	.35	.76	.28	-.09
18	9	.74	.50	.58	.44	+.16	.78	.43	.66	.42	+.12
36	10	.67	.50	.82	.60	-.15	.61	.48	.81	.60	-.20
27	11	.74	.34	.84	.40	-.10	.74	.36	.87	.39	-.13
36	15	.73	.40	.73	.37	00	.54	.37	.61	.38	-.07
14	16	.64	.41	.65	.46	-.01	.70	.42	.70	.53	00
5	19	.48	.37	.44	.45	+.04	.45	.33	.40	.40	+.05
16	23	.66	.39	.64	.32	+.02	.63	.36	.62	.40	+.01
38	24	.56	.46	.58	.48	-.02	.51	.27	.59	.29	-.08
12	25	.51	.52	.51	.63	00	.59	.55	.59	.60	00
24	27	.70	.35	.68	.37	+.02	.64	.31	.66	.35	-.02
4	28	.60	.63	.48	.51	+.12	.52	.55	.36	.49	+.16
4	29	.39	.35	.44	.33	-.05	.38	.33	.46	.32	-.08
5	35	.33	.35	.30	.29	+.03	.40	.49	.35	.43	+.05
37	36	.77	.47	.66	.40	+.11	.71	.43	.53	.43	+.18
6	38	.35	.29	.45	.35	-.10	.41	.30	.51	.38	-.10
28	39	.28	.22	.27	.32	+.01	.41	.31	.41	.32	00
39	40	.33	.26	.29	.21	+.04	.34	.22	.30	.24	+.04
	Σ	11.72		11.73		-.01	11.40		11.63		-.23

The correlations between p-values of the items in the pretest versions and p-values of the corresponding items in the regular test version were r = .91 and ρ = .89 for males and r = .92 and ρ = .90 for females. The corresponding correlations for r_{bis} were r = .83 and ρ = .75 for males and r = .74 and ρ = .74 for females.

The correlations between pretest p-values for males and females were r = 90 and ρ = .85 and for pretest r_{bis} r = .84 and ρ = .86. The correlations between regular test p-values for males and females were r = .80 and ρ = .79 and for the r_{bis} r = .86 and ρ = .87.

The correlation between DIF on pretest and on regular test was r = .95 and ρ = .96.

Item Response Theory

Table 2. *IRT-based item statistics for males and females.*

Item	No		pre	test				reg	ular	test	
pre	reg	ma	le	fem	ale		ma	le	fem	ale	
		b	a	b	a	bdiff	b	a	b	a	bdiff
8	1	-.58	1.25	-.33	1.28	.25	-.44	1.05	-.33	1.06	.11
20	4	-.66	.80	-1.30	.80	-.64	-.28	.55	-.92	.72	-.64
39	5	-1.70	.30	-2.57	.30	-.87	-.59	.49	-1.54	.36	-.95
18	9	-.50	.89	.19	.72	.69	-.93	.66	-.43	.59	.50
36	10	-.08	1.03	-1.17	.98	-1.09	-.15	.81	-1.10	1.04	-.94
27	11	-.92	.51	-1.61	.59	-.69	-1.07	.47	-2.18	.54	-1.11
36	15	-.41	.67	-.71	.52	-.30	.41	.62	-.03	.54	-.44
14	16	-.14	.66	.01	.96	.15	-.78	.57	-.45	.95	.33
5	19	.54	.54	.54	.61	00	.73	.45	.73	.54	00
16	23	-.22	.60	.13	.52	.35	.13	.63	-.22	.58	-.35
38	24	.22	.81	00	.75	-.22	.78	.42	.22	.46	-.56
12	25	.39	1.11	.12	1.06	-.27	-.05	1.05	00	1.30	.05
24	27	.32	1.18	.04	.75	-.28	.70	.98	.37	.91	-.33
4	28	.01	1.57	.65	1.74	.64	.17	.95	.84	1.14	.67
4	29	1.16	.73	1.09	.66	-.07	1.21	.55	1.08	.74	-.13
5	35	1.36	.56	1.94	.45	.58	.70	.87	1.13	.83	.43
37	36	-.77	.75	-.13	.62	-.64	-.60	.65	.33	.68	.93
6	38	1.49	.79	.98	.67	-.52	1.09	.39	.36	.53	-.73
28	39	1.66	1.12	1.57	1.12	-.09	1.24	1.03	1.26	1.48	.02
39	40	1.90	.50	2.58	.45	.68	2.26	.44	2.36	.41	.10
	Σ	3.38		2.02		-1.36	4.54		1.49		-3.05

The correlations between pretest and regular test b-values were r = .88 and ρ = .89 for males, and r = .92 and ρ = .88 for females. The correlations between pretest and regular test a-values were r = .76 and ρ = .66 for males, and r = .77 and ρ = .78 for females.

The correlations between male and female pretest b-values were r = .90 and ρ = .85 and, for pretest a-values r = .90 and ρ = .88. The cor-

relations between male and female regular test b-values were r = .84 and ρ = .83 and, for regular test a-values r = .87 and ρ = .86.

The correlation between pretest and regular test DIF was r = .89 and ρ = .89.

Mantel-Haenszel analysis

Out of the 20 WORD items 16 were flagged as significantly DIF at the .05 level and, out of these, 15 were also significant at the .01 level. Only 4 items were not DIF. 7 of the DIF items belonged to category A, 4 to category B, and 5 to category C.

Item Characteristic Curves of 20 items

In Figures 1 to 20 the ICCs for males and females are shown for the pretest items and for the regular items.

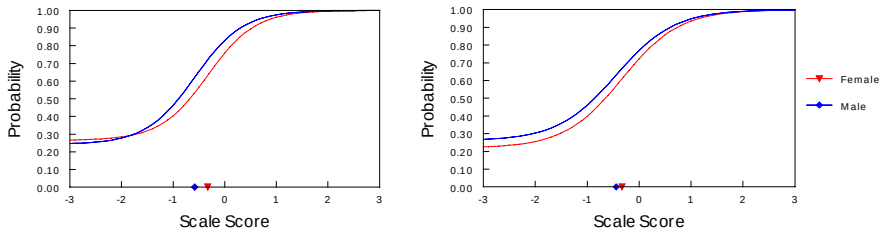

Figure 1. *ICCs for males and females on pretest item No 8 (left) and regular test item No 1 (right).*

In this item one distractor had been changed between pretest and regular test. As may be seen in Figure 1 neither the ICCs nor the gender differences had been much affected.

According to CTT the difficulties for males were p = .76/.71 and for females p = .72/.68; the discrimination indices were, for males, r_{bis} = .61/.54 and, for females, r_{bis} = .60/.55. The very small DIF in favour of males had remained in the regular test.

According to the MH-analysis, item No 1 was an A-item, favouring males (Δ-diff = -.70).

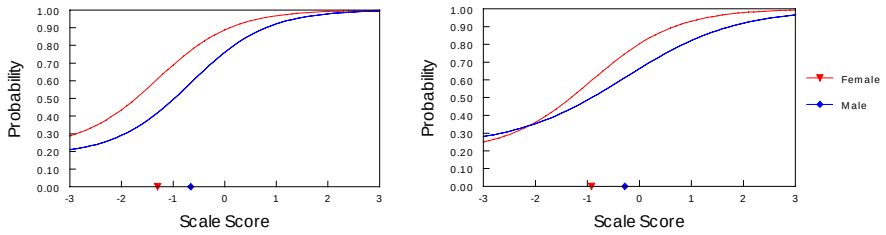

Figure 2. *ICCs for males and females on pretest item No 20 (left) and regular test item No 4 (right).*

On this item two distractors had been changed between pretest and regular test. The item seems to have become somewhat more difficult for both males and females in the regular test and the discrimination has become poorer, but more so for males; the size of the DIF in favour of females was the same but seems to have transferred somewhat towards higher ability levels in the regular test.

CTT also indicates that the item had become more difficult (p_M = .72/.65 and p_F = .84/.76) in the regular test than in the pretest, and also that the discrimination power had become poorer; even more so for males (r_{bis} = .50/.37) than for females (r_{bis} = .49/.45). The DIF in favour of females had remained in the regular test.

According to the MH-analysis, item No 4 was a B-item, favouring females (Δ-diff = 1.35)

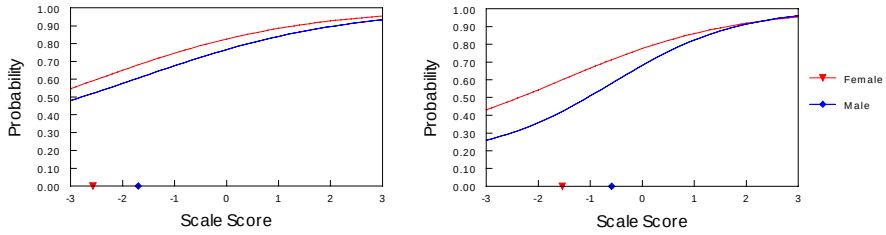

Figure 3. *ICCs for males and females on pretest item No 39 (left) and regular test item No 5 (right).*

On this item two distractors as well as the correct answer had been changed between pretest and regular test and the position of the item in the test booklets also differs much between pretest and regular test. For males the item had become more difficult and also better discriminating; for females the item had become more difficult but the dis-

crimination had remained poor. The small DIF in favour of females on the pretest had increased in the regular test especially at the lower ability levels.

According to CTT this item had become more difficult in the regular test for both males (p_M = .76/.67) and females (p_F = .82/.76) and the discrimination power had improved (r_{bisM} = .22/.35, r_{bisF} = .23/.28). The DIF in favour of females had increased slightly in the regular test.

According to the MH-analysis, item No 5 was a B-item, favouring females (Δ-diff = 1.15).

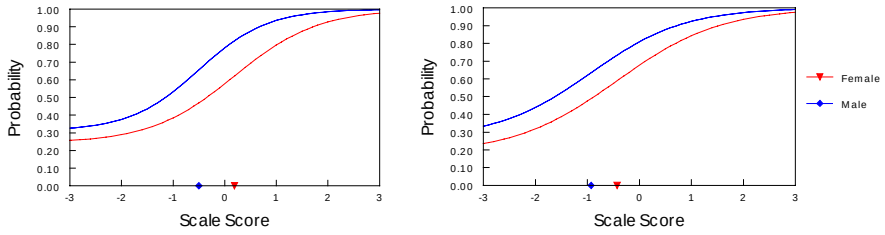

Figure 4. *ICCs for males and females on pretest item No 18 (left) and regular test item No 9 (right).*

On this item no changes had been made between pretest and regular test, but still the difficulty had decreased both for males (b_M = -.50/-.93) and for females (b_F = .19/-.43). The discrimination seems to have become poorer for both males and females and the DIF in favour of males had decreased to a small extent.

CTT also indicates that this item had become easier in the regular test for both males (p_M = .74/.78) and females (p_F = .58/.66), but more so for females. The discrimination had decreased in the regular test, but more so for males (r_{bisM} = .50/.43) than for females (r_{bisF} = .44/.42). The DIF in favour of males had decreased somewhat in the regular test.

According to the MH-analysis, item No 9 was a C-item favouring males (Δ-diff = -1.87).

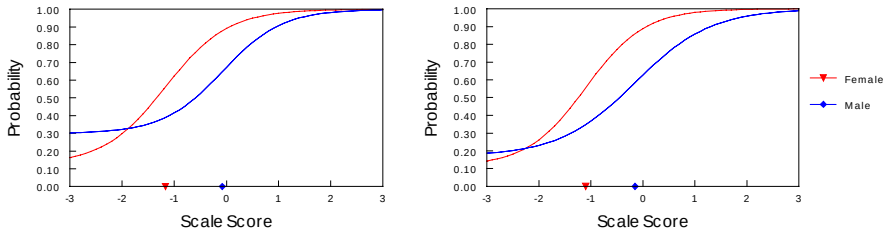

Figure 5. *ICCs for males and females on pretest item No 36 (left) and regular test item No 10 (right).*

On this item one distractor had been changed, and the difficulty level as well as the discrimination power were very similar in the regular test and in the pretest for both males and females. The DIF in favour of females also remained about the same in the regular test. For item No 10, however, there was a slight model data misfit for females.

According to CTT the item had become more difficult in the regular test for males (p_M = .67/.61) while it remained almost the same for females (p_F = .82/.81). The discrimination power remained almost the same for males (r_{bisM} =.50/.48) and exactly the same for females (r_{bisF} = .60). The DIF in favour of females had increased in the regular test.

According to the MH-analysis, item No 10 was a C-item, favouring females (Δ-diff = 2.12).

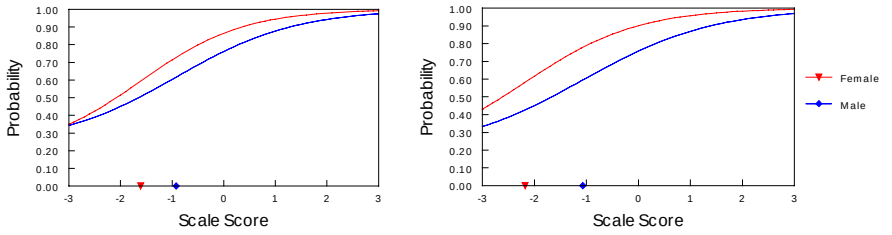

Figure 6. *ICCs for males and females on pretest item No 27 (left) and regular test item No 11 (right).*

On this item, as well, one distractor had been changed between pretest and regular test. The item had become somewhat easier in the regular test, and a little bit more so for females, so the DIF in favour of fema-

les had increased in the regular test. The discrimination had become slightly poorer in the regular test.

According to CTT this item had the same difficulty level in the pretest and the regular test for males (p_M =.74) while it had become somewhat easier in the regular test for females (p_F = .84/.87), hence the DIF in favour of females had increased. The discrimination had hardly changed.

According to the MH-analysis, item No 11 was a C-item favouring females (Δ-diff = 1.78).

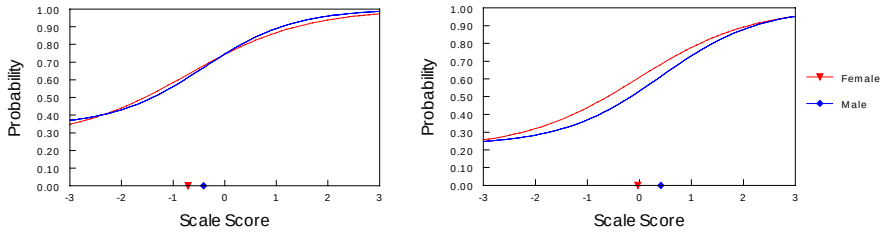

Figure 7. *ICCs for males and females on pretest item No 36 (left) and regular test item No 15 (right*).

On this item one distractor had been changed between pretest and regular test. The item had become more difficult in the regular test for both males and females, but a little bit more so for males. The discrimination had hardly changed at all. The small DIF in favour of females had increased slightly.

According to CTT, as well, this item had become more difficult in the regular test, and more so for males (p_M = .73/.54) than for females (p_F = .73/.61). The discrimination power was about the same in pretest and regular test for both males and females. While there was no DIF in the pretest there was a small DIF in favour of females in the regular test.

According to the MH-analysis, item No 15 was an A-item, favouring females (Δ-diff = .43).

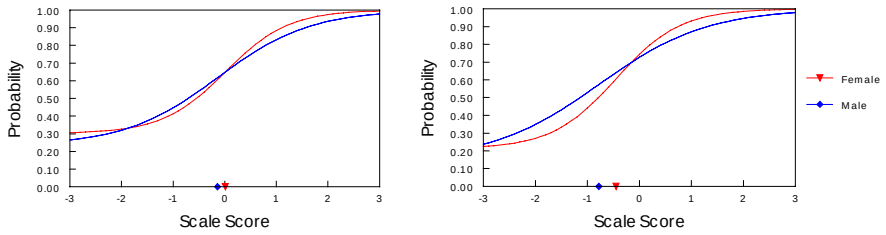

Figure 8. *ICCs for males and females on pretest item No 14 (left) and regular test item No 16 (right).*

On this item two distractors had been changed between pretest and regular test. The item had become easier in the regular test for both males and females but more so for males. The discrimination power had decreased for males but remained the same for females. There was a small DIF in favour of males in the pretest which increased in the regular test.

The p-values had increased for both males and females by about the same amount (p_M = .64/.70 and p_F =.65/.70) while the r_{bis} for males was about the same (r_{bisM} = .41/.42) but had increased for females (r_{bisF} = .46/.53). There was no DIF, neither on the pretest nor on the regular test.

According to the MH-analysis, there was no DIF in item No 16.

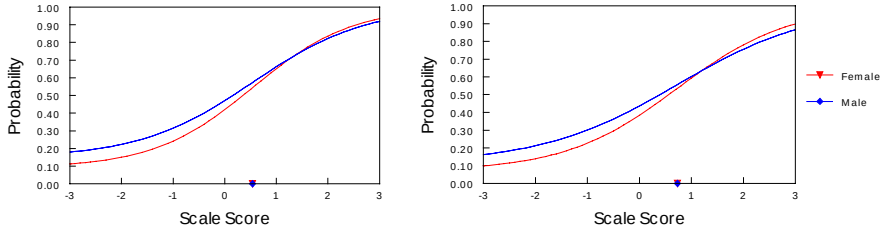

Figure 9. *ICCs for males and females on pretest item No 5 (left) and regular test item No 19 (right).*

On this item one distractor had been changed. The difficulty level was exactly the same for males and females, on the pretest as well as on the regular test, even though the item was somewhat more difficult in the regular test. There was a small gender difference regarding discrimination, however, and the item functioned somewhat better for fe-

males; the discrimination for both genders was somewhat poorer in the regular test than in the pretest. There was also a small model data misfit for both males and females on item No 19.

According to CTT there was a very small difficulty difference in favour of males on the pretest as well as on the regular test (p_M = .48/.45 and p_F = .44/.40). The discrimination was slightly better for females than for males according to CTT as well, but it was poorer in the regular test than in the pretest for both genders.

According to the MH-analysis, item No 19 was an A-item, favouring males (Δ-diff = -.55).

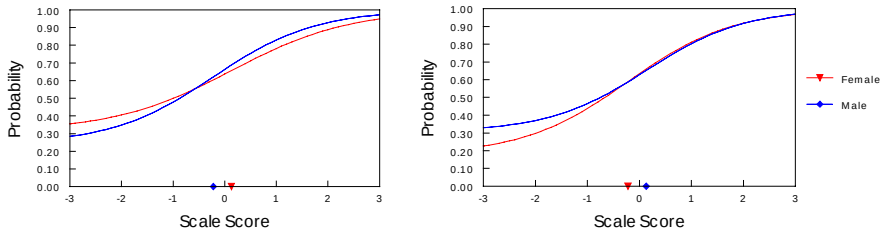

Figure 10. *ICCs for males and females on pretest item No 16 (left) and regular test item No 23 (right).*

On this item no changes had been made betwen pretest and regular test. The item, however, was more difficult in the regular test than in the pretest for males, while the opposite was true for females. The discrimination had increased to a small extent in the regular test for both males and females. On the pretest there was a small DIF in favour of males, which in the regular test had changed to a small DIF in favour of females.

According to CTT the item was imperceptibly more difficult for females in the pretest as well as in the regular test (p_M = .66/.63 and p_F = .64/.62). The discrimination power, however, had decreased slightly for males while it had increased for females between pretest and regular test.

According to the MH-analysis, there was no DIF in item No 23.

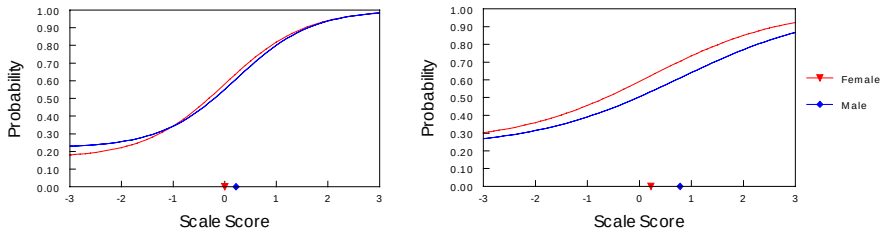

Figure 11. *ICCs for males and females on pretest item No 38 (left) and regular test item No 24 (right).*

On this item no changes had been made between pretest and regular test; despite this the difficulty level had increased for both males and females, but more so for males. The discrimination power had decreased for both males and females between pretest and regular test. The small DIF in favour of females had increased in the regular test.

According to CTT the item had become more difficult for males, but not for females, between pretest and regular test (p_M = .56/.51 and p_F = .58/.59). The discrimination power had decreased for both males and females between pretest and regular test. The small DIF in favour of males on the pretest had disappeared in the regular test.

According to the MH-analysis, item No 24 was a B-item, favouring females (Δ-diff = .91).

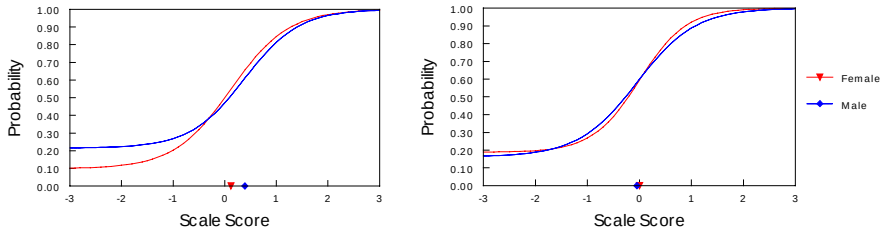

Figure 12. *ICCs for males and females on pretest item No 12 (left) and regular item No 25 (right).*

On this item no changes had been made between pretest and regular test. On the pretest there was a small DIF in favour of females which changed to an imperceptible DIF in favour of males in the regular test, where the item was easier for both males and females. The discrimi-

nation power was very similar for males and females and for pretest and regular test.

According to CTT, the difficulty level was exactly the same for males and females in the pretest (p = .51) but the item became easier in the regular test and slightly more so for females (p_F = .66 and p_M = .64). The discrimination which was better for females in the pretest (r_{bisF} = .63 and r_{bisM} = .52) decreased substantially for both males and females in the regular test (r_{bisM} = .31 and r_{bisF} = .35).

According to the MH-analysis, there was no DIF in item No 25.

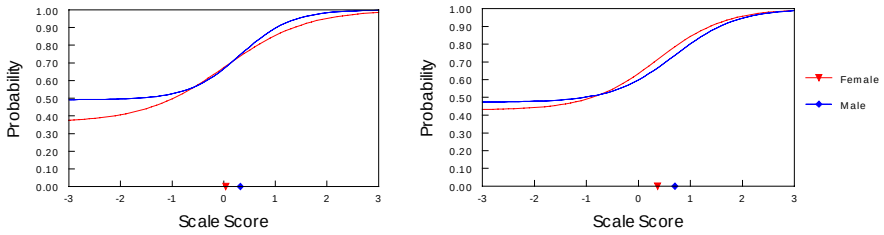

Figure 13. *ICCs for males and females on pretest item No 24 (left) and regular test item No 27 (right).*

On this item no changes had been made between pretest and regular test. The item had become a little more difficult in the regular test for both males and females but the DIF had hardly changed. The discrimination power had decreased to a small extent for males while it had increased to a small extent for females between pretest and regular test.

According to CTT, this item had become more difficult in the regular test than in the pretest, but more so for males; a small DIF in favour of males in the pretest changed to a small DIF in favour of females in the regular test (p_M = .70/.64 and p_F = .68/.66). The discrimination had decreased to a small extent for both males and females.

According to the MH-analysis, item No 27 was an A-item, favouring females (Δ-diff = .34).

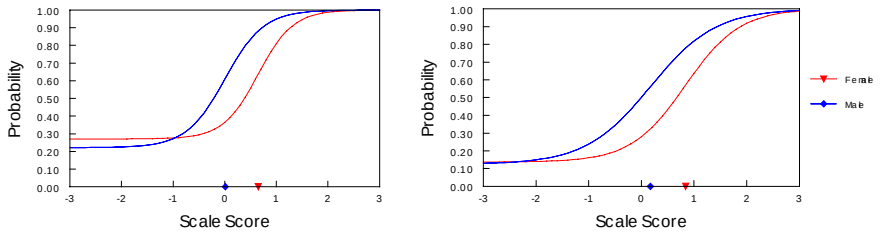

Figure 14. *ICCs of males and females on pretest item No 4 (left) and regular test item No 28 (right).*

On this item one distractor had been changed between pretest and regular test. On the pretest version of this item there was a clear DIF in favour of males, while in the regular version the item became slightly more difficult and the DIF increased. The discrimination, which was very high in the pretest version, decreased somewhat in the regular version for both males and females.

According to CTT, as well, there was a clear DIF in favour of males on the pretest version (p_M = .60, p_F = .48) which increased somewhat on the regular test (p_M = .52, p_F = .36), where the item was also more difficult for both males and females. The discrimination power was also higher for males but decreased slightly in the regular test for both males and females.

According to the MH-analysis, item No 28 was a C-item, favouring males (Δ-diff = -1.59).

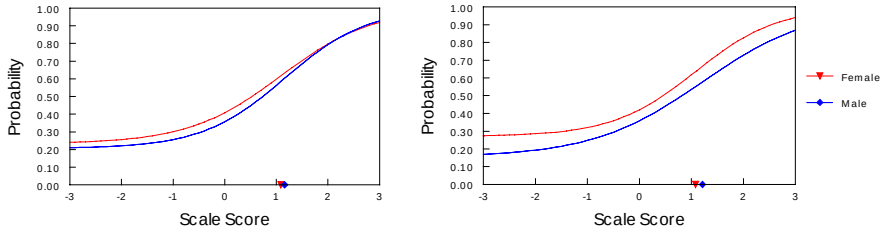

Figure 15. *ICCs for males and females on pretest item No 4 (left) and regular item No 29 (right).*

On this item no changes had been made between pretest and regular test. The difficulty level had increased somewhat for males, but not for females, between pretest and regular test, and hence the small DIF in favour of females on the pretest increased on the regular test. The

discrimination decreased slightly for males while it increased slightly for females.

According to CTT as well, the item became imperceptibly more difficult for males in the regular test while it became a little easier for females (p_M = .39/.38, p_F = .44/.46), hence an imperceptible DIF in favour of females appeared in the regular test. The discrimination power decreased slightly for both males and females.

According to the MH-analysis, item No 29 was an A-item, favouring females (Δ-diff = .86).

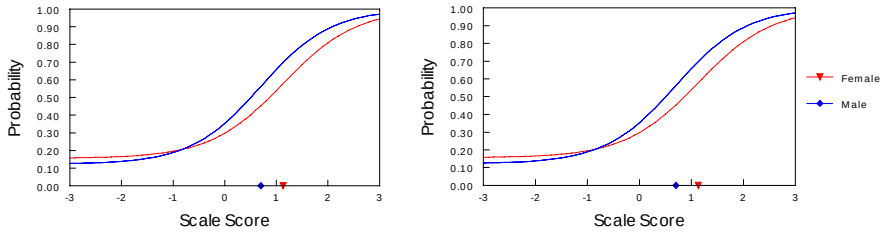

Figure 16. *ICCs for males and females on pretest item No 5 (left) and regular test item No 35 (right).*

On this item one distractor had been changed between pretest and regular test. The difficulty had decreased for both males and females in the regular test and the small DIF in favour of males on the pretest remained on the regular test. The discrimination, which was rather poor in the pretest, had increased for both males and females in the regular test.

According to CTT, as well, the difficulty had decreased for both males and females (p_M = .33/.40, p_F = .30/.35), but more so for males and hence the small DIF in favour of males on the pretest had increased somewhat on the regular test. Moreover, according to CTT, the discrimination had increased for both males and females.

According to the MH-analysis, item No 35 was an A-item, favouring males (Δ-diff = -.38).

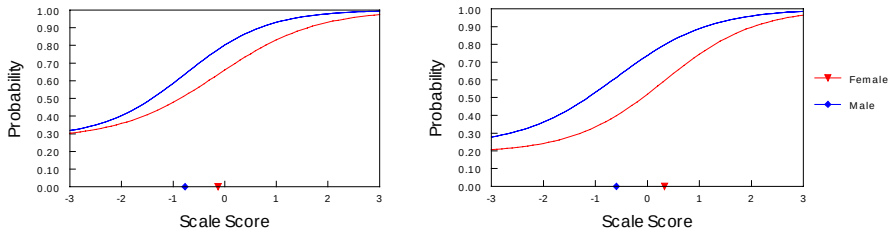

Figure 17. *ICCs for males and females on pretest item No 37 (left) and the regular test item No 36 (right).*

On this item, one distractor had been changed. The item was more difficult in the regular version for both males and females, but somewhat more so for females, and therefore the DIF in favour of males had increased on the regular test. The discrimination power was approximately the same in the pretest and in the regular test.

According to CTT the changes were the same. The item was more difficult in the regular version, and more so for females (p_M = .77/.71, p_F = .66/.53). The DIF in favour of males had increased on the regular test. As for the discrimination there were only minor changes.

According to the MH-analysis, item No 36 was a C-item, favouring males (Δ-diff = -1.91).

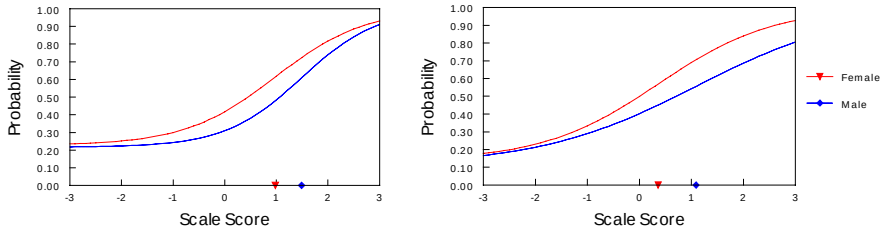

Figure 18. *ICCs for males and females on pretest item No 6 (left) and regular test item No 38 (right).*

On this item, one distractor had been changed. The item was easier for both males and females in the regular test, and the DIF in favour of females had increased a little on the regular test. The discrimination power had decreased, in the regular test, especially for males. There was also a small model data misfit for males on item No 38.

According to CTT, as well, the item was easier in the regular test and here the DIF in favour of females remained exactly the same (p_M =

.35/.41, p_F = .45/.51) on the regular test as it had been on the pretest. The discrimination power had increased to a small extent in the regular test for both males and females.

According to the MH-analysis, item No 39 was a B-item favouring females (Δ-diff = 1.15).

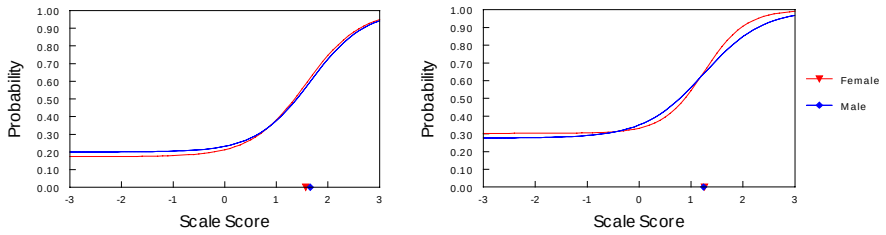

Figure 19. *ICCs for males and females on pretest item No 28 (left) and regular test item No 39 (right).*

On this item, three distractors had been changed between pretest and regular test. The difficulty level was very similar for males and females and was slightly lower in the regular test for both males and females. The discrimination power, which was fairly high for both males and females, increased somewhat for females while it decreased slightly for males on the regular test.

According to CTT, as well, the difficulty level was about the same for males and females on both pretest and regular test, even though the item was somewhat easier in the regular test (p_M = .28/.41, p_F =.27/.41). The DIF in favour of females on the pretest remained on the regular test. According to CTT, however, the discrimination power increased for males in the regular test while for females it remained the same as in the pretest (r_{bisM} = .22/.31, r_{bisF} = .32/.32).

According to the MH-analysis, item No 39 showed no DIF.

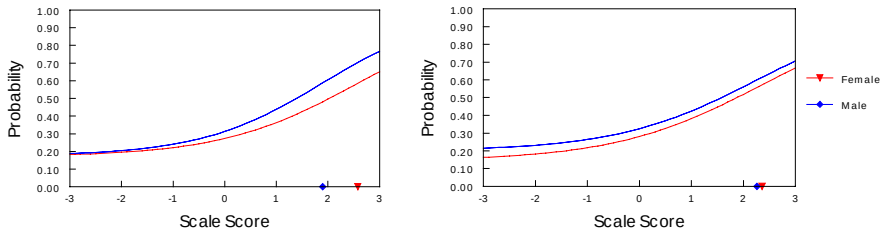

Figure 20. *ICCs for males and females on pretest item No 30 (left) and regular test item No 40 (right).*

On this item, two distractors had been changed from pretest to regular test version. The item was difficult in the pretest version and more difficult for females; in the regular version the difficulty had increased both for males and females and the difference had almost disappeared. The discrimination was rather poor, and remained so in the regular test.

According to CTT, as well, the item was somewhat more difficult for females in the pretest version, and the difference remained in the regular version (p_M = .33/.34, p_F = .29/.30). The discrimination in the pretest version was poor for both males and females, and remained so in the regular test.

According to the MH-analysis, item No 40 was an A-item, favouring males (Δ-diff = -.41).

Discussion

In Figure 21 the IRT based DIF (b-differences) in the regular test is plotted against the CTT based DIF (p-differences) in the regular test.

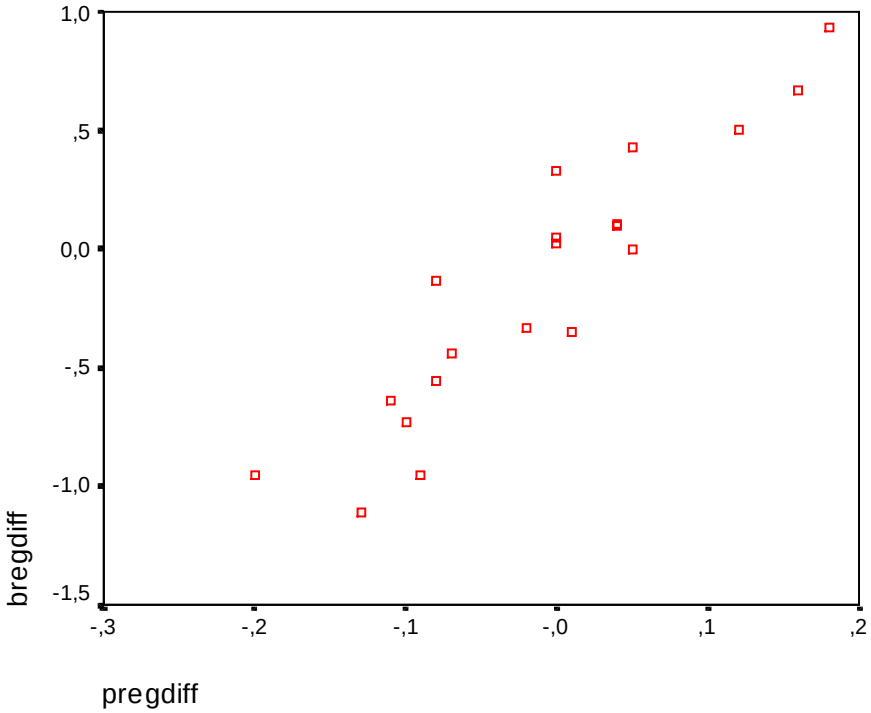

Figure 21. *IRT-based DIF plotted against CTT-based DIF on the regular test.*

The agreement between the two theories regarding DIF was very good; the correlation between IRT- and CTT-based DIF was r = .92. The most deviating items below the line, i.e. items where the DIF in favour of females was unexpectedly high according to IRT in comparison with CTT, were items number 5, 11, and 23. The items deviating above the line, i.e. items where the DIF favouring males according to IRT was higher than the DIF according to CTT, were items number 16, 29, and 35.

Of the 4 items which, according to MH, did not have DIF (16, 23, 25 and 39), none showed any DIF according to CTT either, while two showed a slight DIF according to IRT (16 and 23).

As for the 7 items identified by MH as A-items, the p-differences ranged between .02 and .08, while the b-differences ranged from .00 to .43. The 4 B-items ranged in p-differences between .08 and .11, while the b-differences ranged from .56 to .95. Finally the 5 C-items ranged in p-differences from .12 to .20 and in b-differences from .50

to 1.11. Hence the overlap between the categories was larger for the IRT based differences.

The correlation between Δ-differences and p-differences was r = .98 and the correlation between Δ-differences and b-differences was r = .92.

The overall conclusion is that the three methods appear to give very similar results. There was a somewhat higher agreement between MH and CTT regarding DIF; but then the comparison with IRT was made solely with the estimated b-parameter of each item, and the great advantage of IRT is that information is given along the whole ability continuum and not only for a single point as with CTT.

Hambleton and Rogers (1989) compared the IRT-based area method and the MH method for investigating DIF and they concluded that:

… when the unreliability of the statistics was taken into account, the two methods led to very similar results. Discrepancies between methods were due to the presence of nonuniform DIF (the Mantel-Haenszel method could not identify theses items) and the choice of interval over which DIF was assessed (the IRT method results depended on the choice of interval). (p. 314)

With regard to the stability of the item statistics, however, there seems to be no great differences between CTT and IRT. When the groups of examinees are as large and representative as in this study, CTT seems to work as well as IRT for predicting item difficulty and item discrimination.

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.

Hambleton, R. K., & Rogers, H. J. (1989). Detecting Potentially Biased Test items: Comparison of IRT Area and Mantel-Haenszel Methods. *Applied Measurement in Education, 2 (4),* 313-334.

Hays, W. L. (1963). *Statistics.* London: Holt, Rinehart and Winston.

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.

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