INTERNATIONAL ASSOCIATION
FOR
EDUCATIONAL ASSESSMENT

23RD ANNUAL CONFERENCE JUNE 1997 DURBAN SOUTH AFRICA

EQUITY ISSUES IN EDUCATION AND ASSESSMENT

Do Males and Females With Identical Test Scores Solve Test Items in the Same Way?

Dr Christina Stage

Organised by

INDEPENDENT EXAMINATIONS BOARD

Background

The Swedish Scholastic Aptitude Test 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. A test is made public after its administration 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 with 22 items.

2. **DTM** a subtest measuring the ability to interpret diagrams, tables and maps with 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.

Each correct answer is given one point, and the total number of correct answers represents the raw score of the test. In order to ensure that scores on different administrations of the test are comparable, the raw scores are converted to standardized or normed scores, which have a range of 0.0 to 2.0, the latter being the top result. In the selection process the normed scores are used.

Ever since the SweSAT was introduced 20 years ago there has been a very consistent difference between males and females in average results. Gender differences are not in any way unique to the SweSAT, on the contrary gender differences are found on many standardized tests. In Sweden, however, the equality between the genders is an important issue and there is minimal acceptance of gender differences in test results regardless of the reasons for the differences. Generally the tests are blamed for being unfair or biased.

The gender differences in results on SweSAT have been subject to extensive research with the purpose to describe, understand and explain the differences (see Stage, 1993 for a summary of studies).

Purpose

The main purpose of this study was to examine whether the total SweSAT scores had the same interpretation for males and females. That is do males and females perform about the same on the subtest and even the specific items, or are there differences? A second purpose was to examine whether there are items in the test which should be judged as gender biased.

METHOD

Description of the Test Data and Examinee Samples

The samples used in the studies below were drawn from the data set containing the responses of 82,485 testtakers of the SweSAT administered in spring 1996. Of these testtakers 38,388 were males and 44,097 were females. The results of these groups of examinees on the five subtests and on the total test are shown in Table 1.

Table 1. Mean results of males and females and effect sizes[1] on the five subtests and the total test in spring 1996. The standard deviations are reported within brackets.

Subtest	**Males**	**Females**	**d[1]**
DS	15.32 (6.57)	12.54 (4.64)	.58
DTM	12.88 (3.56)	10.64 (3.53)	.60
ERC	13.66 (3.84)	12.68 (3.91)	.25
READ	12.05 (3.46)	12.25 (3.51)	-.06
WORD	25.72 (6.57)	25.15 (7.01)	.08
Total	79.62 (16.83)	73.26 (17.47)	.36

As may be seen in Table 1 the gender differences on the two quantitative subtests, DS and DTM are substantial. The gender difference on the English subtest is notable as well while the gender differences are small on both the Swedish reading comprehension subtest and on the vocabulary subtest. These are very typical results for the SweSAT over the years.

Differences in average performance between two groups are often mistakenly described as bias against the lower performing group. But just as there is no a priori basis for deciding that differences exist between groups there is no a priori basis for deciding that differences do not exist. Actually the mean difference concept of test bias has been the most uniformly rejected of all criteria of test bias by psychometricians (Reynolds, 1982). Were all differences interpreted as bias, then we would need to challenge on weigh scales and measuring sticks because they too show male-female differences.

All currently accepted definitions of bias contain some way of conditioning the differences to equal ability levels of the compared groups. For example:

[1] The effect size d is the difference between the means divided by the pooled standard deviation

> *An item is unbiased if, for all individuals having the same score on a homogenous subtest containing the item, the proportion of individuals getting the item correct is the same for each population group being considered. (Scheuneman, 1975, p 2)*

Because of the importance of the item bias issue, considerable attention has been devoted to the development and evaluation of methods to detect bias. Although no statistical or judgmental method can detect ”bias” there are several methods available to detect items that are functioning differentially in two groups of interest, for example males and females. These methods are referred to as methods for investigation of differential item functioning (DIF). 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)*

Dorans and Holland (1993) make a distinction between DIF and impact:

> *In contrast to impact, which often can be explained by stable consistent differences in examinee ability distributions across groups, DIF refers to differences in item functioning after groups have been matched with respect to the ability or attribute that the item purportedly measures. DIF is an unexpected difference among groups of examinees who are supposed to be comparable with respect to the attribute measured by the item and test on which it appears. (p 36)*

DIF rather than bias is currently used to describe the empirical evidence which has been collected in the investigation of bias, while the existence or non-existence of bias is a matter of professional judgment using DIF as a statistical indication of differential response patterns.

ANALYSIS I: Comparison of p-values of males and females with the same normed score

Subjects

From the the total number of testtakers spring 1996 all testtakers with a normed score of 1.3 were chosen. A normed score of 1.3 is clearly above average and is also a score which may result in admittance to several educations for which competition is not very strong. In the group of testtakers with a normed score of 1.3 there were 3,287 male examinees and 2,896 female, i.e. 6,183 testtakers had a normed score of 1.3 which in this test corresponded to 87-90 correct answers. The results of these groups of males and females were compared in terms of both subtest and item level performance. The results on the five subtests and on the total test for these groups of examinees are presented in Table 2.

Table 2. Results on the five subtests and the total test of males and females with a normed score of 1.3. The standard deviations are reported within brackets.

Subtest	Males	Females	d
DS	17.17 (2.94)	15.42 (3.15)	.55
DTM	14.21 (2.56)	12.73 (2.61)	.55
ERC	15.33 (2.34)	15.40 (2.31)	-.03
READ	13.31 (2.22)	14.76 (2.05)	-.65
WORD	28.48 (3.92)	30.11 (4.02)	-.40
Total	88.50 (1.13)	88.42 (1.11)	.07

As may be seen in Table 2 there are still considerable gender differences in subtest scores even though the groups of males and females have the same normed scores. The pattern of differences is similar to that of the original groups in as far as males score higher on the quantitative subtests, DS and DTM. The effect sizes on the two quantatiative subtests are still notable, but for these groups of males and females they are counterbalanced by differences on the verbal subtests READ and WORD.

To compare the gender differences at the item level, the items were classified into seven categories on the basis of differences in p-values between males and females:

Category 1: items where $p_F - p_M > .15$

Category 2: items where $.15 \geq p_{F} - p_M \geq .11$

Category 3: items where $.10 \geq p_F - p_M \geq .03$

Category 4: items where $| p_F - p_M | \leq .02$

Category 5: items where $.10 \geq p_M - p_F \geq .03$

Category 6: items where $.15 \geq p_M - p_F \geq .11$

Category 7: items where $p_M - p_F > .15$

The outcome of the classification of the test items is presented in Table 3.

Table 3. The test items divided into categories based on the gender differences in p-values.

	Females outperformed males			No difference	Males outperformed females		
Category Subtest	1	2	3	4	5	6	7
DS				1, 7	2, 3, 4, 5, 8, 10, 11, 12, 13, 14, 15, 17, 19, 21	6, 16,18, 20, 22	9
DTM				1, 4, 6, 9, 15	2, 3, 5, 7, 8, 11, 12, 13, 19, 20	10, 17	14, 16, 18
ERC		5	1, 4, 6, 14, 15, 18	2, 3, 9, 10, 11, 16, 17	7, 8, 12, 13, 19, 20		
READ	7	8, 12, 16	1, 2, 3, 4, 5, 6, 9, 10, 11, 13, 14, 15 20	17, 19	18		
WORD	11, 21, 32, 34, 36	3, 10, 20, 23, 39	7, 8, 13, 17, 18, 25, 27, 28, 30, 31, 40	1, 2, 5, 6, 9, 12, 16, 19, 24, 33, 35	4, 14, 15, 29	37, 38	22, 26
Total number	6	9	30	27	35	9	6

It is remarkable that for these groups of males and females who performed equally well on the total test, i.e. they had the same normed score, there were only 27 items out of 122 for which there were no differences. It is also obvious that the quantitative items were easier for males while the verbal items were easier for females. There are also quite a few items where the differences in p-values between the groups are substantial; on 15 items, all of which were verbal, the differences in favour of females were greater than .10. The same number of items were in favour of males to the same extent, but of these items 11 were from the quantitative subtests.

MANTEL-HAENSZEL ANALYSES

As already mentioned the importance of the item bias issue has led to considerable attention being devoted to the development and evaluation of methods for detecting differentially functioning test items (DIF). The use of the Mantel-Haenszel (MH) statistic has emerged as one of the most popular procedures. Hills (1989) cites the following advantages of the MH method: ”*it provides both a measure of effect size and a test of statistical significance; it can be used with relatively small samples and it is inexpensive to use*.” The MH method also compares the probabilities of a correct response for groups of examinees of the same ability. The MH statistic uses a constant odds ratio (α MH) as an index of DIF. The estimated value of α MH is usually transformed to delta units, MH D-DIF, which are more straightforward to interpret. Details of the MH method may be found in papers by Holland and Thayer (1986, 1988).

ANALYSIS II: MH-Analysis on males and females with a normed score of 1.3

In order to examine whether any items could be revealed as differentially functioning, a MH-analysis was performed on the testtakers with a normed score of 1.3, i.e. the 3,287 males and the 2.896 females who had a total score from 87 to 90 in the test spring 1996.

For each MH-analysis of items in each subtest, the subtest score was used to match males and females.

The outcome of the analysis was that altogether 80 items showed significant DIF at the .05 level and of these 74 were significant at the .01 level as well. Hence a considerable number of items were flagged as significantly DIF. Statistical tests, however, have a well-known and serious flaw: their sensitivity to sample size. This is what Hays (1969) calls the fallacy of evaluating a result in terms of statistical significance alone:

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

At Educational Testing Service (ETS) 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 delta-scale) and whether the difference is statistically significant or not. **A:** if either MH D-DIF is not significantly different from zero or the absolute value is less than one delta unit. **C:** if MH D-DIF exceeds 1.5 in absolute value and is statistically significant. **B:** all cases in between (Dorans & Holland, 1993).

In Table 4 the items flagged as significantly DIF are classified according to the categories used by ETS and divided into those favouring females and those favouring males.

When delta plot was used as a DIF method (Stage, 1985) one problem was that group differences in the ends of the distribution were enlarged, i.e. the method was too sensitive to differences on items which were very easy or very difficult. The same seems to be true with MH D-DIF. If the p-value of an item for one or both groups is larger than .90 (which is the case for some items with these selected groups, with a normed score of 1.3), the method does not accept any difference at all. In Table 4 the items where one or both of the groups had a p-value larger than .90 are marked *. These items should probably be disregarded as DIF.

Table 4. The items with significant DIF for examinees with a normed score of 1.3 classified into ETS´s categories A, B and C.

	Items favouring females			Items favouring males		
Category Subtest	C Large	B Intermediate	A Negligible	A Negligible	B Intermediate	C Large
DS			1, 3, 7, 10, 15, 21	6, 18	8*, 9, 11*	
DTM		6	1, 3, 4, 8, 9, 15, 20	7, 11, 14, 17, 18	16	
ERC		5, 14*	1, 4, 6, 15, 18	7, 8, 10, 12, 13, 17, 19, 20		
READ			7, 9, 11, 12, 13, 16	3, 17, 19, 20	18	
WORD	3, 17*, 21, 23, 32, 36	8*, 10, 11, 20, 34, 39	7, 30, 40	9, 12, 24, 33, 35,	2*, 4, 6, 14,	15, 22, 26, 37, 38
Total number	6	9	27	24	9	5

Of the 15 items classified as B or C favouring females, 11 were found in category one or two in the p-value comparison and three items have p-values higher than .90. Of the 14 items classified as B or C favouring males, six items were found in category six or seven and three items have p-values higher than .90.

Analysis III: MH-analysis on a random sample of examinees using SweSAT scores as the matching variable for males and females.

From the 82,485 testtakers in spring 1996 a random sample of three percent was drawn. The resulting sample consisted of 2,461 persons, 1,112 males and 1,349 females. The results of these groups of males and females were very close to the results of the total groups of examinees. On these groups a conventional MH analysis was performed.

The results from the MH-analysis of the random sample of examinees when the total normed score was used as matching variable resulted in 79 items being flagged as significantly DIF and 70 of these were significant at .01 level. The results are presented according to ETS´s categories of DIF only, where as before A means negligible DIF, B intermediate DIF and C means large DIF. This presentation is found in Table 5.

Table 5. Items classified into ETS´s categories

	Items favouring females			Items favouring males		
Category Subtest	C Large	B Intermediate	A Negligible	A Negligible	B Intermediate	C Large
DS				3, 4, 7, 10, 12, 19	6, 9, 13, 14, 16, 18, 22	2, 8, 11, 20
DTM				1, 2, 9, 10, 17, 19, 20	7, 11, 13	14, 16, 18
ERC		5, 6	4, 14, 15, 18	7, 10, 20		
READ	12	2, 7, 13, 16	5, 6, 8, 9, 10, 11, 14, 15, 20	18		
WORD	21, 23, 32, 36, 39	3, 8, 10, 11, 25, 34, 40	7, 9, 13, 17, 18, 20, 27, 28, 30, 35	15	38	22, 26, 37
Total number	6	13	23	16	11	10

As may be seen in Table 5, when the total normed test score was used as the matching variable, 16 items were classified as C-items, 6 of them favouring females and 10 favouring males. Of the five WORD-items favouring females, four were found in the same category in the MH-analysis on groups with a normed score of 1.3 and three were found in category 1 in the p-value comparison. The READ-item in category C favouring females was classified as a category 2 item in the p-value comparison. The three WORD-items in category C favouring males were all found in the same category in the former MH-analysis and two of them were found in category 7 in the p-value comparison. The three DTM-items in category C, favouring males were all found as category 7 in the p-value comparison and one of them was found in category B in the MH-analysis of examinees with a normed score of 1.3. The four DS-items in category C, favouring males were not refound in the extreme category in any of the other analyses, but two of them were found as B-items in the MH-analysis of groups with a normed score of 1.3.

IRT analysis

Finally, a three parameter logistic item response theory model was adapted to the response data of the random sample of 2,462 examinees. The use of the three parameter item response model is a comprehensive method for detecting item-bias, taking into account not only differences between the groups with respect to item difficulty, but also differences with respect to discriminating power and differences with respect to the pseudo guessing parameter (Angoff, 1993). There is a problem with IRT for determining item-bias. If the fit of the model to the data is not very good there is a risk of confusing model misfit and item-bias. But even though item characteristic curves may not be the best way to identify or detect DIF in test items, they provide very good illustrations.

The range of ability scores for males with a normed score of 1.3 was .28 - .96 and for females the range was .24 - .91.

The DS subtest

In Table 6 a summary is given for the DS subtest of the outcome of the analyses performed. From the p-value comparisons (pv-comp) for males and females with a standardized score of 1.3, the items with differences larger than .10 or categorized as one, two, six or seven are included; from the Mantel-Haenszel analysis of examinees with a normed score of 1.3 (MH-1.3) the items categorized as B or C items are included and the same goes for the Mantel-Haenszel analysis of random samples of males and females with the total normed test score as matching variable (MH-rs).

Table 6. Items in the DS subtest flagged as DIF by different methods.

Category	Female		Male	
Method	1 or C	2 or B	6 or B	7 or C
pv-comp			6, 16, 18, 20, 22	9
MH-1.3			8*, 9, 11	
MH-rs			6, 9, 13,14, 16, 18, 22	2, 8, 11, 20

Item 9 in the DS subtest was showing DIF favouring males according to pv-comp and was flagged as B-item favouring males by MH. The ICCs of item 9 are shown in Figure 1[2].

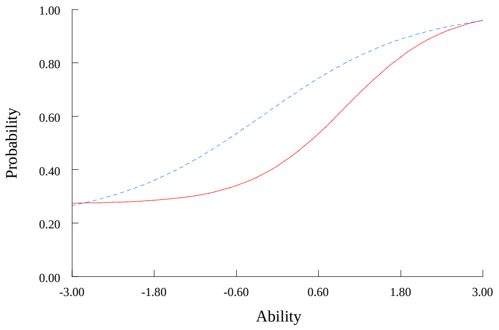

Figure 1. The ICCs of item 9 in the DS subtest.

In figure 1 it may be seen that there is a clear difference between the curves of males and females for item 9 at most ability levels even though the differences are negligible at the extremes. The content of this item was about *index*. For this item the ICCs give support to the item being biased.

[2] Legends in the figures containing ICCs:

Female Male

Items 2 and 8 were flagged as C-items favouring males by MH-rs. The ICCs of these items are shown in Figure 2.

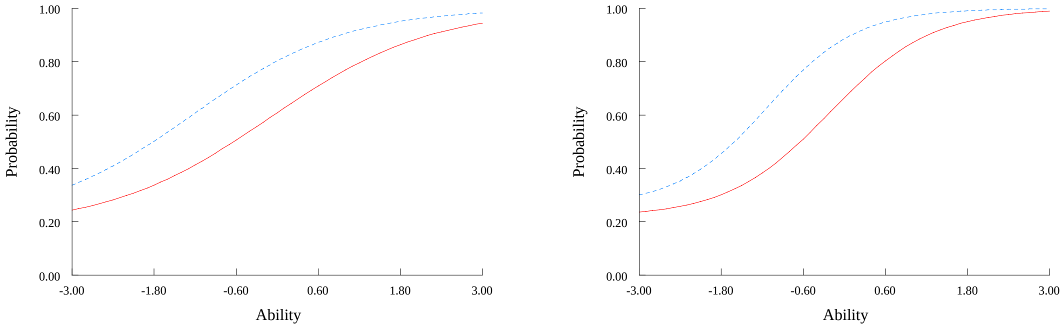

Figure 2. The ICCs of items 2 and 8 in the DS subtest.

For both items 2 and 8 there seem to be a clear difference between males and females in the probability for correct response for all ability levels but the very high. In item 2 the question was *What part of Sweden´s population lives in the Northern part*. In item 8 the question was about *number of people taking part in a competition*

Items 11 and 20 were flagged as C-items favouring males by the MH-rs analysis. The ICCs of these items are shown in Figure 3.

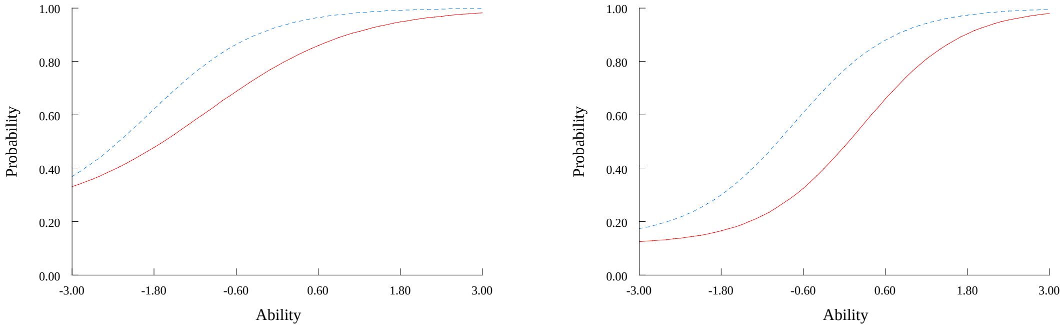

Figure 3. The ICCs of items 11 and 20 in the DS subtest.

For items 11 and 20 as well there seem to be differences between males and females regarding probability for a correct answer. Item 11 was about *number of persons in a queue* and item 20 was about *number of eggs laid by hens.*

The DTM subtest

In Table 7 a summary of the outcome of the analyses of the DTM subtest is presented.

Table 7. Items flagged as DIF on the DTM subtest

Category	Female		Male	
Method	1 or C	2 or B	6 or B	7 or C
pv-comp			10, 17	14, 16, 18
MH-1.3		6	16	
MH-rs			7, 11, 13	14, 16, 18

In the DTM subtest item 6 was categorized as B-item favouring females by the MH-1.3 analysis. The ICCs of this item are shown in Figure 4.

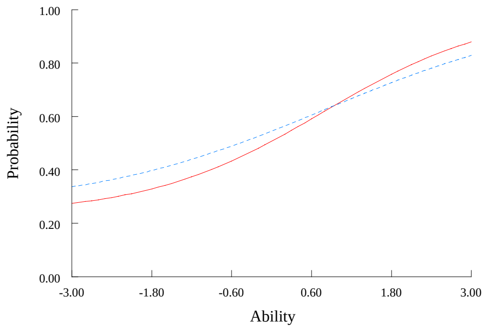

Figure 4. The ICCs of item 6 in the DTM subtest.

The differences between the curves for item 6 seem to be very small and this item can hardly be judged as biased according to the ICCs. Item 6 was about *Sweden´s export of food* and the answer was found in a table. The p-values for males and females on this item were p_M = .55 and p_F = .52 for the random sample and p_M = .58 and p_F = .60 for the testtakers with a normed score of 1.3.

Items 14, 16 and 18 were flagged as DIF, favouring males by MH-rs as well as by the pv-comp. The ICCs of these items are shown in Figures 5 and 6.

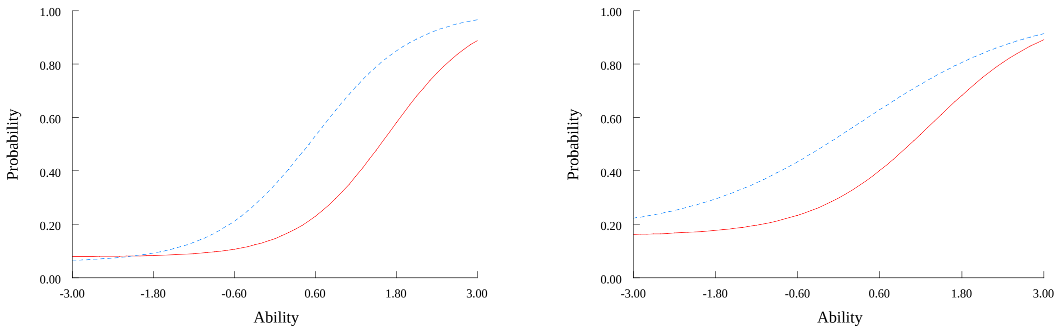

Figure 5. The ICCs of items 14 and 16 in the DTM subtest.

The differences between the ICCs for males and females are fairly large in the middle ability levels for both item 14 and 16. Both items seem to be biased against females according to the ICCs. Item 14 was about *meteorology (percent cloudiness)*. The content of the item 16 was about *the rate of exchange* and the answer was found in a diagram.

The ICCs of item 18 in the DTM subtest are shown in Figure 6.

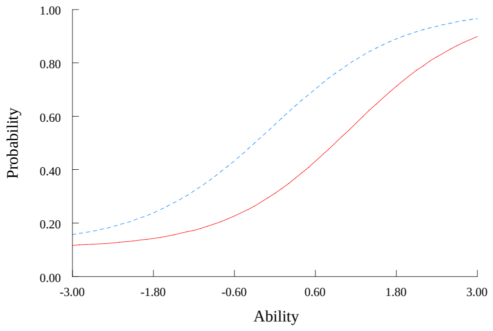

Figure 6. The ICCs of item 18 in the DTM subtest.

The differences between the ICCs for males and females are fairly large in the middle ability level for item 18. The item seems to be biased against females according to the ICCs as well. Item 18 was about the *number of Christians and Catholics* and the answer was found by combining information given in a table and a diagram.

The ERC subtest

In Table 8 a summary is given for the ERC subtest.

Table 8. Items flagged as DIF in the ERC subtest.

Category	Female		Male	
Method	1 or C	2 or B	6 or B	7 or C
pv-comp		5		
MH-1.3		5, 14*		
MH-rs		5, 6		

In the ERC subtest no items were flagged as large DIF. Item 5, however, was flagged as intermediate DIF, favouring females, by all analyses and item 6 by MH-rs. The ICCs of these items are shown in Figure 7.

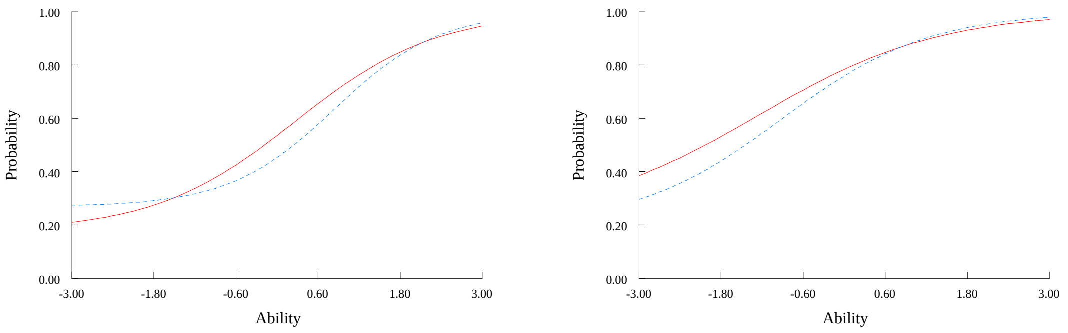

Figure 7. The ICCs of items 5 and 6 in the ERC.

There are very small differences between the curves in Figure 7. The question in items 5 and 6 were to the same text and went *What are Doris Lessing´s feelings for her mother later in life*? and *What is suggested about Doris Lessing´s career in the last paragraph*. The support by the ICCs for these items being biased is very small.

The READ subtest

In Table 9 a summary is given of the outcome of the analyses of the READ subtest.

Table 9. Items flagged as DIF in the READ subtest.

Category	Female		Male	
Method	1 or C	2 or B	6 or B	7 or C
pv-comp	7	8, 12, 16		
MH-1.3			18	
MH-rs	12	2, 7, 13, 16		

Item 7 and item 12 in the READ subtest were flagged as large DIF favouring females by pv-comp and MH-rs respectively. The ICCs of these items are shown in Figure 8.

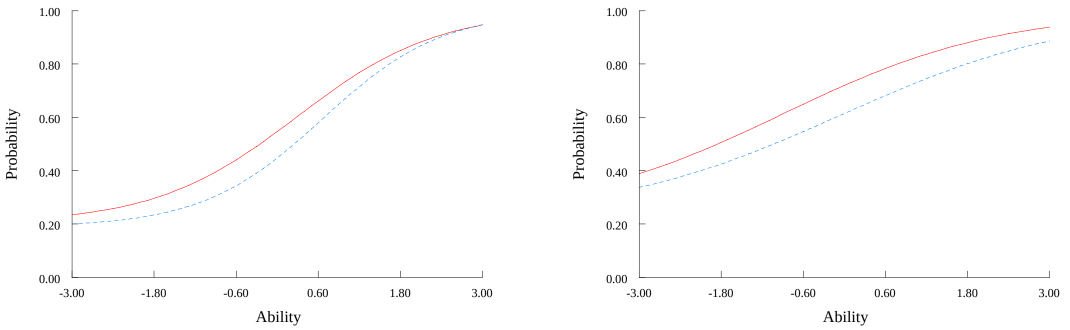

Figure 8. The ICCs of items 7 and 12 in the READ.

As may be seen in Figure 8 there are differences between the curves of males and females except at the very high ability level for item 7. The differences between the curves are not very large, however. Item 7 was about *changes of the Oidipal conflict*. In item 12 the question was to a text about *a children´s disease* and read *What would be a suitable title of the text?* Here as well, there is some small support by the ICCs for the items being biased against males.

Item 18 was flagged as intermediate DIF, favouring males by MH-1.3. The ICCs of this item are shown in Figure 9.

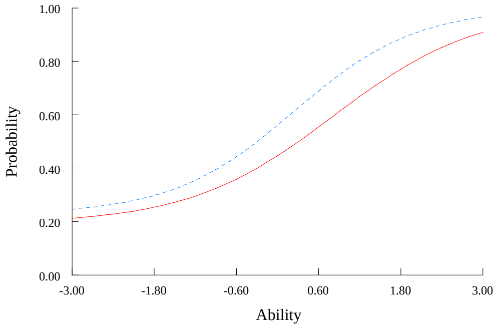

Figure 9. The ICCs of item 18 in the READ subtest.

For item 18 the curves differ moderately and the curve for males is above the curve for females along the whole ability continuum. The question was to a text about *Eutrofication* and read *What was the result of the introduction of biological drain-cleaning?* The ICCs give some small support for the item being based against women.

The WORD subtest

In Table 11 a summary is given of the outcome of the analyses of the WORD subtest.

Table 11. Items flagged as DIF in the WORD subtest.

Category	Female		Male	
Method	1 or C	2 or B	6 or B	7 or C
pv-comp	11, 21, 32, 34, 36	3, 10, 20, 23, 39	37, 38	22, 26
MH-1.3	3, 17*, 21, 23, 32, 36	8*, 10, 11, 20, 34, 39	2*, 4, 6, 14	15, 22, 26, 37, 38
MH-rs	21, 23, 32, 36, 39	3, 8, 10, 11, 25, 34, 40	38	22, 26, 37

In the WORD subtest there were many items flagged as large and intermediate DIF. Seven items were flagged as large DIF favouring females (disregarding item 17); out of these seven items, three were the same according to all analyses, items 21, 32 and 36. The ICCs of these items are shown in Figures 10 and 11.

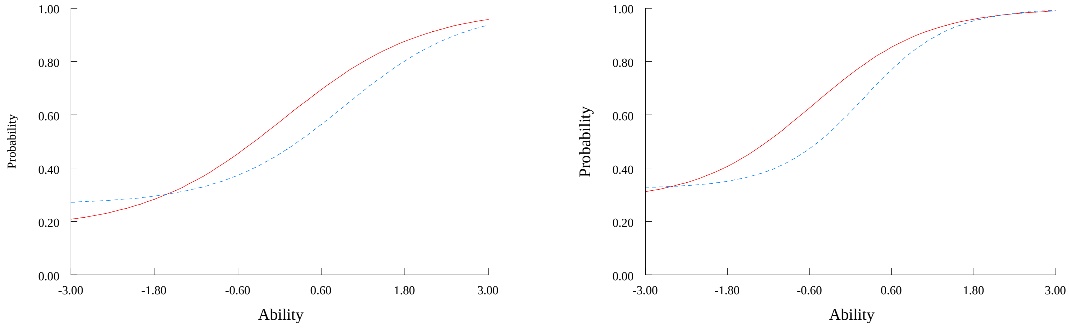

Figure 10. The ICCs of items 21 and 32 in the WORD subtest.

The ICCs of both item 21 and 32 indicate that the items are easier for females at least in the medium ability levels. The words also seem to be somewhat female oriented *past* (förliden) and *adorn* (utsira).

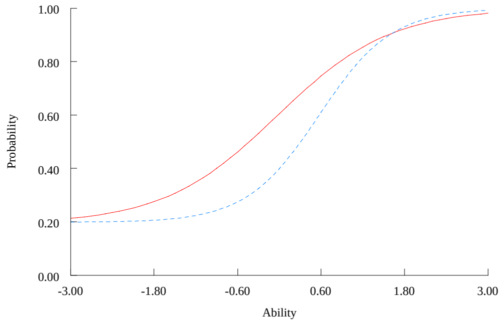

Figure 11. The ICCs of item 36 in the WORD subtest.

The ICCs of item 36 as well indicate that the item is easier for females at least in the medium ability level. This word also seems to be female oriented: *brocade* (brokad).

Item 23 was flagged as severe DIF favouring females by all MH-analyses and was classified as category 2 by pv-comp. The ICCs of item 23 are shown in Figure 12.

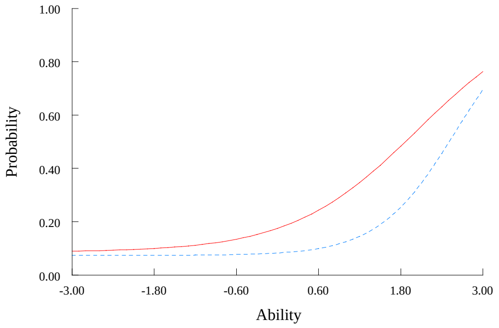

Figure 12. The ICCs of item 23 in the WORD subtest.

Item 23 also seem to be a female item according to the ICCs. The word was *reserved* (avmätt).

Items 11 and 34 were flagged as large DIF, favouring females by pv-comp. The ICCs of these items are shown in Figure 13.

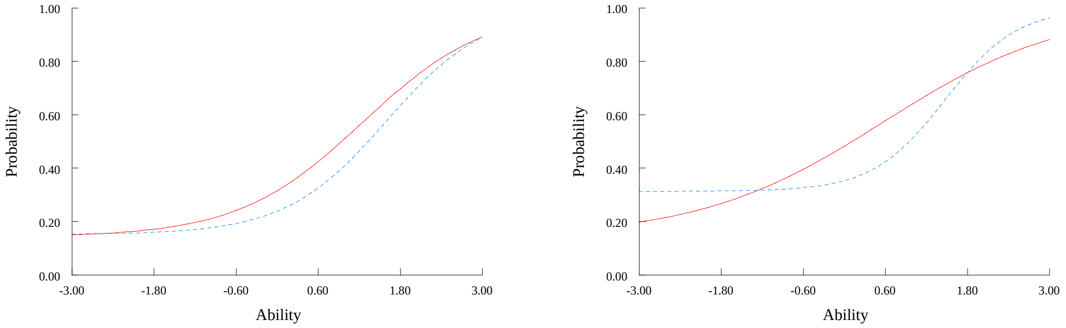

Figure 13. The ICCs of items 11 and 34 in the WORD subtest.

The ICCs of items 11 and 34 are a bit more complicated than the earlier ones especially 34 where the item seem to be easier for males at the very low and the very high ability levels while it is easier for females of medium ability. The words were *there is a rumour* (det glunkas) and *settle* (lägra sig).

Item 39 was flagged as large DIF favouring females by MH-rs and as intermediate DIF by the other methods. The ICCs of this item are shown in Figure 14.

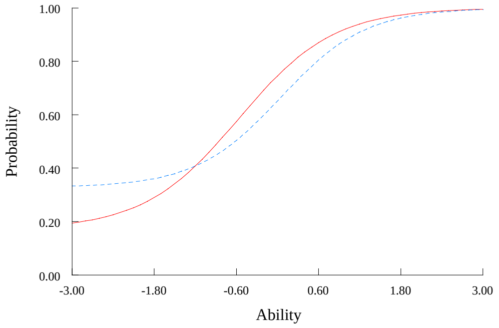

Figure 14. The ICCs of item 39 in the WORD subtest.

The difference between the curves of males and females is not very large, but there is a difference. The word of this item was *cliché* (klyscha).

Item 3 was flagged as large DIF by one analysis only, the MH-1.3. The ICCs of this item are found in Figure 15.

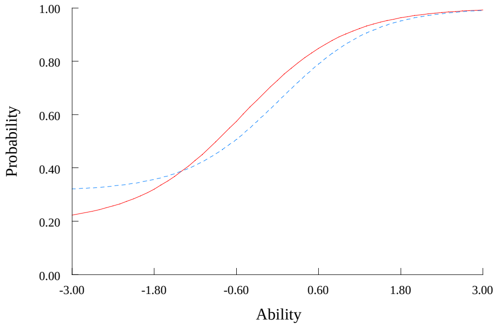

Figure 15. The ICCs of item 3 in the WORD subtest.

The difference between the ICCs of males and females is not very large on this item either, but it seems to be slightly in favour of females. The word of this item was *gorged* (däst).

Altogether five different items were flagged as severely DIF favouring males. Two of these items were the same by all methods, namely items 22 and 26. The ICCs of these items are shown in Figure 16.

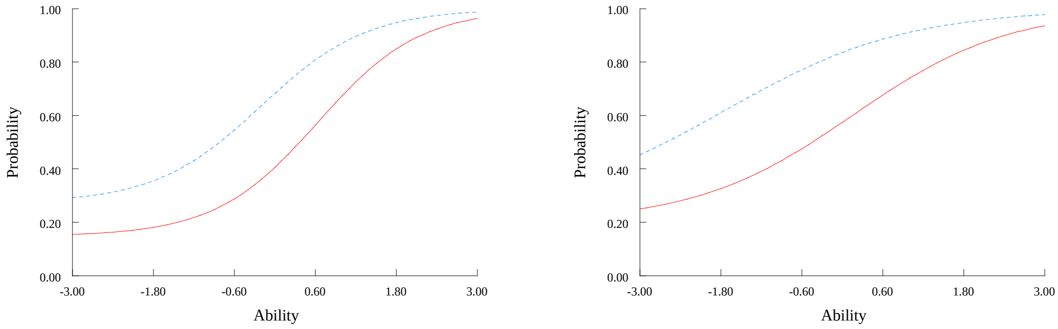

Figure 16. The ICCs of items 22 and 26 favouring males.

As may be seen in Figure 16 the ICCs support strongly the conclusion that items 22 and 26 favour males. The words in these items were *reprisals* (repressalier) and *embargo* (embargo).

Items 37 and 38 were both flagged as large DIF by MH-1.3. Item 37 was flagged as large DIF by MH-rs as well, while item 38 was flagged as intermediate DIF by this method. The ICCs of these items are shown in Figure 17.

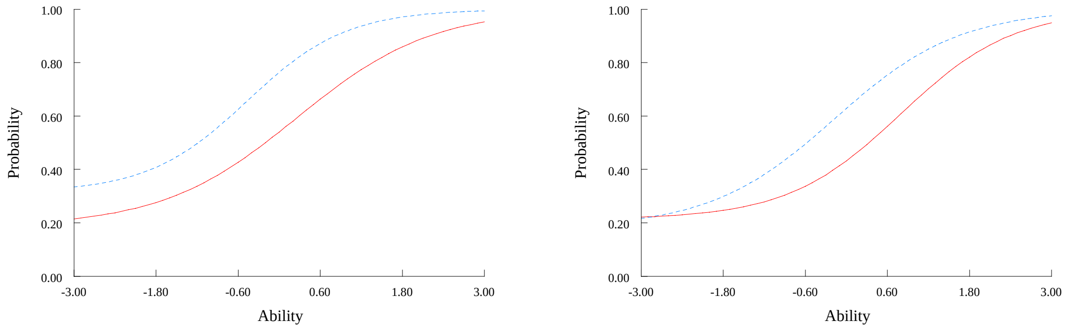

Figure 17. The ICCs of items 37 and 38 in the WORD subtest.

Also for items 37 and 38 the ICCs give support to the conclusion that the items are favouring males. The words in these items were *quisling* (Quisling) and *evaluate* (evaluera).

Item 15 was flagged as severely DIF, favouring males by MH-1.3. The ICCs of this item are shown in Figure 18.

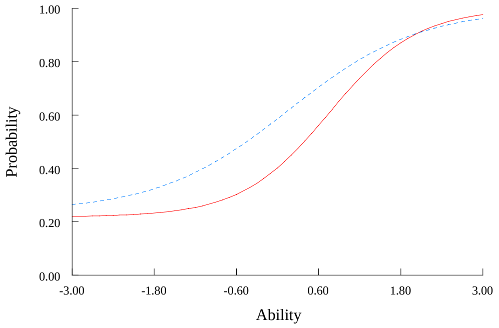

Figure 18. The ICCs of item 15 in the WORD subtest.

For item 15 as well the ICCs give support for the item being male. The word of this item was *hausse* (hausse).

DISCUSSION

The main purpose of this study was to examine whether the total SweSAT score had the same interpretation for males and females. The examination was performed by comparing the results on subtest and item level of groups of males and females who had the same normed score on the test. The normed score 1.3 was chosen as it is a result clearly above average and it is also a competitive result for many educations.

The comparison of males and females with a normed score of 1.3 demonstrated that the composition of the score was very different for the two groups. The effect sizes of differences on the subtests varied from .03 to .65. According to Cohen (1987) an effect size of .20 or less is considered very small, a difference between .2 and .5 is still considered small, but worth noting, differences between .5 and .8 are considered medium in size and differences above .8 are considered large. For three out of the five subtests the effect sizes were of medium sizes, i.e. between .5 and .8 on the DS and DTM subtests in favour of males and on the READ subtest in favour of females. On the WORD subtest the effect size was small, but noticeable and in favour of females. It was only on the ERC subtest that the effect size between males and females was very small. For these groups of males and females with the same overall performance on the test the results on subtest level still followed the pattern found by Maccoby and Jacklin in 1974 (Maccoby & Jacklin, 1975). Females outperform males in verbal ability while males outperform females in quantitative ability.

The alternative approach used was to compare p-values of males and females. On test item level there were significant differences in p-values on 95 items out of the total number of 122 and on 30 of these items the differences in p-values were larger than .10. These findings further support that testtakers with the same normed score evidently have achieved this score in very different ways. The greatest differences in favour of males were found on quantitative test items while the greatest differences in favour of females were found on verbal test items.

Even though the conclusion from the test results of these males and females with the same normed score is that they have the same ability as measured by the test it is evident that there are substantial differences between their respective performance.

The second aim of this study was to examine whether there were items in the test which should be judged as gender biased. An MH analysis was performed on the males and females who had a normed score of 1.3 with the subtest scores as matching variables. The outcome of this analysis was that only items from the subtest WORD were flagged for giving large DIF, six items were flagged as large DIF favouring females and five favouring males.

The MH-analyses on random groups of males and females and with the total normed score as matching variable gave a somewhat different picture. In this analysis four DS items were flagged as large DIF favouring males, three DTM items were flagged as large DIF favouring males and only three WORD items were flagged as large DIF favouring males. On the other hand one READ item and five WORD items were flagged as large DIF favouring females.

A critical difficulty in interpreting results from DIF studies is that the conclusion of bias does not directly follow from the statistical results. Finding DIF in an item does not necessarily imply that the item is biased. Item bias refers to an informed judgment about an item which takes into account the purpose of the test, the relevant experiences of the examinees taking the test and statistical information about the item. Hence one must still judge whether the basis of the differences on items is irrelevant to the construct measured by the test and therefore biased or whether it is relevant to the construct and therefore not an issue of bias.

In the DS subtest there was one problematic item. Item nine in the DS subtest seemed to be genuinely biased according to ICCs as well as to the pv comparison. This item was about index and during the twenty years the SweSAT has existed there has always been a big difference in results between males and females on items containing indices. Still the concept ”index” is something that all students in higher education should know about. Items 2, 8, 11 and 20 were flagged as large DIF according to the MH-rs analysis and the DIF was supported by the ICCs as well. For these items, however, it is very difficult to find any reason for the DIF.

In the DTM subtest items 14, 16 and 18 were possibly biased in favour of males. To answer item 16 you should read a diagram about exchange rates and realize that an increase from US$ 1 to US$ 2 corresponds to an increase of 100 percent in order to find the correct answer. In order to answer item 14 correctly you should find in a table for how long time the cloudiness had been more than 25 percent but less than 75 percent. To find the correct answer to item 18 you had to estimate how many percent 11 millions is out of 558 millions. Common to these three items was that you should be able to handle the concept of percent, which is another area were males usually have performed better than females.

The dilemma facing the test constructors in cases like the examples above from the DS and DTM subtests has been described by Nancy Cole (1997) in the following way:

> *By manipulating the content we could mute differences somewhat …… The problem with this manipulation arises if content of less importance replaces content of more importance as would presumably be the case when the ”fix” is driven by a goal of no difference rather than a goal of important content. If the importance of content is reduced, it would harm the meaningfulness and usefulness of the test. The skills or content that are most important have always and should always drive the make up of a test. The preeminence of the knowledge and skills is an essential technical characteristic of tests on which public confidence is largely based. (p 24)*

In the ERC subtest there were no really problematic items.

In the READ subtest there were two items which were flagged as large DIF, favouring females by one method each. For both items the ICCs gave some support to different probabilities for correct answers for males and females. Judging from past experience the content of these items also seemed to be somewhat feminine: psychology and a children´s disease.

In the WORD subtest there were a lot of items flagged as large DIF, regarding which a judgment should be made whether or not the items are biased. The items flagged as DIF in the WORD subtest are still more difficult to judge. For quite a few items all analyses flagged for large DIF but does that mean that the items also are biased? Words which can be used in a vocabulary test at advanced level must be rather infrequent; some of these infrequent words are more familiar to females and some are more familiar to males, but as long as they are proper words which exist in the language, how can they be judged as biased? The problem is to find a proper balance and hitherto the WORD subtest has been compiled in a way that has balanced male and female words.This could, however, be called in question as well. Just as there is no reason to presume that there are differences between males and females there is perhaps no reason to presume that there are no differences.

These results raises the question of the meaning of the test score. The number of correct answers of an examinee decides her/his normed score. How should one act upon the fact that different groups on the average tend to achieve a given score in clearly different ways. In the case of SweSAT the score is treated as an unambiguous result.

A second and related problem, has to do with the criterion of ability. Ideally, the raw score and a normed score is a measure of the examinee´s ability. However, tests are often suspected of being biased, i.e. the score may not give a true measure of examinee ability. The item-bias methods originally were developed because of the problem to find a good enough criterion for test-bias studies. Item-bias then was an attempt to define test-bias in the absence of a criterion and the concept item-group interaction was common (Echternacht, 1974). However, item-group interaction also deals with group differences. DIF which is the commonly accepted term today is defined as differences in probability for correct answers between individuals with the same ability, but belonging to different groups. Hence an accepted definition and measure of ability is strongly needed.

The ideal matching variable would be a perfectly reliable, unbiased measure of the developed ability the test is intended to measure. Such a measure obviously is unavailable and test scores are generally the closest approximation that is available (Linn, 1993). The comparability of the groups is achieved by matching them on the basis of a measure of test performance. Critics have a point when they argue that there is a kind of circularity. To trust the procedures you need to have some general level of acceptance of the overall test.

Although the definition of bias is rather simply stated, the determination is a rather complex process. The main problem in DIF studies as well as in bias studies is to find a criterion which is reliable, valid and unbiased. When the total test score is used as criterion the assumption is made that this score is unbiased per se, something which is questioned by those who criticize the SweSAT for favouring males. When subtest score is used as criterion this problem is still worse regarding the quantitative subtests.

Bias is sometimes defined as differential validity (Cole & Moss, 1989) and for every validation question there is a related bias question concerning differences in interpretation for different groups of concern. It is important to remember that it is the interpretation of a test score that is possibly biased, not the test score per se.

REFERENCES

Angoff, W.H. (1993) Perspectives on Differential Item Functioning Methodology, in: P.W. Holland. & H. Wainer (Eds.) *Differential Item Functioning*, pp. 3-23. Hillsdale, New Jersey, Lawrence Erlbaum Associates.

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, New Jersey, Lawrence Erlbaum Associates.

Cole and Moss

Cole, N. (1997) *The ETS Gender Study: How Females and Males Perform in Educational Settings.* Monograph, Educational Testing Service, New Jersey.

Echternacht, G. (1974) A Quick Method for Determining Test Bias, *Educational and Psychological Measurement*, 34, pp. 271-280.

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

Hills, J.R. (1989) Screening for Potentially Biased Items in Testing Programs, *Educational Measurement Issues and Practice*, 8, pp. 5-11.

Holland, P.W. & Thayer, D.T. (1986) *Differential item functioning and the Mantel-Haenszel procedure*. Technical Rep, No. 86-69. Princeton, NJ, Educational Testing Service.

Holland, P.W. & Thayer, D.T. (1988) Differential item performance and the Mantel-Haenszel procedure, in: H. Wainer. & H. Braun (Eds.) *Test validity*, (pp. 129-145). Hillsdale, NJ, Lawrence Erlbaum Associates.

Linn, R.L. (1989) Current Perspectives and Future Direction, in: R.L. Linn (Ed.) *Educational Measurement, Third Edition* (pp. 1-10). New York, American Council on Education, Macmillan Publishing Company.

Maccoby, E.E. & Jacklin, C.N. (1974). *The Psychology of Sex Differences.* London, Oxford University Press.

Scheuneman, J. (1975) *A New Method of Assessing Bias in Test Items*. Paper presented at the annual meeting of the American Educational Research Association. Washington, April, 1975.

Stage, C. (1985) *Gruppskillnader i provresultat. Uppgiftsinnehållets betydelse för resultatskillnader mellan män och kvinnor på prov i ordkunskap och allmänorientering.* Avhandling för doktorsexamen. Umeå, Umeå universitet, Pedagogiska institutionen.

Stage, C. (1993) *Gender Differences on the SweSAT. A Review of Studies since 1975.* Educational Measurement, 7. Umeå, Umeå University, Department of Educational Measurement.

Cole & Moss, 1989. p 201-219)

Echternacht, G. (1974). A Quick method for Determining Test Bias. Educational and Psychological Measurement, 1974, 34, 271-280

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Maccoby, E.E. & Jacklin, C.N. (1974). The Psychologyof Sex Differences. London: Oxford University Press

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

Hambleton, R.K. (1995) Meeting the Measurement Challenges of the 1990s and Beyond. New Assessment Models and Methods, in: T. Oakland. & R.K. Hambleton (Eds.) *International Pespectives on Academic Assessment*, (pp. 83-104). Boston, Kluwer.

Hattie, J. (1985) Methodology Review: Assessing Unidimensionality of Tests and Items, *Applied Psychological Measurement*, 9, pp. 139-164.

Scheuneman, J.D. (1982) A Posteriori Analyses of Biased Items, in: R.A. Berk (Ed.) *Handbook of Methods for Detecting Test Bias*, (pp. 180-198). Baltimore, Johns Hopkins University Press.