UNIVERSITY OF MASSACHUSETTS
AMHERST

152 Hills House South
Box 34140
Amherst, MA 01003-4140
(413) 545-0262
FAX: (413) 545-4181

Laboratory of Psychometric and
Evaluative Research

An Attempt to Fit IRT Models to the DS Subtest in the SweSAT

Christina Stage

The author is grateful to Professor Ronald, K. Hambleton for providing a careful review and constructive suggestions for this study.

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 Fall. The number of examinees in the Spring administration is usually 75,000 and in the Fall administration the number is normally 55,000. After each administration that particular test is made public and therefore a new version must be constructed for each administration. A test result is valid for five years and hence it is important that results from different administrations are comparable.

The test was changed in 1996 and now consists of 122 multiple-choice items, divided into five subtests:

1. DS a data sufficiency subtest consisting of 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 which consists of 20 items
5. WORD a vocabulary subtest consisting of 40 items.

On the DS, DTM and WORD sub-tests there are five response alternatives, one of which is correct. On the ERC and READ subtests there are four response alternatives, one of which is correct.

Since the SweSAT was first taken into use in 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. The SweSAT is of high quality and is generally well accepted. There are, however, some shortcomings with classical test theory. One shortcoming is that item difficulty and item discrimination indices are group dependent; the values of these characteristics depend on the examinee group in which they have been obtained. Another shortcoming is that observed and true test scores are test dependent. Observed and true scores rise and fall with changes in test difficulty. A third shortcoming has to do with the assumption of equal errors of measurement for all examinees. Unfortunately, the ability estimates are in fact less precise both for low and for high ability students than for students of average ability.

During the last decades a new measurement system, item response theory (IRT), has been developed and has become an important complement to classical test theory in the design, construction and evaluation of tests. Within the framework of IRT it is possible to obtain item characteristics which are *not* group dependent, ability scores whic are *not* test dependent and a measure of precision for each ability level.

According to Hambleton et. al. (1991):

> *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. This function specifies that as the level of the trait increases, the probability of a correct response increases. (p 7)*

There are several different IRT models but they all have in common that they use a mathematical function to specify the relationship between observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on the test. In any practical application of latent trait models one must specify the mathematical form of the item characteristic curves and obtain estimates of the item parameters needed to describe the curves. In the three parameter model these parameters are a) item difficulty, b) item discrimination and c) a parameter for guessing. In the two parameter model no guessing is supposed to exist, while in the one parameter model item discrimination is assumed to be the same for all items.

Once a latent trait model is specified, the precision with which it estimates examinee ability can be determined for different ability levels. The information varies with ability level, which makes it possible to determine the standard error of estimate for different ability levels. The item information function gives information of the usefulness of the item in measuring ability at a particular ability level.

Presently IRT is receiving increasing attention from test agencies in test-design, test-item selection, in addressing item-bias and in equating and reporting test scores. The potential of IRT for solving these kinds of problems is substantial. In order to achieve the possible advantages of using an IRT model it is, however, essential that there is a fit between the item response model and the test data of interest.

If IRT model fit can be established, such models would be especially useful to the SweSAT program in equating forms from one administration to the next. IRT equating is less cumbersome to carry out than classical equating procedures and considerably more flexible in its application. It would also be possible to improve the test design and the investigations of item-bias.

The study reported here is a first attempt to investigate the possibility to fit an IRT model to SweSAT data.

The specific purpose of this study is to investigate the fit of two popular IRT models to the student response data from the DS subtest of the SweSAT. Specifically, fit will be investigated by three kinds of evidence: evidence addressing model assumptions, model parameter invariance, and actual fit between particular models and the student response data.

METHOD

Sample

From the 82,506 test-takers in the administration of the SweSAT in Spring 1996 a random sample of three percent was drawn. The resulting sample consisted of 2,461 test-takers. The results of these test-takers on the DS subtest is the data which will be analysed in different ways.

Analyses

Classical item analysis

The classical item analysis of the subtest DS showed a range of p-values from .40 to .81 and a range of biserial correlations from .25 to .70. The reliability coefficient alpha was r = .82.

The range of the biserial correlations indicates that there is a substantial variation in the discrimination of the items in the test. Sometimes though the range may be deceptive because of a couple of ”outliers”. Also, high biserial correlations are sometimes associated with very easy items. These item discrimination indices do not really reveal effective items. The distribution of biserial correlations and the corresponding p-values and vice versa is presented in table 1.

Table 1. Range of p-values and corresponding biserial correlations in the subtest DS

p-values		**corr. r_{bis}**	**r_{bis}**		**corr. p-values**
range	**n**		**range**	**n**	
$\leq$ .50	4	.25 - .51	.25 - .30	2	.40 - .64
.51 - .60	4	.38 - .55	.31 - .40	3	.40 - .80
.61 - .70	7	.28 - .70	.41 - .50	4	.49 - .73
.71 - .80	4	.49 - .69	.51 - .60	11	.40 - .82
$\geq$ .81	3	.40 - .60	.61 - .70	2	.60 - .74

The values in Table 1 give support to the assumption that there really is true variation in the discriminatory power of the items in the test. There does not seem to be any connection between very easy items and high biserial correlations. The conclusion is that there seems to be a need for an item discrimination parameter and, therefore, the one parameter IRT model seems unsuitable for these results.

To make a rough examination of whether guessing was a factor in the test, the test-takers with the lowest results were studied. All test-takers with a total score of less than 8 were selected, giving a total of 193. The results of these test-takers were examined on the eight most difficult items of the test. The resulting p-values for this group of 193 test-takers on these eight items were:

p = .11, .30, .08, .14, .20, .13, .11 and .17.

This result indicated that guessing cannot be excluded and, therefore, the two-parameter-model appears unsuitable.

An assumption common to most IRT models is that the set of test items is unidimensional. There are various methods for determining unidimensionality, as well as different definitions of the concept. The crucial meaning in this context is that only one ability is measured by the items in the test, i.e. all the items in the test measure just one thing in common. According to Hattie (1985) the most widely used index of unidimensionality has probably been coefficient alpha (internal consistency), which was r = .82 for this subtest. There are, however, several problems with the use of alpha as an index of unidimensionality. A more appropriate method for assessing the unidimensionality is factor analysis (Hambleton & Rovinelli, 1986).

Factor analysis

For this sample of 2,641 examinees an unrotated factor analysis resulted in three factors with eigenvalues 4.77, 1.21 and 1.09 respectively. The variance explained by the first factor was 21.7 percent, the variance explained by the second factor was 5.5 percent and the variance explained by the third factor was 5.0 percent. All items, however, had substantial loadings on the first factor (between .24 and .64). A plot of the eigenvalues is shown in Figure 1.

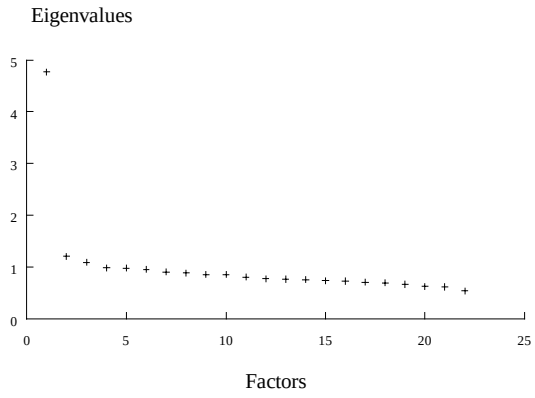

Figure 1. Plot of Eigenvalues

Clearly, there is a dominant first factor, which accounts for 21.7 per cent of the total variance. Even if there may be a second and a third factor they do not seem to be much different from the remaining factors. Assuming a single factor does not seem unreasonable from these results, so the assumption was made that the subtest DS is unidimensional.

The one-parameter logistic model

In spite of the wide range of biserial correlations as well as the clear indication of guessing existing in the test, an attempt was made to fit the one-parameter logistic model to the data. If sufficient fit to the data can be established it is advantageous to use a simple model, as fewer estimations are needed. The resulting approximate chi-square statistics for the goodness of fit,

however, showed that for 20 of the 22 items there was a significant misfit between the one parameter model and the data.

The approximate chi-square values for each item are presented in Table 3, page 8, where a discussion of the usefulness of statistical significance tests in assessing goodness of fit, also may be found.

The three-parameter logistic model

The next step was to estimate three item parameters and the ability parameters for the three-parameter logistic model. A summary of the results with the three-parameter model was that the approximate chi-square-statistics for the goodness of fit were < 10 for 3 items, < 20 for 16 items and < 30 for 20 items. The remaining two items had chi-square statistics of 30.5 and 37.6 respectively. The approximate chi-square statistics for each item are found in Table 3, page 8.

The three-parameter logistic model with item parameters estimated on a smaller sample

To examine the consistency of the item parameter estimates over different samples and sample sizes a random sample of 1000 examinees was drawn from the same population.

A summary of the results was that the approximate chi-square statistics were < 10 for 11 items, < 20 for 19 items and < 30 for all the 22 items. The chi-square statistics for each item are presented in Table 3, page 8.

Plots of estimated item parameters

In Figure 2 a plot is shown of the b-values estimated in the three paramater analyses on the two different sample sizes. As may be seen in Figure 2 a plot of the b-values from the two analyses gave an almost perfect straight line; the correlation between the b-values was r = . 99

In Figure 2 a plot of the a-values from the two analyses may also be found; the correlation between the a-values was r = .95.

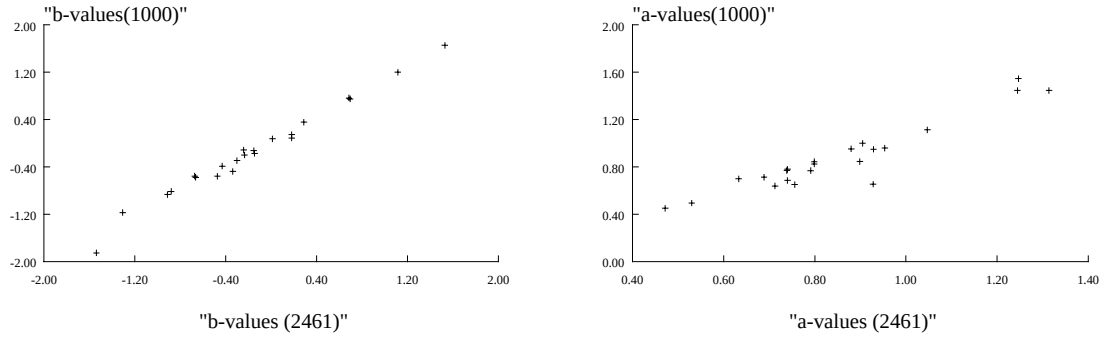

Figure 2. Plots of b-values (left) and a-values (right) estimated on samples of 1000 examinees and 2,461 examinees respectively

The plots in Figure 2 show that a sample of 1000 examinees seems to be sufficient to produce stable item parameter estimates.

In Figure 3 the b-values are plotted against the p-values from the classical item analysis and the a-values are plotted against the biserial correlations from the classical item analysis. The correlation between the b-values, from the three-parameter model and the p-values was r = -.93 (while the correlation between the b-values from the one-parametermodel and p-values was r = -.99). The correlation between the a-values and the biserial correlations was r = .72.

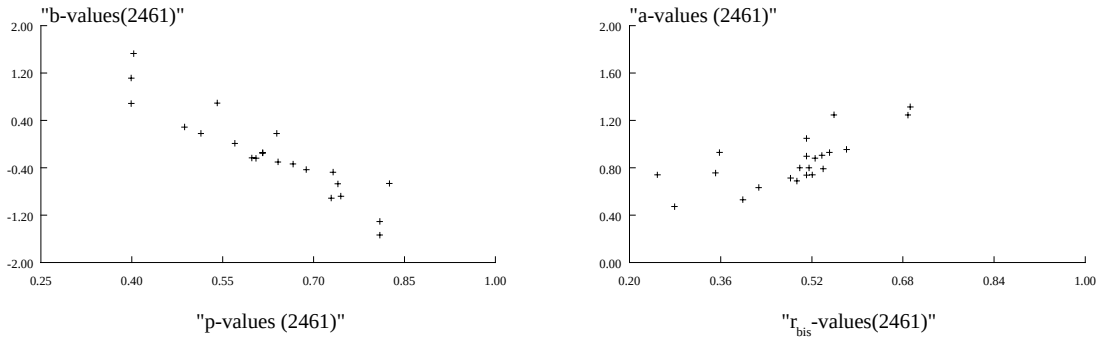

Figure 3. Plots of item parameter estimates from the three parameter model and item indices from classical item analysis

Goodness of fit analysis with 12 ability levels

An assessment of the model-data fit of the three-parameter model was performed with the Resid program (Rogers, 1994). 12 ability levels were chosen and the observed and predicted proportions of examinees in each interval was compared. A summary of the results given by Yen´s chi-square fit statistics is that the chi-square statistics were < 10 for five items, < 20 for 21 items and chi-square = 23.12 for the last item. The Yen´s chi-square fit statistics for each item are given in Table 3, page 8.

Goodness of fit analysis with eight ability levels

In the analysis with 12 ability levels the highest ability level interval was empty and the second highest interval contained only examinees with full scores; 12 ability levels was evidently too many for a test as short as 22 items and with a relatively homogenous score distribution. The same analysis was therefore rerun, but this time with only eight ability levels. A summary of the results given by Yen´s chi-square fit statistics in this analysis was that the chi-square statistics were < 10 for 14 items and < 20 for all 22 items. The chi-square statistics for each item are presented in Table 3, page 8.

Standardized residuals

Residuals provide comparisons of predicted performance results with actual performance results. The raw residuals are the differences between expected and observed performance on an item at a specified performance level. Standardized residuals take into account the sampling errors associated with each expected performance level and the number of examinees at that particular level of performance.

In Table 2 a summary of the standardized residuals from the two goodness of fit analyses is given.

Table 2. Summary of absolute-valued standardized residuals

res.	**12 ab. levels**	**8 ab. levels**
I 0 - 1 I	63.64	72.16
I 1 - 2 I	29.75	22.16
I 2 - 3 I	6.20	5.11
I > 3 I	.41	.57

These standardized residuals are roughly normally distributed (with mean = 0 and standard deviation = 1), which provides strong evidence for the model fit of the three parameter model to the DS data.

CONCLUSIONS

In Table 3 the chi-square statistics for each of the 22 items can be compared between the different analyses performed. The reason that most items get significant misfit with the one-parameter model is probably that an item discrimation parameter as well as a pseudo guessing parameter is needed for these data. The columns of main interest in Table 3 are number two, three, four and five.

Statistical significance tests of model fit should, however, never be the sole reason for rejecting or accepting a model. In Table 3 below the importance of the sample size is clearly demonstrated. When the sample size was 1000 only four of the 22 items were judged as misfitting, while 11 items were judged as misfitting when the sample size was 2,461, even though, as could be seen in Figure 2, the estimates of the item parameters were very similar.

All the same it can be concluded from the analyses performed that the one-parameter model does not fit the data very well. The classical item analysis also suggested the desirability of a three parameter model to fit the data. It can also be concluded, however, that an assumption of unidimensionality seems very reasonable. The evidence for a three parameter model fit is very good.

Table 3. Chi-square statistics from different analyses

Item	1-par	3-par (N=2461)	3-par (N=1000)	Res 12 ab	Res 8 ab
1	9.7	19.1*	7.6	10.5	8.1
2	17.1*	12.6	9.6	14.9	9.5
3	17.8*	20.5*	12.1	13.3	10.1
4	25.9**	30.5**	22.9**	11.4	8.3
5	28.3**	24.6**	21.3**	19.1*	17.6*
6	.8.8	16.0*	11.0	12.1	6.2
7	33.5**	37.6**	23.6*	19.2*	15.4
8	142.8**	25.2*	7.6	16.9*	10.2
9	44.7**	15.9*	6.4	11.9	9.6
10	20.4*	5.5	5.3	7.2	4.6
11	53.1**	12.8	5.7	16.3*	10.6
12	25.9**	8.9	7.2	10.3	2.7
13	52.1**	15.0	12.1	9.6	3.4
14	38.6**	13.4	15.0	10.4	2.5
15	78.8**	26.5**	15.0	23.1*	17.0*
16	51.9**	15.9*	9.7	17.6*	11.8
17	107.0*	12.3	9.6	18.4*	12.0
18	20.3*	14.7	10.0	7.9	6.3
19	38.6**	15.4	8.6	8.9	5.4
20	146.3**	19.1*	17.6*	7.1	4.0
21	27.4**	15.9	14.9	11.2	7.3
22	52.3**	7.4	5.2	12.6	8.6

Even though there seems to be a reasonably good fit between the three parameter model and the test data, there are some problematic items. Items four, five, seven, fifteen and twenty will be studied more closely.

The item response function and item information as well as the item fit for item four are presented in Figure 4.

In Figure 5 the same information is presented for item five, which was another problematic item according to the chi-square statistics in Table 3.

Another problematic item was item seven and the same information about this item is presented in Figure 6.

The next problematic item was item 15 and the same information about this item is presented in Figure 7.

The final problematic item according to the chi-square statistics was item 20 and the information about this item is presented in Figure 8.

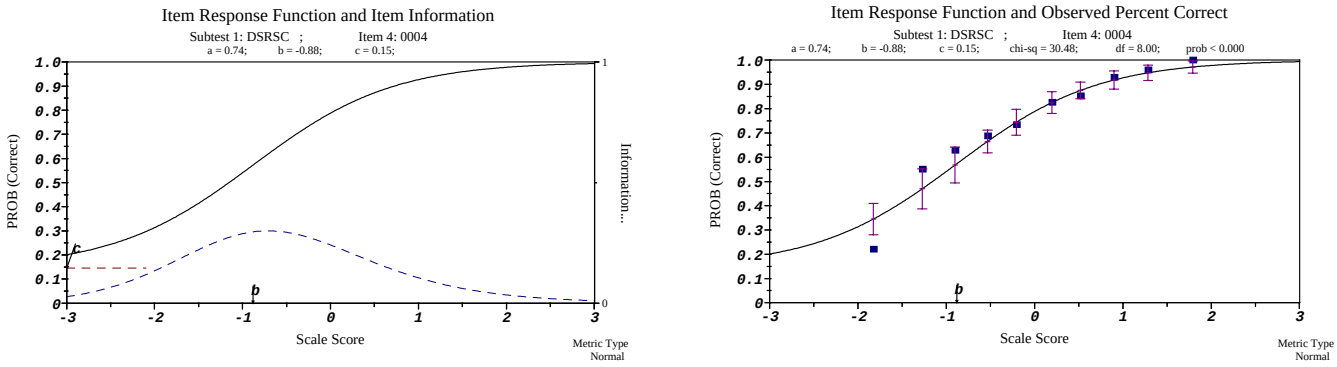

Figure 4. Item response function, item information and item fit for item four in the DS subtest

It may be seen from Figure 4 that item four provides the most information about examinees somewhat below average ability. The main misfit seems to be for low ability examinees, while the fit for examinees above the average seems fairly good.

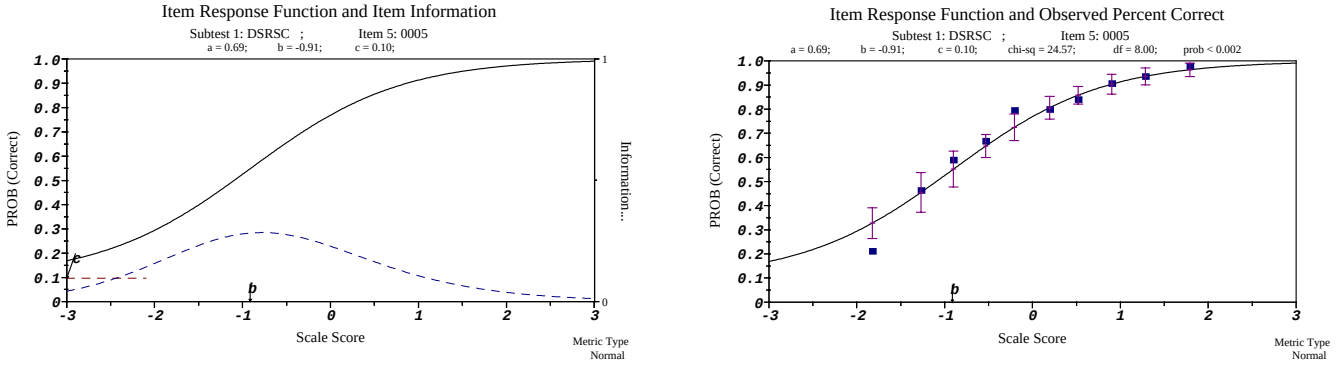

Figure 5. Item response function, item information and item fit for item five in the DS subtest

Item five is very similar to item four; the most information is given for examinees who are about one standard deviation below the mean, and the greatest misfit is also found for examinees below average ability.

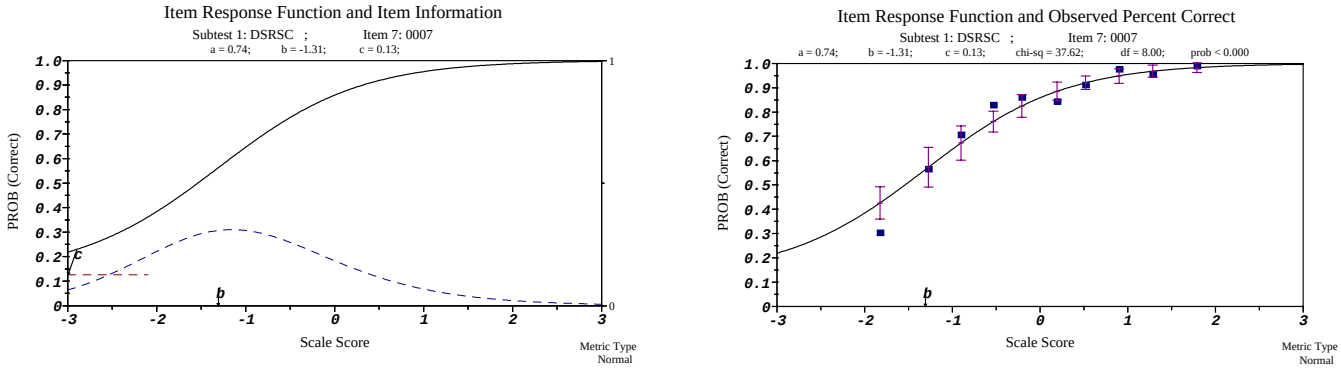

Figure 6. Item response function, item information and item fit for item number seven in the DS subtest

Item seven has the same general appearance as items four and five. The main information is about low ability examinees and the misfit is greatest below the average.

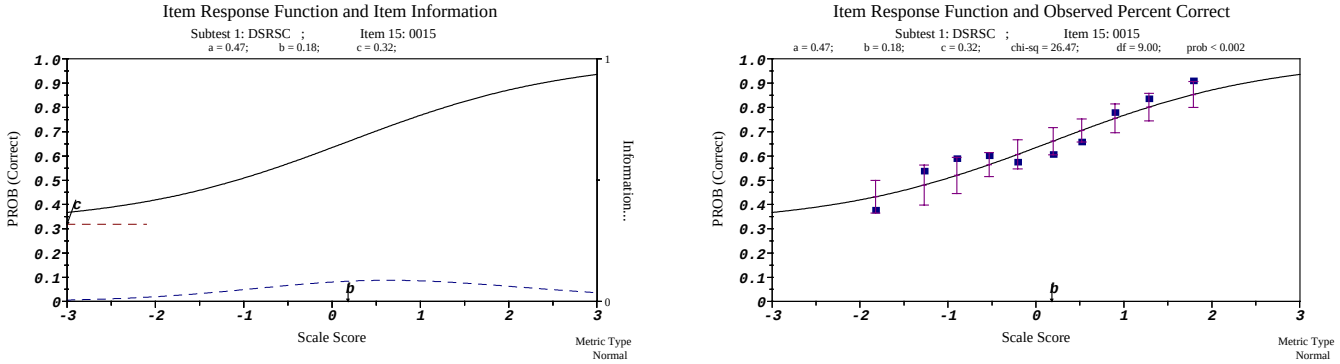

Figure 7. Item response function, item information and item fit for item 15 in the DS subtest

Item 15 is quite different from items four, five and seven. This item seems to be a very poor item; the information is low on all ability levels and the discriminatory power is also very low. The misfit of the item also covers the whole ability range. This item had a low biserial correlation in the classical item analysis as well. The conclusion is that this item should never have been included in the test.

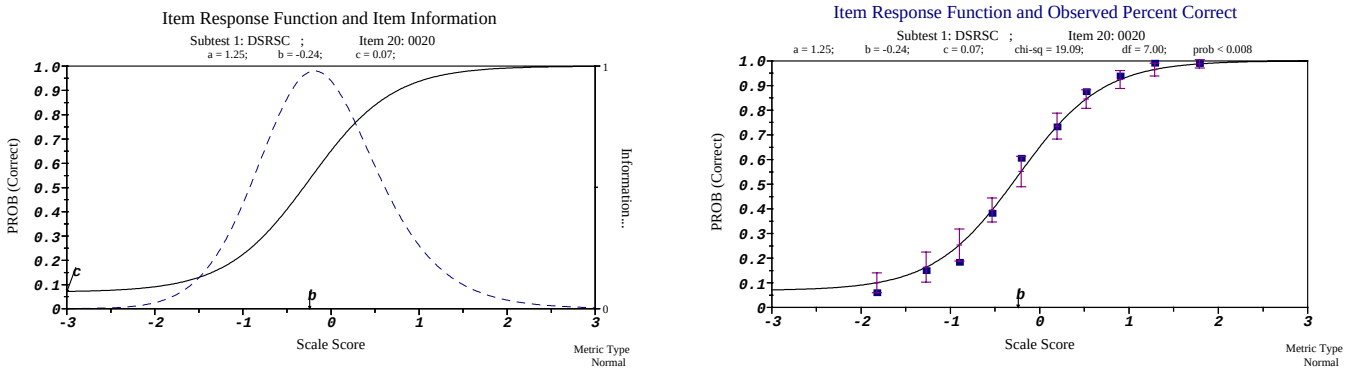

Figure 8. Item response function, item information and item fit for item 20 in the DS subtest

Item 20 again looks quite different from the other problematic items. The information provided by this item is very high and it is concentrated just below the mean of the ability scale. The discriminatory power of the item is consequently also very high. The fit at the higher ability levels seems to be reasonable and it is mainly in the low end of the ability scale that the misfit is existing.

In Figure 9, finally, the total test information function is presented. The information given by a test at different ability levels is the sum of the item information functions at the same abilities. As may be seen from Figure 9 the standard error is inversely related to the information at each ability level, i.e. the standard error is different at different ability levels (in contrast to the standard error of measurement in classical test theory, which is assumed to be the same for all score levels).

The reliability index was r = .84 which may be compared with coefficient alpha in the classical analysis which was r = .82. It may be seen from the test information function, however, that the errors are much smaller around average ability than for low or high ability levels, but this finding is to be expected with the current approach to test design.

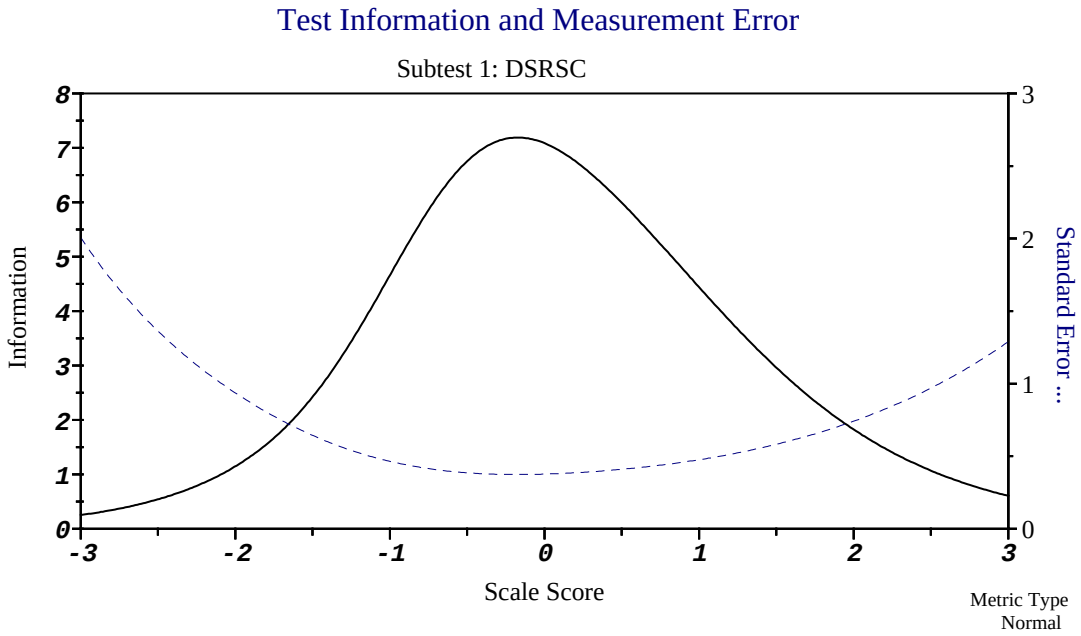

Figure 9. Test information function for the subtest DS

The results from this first attempt to fit an IRT model to response data from the SweSAT are, on the whole, very promising. The model assumption regarding unidimensionality seems to be met to a reasonable degree at least in this subtest. There seems to be item differences regarding discrimination power and there also seems to be guessing going on in the subtest, making one parameter and two parameter models unsuitable. A three parameter logistic model, however, seems to fit the data very well. The few items which did not fit this model very well, did not seem to be any serious threat to model usefulness or validity.

The next step will be to investigate the remaining four subtests in the SweSAT in the same way. If all five subtests may be fitted to the same model, the test as a whole will be analysed as a final step before it can be decided whether the SweSAT program may be improved by the use of IRT.

REFERENCES

Hambleton, R. K. & van der Linden, W. J. (1982). Advances in Item Response Theory and Applications: An Introduction. *Applied Psychological Measurement, 6,* pp. 373-378.

Hambleton, R. K. (1983) (Ed.) *Applications of Item Response Theory*. Educational Research Institute of British Columbia.

Hambleton, R. K. & Cook, L. L. (1983).Robustness of Item Response Models and Effects of Test Length and Sample Size on the Precision of Ability Estimates. In Weiss, D. (Ed.), *New Horizons in Testing.* pp. 31-49. New York: Academic Press.

Hambleton, R. K. & Rovinelli, R. J. (1986). Assessing the Dimensionality of a Set of Test Items. *Applied Psychological Measurement*,10, 2. pp.287-302.

Hambleton, R. K. & Rogers, H. J. (1990). Using Item Response Models in Educational Assessments. In Schreiber, W. H. & Ingenkamp, K. (Eds.), *International Developments in Large-Scale Assessment.* pp. 155-184. England: NFER-Nelson 334.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In Linn, R. L. (Ed.), *Educational Measurement* (3rd ed.: pp. 147-200) New York, Macmillan.

Hambleton, R. K. & Rogers, H. J. (1990). Using Item Response Models in Educational Assessments. In Schreiber, W. H. & Ingenkamp, K. (Eds.), *International Developments in Large-Scale Assessment* (pp. 155-184). England: NFER-Nelson 334.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury: Sage.

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

Hattie, J. (1985). Methodology Review: Assessing Unidimensionality of Tests and Items. *Applied Psychological Measurement*. 9, 2. pp. 139-164.