THE APPLICABILITY OF ITEM RESPONSE MODELS TO THE SweSAT

A Study of the WORD Subtest

Christina Stage

Em No 26, 1997

ISSN 1100-696X
ISRN UM-PED-EM--26--SE

This study is the fifth 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, ERC and READ subtests (Stage, 1996, 1997a, 1997b and 1997c). 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 WORD subtest.

METHOD

Sample

In the DS, DTM, ERC and READ 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 WORD subtest are the data which will be analysed in different ways.

Classical item analysis

The classical item analysis of the WORD subtest gave a range of p-values from .16 to .96 and a range of biserial correlations from .02 to .64. The mean of the test was M = 25.39 and the standard deviation was s = 6.9. The reliability, coefficient alpha, was r = .85.

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 40 items.

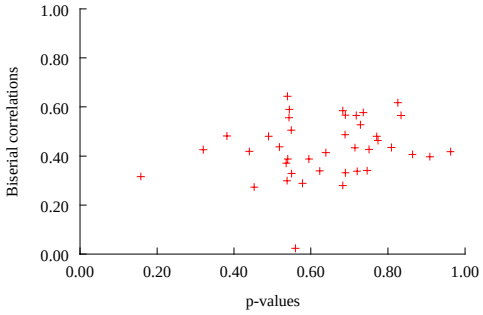

Figure 1. Biserial correlations plotted against p-values of the 20 items in the WORD subtest.

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 16 of the 40 possible points were selected. This gave a number of 210 examinees. The results of these 210 examinees on the

most difficult items in the test were studied. Six items had p-values lower than .50 and the p-values of this poor group on these six difficult items were:

p = .13 .01 .14 .01 .11 .22

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 = .85 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 nine factors with eigenvalues 6.1, 1.4, 1.2, 1.2, 1.1, 1.1, 1.0, 1.0 and 1.0 respectively. The variance explained by the first factor was 15.4 percent, the variance explained by the second factor was 3.6 percent, by the third factor 3.1 percent, by the fourth factor was 2.9 and by the remaining factors 2.8, 2.7 or 2.6 percent. A plot of the eigenvalues is shown in Figure 2.

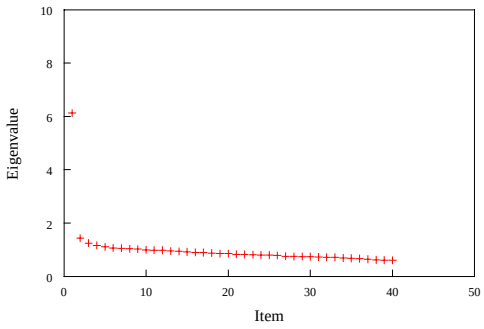

Figure 2. Plot of eigenvalues for the WORD subtest.

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 nine of the 40 items had loadings of at least .30 on the first factor and only one item (7) 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 WORD 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 40 the sample of examinees was large. The number of intervals used was 10 for most of the items (24), nine for 11 items, eight for three items and seven for two items.

The outcome of the goodness of fit analysis was that for eight of the items in the subtest there was a model data misfit which was significant at .05 level and for one of these eight 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 = .87, 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 34 items the differences between observed and predicted results were insignificant. For six items the differences were significant at .05 level and for one of these six items the difference was 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.

residuals	**number**	**percent**
| 0-1 |	225	70.31
| 1-2 |	84	26.25
| 2-3 |	10	3.13
| >3 |	1	.31

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 .74. In Figure 3 the estimated b-values are plotted against the p-values.

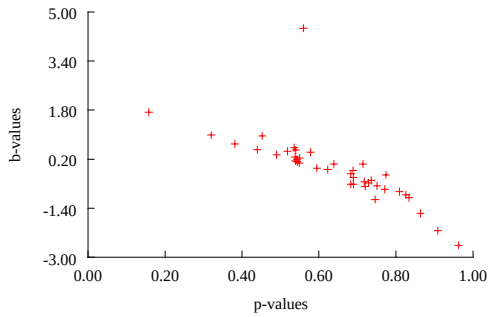

Figure 3. Estimated b-values plotted against p-values.

There is one very deviating item (7) with a SR = 5.6, this is also the item which had the lowest loading on the first factor in the factor analysis. Item 16 has a SR = 1.1, all other items have SRs less than 1.0. With one or possibly two exceptions the estimated difficulty parameters seem to correspond very well to the the observed difficulty indices of the 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.

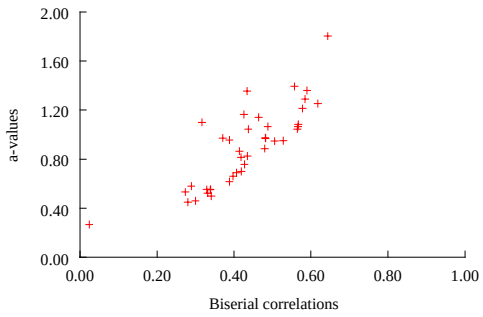

Figure 4. Estimated a-values plotted against corresponding biserial correlations.

Three items (4, 18, 23) had SRs between 2 and 2.5, four items (7, 11, 21, 36) had SRs between 1 and 1.5 and two items (5, 14) had SRs = -1.1. The correlation between the estimated a-values and corresponding biserial correlations was r = .82. 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 = .97.

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.

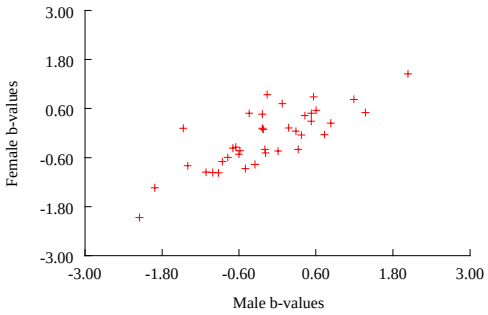

Figure 5. The b-values estimated on the female group plotted against b-values estimated on the male group.

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 = .85. 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). For 28 items the SRs are smaller than |1|, for 11 items the SRs are smaller than |2| and for one item (34) the SR is 2.3.

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

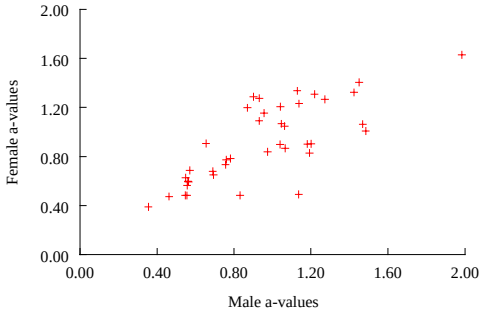

Figure 6. The a-values estimated on female examinees plotted against the a-values estimated on male examinees.

The discrimination parameters also seem to be fairly congruent even though they are estimated on different samples. For 29 items the SRs were smaller than |1| and for one item only (34) the SR was larger than |2| (-2.6). The correlation between a-values estimated on the two groups was r = .77. This invariance of parameter estimations can be taken as strong support for model data fit.

The goodness of fit analysis for the female results gave as result that for two items there was a significant model data misfit at .05 level and for one of these items there was a misfit at .01 level. For males there was a significant misfit for two items at .05 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	14.90	7.42	9.10	5.90
2	12.80	12.22*	8.30	2.10
3	11.70	5.00	8.90	5.20

4	6.80	9.24	5.70	1.50
5	16.90*	11.69*	12.40	11.40
6	21.50**	1.73	9.80	11.20
7	13.20	5.63	17.00*	13.90
8	7.70	1.35	3.40	6.70
9	15.20	5.76	5.90	11.40
10	9.80	7.66	8.00	8.30
11	16.80	5.09	11.70	10.40
12	9.40	3.03	7.10	3.40
13	11.90	6.87	9.70	14.60
14	18.80*	12.44*	12.50	8.50
15	10.00	3.91	8.10	7.10
16	6.10	10.15	2.20	1.20
17	10.60	2.35	6.30	5.40
18	20.00*	7.67	10.30	7.20
19	17.60*	6.05	15.30	15.10
20	14.60	3.33	7.40	12.40
21	10.10	6.47	5.10	10.90
22	12.70	3.15	9.10	7.00
23	16.80	28.19**	9.30	20.40*
24	8.70	4.28	4.80	8.20
25	10.10	6.46	12.40	10.20
26	10.50	6.08	4.80	8.20

27	10.00	3.69	5.80	7.80
28	10.00	3.51	5.70	7.50
29	5.50	11.26*	6.00	11.50
30	16.40*	3.81	12.60	10.80
31	19.10*	14.91*	23.90**	1.70
32	7.80	6.59	7.30	6.80
33	14.70	5.28	8.60	14.70
34	8.70	8.07	8.00	20.70*
35	10.90	5.21	6.80	6.60
36	6.20	6.95	10.20	14.90
37	7.10	4.25	10.40	3.20
38	7.80	8.40	3.60	3.00
39	16.60*	6.97	12.40	7.60
40	12.40	13.50	12.90	4.40

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. 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 40 items in the WORD subtest are presented in Figure 7 to Figure 46. 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 46 there is a representation of the model data fit for each item.

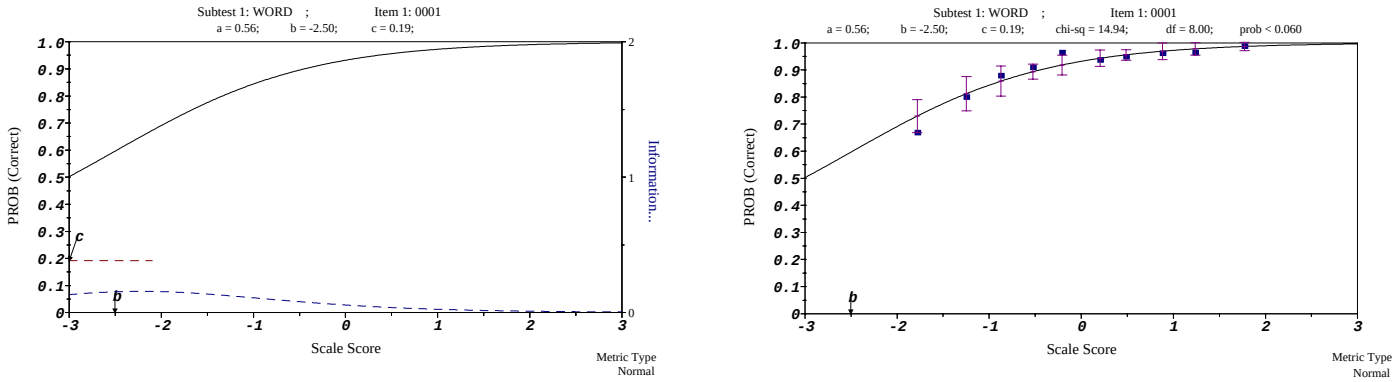

Figure 7. Response curve, information and model fit for item one in the WORD subtest.

For item one the model data fit was acceptable according to all analyses performed. The item is very easy, however, and the information provided is very low and also located at very low ability levels.

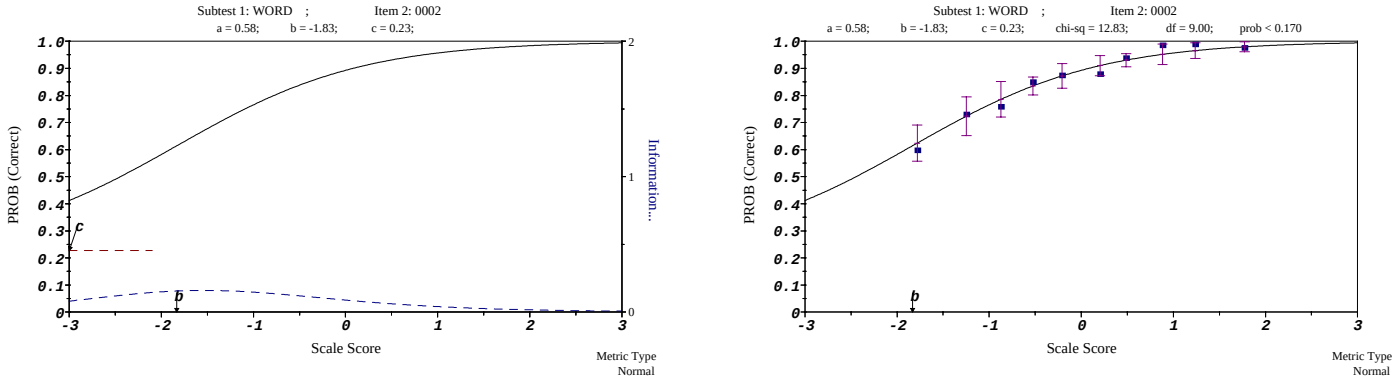

Figure 8. Response curve, information and model fit for item two in the WORD subtest.

For item two there was significant model data misfit according to the Resid test. This item as well is very easy and the information provided is poor.

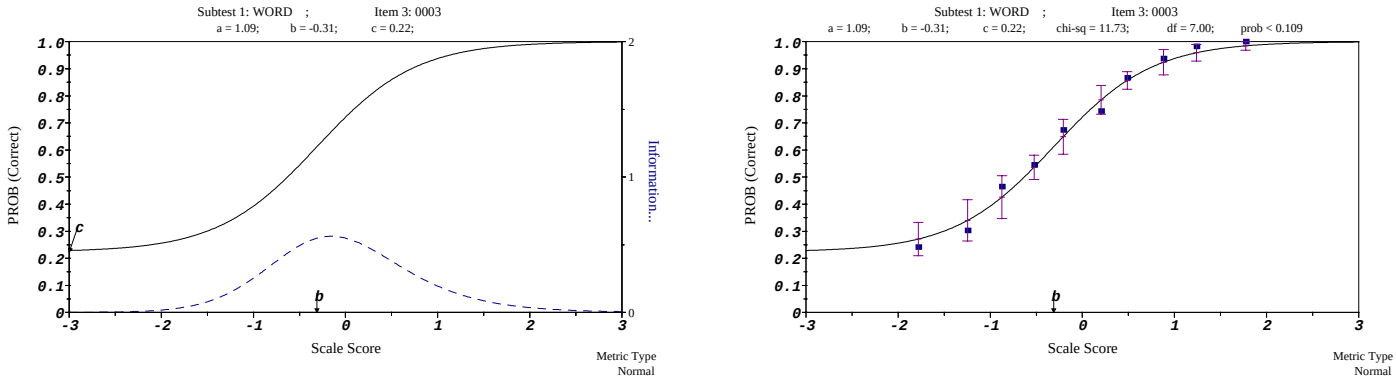

Figure 9. Response curve, information and model fit for item three in the WORD subtest.

For item three the model data fit was acceptable according to all analyses. The discrimination and hence the information is fairly good and located around medium ability level. This item contributes to the test.

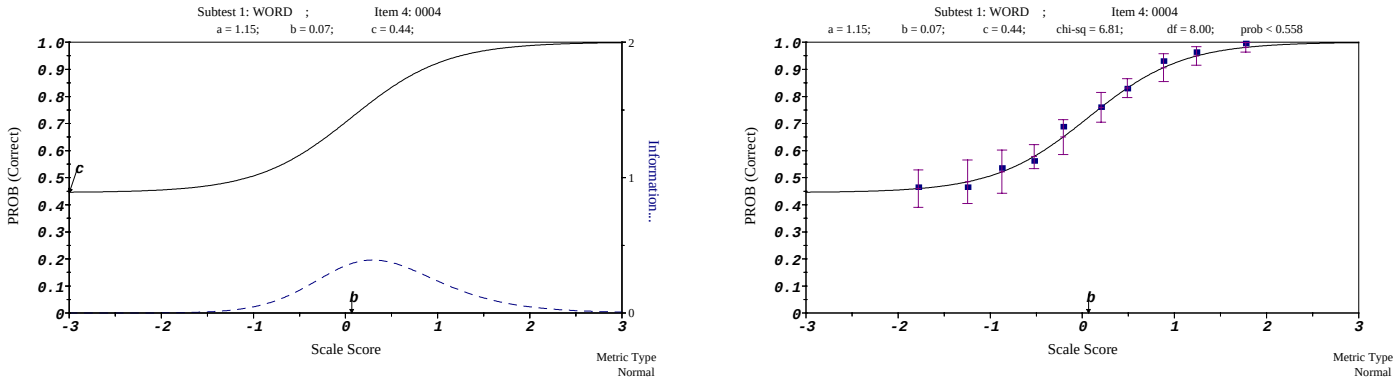

Figure 10. Response curve, information and model fit for item four in the WORD subtest.

For item four as well the model data fit was acceptable according to all analyses. The discrimination and information are fairly good and located around medium ability level. This is a useful item as well.

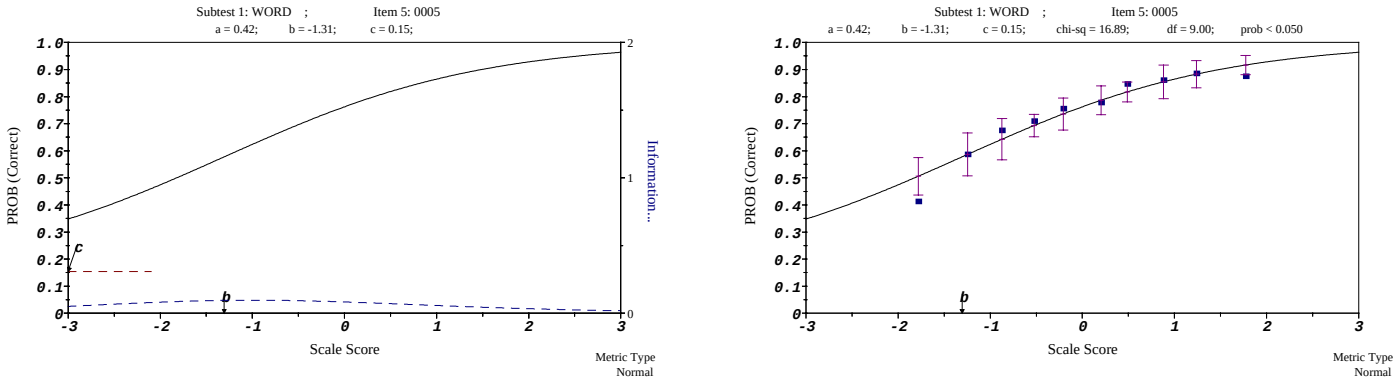

Figure 11. Response curve, information and model fit for item five in the WORD subtest.

For item five there is significant misfit according to the Bilog as well as the Resid test. The discirmination and information are poor as well.

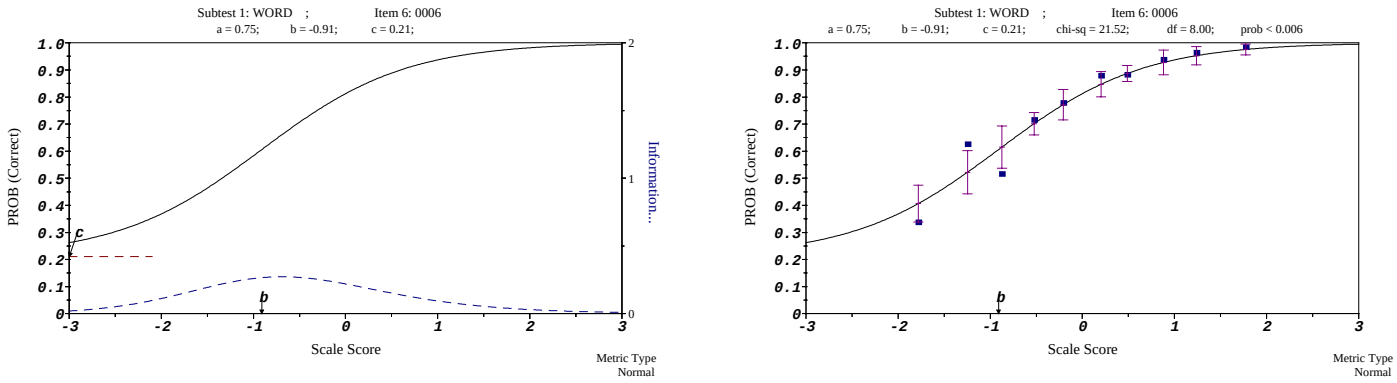

Figure 12. Response curve, information and model fit for item six in the WORD subtest.

For item six there is significant misfit according to the Bilog test. The misfit also seems to be located at low ability levels where the main information is given.

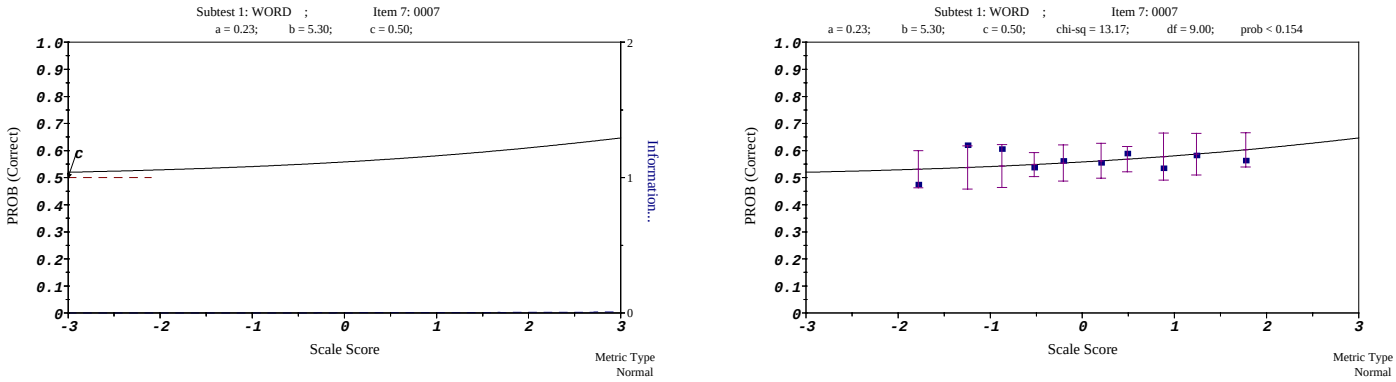

Figure 13. Response curve, information and model fit for item seven in the WORD subtest.

Item seven seems to be very problematic. There is significant misfit only according to the Bilog test for females but there is no discrimination and no information provided by this item. Item seven also had the lowest loading on the first factor in the factor analysis (p 4). This item does not seem to belong to the test at all.

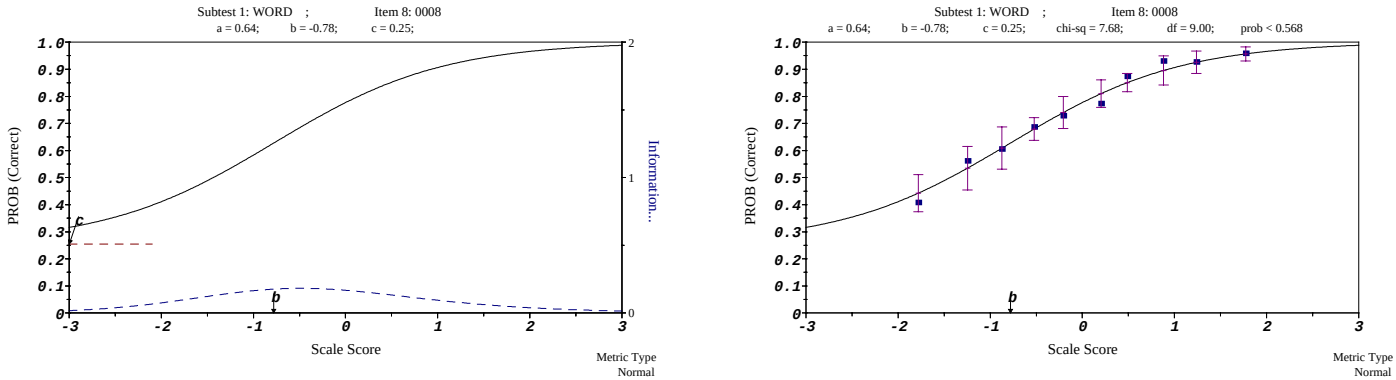

Figure 14. Response curve, information and model fit for item eight in the WORD subtest.

For item eight the model data fit is acceptable. The discrimination and information are somewhat on the low side but on the whole this item may be useful.

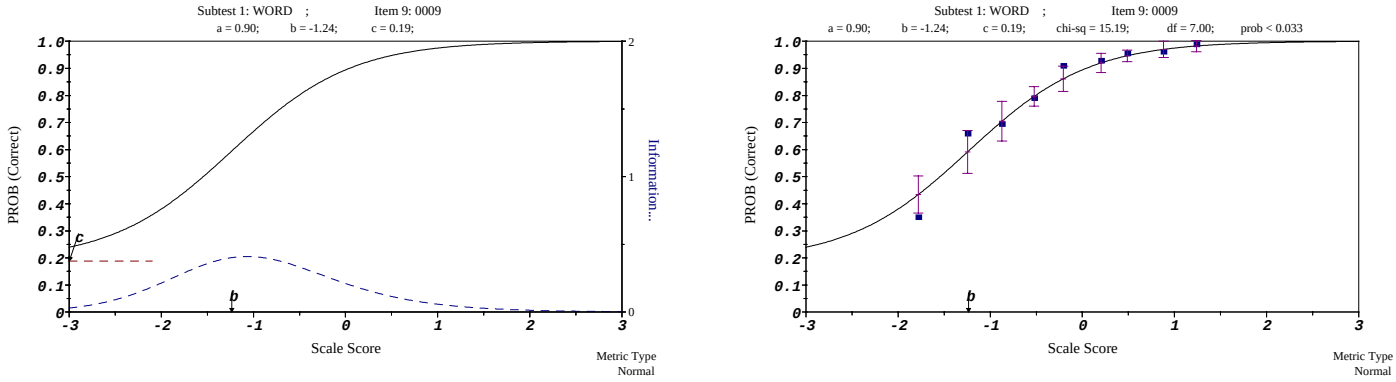

Figure 15. Response curve, information and model fit for item nine in the WORD subtest.

For item nine as well the model data fit was acceptable according to all analyses. On the graph at the left in Figure 15, however it can be noted that the fit is rather poor at low ability levels where also the main information is located.

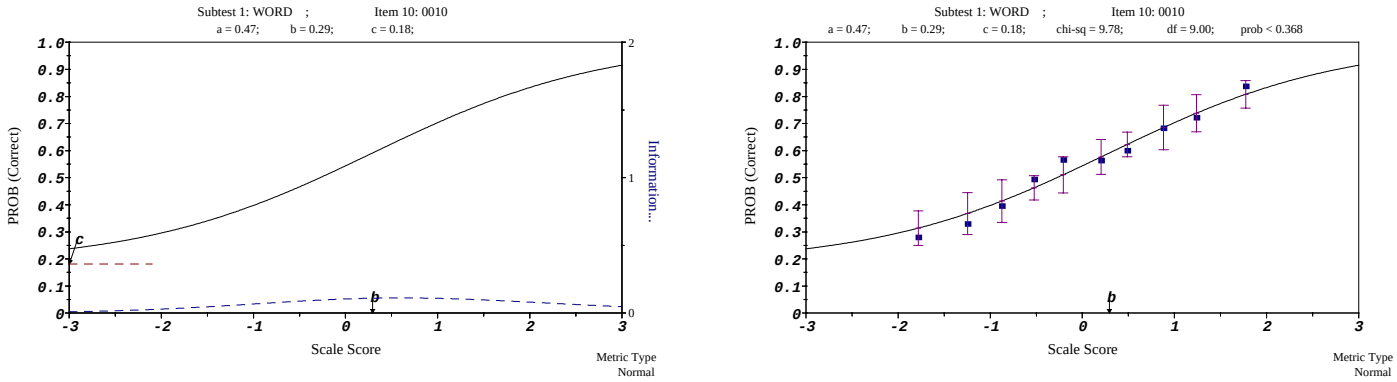

Figure 16. Response curve, information and model fit for item 10 in the WORD subtest.

For item ten as well the fit was acceptable according to all the statistical analyses. The discrimination and information are rather poor, however.

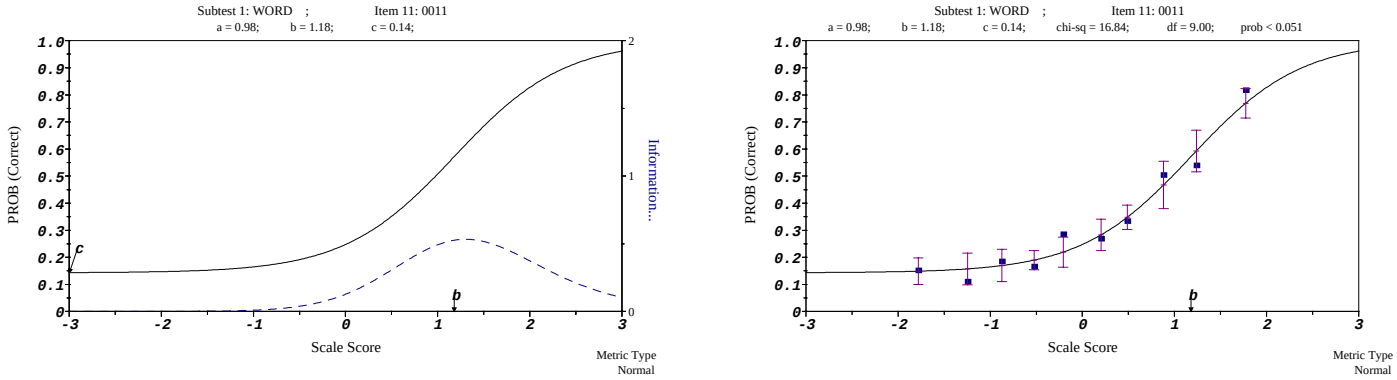

Figure 17. Response curve, information and model fit for item 11 in the subtest WORD.

For item 11 the model data fit is acceptable according to the statistical analyses. This item has good discrimination and information and the information is located at high ability levels where it is most needed.

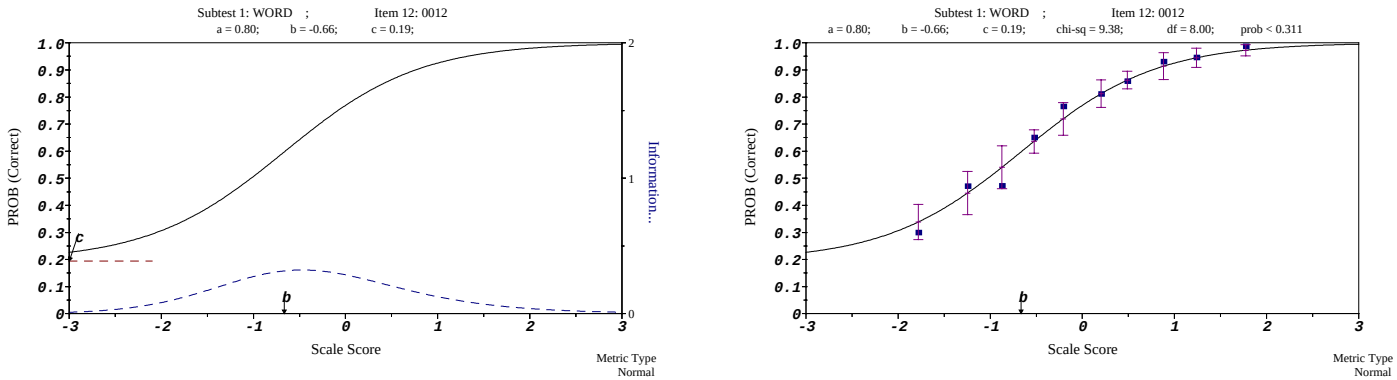

Figure 18. Response curve, information and model fit for item 12 in the subtest WORD.

Also for item 12 the model data fit was acceptable. The discrimination and information are acceptable too even though the information is located mainly below medium ability.

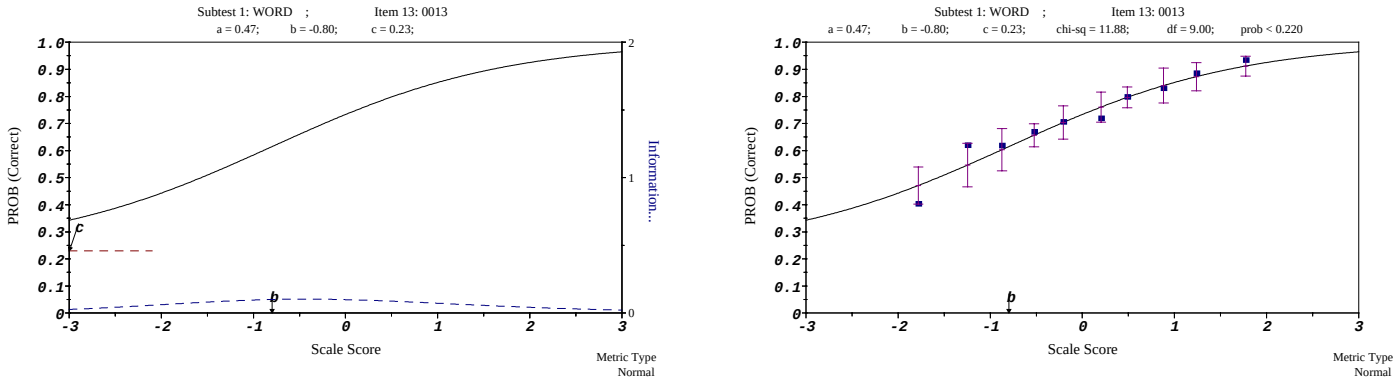

Figure 19. Response curve, information and model fit for item 13 in the WORD subtest.

For item 13 as well there is acceptable model data fit. The discirmination and information are fairly low but the information is distributed aver a wide range of ability and so the item may be useful.

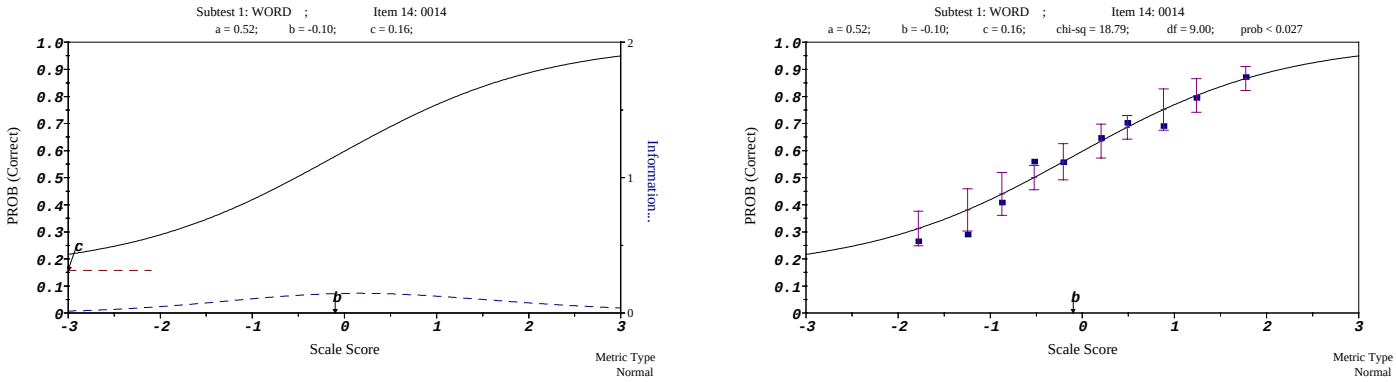

Figure 20. Response curve, information and model fit for item 14 in the WORD subtest.

For item 14 there is significant model data mifit according to the Bilog test as well as the Resid test.

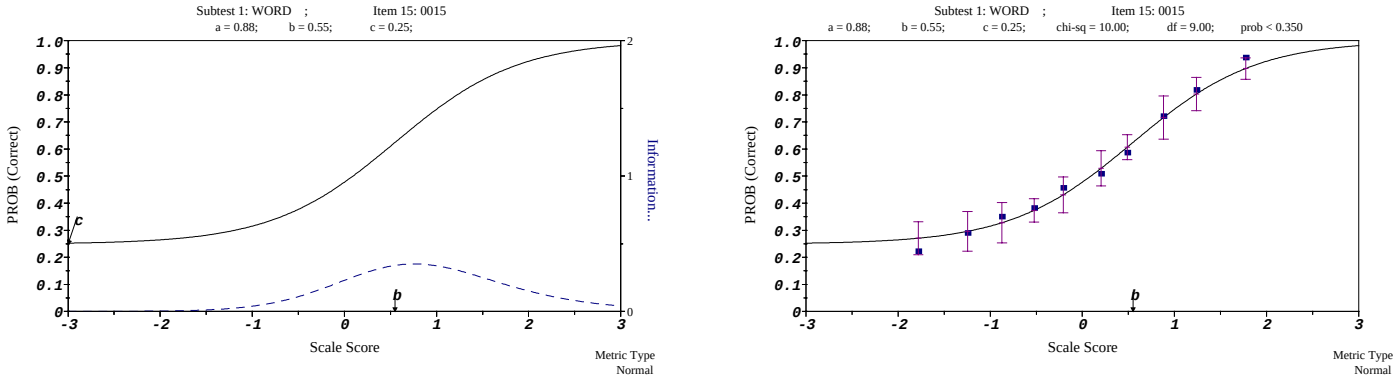

Figure 21. Response curve, information and model fit for item 15 in the WORD subtest.

For item 15 the model data fit is acceptable. The discrimination and information are rather good and the information is mainly above medium ability, which is very useful.

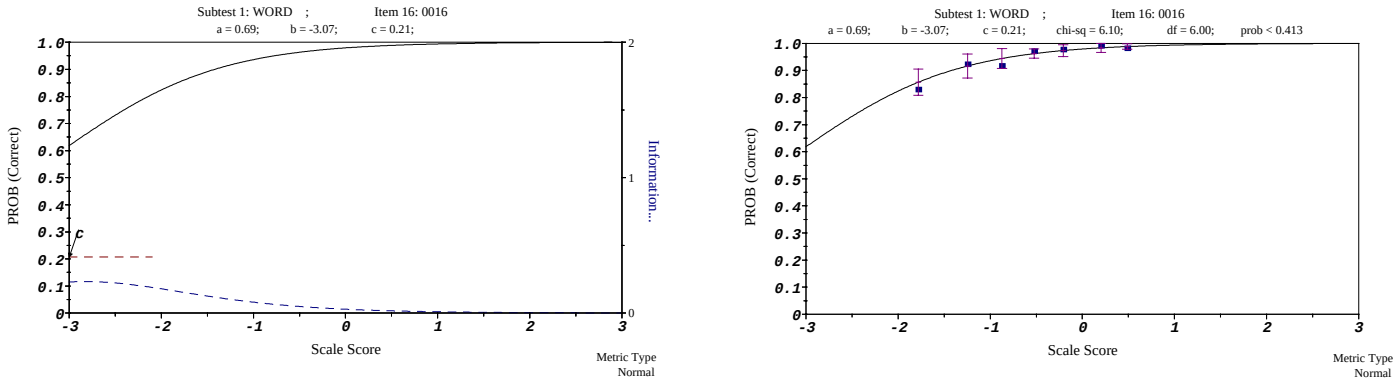

Figure 22. Response curve, information and model fit for item 16 in the WORD subtest.

For item 16 the model data fit is acceptable according to the statistical tests. The discrimination and information, however are at very low ability levels where they are of little use.

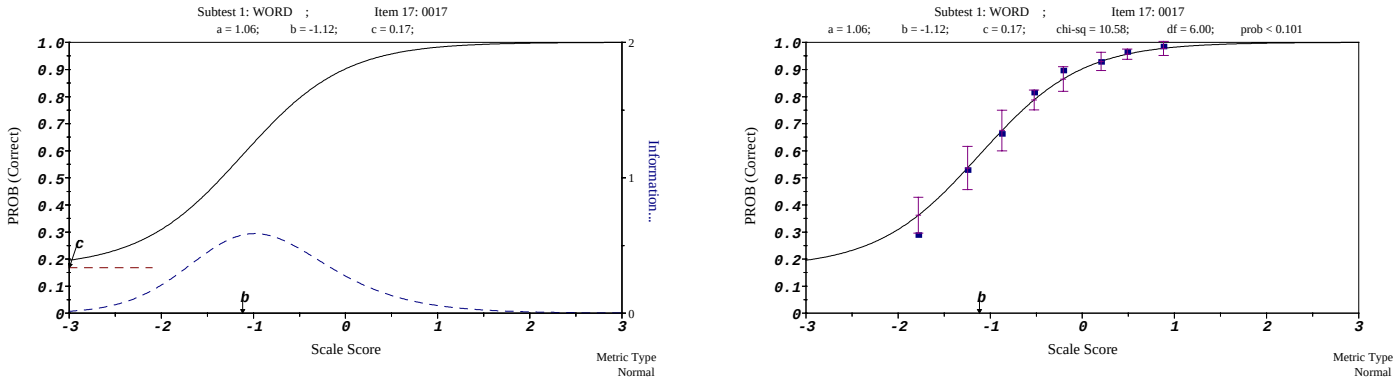

Figure 23. Response curve, information and model fit for item 17 in the WORD subtest.

For item 17 the fit is acceptable. The discrimination and information are pretty good even though they are located below medium ability.

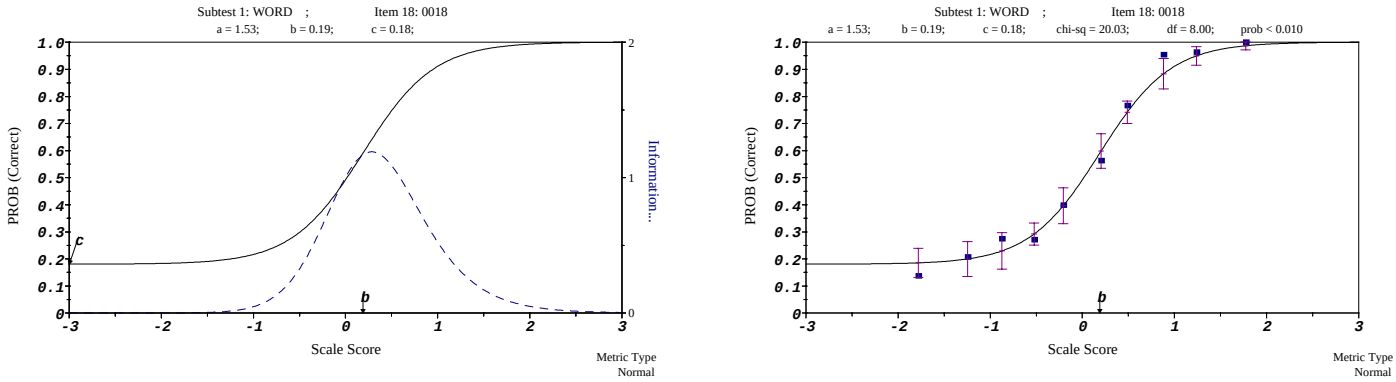

Figure 24. Response curve, information and model fit for item 18 in the WORD subtest.

For item 18 there is significant model data misfit according to the Bilog test. The discrimination and information are good though.

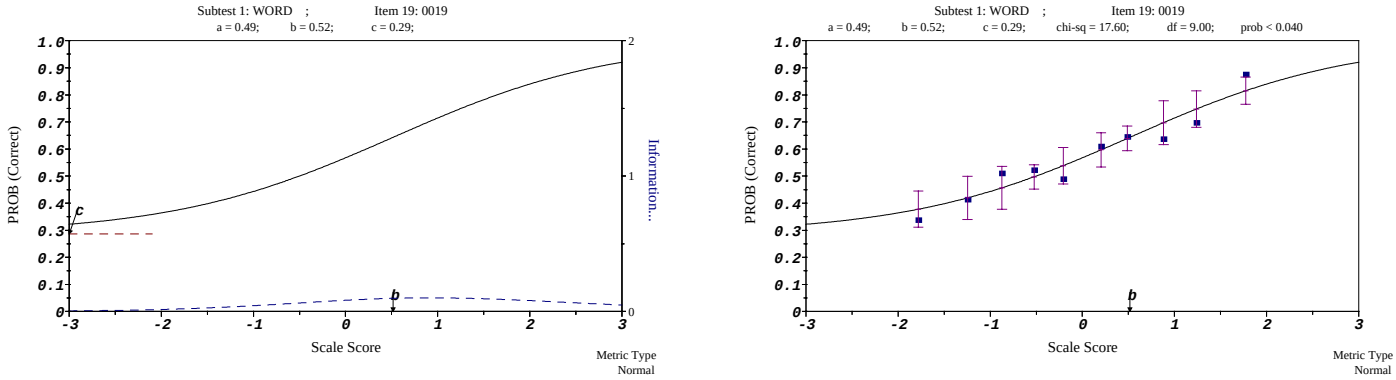

Figure 25. Response curve, information and model fit for item 19 in the WORD subtest.

For item 19 there is also significant model data misfit according to the Bilog test. For this item the discrimination and information are very low as well.

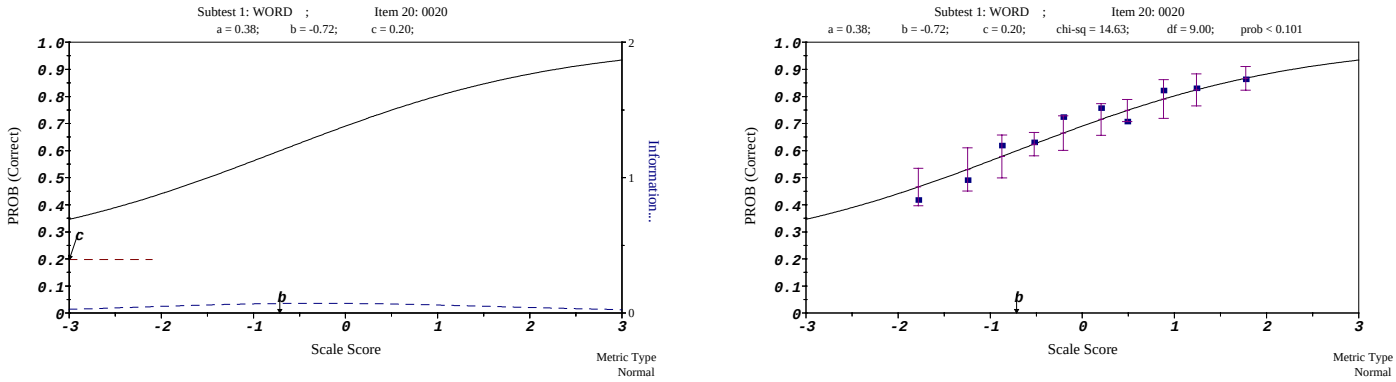

Figure 26. Response curve, information and model fit for item 20 in the WORD subtest.

For item 20 the model data fit is acceptable, but the dicrimination and information are very poor.

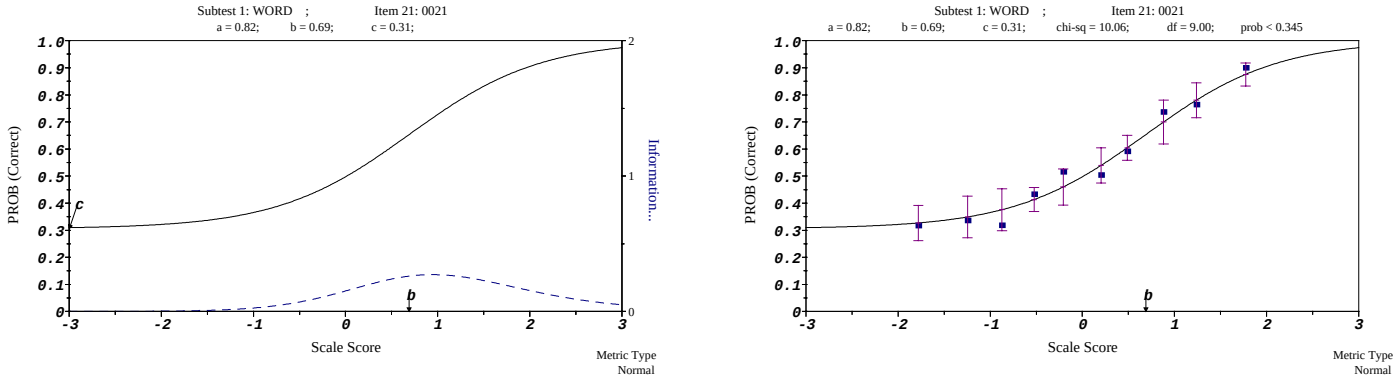

Figure 27. Response curve, information and model fit for item 21 in the WORD subtest.

For item 21 the model data fit is acceptable; the discrimination and information are acceptable too and the information is mainly provided above medium ability.

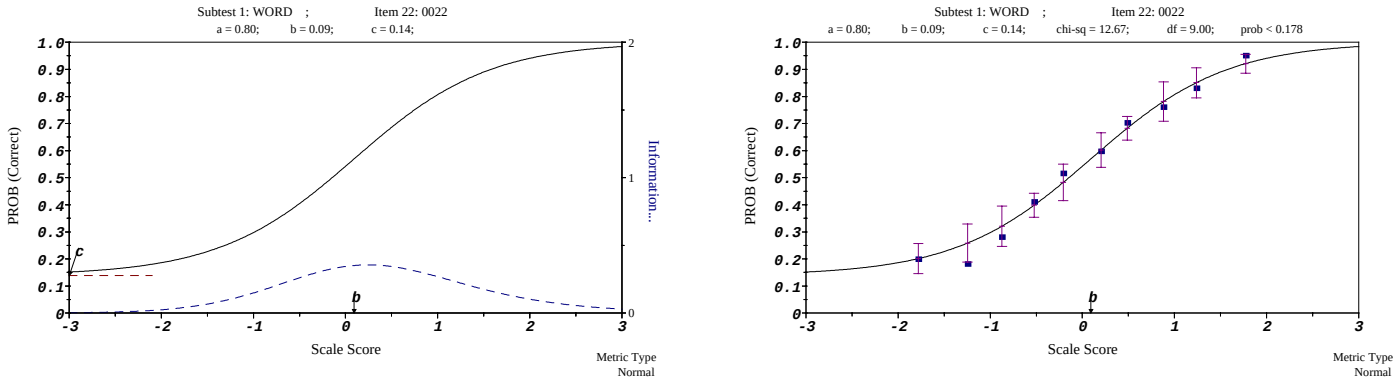

Figure 28. Response curve, information and model fit for item 22 in the WORD subtest.

For item 22 also the model data fit is acceptable and so are the discrimination and information. The information is mainly provided around medium ability.

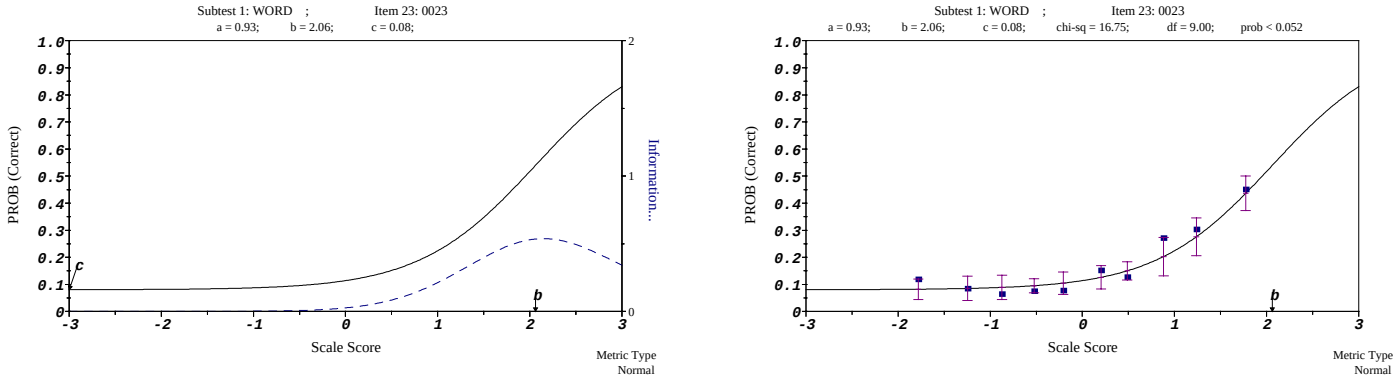

Figure 29. Response curve, information and model fit for item 23 in the WORD subtest.

For item 23 there is significant model data misfit according to the Resid test and the Bilog test for males.

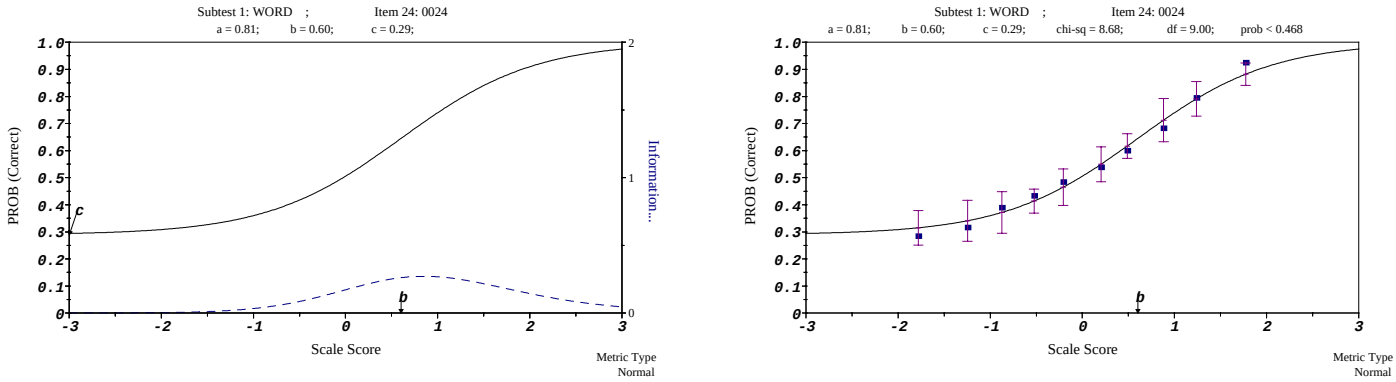

Figure 30. Response curve, information and model fit for item 24 in the WORD subtest.

For item 24 there is acceptable model data fit; the discrimination is acceptable and so is the information, which is mainly provided above medium ability level.

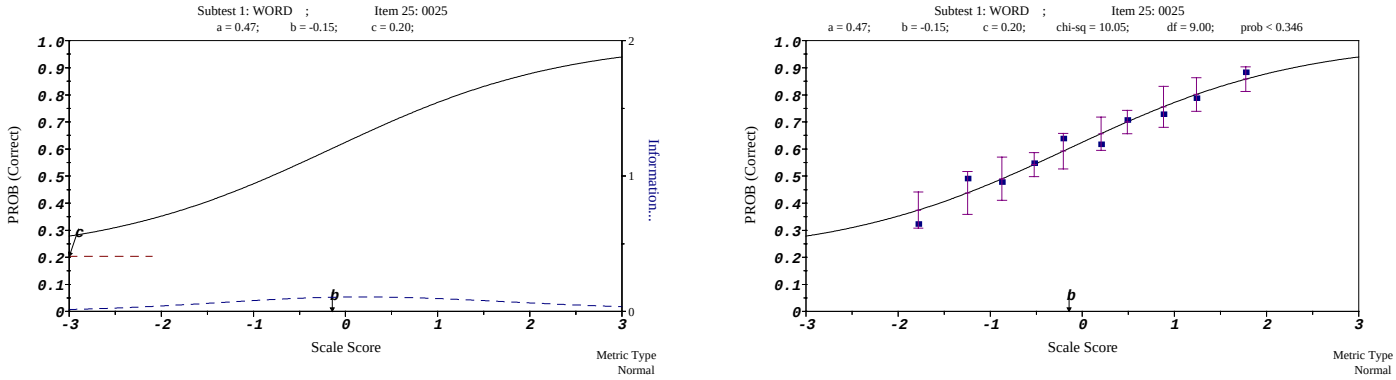

Figure 31. Response curve, information and model fit for item 25 in the WORD subtest.

For item 25 the model data fit is acceptable the discrimination and hence the information are very poor, however.

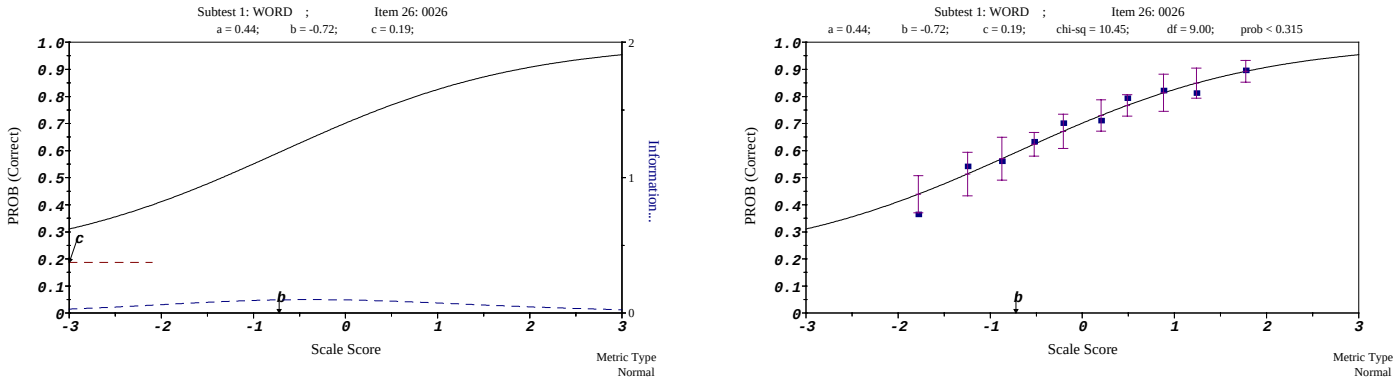

Figure 32. Response curve, information and model fit for item 26 in the WORD subtest.

For item 26 as well the model data fit is acceptable but the discrimination and information are very poor.

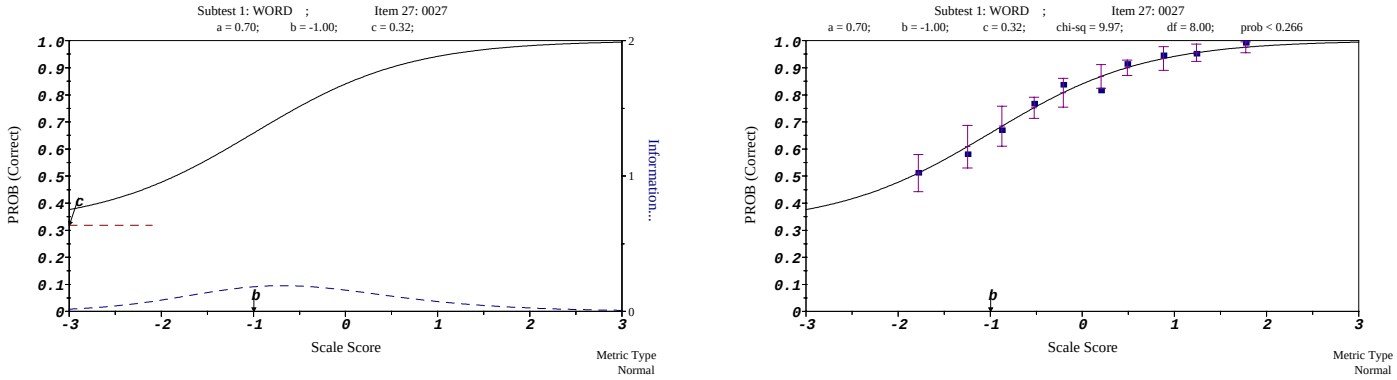

Figure 33. Response curve, information and model fit for item 27 in the WORD subtest.

For item 27 the model data fit is also acceptable and for this item the discrimination and informationare are acceptable as well, even though the information is located below medium ability.

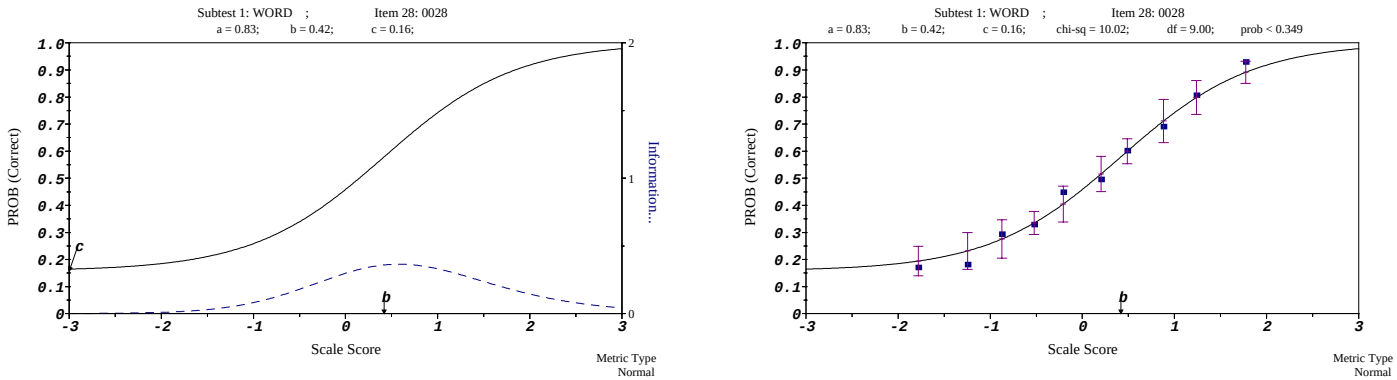

Figure 34. Response curve, information and model fit for item 28 in the WORD subtest.

For item 28 the model data fit is acceptable and so are the dicrimination and information. For this item also the information is located above medium ability.

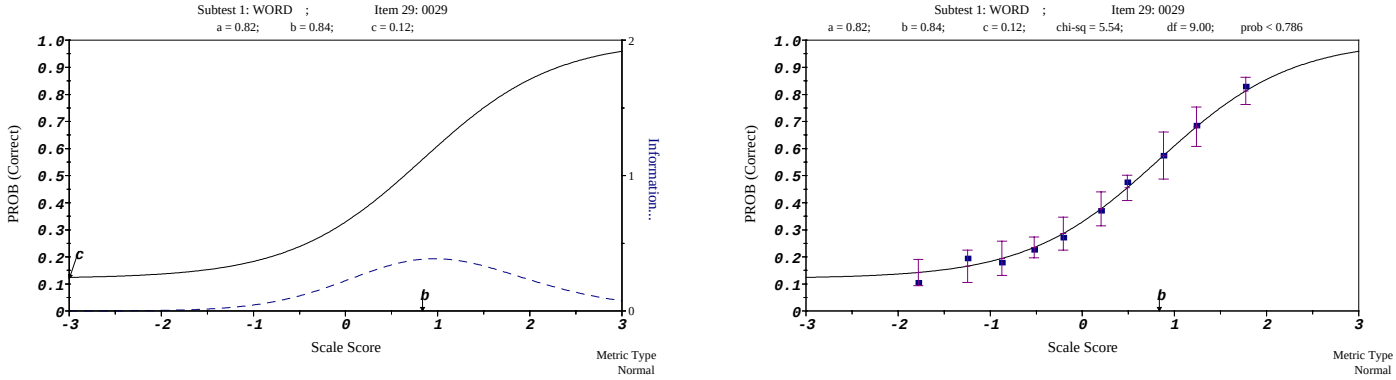

Figure 35. Response curve, information and model fit for item 29 in the WORD subtest.

For item 29 there is significant misfit according to the Resid test. The discrimination and information are good, however, and located above medium ability.

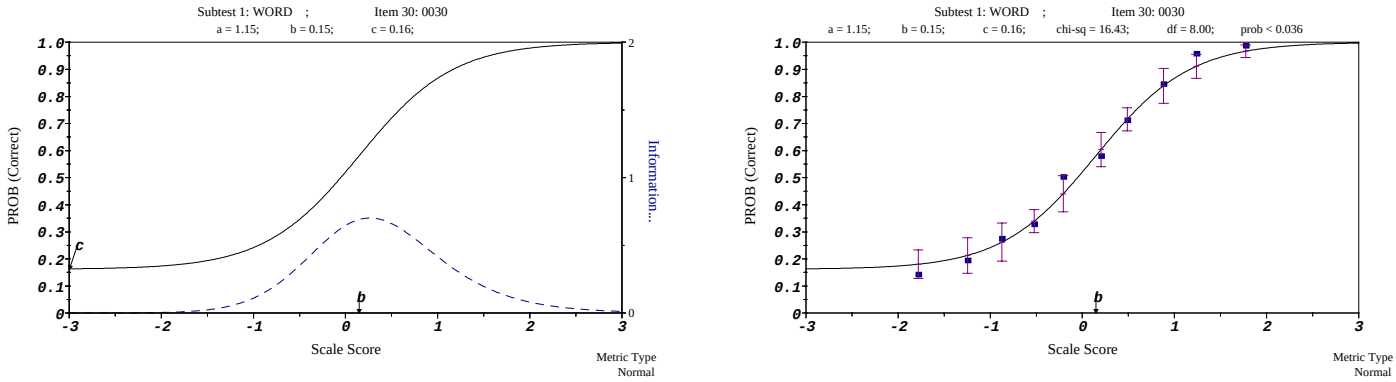

Figure 36. Response curve, information and model fit for item 30 in the WORD subtest.

For item 30 there is acceptable model data fit according to all analyses; the discrimination and information are good and the information is located somewhat above medium ability level.

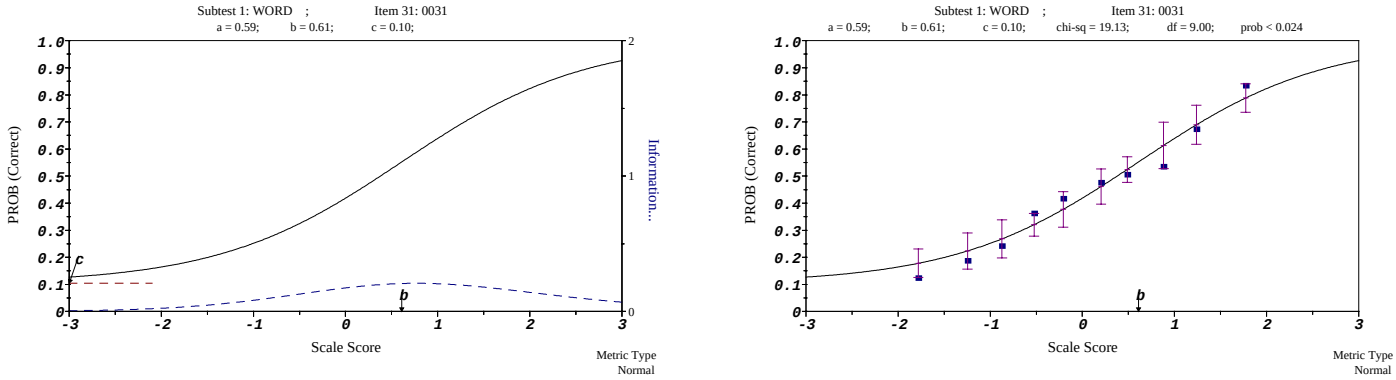

Figure 37. Response curve, information and model fit for item 31 in the WORD subtest.

For item 31 there is significant misfit according to three of the four statistical tests. The discrimination and information are rather poor as well.

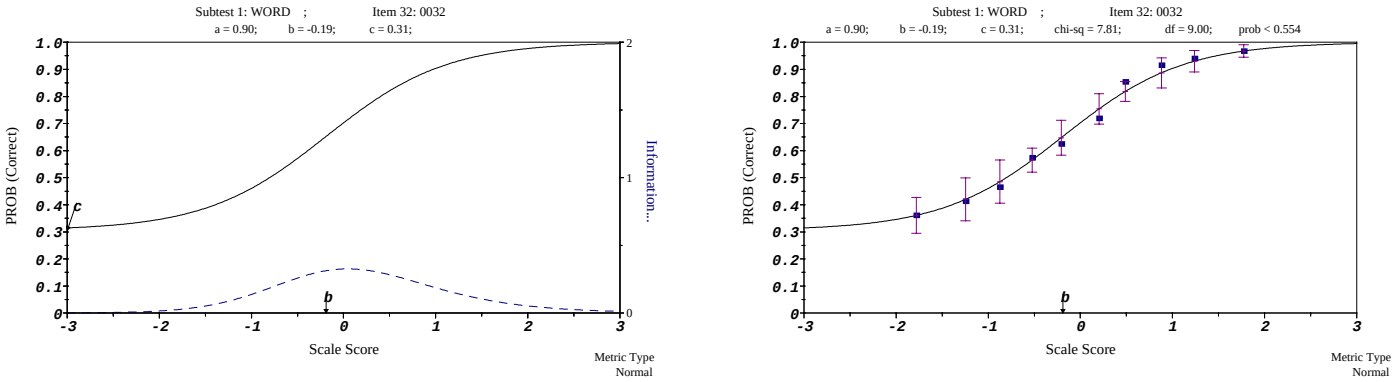

Figure 38. Response curve, information and model fit for item 32 in the WORD subtest.

For item 32 there is acceptable model data fit and also the discrimination and information are acceptable. The information is located around medium ability level.

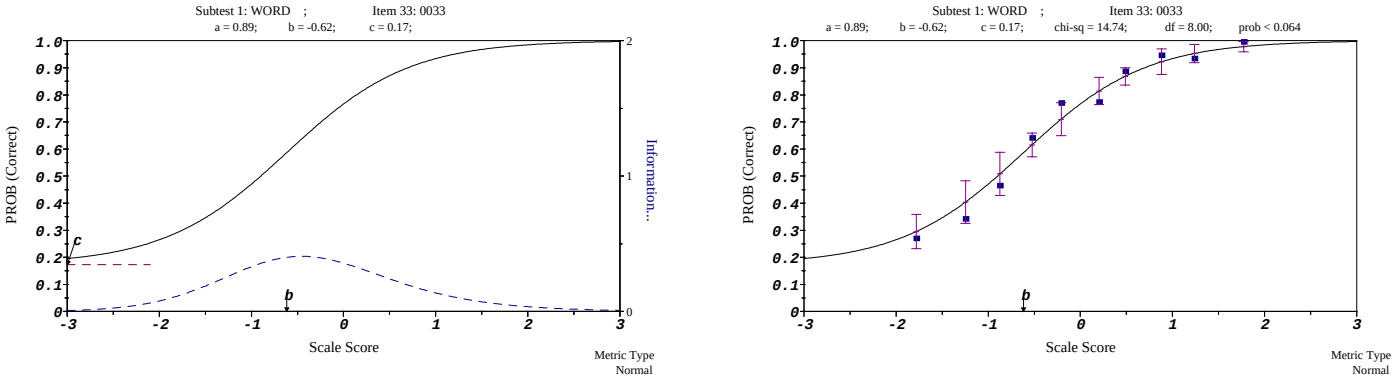

Figure 39. Response curve, information and model fit for item 33 in the WORD subtest.

For item 33 as well there is acceptable model data fit. The information is acceptable as well even though it is located somewhat below medium ability level.

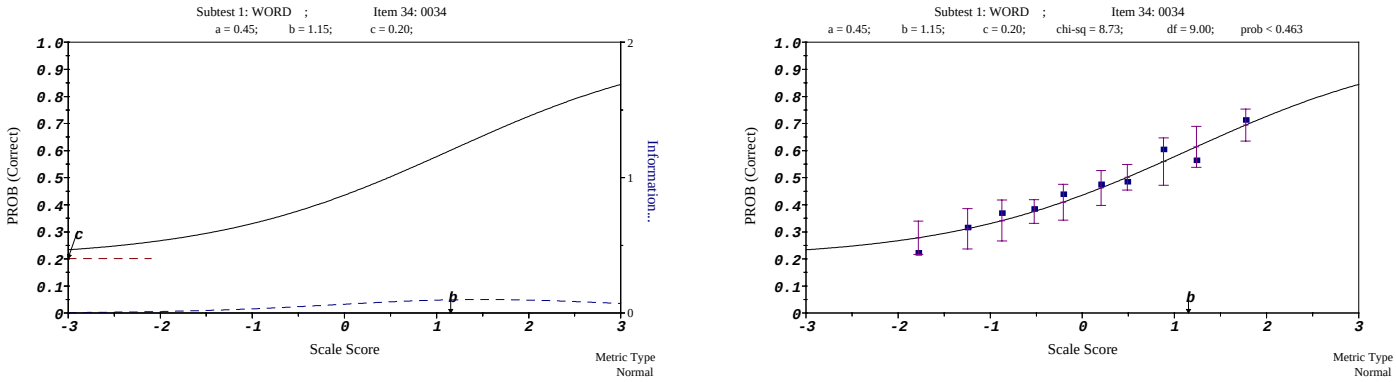

Figure 40. Response curve, information and model fit for item 34 in the WORD subtest.

For item 34 there is significant misfit for the male group. This item also provides very poor information.

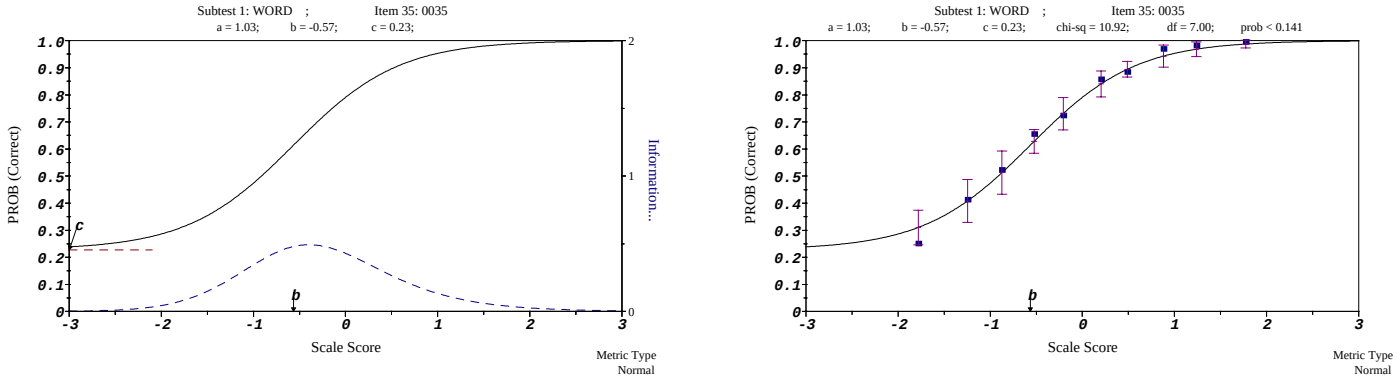

Figure 41. Response curve, information and model fit for item 35 in the WORD subtest.

For item 35 the model data fit is acceptable; the information is rather good even though it is located somewhat below medium ability level.

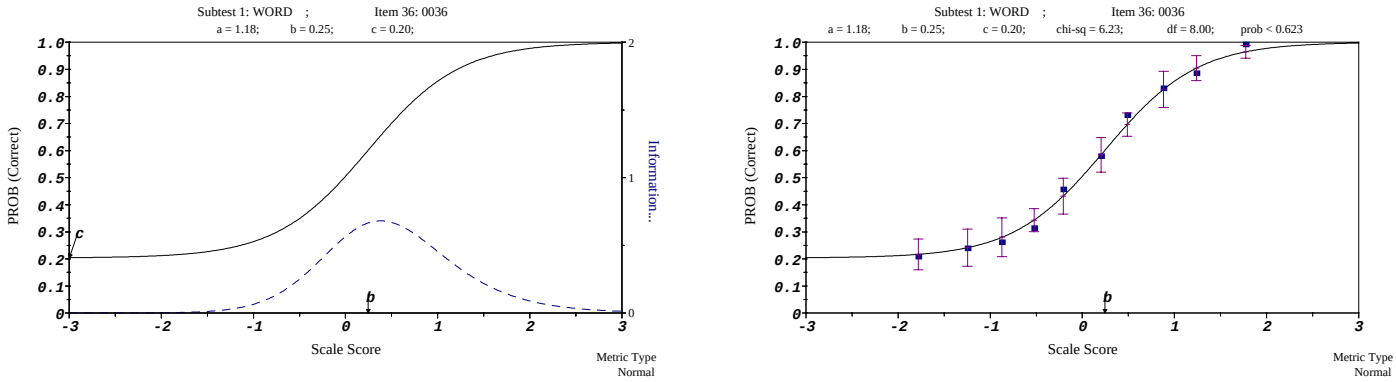

Figure 42. Response curve, information and model fit for item 36 in the WORD subtest.

For item 36 as well the model data fit is acceptable. The information is good and located somewhat above medium ability level.

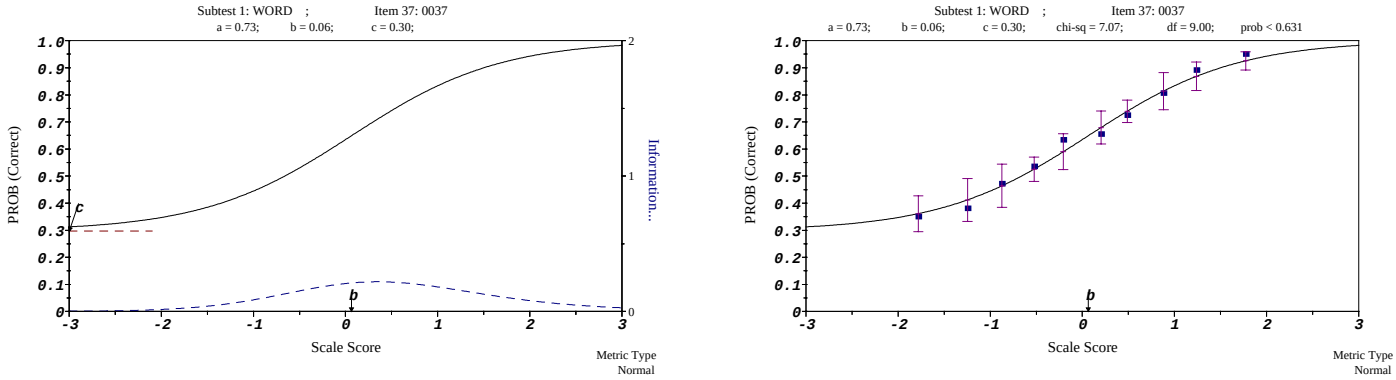

Figure 43. Response curve, information and model fit for item 37 in the WORD subtest.

For item 37 the model data fit is acceptable; the information is somewhat low but located just above medium ability level.

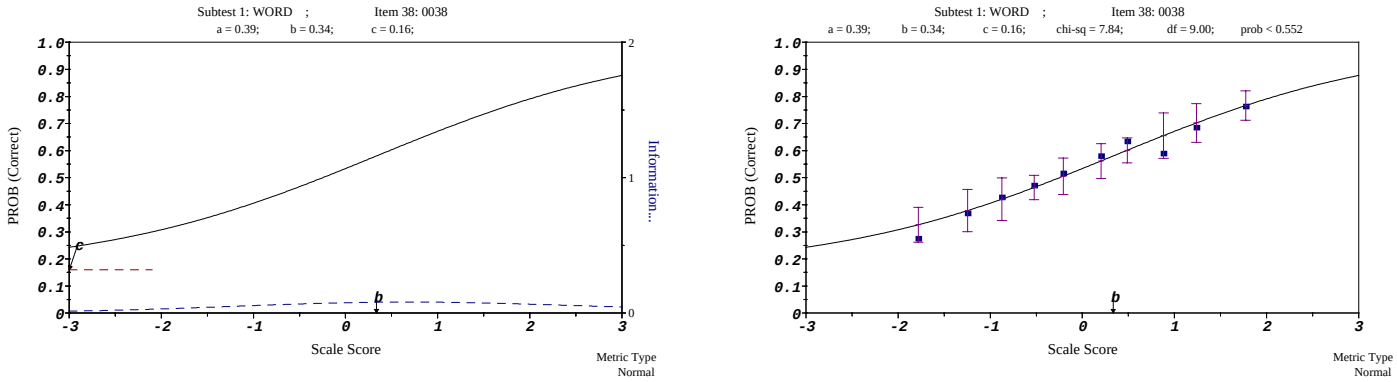

Figure 44. Response curve, information and model fit for item 38 in the WORD subtest.

For item 38 the model data fit is acceptable but the information is poor.

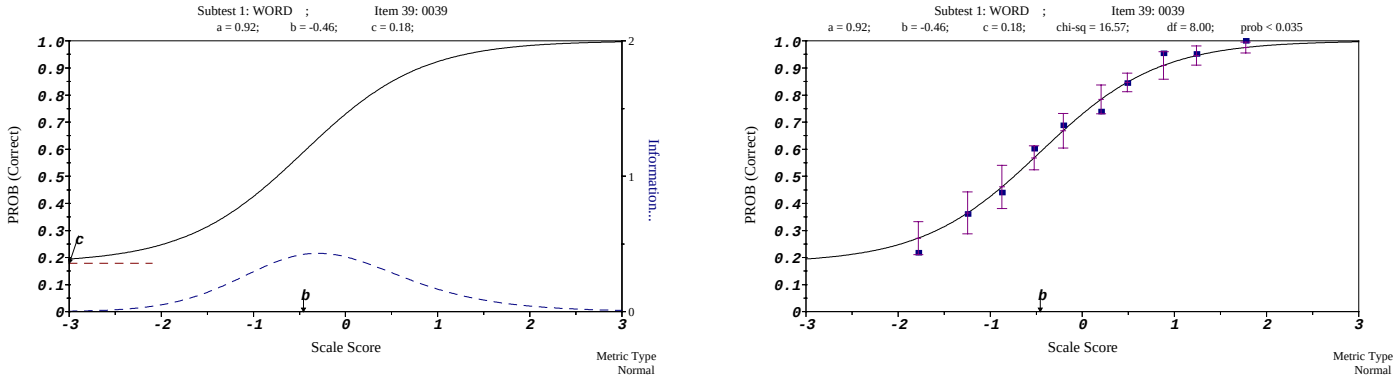

Figure 45. Response curve, information and model fit for item 39 in the WORD subtest.

For item 39 there is significant model data misfit according to the Bilog test. The information is acceptable but located a bit below medium ability.

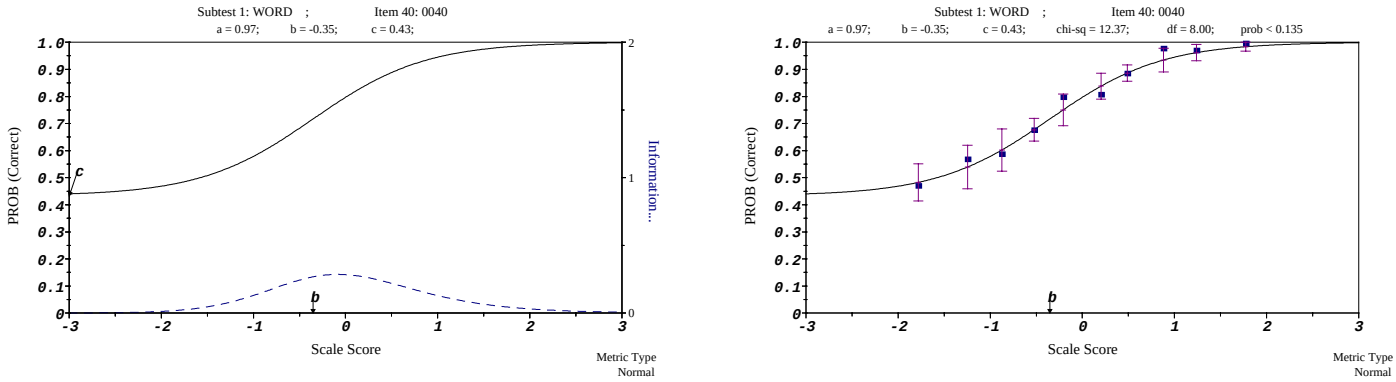

Figure 46. Response curve, information and model fit for item 40 in the WORD subtest.

For item 40 finally the model data fit is acceptable and so is the information even though it is located somewhat below medium ability.

For 27 of the 40 items the model data fit was acceptable according to all the statistical tests. Some of these items were not very effective though as they were either too easy or had too low discrimination power. In Figure 47 the total test information function is presented. The information given by a test at different ability levels is the sum of the information provided by the individual items at the same ability levels.

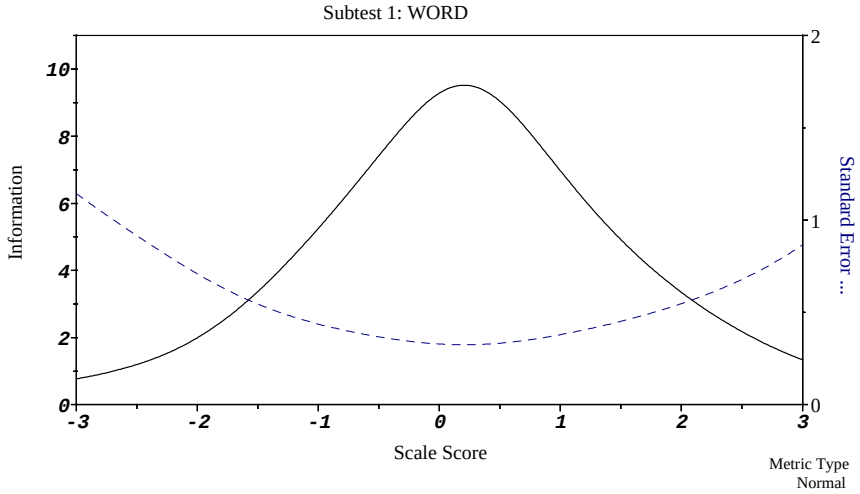

Figure 47. Test information curve and measurement error in the WORD subtest.

As may be seen from Figure 47 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.

As the number of items in the WORD subtest is great compared to the other subtests in SweSAT an additional analyis was made. Out of the 40 items 20 items were selected. Those items were selected which had acceptable model data fit according to the statistical tests and at the same time provided acceptable information.

The resulting test consisted of items number: 3, 4, 8, 11, 12, 13, 15, 17, 21, 22, 24, 27, 29, 30, 32, 33, 35, 36, 37, 40. The mean of this ”new” test was 12.83 and the reliability coefficient alpha was .79. The reliability of this shortened test estimated by the Spearman-Brown formula would be .74.

An unrotated factor analysis of these 20 items resulted in four factors with eigenvalues 4.1, 1.1, 1.0 and 1.0. The variance explained by the first factor was 20.5 percent, by the second 5.5 percent, by the third 5.3 and by the fourth 5.1. All 20 items had loadings larger than .30 on the first factor. Evidently this small test was more homogenous than the original one of 40 items and it seems very reasonable to assume one single factor with these test data.

A comparison between the estimated IRT parameters and the item indices from the classical test theory gave a correlation of -.93 between estimated b-values and computed p-values and a correlation of .69 between estimated a-values and computed biserial correlations.

A comparison between the item parameters estimated on 40 items and the same parameters estimated on 20 items gave a correlation between b-values of .996 and a correlation between a-values of .987. The reliability of this shortened test according to IRT was .82. The reliability estimated by the Spearman-Brown formula would be .77, which was exactly the result for the first 20 items of the 40 items test.

Concerning reliability this short test was evidently better than a randomly shortened test from the same 40 items would have been. As for the model data fit,however, even though all the items had acceptable model data fit in the test of 40 items there was considerable misfit for the same items in the test of 20 items. According to the Bilog chi-square test, there was statistically significant misfit for 12 of the 20 items at .05 level and for eight of these 12 items there was significant misfit at .01 level. The average information of the 40 items test was .9689 (see Figure 47) and for this selected 20 items test the average information was .6581 (see Figure 48), while the first 20 items gave an average information of .5983.

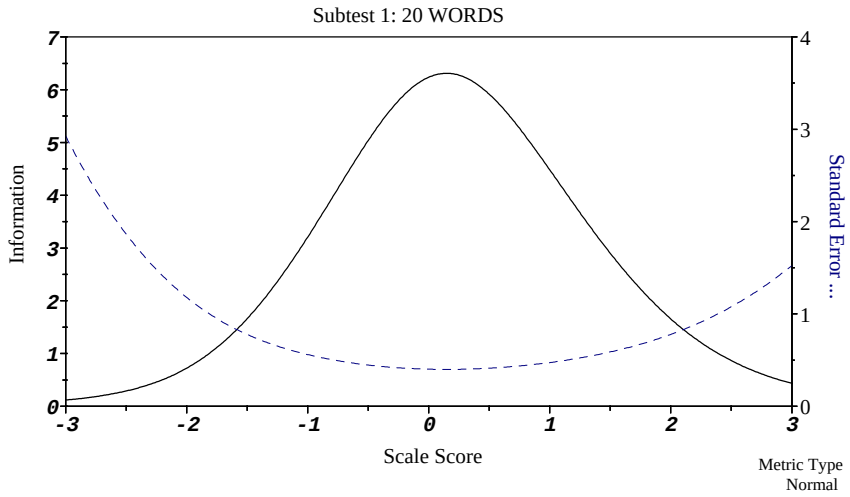

Figure 48. Test information curve and measurement error in the selected 20 items WORD subtest.

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 WORD subtest are quite promising. Without doubt the results from the classical test theory gave support for the need of a three parameter model. The statistical tests were not too discouraging either since the model data fit was acceptable for most items. Also the separate parameter estimates for males and females corresponded fairly well. Some items were revealed as less good, however, but when the 20 ”best” items were selected and anlysed the results from the statistical test were less encouraging. There was significant model data misfit for several items in the 20 item test whereas for the same items there was acceptable model data fit in the 40 item test. The main reason for this, however, is probably that the ability estimates are less sure with a shorter test. The reliability of these selected 20 items was higher than would have been expected for 20 randomly chosen items from the 40 item test.

After these separate investigations of the five subtests in the SweSAT (see also Stage, 1996, 1997a,b,c) it seems fruitful and important to investigate different combinations as well as the total test. The use of IRT would make it possible to compile the subtests in a more efficient way than the present. The use of IRT would also simplify and improve the equating of separate administrations of the SweSAT. Finally the use of IRT would be necessary if some day a version of the SweSAT or maybe the regular test should be computer adaptively administered.

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.

Stage, C. (1997c) *The Applicability of Item Response Models to the SweSAT. A Study of the READ Subtest* (Educational Measurement No. 25). Umeå, Umeå University, Department of Educational Measurement.