THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SweSAT

A Study of the READ Subtest

Christina Stage

Em No 25, 1997

ISSN 1100-696X
ISRN UM-PED-EM--25--SE

This study is the fourth in a series of studies with the aim of fitting an IRT model to the Swedish Scholastic Aptitude Test (SweSAT).

The 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 that particular 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.

Ever 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 the classical test theory.

During the last decades a new measurement system, item response theory (IRT), has been developed and it has become an important complement to classical test theory in the design, construction and evaluation of tests. The potential of IRT for solving different kinds of test problems is substantial. It is essential, however, in order to achieve the possible advantages from an IRT model, that there is fit between the model and the test data of interest.

In earlier studies attempts were made to fit IRT models to the DS, DTM and ERC subtests (Stage, 1996, 1997a and 1997b). The conclusion from those studies was that the three parameter logistic IRT model fitted the data reasonably well.The specific purpose of this study is to investigate the fit of the same model to the READ subtest.

METHOD

Sample

In the DS, DTM and ERC studies a random sample of three percent of the 82,506 examinees who took part in the SweSAT in spring 1996 was used. The same random sample has been used in this study. This sample consisted of 2,461 testtakers; 1,349 females and 1,112 males. The results of these examinees on the READ subtest are the data which will be analysed in different ways.

Classical item analysis

The classical item analysis of the READ subtest gave a range of p-values from .34 to .84 and a range of biserial correlations from .21 to .45. The mean of the test was M = 12.15 and the standard deviation was s = 3.51. The reliability, coefficient alpha, was r = .68.

The range of biserial correlations indicates that there is a substantial variation in the discrimination power of the items in the test. Sometimes, though, the range may be deceptive because of a couple of ”outliers”. Moreover high biserial correlations are sometimes associated with very easy items. These discrimination indices do not really reveal effective items. In Figure 1 the p-values are plotted against the biseral correlations for the 20 items.

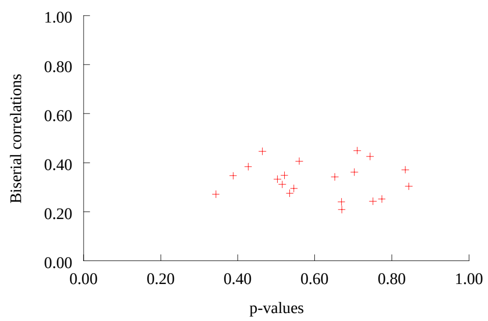

Figure 1. Biserial correlations plotted against p-values of the 20 items in the READ subtest.

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

To examine whether guessing had taken place in the test, the testtakers with the lowest results were studied. All testtakers with a total score less than eight of the 20 possible points were selected. This gave a number of 379 examinees. The results of these 379 examinees on the most difficult items in the test were studied. Four items had p-values lower than .50 and the p-values of this poor group on these four difficult items were:

p = .17 .16 .12 .15

This result indicated that guessing can hardly be excluded, and therefore a two parameter model also appears unsuitable to fit the data.

Factor analysis

An assumption common to all IRT models is that the set of test items is unidimensional. A crude measure of unidimensionality is coefficient alpha, as this coefficient is a measure of the internal consistency of the items in a test. The coefficient alpha was r = .68 for this subtest. A more appropriate method, however, for assessing the unidimensionality of a test is factor analysis (Hambleton & Rovinelli, 1986).

For this sample of 2,461 examinees an unrotated factor analysis resulted in five factors with eigenvalues 2.9, 1.1, 1.0, 1.0 and 1.0 respectively. The variance explained by the first factor was 14.5 percent, the variance explained by the second factor was 5.7 percent, by the third factor 5.2 percent, by the fourth factor 5.1 percent and the variance explained by the fifth factor was also 5.1 percent. A plot of the eigenvalues is shown in Figure 2.

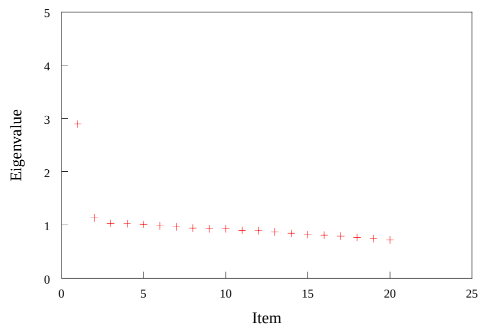

Figure 2. Plot of eigenvalues for the READ subtest.

In Figure 2 it is shown that there is a dominant first factor in the subtest and according to Hambleton and Rovinelli (1986):

> *The number of ”significant” factors is determined by looking for the ”elbow” in the plot. The number of eigenvalues to the left of the elbow is normally taken to be the number of significant factors underlying test performance. (p 289)*

All but five of the 20 items have loadings larger than .30 on the first factor and no item had a loading below .20. Even though it would have been better if the amount of variance explained by the first factor had been greater it is not implausible or unreasonable to assume a single factor with the test data.

The three parameter logistic IRT model

An attempt was made to fit the results of the READ subtest to the three parameter logistic IRT model by means of the program BILOG (Mislevey & Bock, 1990). When the number of items is 20 or greater, approximate chi-square statistics for the goodness of fit of each item are included as output of the program. For this purpose, the cases in the calibration sample are sorted into successive intervals of the latent continuum according to the estimates of their ability rescaled to mean = 0 and standard deviation = 1. This gives a reasonable test of fit if the number of items is large enough to make an assignment of cases accurate, and if the sample size is large enough to retain three or more intervals.

In this study the number of items was exactly 20 which is the lower limit and which might in fact be on the low side. The sample of examinees, on the other hand, was large. The number of intervals used was ten for most of the items, and nine for six items.

The outcome of the goodness of fit analysis was that for nine of the items in the subtest there was a model data misfit which was significant at .05 level and for seven of these nine items the misfit was significant at .01 level. The chi-square statistics of each item are presented in Table 2 (p 9).

The reliability index reported was r = .72, which is somewhat higher than the coefficient alpha on the same test.

Goodness of fit analysis with eight ability levels

Another goodness of fit analysis was made by means of the program RESID (Rogers, 1994). In carrying out this analysis, examinees were first sorted into ability categories. The number of ability levels was specified to eight and the observed proportions of examinees in each ability category, answering the item correctly, were calculated. Expected proportions correct for each ability interval were obtained by computing the probability of success on the item on each ability level. Residual values (observed - expected) and standardized residuals were then computed. The program also contains chi-square fit statistics as output.

The outcome of the RESID analysis was that for 15 items the differences between observed and predicted results were insignificant. For five items the differences were significant at .05 level and for two of these five items the differences were significant at .01 level. The chi-square statistics of each item are presented in Table 2 (p 9).

Residuals provide a comparison between predicted and actual performance. Raw residuals are the differences between expected and observed performance on an item at a specified performance level. Standardized residuals (SRs) take into account the sampling error associated with each performance level as well as the number of examinees at that particular level of performance. When the model fits the data the SRs might be expected to be small and randomly distributed about 0. Within the framwork of regression theory it is common to assume that the distribution of SRs is approximately normal. In Table 1 a summary of the SRs from this goodness of fit analysis is given.

Table 1. Summary of absolute-valued standardized residuals.

residuals	**number**	**percent**
| 0-1 |	110	68.75
| 1-2 |	39	24.38
| 2-3 |	11	6.88
| >3 |	0	0.00

The results in Table 1 show that the distribution of SRs is very close to the normal distribution which is strong support for model data fit.

Comparison between estimated IRT parameters and item indices from classical test theory

In IRT the parameter, b, is an item difficulty parameter, which in the one parameter model is the point on the scale where the probability of a correct response to an item is 0.5. In a three parameter model, where a pseudo guessing parameter (c) is included, the b parameter is the value where the probability for a correct response is 0.5 + c. The correlation between the b-values estimated by the three parameter logistic IRT model and the p-values achieved by classical test theory was r = -0 .98. In Figure 3 the estimated b-values are plotted against the p-values.

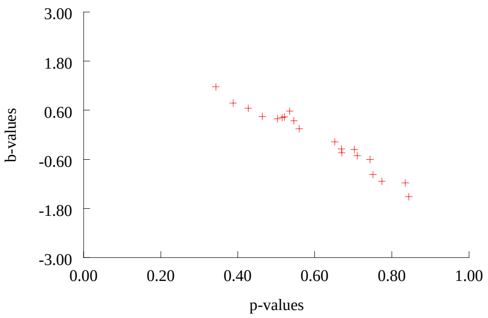

Figure 3. Estimated b-values plotted against p-values.

On the whole the estimated difficulty parameters seem to correspond very well to the the observed difficulty indices of the items, there are no deviating items.

In IRT the discrimination parameter is called the a-parameter. The a-parameter is proportional to the slope of the item characteristic curve at the point b on the ability scale. The usual range for item discrimination parameters is (0 - 2). A plot of the estimated a-values and the corresponding biserial correlations is found in Figure 4.

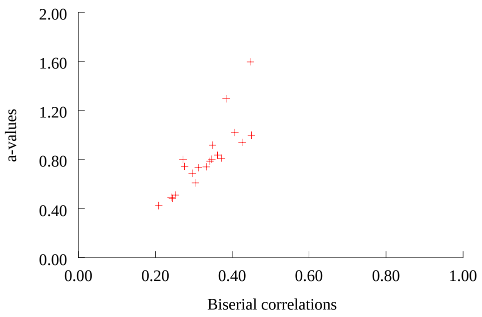

Figure 4. Estimated a-values plotted against corresponding biserial correlations.

The items deviating most are item 15 with SR = 2.52, item four with SR = 2.06, item 13 with SR = -1.33, item 19 with SR = -1.24 and item three with SR = 1.20, all the other items have SR:s less than 1.0. The correlation between the estimated a-values and corresponding biserial correlations was r = .84. There seems to be a reasonable correspondance between the estimated discrimination parameters and the oberved discrimination indices as well.

The correlation between the observed test scores and the abilities estimated by the IRT model was r = .96.

Parameters estimated separately for males and females

One feature of IRT which is regarded as very valuable is the invariance of item parameters. In the classical test theory item difficulty and discrimination indices are dependent on the group in which they have been obtained. In IRT, on the other hand, the item parameters are invariant across ability subpopulations. This invariance only holds, however, when there is good fit of the model to the data. It is important to determine whether invariance holds, since all application of IRT capitalizes on this property. If two samples of different ability are drawn from the same population and item parameters are estimated in each sample, the congruence between the two sets of estimates of each item parameter can be taken as an indication of the degree to which invariance holds. The degree of congruence can be assessed by the correlation between the two sets of estimates of the item parameters or by studying the corresponding scatter plot. (Hambleton, Swaminathan & Rogers, 1991).

The sample used in this study was divided into males and females, giving two samples of respectively 1,112 and 1,349. These samples were run through BILOG, which gave separate parameter estimates for the two groups. In Figure 5 the difficulty parameters b estimated on the female group are plotted against the same item parameters estimated on the male group.

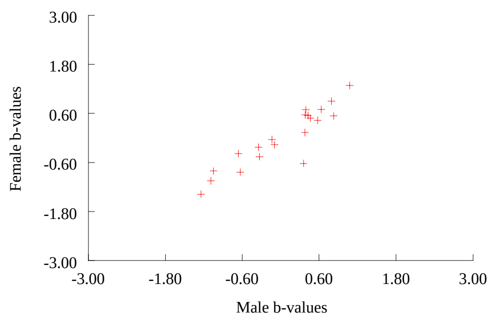

Figure 5. The b-values estimated on the female group plotted against b-values estimated on the male group.

If the estimates are invariant, the plots for the subgroups should be linear with the amount of scatter reflecting errors due to measurement errors and sampling. The correlation between b-values estimated on male and female examinees is r = .92. The mean for males on the subtest was M = 12.09 and for females M = 12.21. Since the difficulty estimates based on the two samples lie on a straight line, with some scatter, it can be concluded that the invariance property of the item parameters holds. Some degree of scatter can be expected because of the use of samples; a large amount of scatter would indicate lack of invariance which might be caused either by model data misfit or poor item parameter estimation (Hambleton, Swaminathan & Rogers, 1991). Only two items had SRs larger than 1; item 12 with SR = -3.24 and item 18 with SR = 1.12.

In Figure 6 the a-values estimated on female examinees are plotted against the a-values estimated on male examinees.

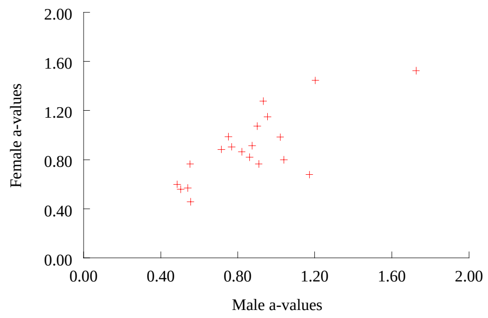

Figure 6. The a-values estimated on female examinees plotted against the a-values estimated on male examinees.

The discrimination parameters also seem to be very congruent even though they are estimated on different samples. The correlation between a-values estimated on the two groups was r = .76. This invariance of parameter estimations can be taken as strong support for model data fit. For five items (4, 8, 12, 16 and 18) the SRs were larger than 1.

The goodness of fit analysis for the female results gave as result that for eight items there was a significant model data misfit at .05 level and for six of these items there was a misfit at .01 level. For males there was a significant misfit for three items at .01 level. The chi-square statistics for each item for both males and females are presented in table 2 (p 9).

Statistical goodness of fit analyses

In Table 2 the outcome of the statistical goodness of fit analyses is presented. In the second column the chi-square values for each item from the BILOG test are given. In column three the chi-square values for each item from the RESID analysis are given; in column four the outcome from the BILOG test of the parameters estimated on the female group is presented and in column five the chi-square values from the BILOG test of parameters estimated on the male group are given.

Table 2. The Chi-square statistics from different analyses and for different samples.

Item	$IRT_{total\ group}$	$RESID_{total\ group}$	$IRT_{females}$	IRT_{males}
1	14.50	3.96	9.00	11.80
2	17.20*	12.63*	22.30**	8.20
3	13.60	17.56**	11.70	8.30
4	16.50	4.87	7.80	10.60
5	21.90**	7.12	17.60**	5.90
6	24.10**	3.24	12.90	14.30
7	16.30	4.08	11.50	9.20
8	15.20	13.16*	10.00	8.30
9	29.80**	6.42	26.60**	11.20
10	21.70*	3.99	8.50	22.80**
11	10.20	5.51	6.80	5.80
12	8.70	10.75	12.10	4.20
13	51.30**	12.91*	38.40**	20.70**
14	16.80	3.97	12.20	6.50
15	39.60**	16.80**	23.00**	13.50
16	22.00**	4.02	22.30**	11.60
17	13.40	3.51	20.30*	14.80
18	11.20	9.82	14.50	5.40
19	33.00**	3.47	19.50*	25.90**
20	13.00	9.75	8.40	13.80

The difference between the results from the BILOG and the RESID analyses may be caused by the fact that the number of ability levels differs. In the Bilog test ability intervals of ten or nine points are used, in the RESID test all ability intervals are eight points. Also, 20 items may be too few to get stable ability estimates, and this could explain why there was a misfit for several items. There is also a noticeable difference between the number of items with significant misfit when the parameters are estimated on the whole group in comparison with separate male and female groups and the reason for this is probably the difference in sample sizes.

Statistical tests have a wellknown 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)*

Almost any departure from the model under consideration will lead to rejection of the null hypothesis of model data fit if the sample size is sufficiently large. If, on the other hand, sample sizes are small, even large model data discrepancies may not be detected due to the low statistical power associated with the significance tests (Hambleton, Swaminathan & Rogers, 1991).

Hambleton et al. (1991) give the following recommendations regarding assessment of model data fit:

> *In assessing model-data fit, the best approach involves a) designing and conducting a variety of analyses designed to detect expected types of misfit, b) considering the full set of results carefully, and c) making a judgment about the suitability of the model for the intended application. Analyses should include investigations of model assumptions, of the extent to which desired model features are obtained, and of differences between model predictions and actual data. Statistical tests may be carried out, but care must be taken in interpreting the statistical information. The number of investigations that may be conducted is almost limitless. (p 74)*

Graphical model data fit

The item response curves estimated by BILOG for the 20 items in the READ subtest are presented in Figure 7 to Figure 26. In the same Figures the item information functions for each item is included. Item information functions display the contribution items make to ability estimation at points along the ability continuum. The size of this contribution depends, to a great extent on an item´s discrimination power. Where on the ability scale that the information contribution of the item is realized depends on the item´s difficulty. To the right in Figures 7 to 26 there is a representation of the model data fit for each item.

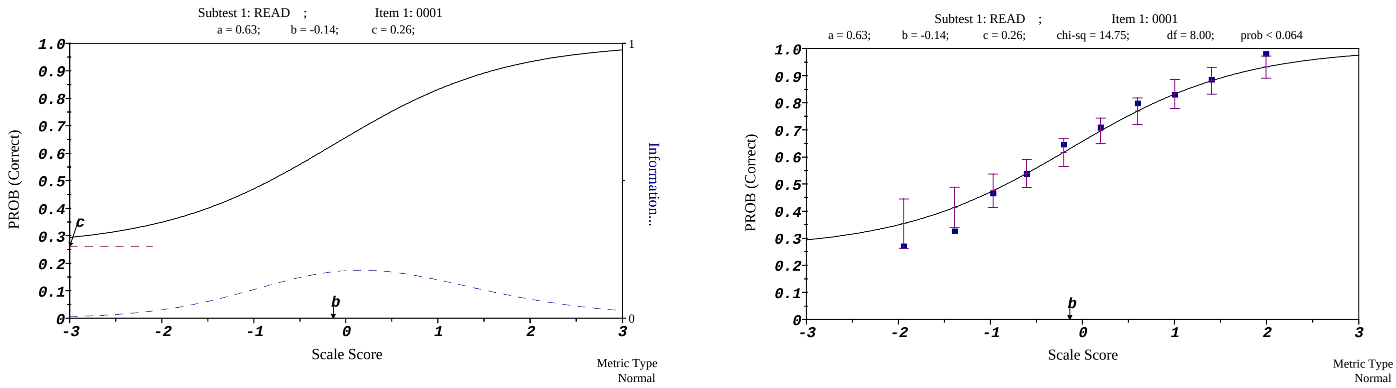

Figure 7. Response curve, information and model fit for item one in the READ subtest.

As may be seen in Figure 7, item one is unproblematic. There is no significant misfit, the discrimination is acceptable and the main information is located at medium difficulty level.

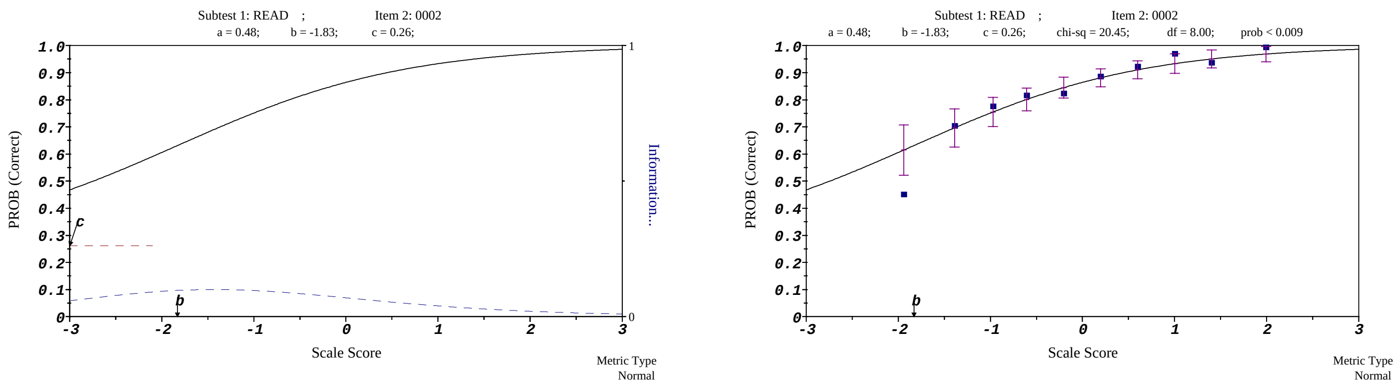

Figure 8. Response curve, information and model fit for item two in the READ subtest.

Item two, on the other hand, does not contribute much to the test. There is significant misfit according to the Bilog test as well as the Resid test. The misfit seems to be mainly at low ability levels, but that is where the information is located as well. On the whole this item is rather useless.

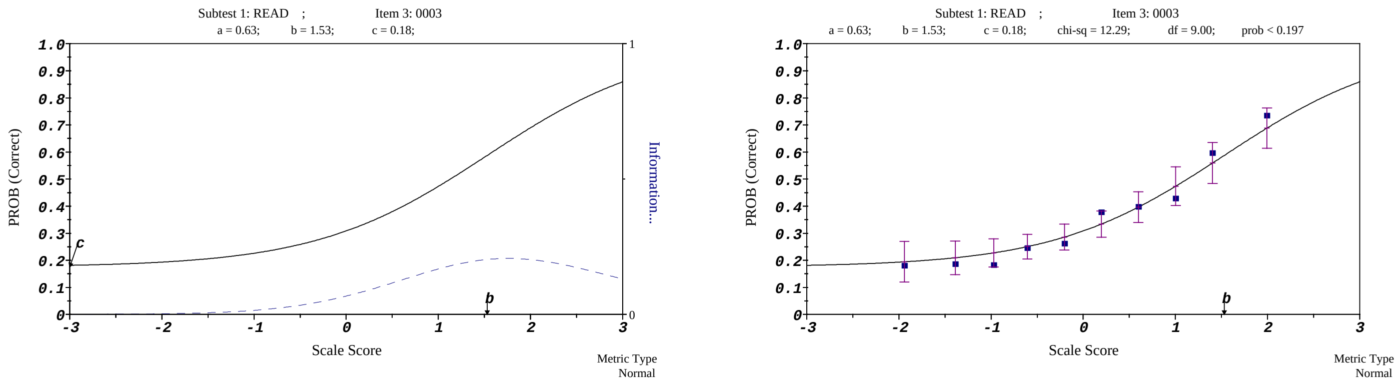

Figure 9. Response curve, information and model fit for item three in the READ subtest.

For item three there was significant misfit according to the Resid test but not the Bilog test. This is a very useful item, however, since the information is located at the very high ability levels, where it is most needed.

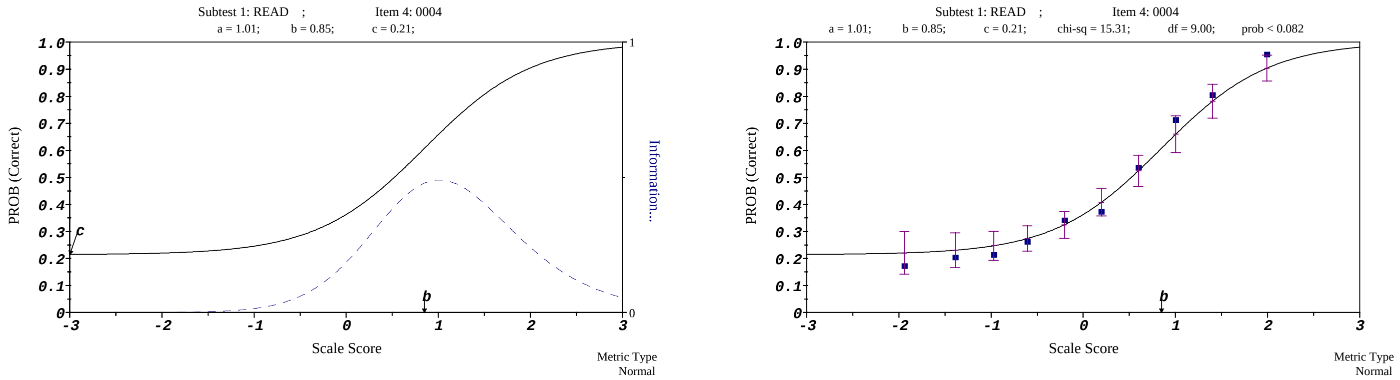

Figure 10. Response curve, information and model fit for item four in the READ subtest.

Item four seems to be a very good item; there is no significant misfit, there is good discrimation power and the information is given at high ability levels.

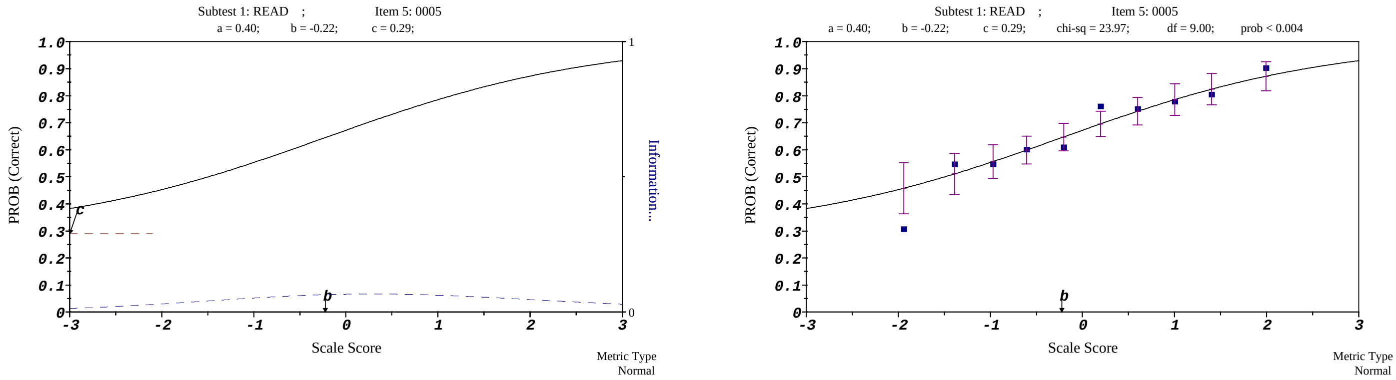

Figure 11. Response curve, information and model fit for item five in the READ subtest.

Item five again is a more problematic item. There is significant misfit according to the Bilog test, but not the Resid test. The misift seems to be mainly at low ability levels, but also the information provided is very poor.

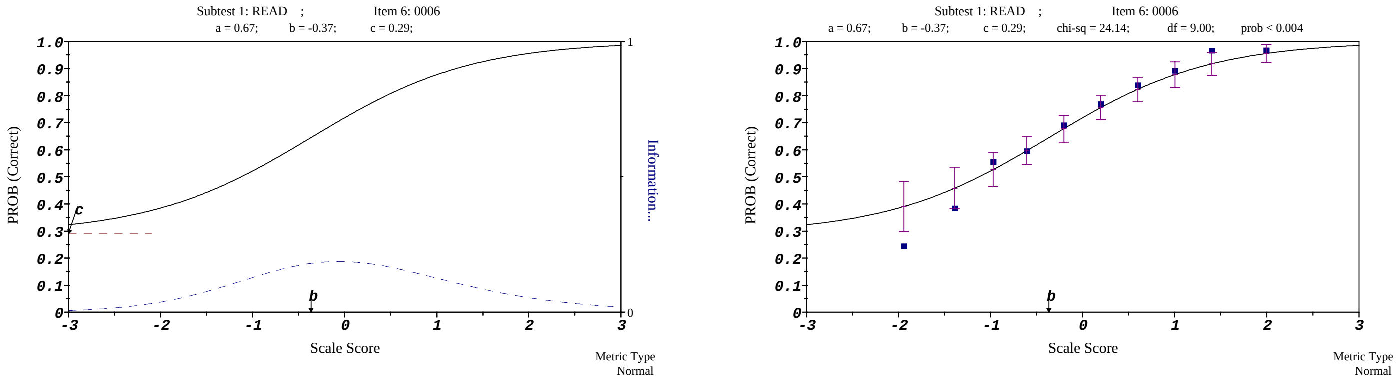

Figure 12. Response curve, information and model fit for item six in the READ subtest.

For item six also there is also significant misfit according to the Bilog test but not the Resid test. The misfit seems to be at very low ability levels, however, and since the information is acceptable and located at the medium ability level, the item may be regarded as fairly useful.

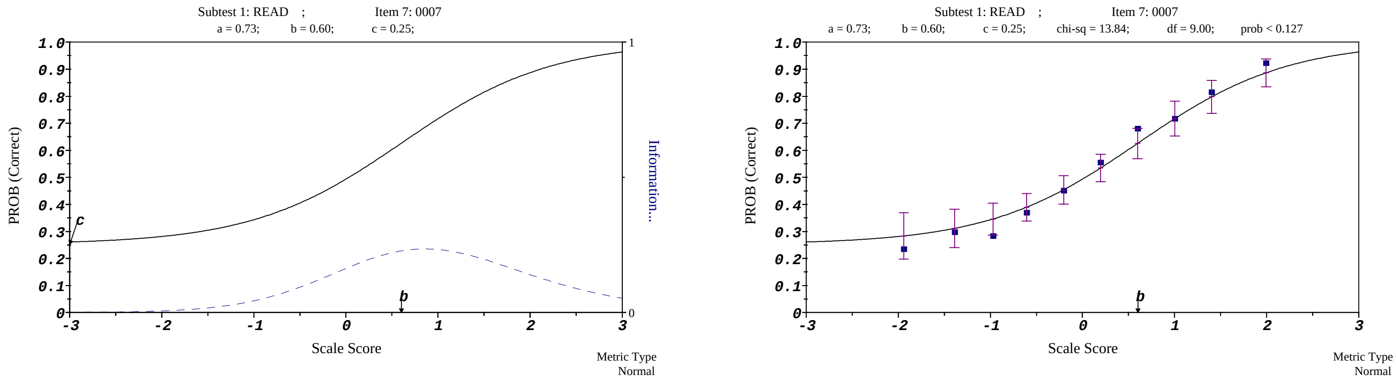

Figure 13. Response curve, information and model fit for item seven in the READ subtest.

Item seven seems to be a very good one. There is no significant misfit; the information is fairly good and it is located above medium ability.

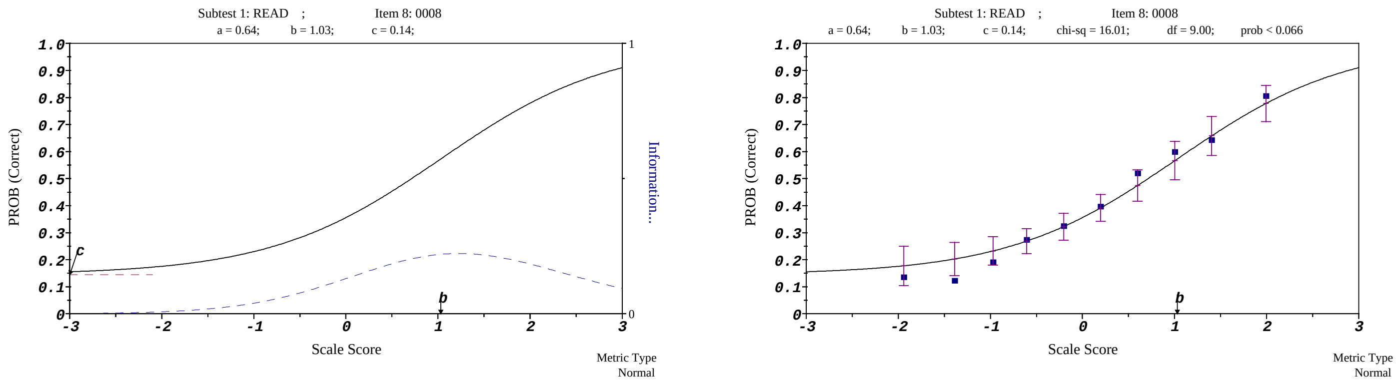

Figure 14. Response curve, information and model fit for item eight in the READ subtest.

For item eight there is significant misfit at .05 level according to the Resid test. As may be seen to the left in Figure 14 the misfit seems to be mainly at low ability levels and since the information is provided at high ability levels there is no big problem.

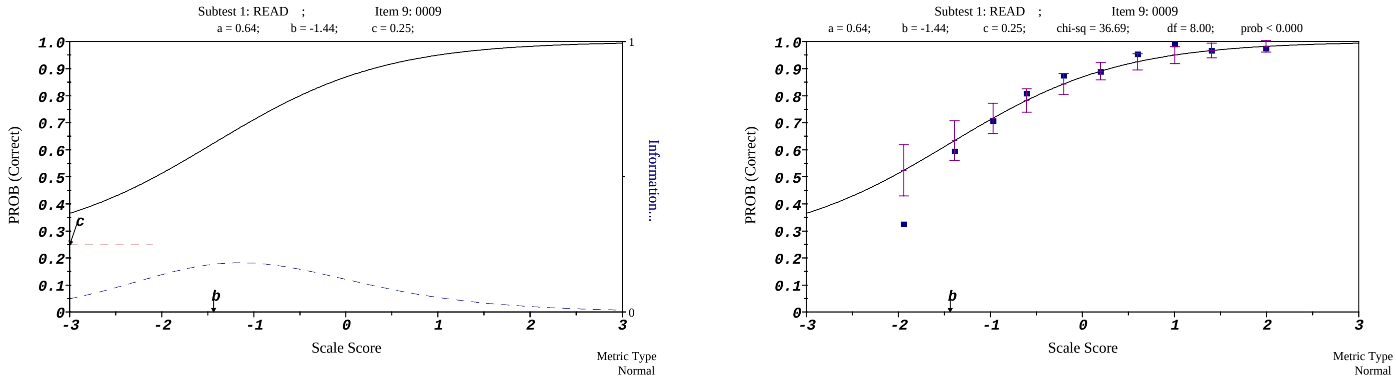

Figure 15. Response curve, information and model fit for item nine in the READ subtest.

For item nine there is significant misfit according to the Bilog test and the misfit seems to be at low ability levels. As the main information is also at low ability levels this item is not very useful.

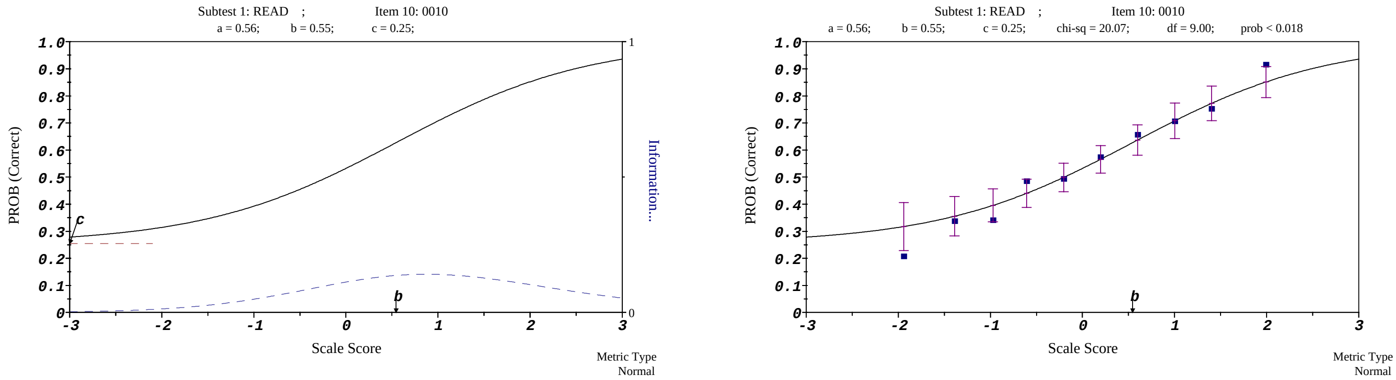

Figure 16. Response curve, information and model fit for item 10 in the READ subtest.

For item 10 there is again significant misfit according to the Bilog test. For this item the misfit seems to be both at very high and very low ability levels, but not in between. The information is rather poor, but located at a useful level somewhat above medium ability.

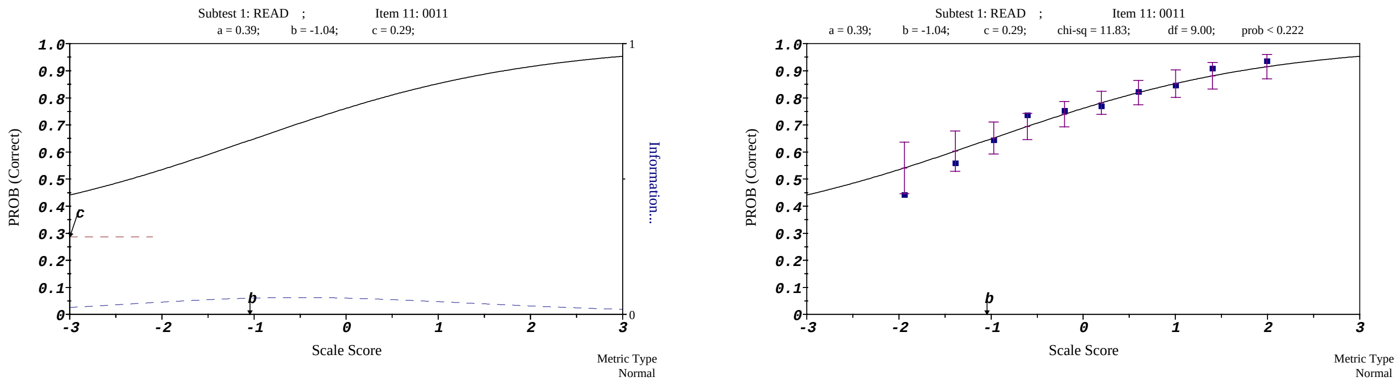

Figure 17. Response curve, information and model fit for item 11 in the subtest READ.

For item 11 there is acceptable model data fit according to all analyses. The discrimination and hence the information is very poor, however.

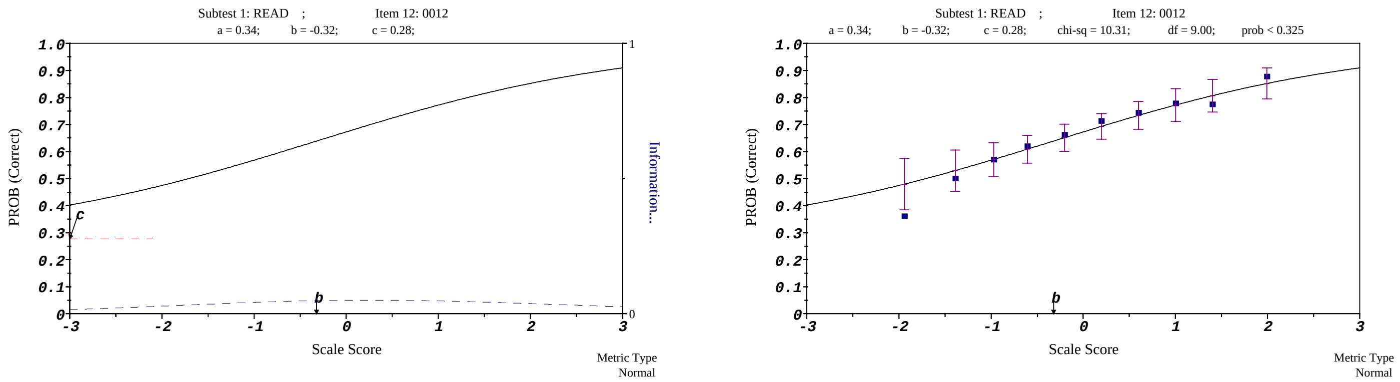

Figure 18. Response curve, information and model fit for item 12 in the subtest READ.

Item 12 is very similar to item 11. There is no significant model data misfit, but the information is poor.

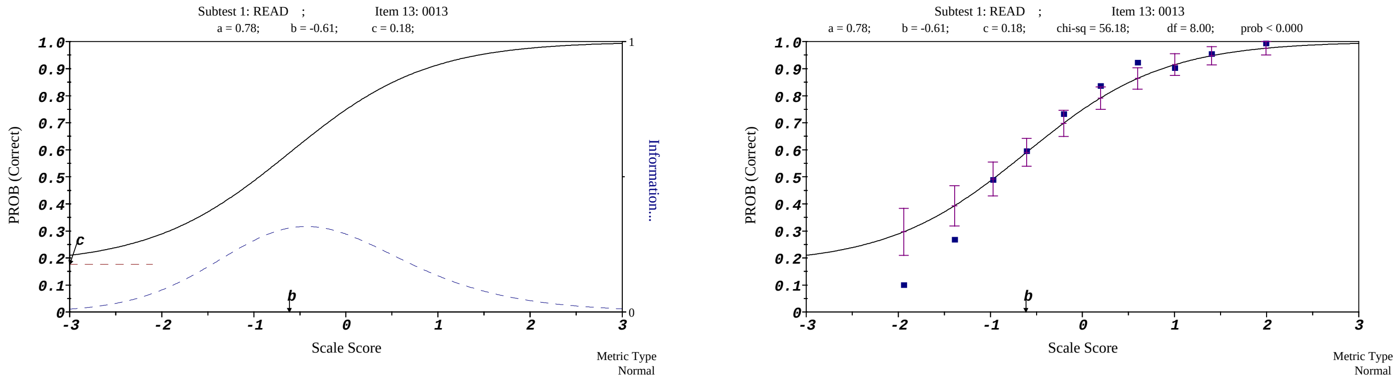

Figure 19. Response curve, information and model fit for item 13 in the READ subtest.

For item13 there is significant misfit according to all analyses. As may be seen in Figure 19 the misift seems to be very serious only at the very low ability levels, however.

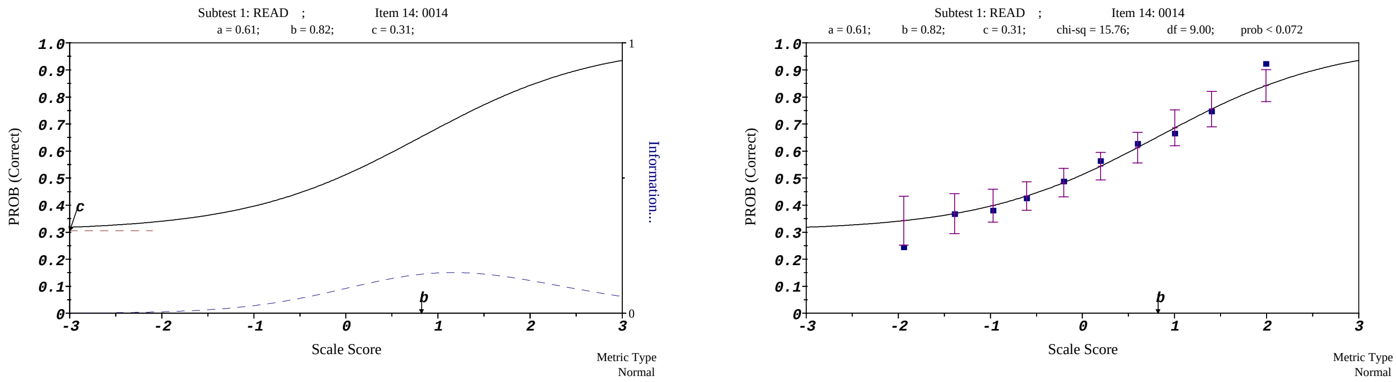

Figure 20. Response curve, information and model fit for item 14 in the READ subtest.

For item 14 the model data fit is acceptable according to all analyses. The information is not very high, but it is located at high ability levels, where most needed.

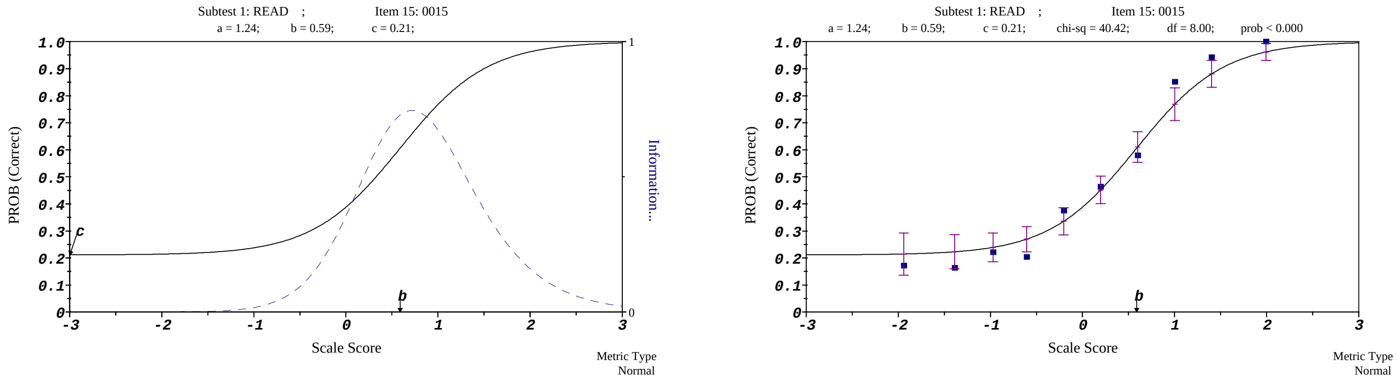

Figure 21. Response curve, information and model fit for item 15 in the READ subtest.

For item 15 there is significant misfit according to the Bilog test as well as the Resid test. As amy be seen in Figure 21 the misfit is very bad only at the highest ability levels though.

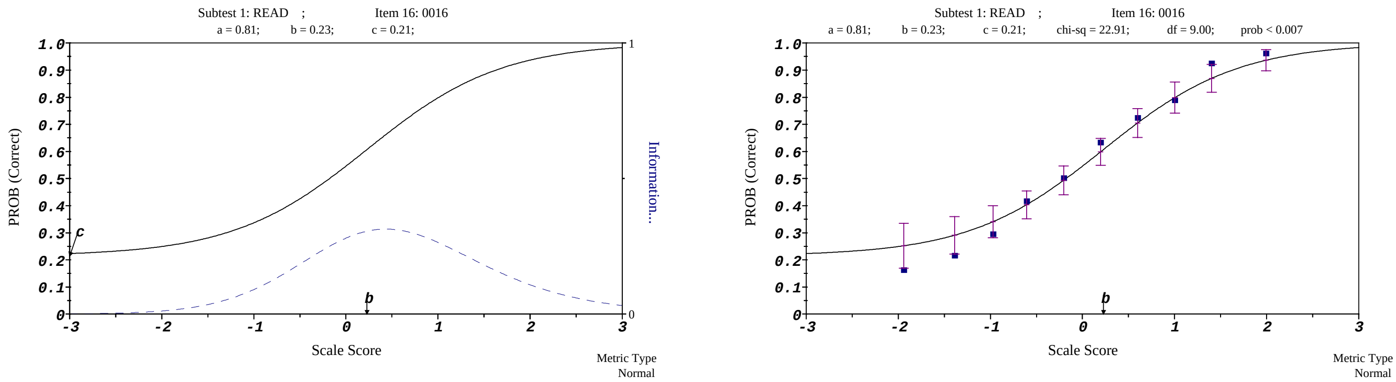

Figure 22. Response curve, information and model fit for item 16 in the READ subtest.

For item 16 there is significant misfit according to the Bilog test but the misfit is serious only at the very low ability levels. The information is provided mainly just above medium ability.

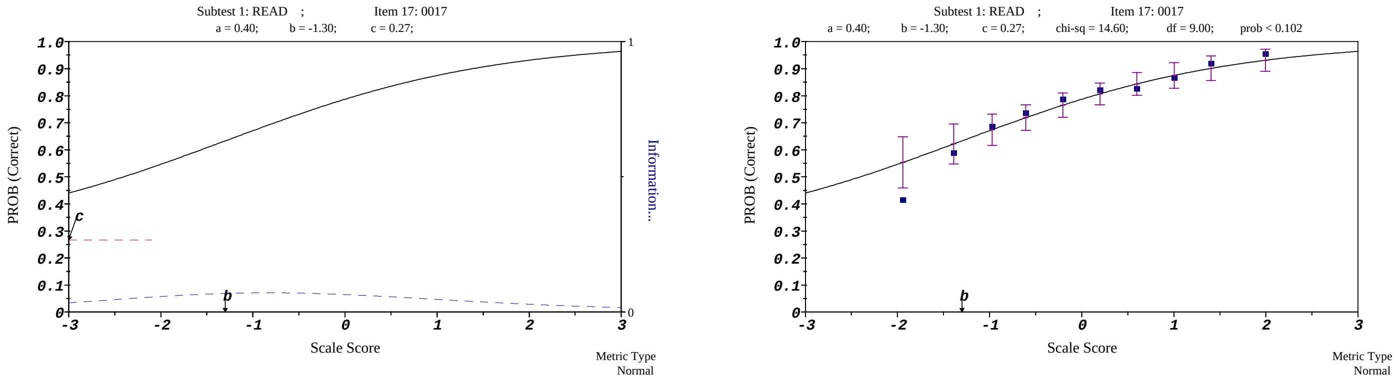

Figure 23. Response curve, information and model fit for item 17 in the READ subtest.

For item 17 there is significant misfit for females and the misfit seems to be located at low ability levels. This item has poor discrimination power and information, however.

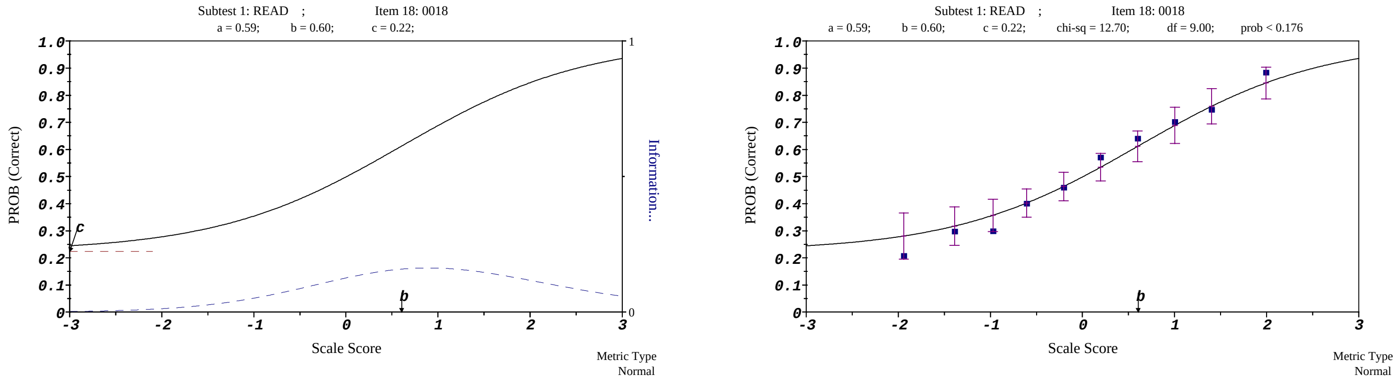

Figure 24. Response curve, information and model fit for item 18 in the READ subtest.

For item 18 the model data fit is acceptable; the information is not very high, but it is mainly located above medium ability.

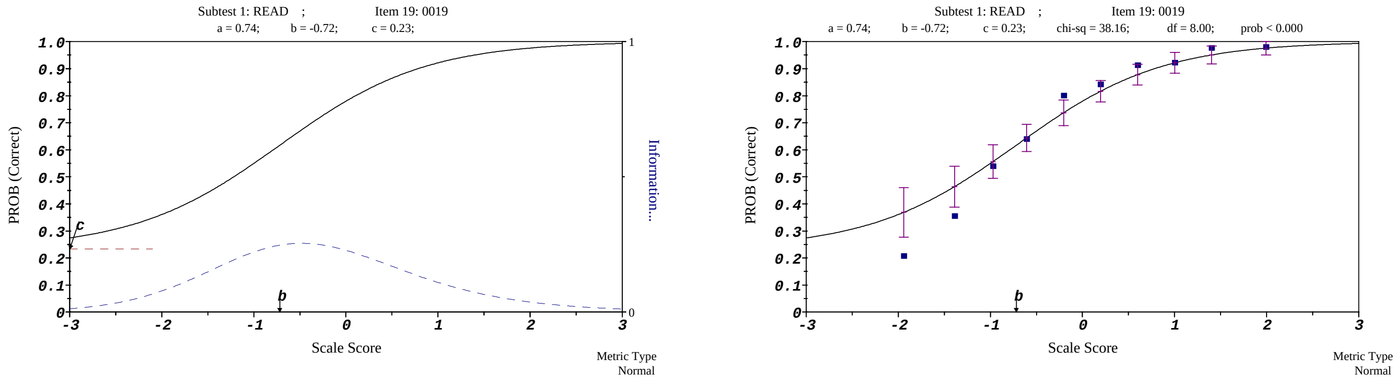

Figure 25. Response curve, information and model fit for item 19 in the READ subtest.

Item 19 is somewhat problematic; the model data fit is very bad at low ability levels and that is also where the main information is located.

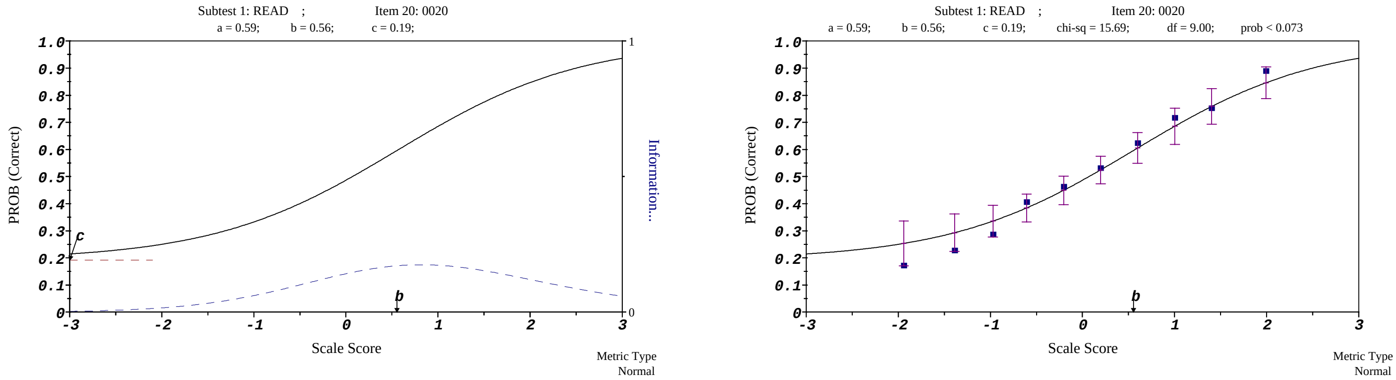

Figure 26. Response curve, information and model fit for item 20 in the READ subtest.

For item 20 the model data fit is acceptable according to all analyses. The information is not very high but it is located somewhat above medium ability.

In Figure 27, 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 curves at the same ability levels.

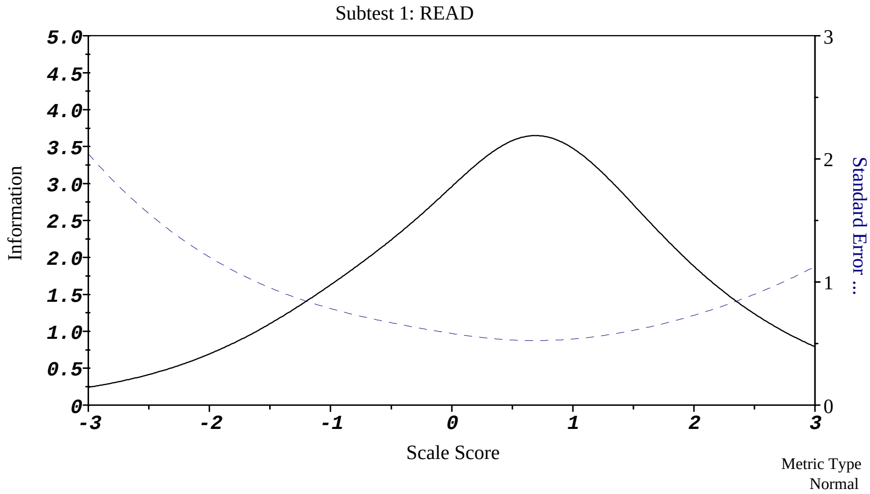

Figure 27. Test information curve and measurement error in the READ subtest.

As may be seen from Figure 27 the standard error is inversely related to the information at each ability level, i.e. the standard error is different at different ability levels. From the test information function it may also be seen that the errors are much smaller around average ability than for both low and high ability levels. This finding is to be expected with the current approach to test design, even though the standard error of measurement in classical test theory is assumed to be the same for all score levels.

Concluding remarks

In many IRT applications reported in the literature, model data fit and the consequences of misfit have not been investigated adequately. As a result, less is known about the appropriateness of particular IRT models for various applications than might be assumed from the voluminous IRT literature. A further problem with many IRT goodness of fit studies is that too much reliance has been placed on statistical tests of model fit (Hambleton et al., 1991).

The results from this attempt to fit a three parameter logistic IRT model to the response data from the READ subtest are not completely clear-cut. Without doubt the results from the classical test theory gave support for the need of a three parameter model. The statistical tests were somewhat discouraging as too many items turned out to have significant model data misfit. On the other hand the separate estimates for males and females were encouraging, since the estimates for the two groups corresponded fairly well. The main problem may be that the number of items in the READ subtest is too low to make reasonable ability estimates. It seems worthwhile, however, to investigate the remaining subtests in the SweSAT in the same way. It also seems reasonable to investigate different combinations of subtests before the final decision on whether the SweSAT program may be improved by using IRT can be made.

REFERENCES

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

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

Stage, C. (1996) *An Attempt to Fit IRT Models to the DS Subtest in The SweSAT* (Educational Measurement No. 19). Umeå, Umeå University, Department of Educational Measurement.

Stage, C. (1997a) *The Applicability of Item Response Models to the SweSAT. A Study of the DTM Subtest* (Educational Measurement No. 21). Umeå, 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, Umeå University, Department of Educational Measurement.