# Psychometric Analyses of the 2006 MCAS High School Technology/Engineering Test[1,2]

**Yue Zhao and Ronald K. Hambleton**
**University of Massachusetts Amherst**

**February 16, 2008**

[1]Financial support for completing this research came from Measured Progress and the Massachusetts Department of Education. The authors are responsible for the contents of the report and for any errors or misinterpretations of the data that may remain.
[2]**Center for Educational Assessment Research Report No. 648**. Amherst, MA: University of Massachusetts, Center for Educational Assessment.

## 1. Goal of the Psychometric Analyses

The primary goal of our work has been to provide readers with a number of worthwhile psychometric analyses of the 2006 MCAS high school Technology/Engineering Test (T/E). These analyses provide more detail on the Technology/Engineering Test than it was possible to provide in the summary report prepared by Hambleton, Zhao, Smith, Lam, and Deng (2008). These analyses include (1) an item analysis, (2) descriptive statistics on the test scores including break-outs for several subgroups of students, (3) classical reliability analyses for the test scores organized by item format, and for the total test, (4) investigations of test dimensionality, (5) item response theory (IRT) item calibrations obtained from fitting the three-parameter logistic model to binary-scored items and the graded response model to polytomously-scored items, (6) various item and test level model fit findings, (7) test information and conditional standard errors, and (8) the identification of differentially functioning test items.

## 2. Description of the Technology/Engineering Test

The 2006 MCAS Grade 9/10 Technology/Engineering Test consisted of 45 items assessing six standards (sometimes called “curriculum strands”): More about the curriculum strands can be found in the *Massachusetts Science and Technology/Engineering Curriculum Framework* (2006). The test was administered in a 2-day session in May of 2006. Each session included multiple-choice and open-response questions. More information about the curriculum and the test items can be found at **www.doe.mass.edu**.

Table 2.1 presents the number of items, by item type, and the total number of items and score points. The 2006 MCAS Technology/Engineering Test included 45 items, 40 of which were multiple-choice items (dichotomously scored) and five of which were open-response items (polytomously-scored, 0 to 4). The maximum score for the multiple-choice items was 1 point and the maximum for the open-response items was 4 points, so that the test has a minimum raw score of 0 points and a maximum score of 60 points. The open-response items were item numbers 11, 25, 26, 32, and 39.

**Table 2.1 Test Information**

| Item Type | Number of Items | Number of Points |
|---|---|---|
| Multiple-Choice | 40 | 40 |
| Open-Response | 5 | 20 |
| Total | 45 | 60 |

## 3. Classical Item Analysis

For all analyses, examinees were excluded if their raw scores were equal to zero or left blank. Thus, the sample size was 2461 for the analyses, whereas the original sample size was 2695.

All items were evaluated in terms of classical item difficulty and item discrimination. Item difficulty (or *p*-value) was measured by averaging the points across all students who were presented the item. For dichotomously-scored items, such as multiple choice items in the test, the item difficulty index is the proportion of students who answer an item correctly. For polytomously-scored items, the item difficulty index can be calculated as the mean score on an item divided by the total score points of the item. In the Technology/Engineering Test, the *p*-values ranged from 0.21 to 0.83, with a mean of 0.49 and a standard deviation of 0.14. It is clear that the difficulty indices range from near-chance performance to moderately easy for the examinees. The distribution of *p* values for the 45 items is reported in Table 3.1 and shown graphically in Figure 3.1.

Item discrimination index (*r*-value) refers to item-test correlations in classical test theory, which can be interpreted as a measure of item construct consistency since they measure how closely an item assesses the same knowledge and skills as other items. For dichotomous items, the statistic is commonly called a point-biserial correlation; for polytomous items, the item discrimination index is simply the value of the Pearson product-moment correlation. In theory, the *r* values range from –1 to +1, but usually range from 0.2 to 0.6 in practice. In the Technology/Engineering Test, the *r*-values ranged from 0.08 to 0.57, with a mean of 0.34 and standard deviation of 0.11. As Table 3.2 and Figure 3.2 show, the *r* values on the 45 items are distributed widely and the five polytomous items have *r*-values from 0.5 to 0.6, which are the highest, as expected. The wide range of both p and r values strongly influenced our decision to move forward with the three-parameter logistic test model and the graded response model when fitting an IRT model (Hambleton, Swaminathan, & Rogers, 1991).

A distractor analysis was not carried out because the information was not available to us on the data files.

**Table 3.1 Distribution of Classical Item Difficulty Indices**

| Group | Range | Frequency |
|---|---|---|
| 1 | 0.000-0.100 | 0 |
| 2 | 0.101-0.200 | 0 |
| 3 | 0.201-0.300 | 5 |
| 4 | 0.301-0.400 | 5 |
| 5 | 0.401-0.500 | 15 |
| 6 | 0.501-0.600 | 11 |
| 7 | 0.601-0.700 | 4 |
| 8 | 0.701-0.800 | 4 |
| 9 | 0.801-0.900 | 1 |
| 10 | 0.901-1.000 | 0 |

**Table 3.2 Distribution of Classical Item Discrimination Indices**

| Group | Range | Frequency |
|---|---|---|
| 1 | 0.000-0.100 | 2 |
| 2 | 0.101-0.200 | 3 |
| 3 | 0.201-0.300 | 8 |
| 4 | 0.301-0.400 | 19 |
| 5 | 0.401-0.500 | 8 |
| 6 | 0.501-0.600 | 5 |
| 7 | 0.601-0.700 | 0 |
| 8 | 0.701-0.800 | 0 |
| 9 | 0.801-0.900 | 0 |
| 10 | 0.901-1.000 | 0 |

**Table 3.3 Summary of Classical Item Difficulty and Item Discrimination Indices, Reported by Item Format**

| Item Difficulty | | | | | | Item Discrimination | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MCQ | | Performance | | Total | | MCQ | | Performance | | Total | |
| Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| 0.51 | 0.14 | 0.36 | 0.11 | 0.49 | 0.14 | 0.32 | 0.09 | 0.55 | 0.02 | 0.34 | 0.11 |

**Figure 3.1 Histogram Showing the Distribution of Classical Item Difficulty Indices**

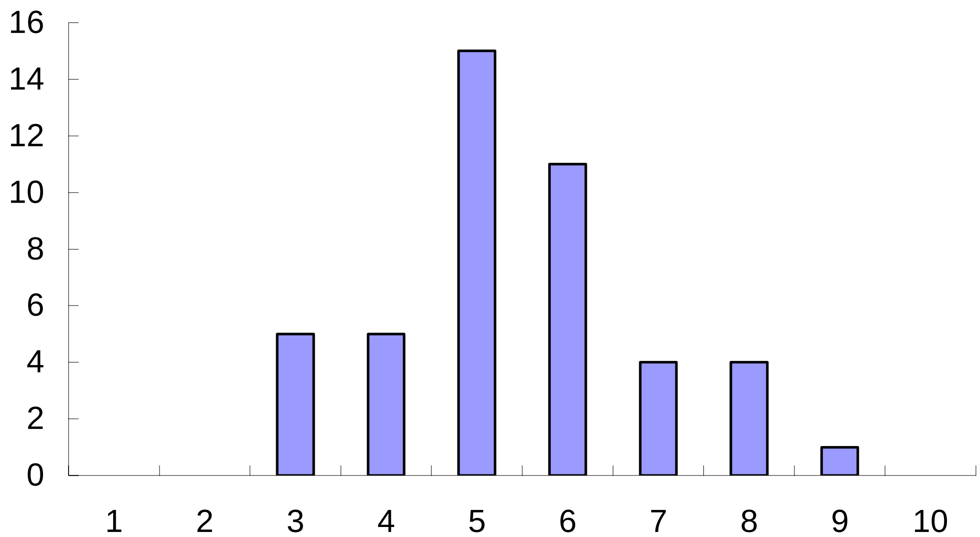



**Figure 3.2 Histogram Showing the Distribution of Classical Item Discrimination Indices**

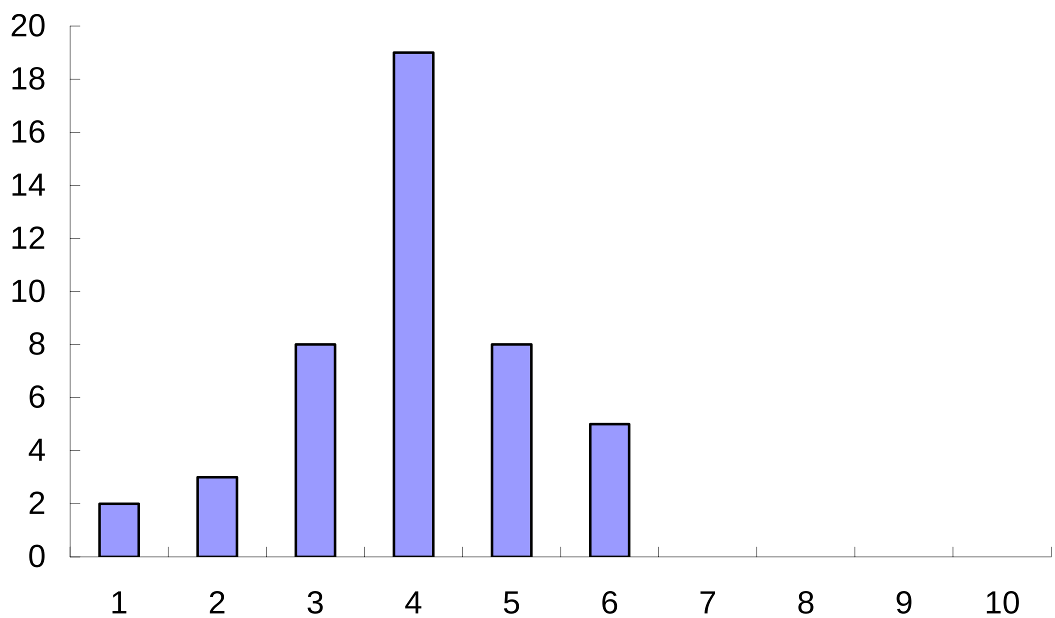


**Table 3.4 Classical Item Statistics (N=2461)**

| Item Order | Item Type | Item Mean | SD | p | r |
|---|---|---|---|---|---|
| 1 | MC | 0.83 | 0.38 | 0.83 | 0.33 |
| 2 | MC | 0.72 | 0.45 | 0.72 | 0.29 |
| 3 | MC | 0.35 | 0.48 | 0.35 | 0.19 |
| 4 | MC | 0.52 | 0.50 | 0.52 | 0.43 |
| 5 | MC | 0.49 | 0.50 | 0.49 | 0.33 |
| 6 | MC | 0.72 | 0.45 | 0.72 | 0.40 |
| 7 | MC | 0.27 | 0.45 | 0.27 | 0.17 |
| 8 | MC | 0.39 | 0.49 | 0.39 | 0.23 |
| 9 | MC | 0.62 | 0.49 | 0.62 | 0.41 |
| 10 | MC | 0.61 | 0.49 | 0.61 | 0.20 |
| 11 | OR | 1.15 | 1.29 | 0.29 | 0.54 |
| 12 | MC | 0.24 | 0.43 | 0.24 | 0.24 |
| 13 | MC | 0.35 | 0.48 | 0.35 | 0.08 |
| 14 | MC | 0.57 | 0.50 | 0.57 | 0.30 |
| 15 | MC | 0.38 | 0.49 | 0.38 | 0.23 |
| 16 | MC | 0.45 | 0.50 | 0.45 | 0.28 |
| 17 | MC | 0.54 | 0.50 | 0.54 | 0.36 |
| 18 | MC | 0.61 | 0.49 | 0.61 | 0.36 |
| 19 | MC | 0.42 | 0.49 | 0.42 | 0.29 |
| 20 | MC | 0.59 | 0.49 | 0.59 | 0.45 |
| 21 | MC | 0.41 | 0.49 | 0.41 | 0.31 |
| 22 | MC | 0.43 | 0.50 | 0.43 | 0.31 |
| 23 | MC | 0.54 | 0.50 | 0.54 | 0.32 |
| 24 | MC | 0.76 | 0.43 | 0.76 | 0.43 |
| 25 | OR | 1.74 | 1.25 | 0.43 | 0.55 |
| 26 | OR | 1.57 | 1.12 | 0.39 | 0.57 |
| 27 | MC | 0.78 | 0.42 | 0.78 | 0.41 |
| 28 | MC | 0.53 | 0.50 | 0.53 | 0.31 |
| 29 | MC | 0.49 | 0.50 | 0.49 | 0.30 |
| 30 | MC | 0.23 | 0.42 | 0.23 | 0.08 |
| 31 | MC | 0.45 | 0.50 | 0.45 | 0.32 |
| 32 | OR | 0.82 | 1.04 | 0.21 | 0.52 |
| 33 | MC | 0.50 | 0.50 | 0.50 | 0.43 |
| 34 | MC | 0.48 | 0.50 | 0.48 | 0.34 |
| 35 | MC | 0.48 | 0.50 | 0.48 | 0.28 |
| 36 | MC | 0.58 | 0.49 | 0.58 | 0.38 |
| 37 | MC | 0.68 | 0.47 | 0.68 | 0.37 |

| | | | | | |
|---|---|---|---|---|---|
| 38 | MC | 0.52 | 0.50 | 0.52 | 0.43 |
| 39 | OR | 1.91 | 1.35 | 0.48 | 0.57 |
| 40 | MC | 0.58 | 0.49 | 0.58 | 0.48 |
| 41 | MC | 0.52 | 0.50 | 0.52 | 0.34 |
| 42 | MC | 0.40 | 0.49 | 0.40 | 0.30 |
| 43 | MC | 0.41 | 0.49 | 0.41 | 0.32 |
| 44 | MC | 0.45 | 0.50 | 0.45 | 0.24 |
| 45 | MC | 0.45 | 0.50 | 0.45 | 0.37 |

## 4. Reliability Analyses and Basic Statistics

The test score distribution for the 2461 examinees had a mean score of 27.5 with a standard deviation of 10.6. The 40 multiple-choice items have a mean 20.4 (40 point maximum) and a standard deviation 7.1; and the five open-response items have a mean 7.2 (20 point maximum) and a standard deviation 4.3, as shown in Table 4.1 and Figure 4.1. Clearly, the open-response items were relatively more difficult for students than the multiple-choice items.

The descriptive statistics of test scores were computed separately in each of the gender and the ethnic groups. Regarding gender (see Table 4.2), the female group has a mean of 27.0 and standard deviation of 9.4 and the male group has a mean of 28.1 and standard deviation of 11.1. Males performed a little better, and were more variable that the females in the test sample. For the ethnic groups, the means and standard deviations were reported in Table 4.3 and we could see that the White (W) group performed substantially better than any of the other groups (except for the Asian sample, and this group was very small).

With respect to reliability (see Table 4.1), it was calculated for the test as a whole using Cronbach's coefficient alpha, as well as for the multiple-choice and the open-response items, separately. There was a high overall reliability (α = 0.87) and the reliabilities for the different item types were a bit lower (MCQ: α = 0.84; open-response: α = 0.75).

**Table 4.1 Test Score Descriptive Statistics**

| Items | N | Sample Size | Mean | SD | Reliability (Coefficient.Alpha) |
|---|---|---|---|---|---|
| Total | 45 | 2641 | 27.54 | 10.64 | 0.87 |
| MCQ | 40 | 2641 | 20.35 | 7.14 | 0.84 |
| Performance | 5 | 2641 | 7.19 | 4.30 | 0.75 |

**Figure 4.1 Test Score Distribution for the Total Group of Students**

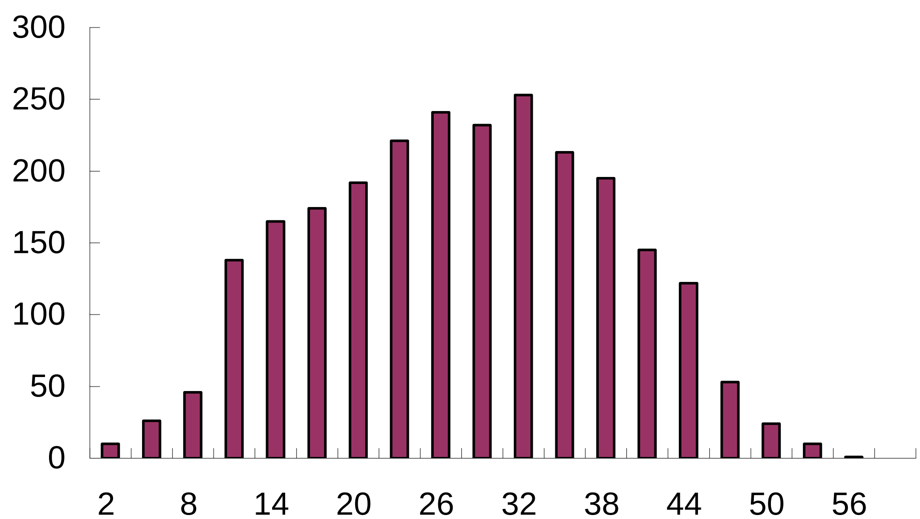



**Table 4.2 Test Score Descriptive Statistics, Reported by Gender**

| | N | Percent (%) | Mean | SD |
|---|---|---|---|---|
| Missing | 97 | 3.9 | 21.18 | 9.11 |
| Female | 737 | 29.9 | 27.03 | 9.43 |
| Male | 1627 | 66.1 | 28.14 | 11.10 |
| Total | 2461 | 100.0 | 27.54 | 10.64 |

**Table 4.3 Test Score Descriptive Statistics, Reported by Ethnic Group**

| | N | Percent (%) | Mean | SD |
|---|---|---|---|---|
| Missing | 98 | 4.0 | 21.34 | 9.20 |
| Asian | 59 | 2.4 | 28.39 | 10.78 |
| Black | 172 | 7.0 | 19.32 | 8.78 |
| Hispanic | 246 | 10.0 | 21.36 | 8.76 |
| NativeAmerican | 10 | 0.4 | 18.00 | 9.99 |
| White | 1876 | 76.2 | 29.45 | 10.28 |
| Total | 2461 | 100.0 | 27.54 | 10.64 |

It was clear from the sample sizes and the test scores that our ethnic DIF analyses would be very limited—because of non-overlapping score distributions and the small samples even in the two largest minority groups. On the other hand, gender DIF would be possible to study.

## 5. Test Dimensionality Analysis

An initial check of test dimensionality was obtained by considering the correlation between MCQ and open-response test scores. Often multiple item formats in a test provide a way for assessing many different skills, and so the potential is present for introducing test multidimensionality, a condition that would undermine the unidimensionality assumption which is made in all of the common applications of IRT. As shown in Table 5.1, the correlation between multiple-choice scores and open response scores is 0.71, and after correcting for the unreliability of each score, the estimated correlation between true MCQ and open-response scores was .89. This high correlation suggests that multidimensionality is not present to any great extent because of the use of multiple item formats in the test.

Further, eigenvalues and eigenvectors were calculated based on the $45 \times 45$ item correlation matrix, and the first ten largest eigenvalues based on the total sample of 2461 examinees are reported in Table 5.2 and shown graphically in Figure 5.1. A dominant first factor is clearly present. The table show a large first eigenvalue, which suggests that there is one dominant factor or dimension since the first eigenvalue exceeded the second one by a ratio of more than 5:1 and the first factor accounted for more than 20% of the variability. These are standard checks on the unidimensionality of a test.

To prove whether the first factor is distinguished from the other factors, a parallel analysis was conducted. The parallel analysis provides for a comparison of the actual eigenvalues with a baseline of eigenvalues using simulated data which are produced by generating random normal deviates of item responses. Since the simulated data are randomly generated, the eigenvalues too are random, and the largest one provides an indication of how big an eigenvalue can be from a random process. It provides a baseline for distinguishing real from random factors. From Figure 5.2, the first eigenvalue is 1.41 based on the parallel analysis, and there are two eigenvalues of the actual data (10.76 and 2.04) which are larger than 1.41. It suggested that there are two factors in the test, one which one is dominant and the other one appears to be very small.

As a final check on test dimensionality, we used LISREL to fit a one-factor model to the available item response data. Table 5.3 presents the factor loadings and in Appendix A, the same information is displayed graphically. The evidence for a single factor underlying the data is clear. Factor loadings on the single factor are moderate to high for all of the items in the test (except for three items, using .30 as a criterion for interpreting the factor loadings).

**Table 5.1 Correlations Among Test Scores**

| | Total | MCQ | Open-Response |
|---|---|---|---|
| Total | 1.00 | | |
| MCQ | 0.96 | 1.00 | |
| Open-Response | 0.88 | 0.71 | 1.00 |

**Note:**
"Total" refers to total scores based on all items;
"MCQ" refers to total scores based on MCQ items only;
"Open-Response" refers to total scores based on open-response items only.

**Table 5.2 Largest 10 Eigenvalues for the 45 Test Items**
**(2,461 students, excluding students with a missing or zero test score)**

| Rank | Eigenvalue | Variance Accounted For |
|---|---|---|
| 1 | 10.76 | 24% |
| 2 | 2.04 | 5% |
| 3 | 1.41 | 3% |
| 4 | 1.36 | 3% |
| 5 | 1.17 | 3% |
| 6 | 1.12 | 2% |
| 7 | 1.11 | 2% |
| 8 | 1.08 | 2% |
| 9 | 1.03 | 2% |
| 10 | 1.01 | 2% |

**Figure 5.1 Eigenvalue Plot (2461 Students)**

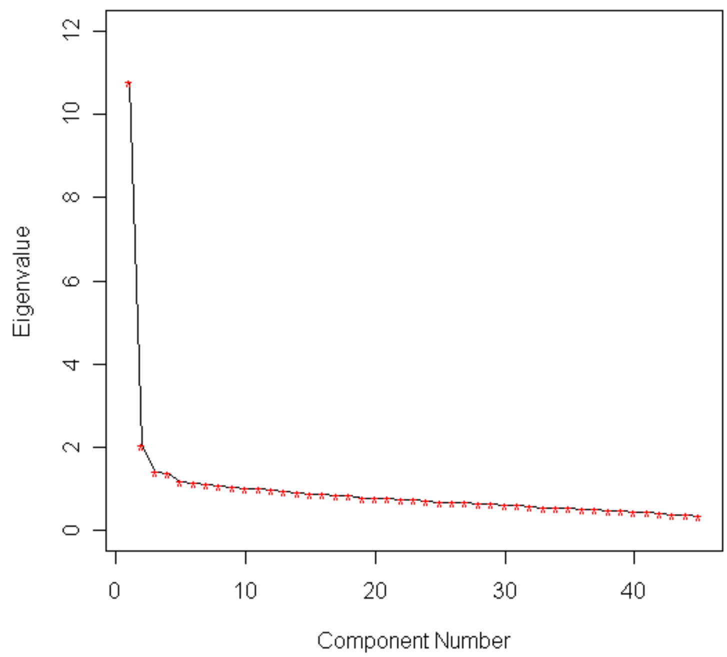



**Figure 5.2 Parallel Analysis of the 45 Item Using Random Normal Deviates with p-values Controlled (The average of the largest eigenvalue was 1.41.)**

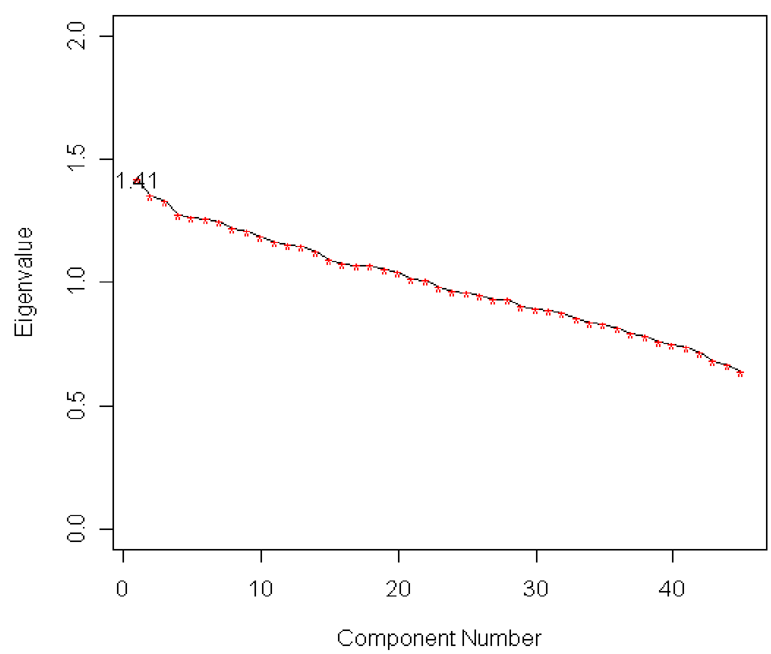


**Table 5.3 Factor Loadings for a One Factor Model (Obtained Using LISREL)**

| Item | Factor Loading |
|---|---|
| 1 | 0.62 |
| 2 | 0.49 |
| 3 | 0.37 |
| 4 | 0.64 |
| 5 | 0.47 |
| 6 | 0.65 |
| 7 | 0.29 |
| 8 | 0.36 |
| 9 | 0.60 |
| 10 | 0.37 |
| 11 | 0.73 |
| 12 | 0.38 |
| 13 | 0.13 |
| 14 | 0.46 |
| 15 | 0.36 |
| 16 | 0.42 |
| 17 | 0.54 |
| 18 | 0.61 |
| 19 | 0.47 |
| 20 | 0.68 |
| 21 | 0.47 |
| 22 | 0.46 |
| 23 | 0.50 |
| 24 | 0.64 |
| 25 | 0.68 |
| 26 | 0.73 |
| 27 | 0.71 |
| 28 | 0.49 |
| 29 | 0.50 |
| 30 | 0.16 |
| 31 | 0.44 |
| 32 | 0.70 |
| 33 | 0.68 |
| 34 | 0.51 |
| 35 | 0.48 |
| 36 | 0.60 |
| 37 | 0.58 |
| 38 | 0.64 |

| | |
|---|---|
| 39 | 0.71 |
| 40 | 0.76 |
| 41 | 0.56 |
| 42 | 0.50 |
| 43 | 0.53 |
| 44 | 0.44 |
| 45 | 0.50 |

In summary, the findings are clear that the item response data are strongly unidimensionality and are likely to be fit by a unidimensional IRT model. The results of our efforts to fit the data with an IRT model follow next.

## 6. Item Calibrations and Model Fit

Item parameters were calibrated with the PARSCALE software, and the estimates are reported in Table 6.1. The 3p model was fit to the binary-scored items, and the graded response model was fit to the polytomously-scored items. PARSCALE also provides an item level fit (chi-square) statistic for each item to serves as evidence of model fit. While we don't like these statistics very much because of their dependence on sample size, with this test, the sample size was not overly large, and so the item fit statistics are less problematic. They showed in this instance that model fit (at the .01 level) was good except for a small number of test items: 25, 26, 37, and 39. Among the four items, one (item 37) was dichotomously scored, and the other three were polytomously- scored. However, these statistics still might be biased due to sample size or small cell frequencies in certain proficiency intervals. Thus, we produced two types of fit plots to investigate further: Residual plots and probability plots (see Appendix B), which were generated by the computer program ResiFIT (prepared by the first author). The residual plots highlighted items 9, 10, 27, 37, and 39 as being problematic. The more compelling evidence for model fit comes from the residual analyses. For several of the test items it appears that the misfit is associated with lower performing examinees.

Across the full set of items, the distribution of standardized residuals was produced and the statistics shown in Figure 6.1 suggest a near normal a normal distribution with a mean of -0.07 and standard deviation of 0.92. This finding suggests that at the test level the fit of the models to the data is quite good.

Alternatively, test level fit was assessed by assuming model parameter estimates to be correct, and then predicting the actual test score distribution using some software prepared by Ning Han. Figure 6.2 shows the actual and the predicted score distributions and they are very close, suggesting model fit is excellent. But these distributions are a big ragged because of small sample sizes and so it is usually better to compare expected and actual cumulative relative frequency distributions. Figure 6.3 shows the two distributions being nearly identical, a finding that strongly supports model fit. Based on Figures 6.2 and 6.3, we could conclude that the three-parameter logistic model and graded response model fit the data very well.

**Table 6.1 2006 MCAS Grades 9/10 Technology/Engineering Test Item Parameter Estimates**

| Item | A | b | c | b1 | b2 | b3 | b4 |
|---|---|---|---|---|---|---|---|
| 1 | 0.78 | -0.94 | 0.37 | | | | |
| 2 | 0.79 | -0.03 | 0.43 | | | | |
| 3 | 1.18 | 1.54 | 0.26 | | | | |
| 4 | 0.93 | 0.24 | 0.15 | | | | |
| 5 | 1.46 | 0.82 | 0.31 | | | | |
| 6 | 0.90 | -0.39 | 0.28 | | | | |
| 7 | 1.26 | 1.70 | 0.20 | | | | |
| 8 | 0.53 | 1.42 | 0.19 | | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 9 | 0.99 | 0.10 | 0.26 | | | | |
| 10 | 0.34 | -0.23 | 0.18 | | | | |
| 11 | 0.89 | 1.09 | 0.00 | 1.11 | 0.68 | -0.07 | -1.72 |
| 12 | 0.87 | 1.68 | 0.13 | | | | |
| 13 | 1.09 | 2.34 | 0.32 | | | | |
| 14 | 0.51 | -0.01 | 0.14 | | | | |
| 15 | 0.80 | 1.42 | 0.24 | | | | |
| 16 | 0.66 | 0.93 | 0.22 | | | | |
| 17 | 1.12 | 0.56 | 0.29 | | | | |
| 18 | 0.73 | -0.02 | 0.21 | | | | |
| 19 | 1.18 | 1.02 | 0.27 | | | | |
| 20 | 1.29 | 0.21 | 0.26 | | | | |
| 21 | 0.69 | 0.94 | 0.16 | | | | |
| 22 | 0.93 | 0.93 | 0.23 | | | | |
| 23 | 0.57 | 0.23 | 0.16 | | | | |
| 24 | 0.84 | -0.88 | 0.13 | | | | |
| 25 | 0.83 | 0.50 | 0.00 | 1.43 | 1.00 | -0.17 | -2.26 |
| 26 | 0.88 | 0.64 | 0.00 | 1.78 | 0.85 | -0.65 | -1.97 |
| 27 | 0.93 | -0.75 | 0.26 | | | | |
| 28 | 0.65 | 0.44 | 0.21 | | | | |
| 29 | 1.01 | 0.85 | 0.29 | | | | |
| 30 | 1.23 | 2.23 | 0.20 | | | | |
| 31 | 0.56 | 0.58 | 0.10 | | | | |
| 32 | 0.87 | 1.59 | 0.00 | 1.51 | 0.51 | -0.47 | -1.55 |
| 33 | 0.91 | 0.29 | 0.13 | | | | |
| 34 | 0.57 | 0.28 | 0.07 | | | | |
| 35 | 0.47 | 0.46 | 0.11 | | | | |
| 36 | 0.65 | -0.21 | 0.08 | | | | |
| 37 | 0.64 | -0.70 | 0.09 | | | | |
| 38 | 0.87 | 0.19 | 0.13 | | | | |
| 39 | 0.91 | 0.26 | 0.00 | 1.18 | 0.71 | 0.14 | -2.03 |
| 40 | 1.07 | 0.01 | 0.15 | | | | |
| 41 | 0.61 | 0.25 | 0.12 | | | | |
| 42 | 0.59 | 0.91 | 0.13 | | | | |
| 43 | 0.62 | 0.76 | 0.11 | | | | |
| 44 | 0.41 | 0.78 | 0.11 | | | | |
| 45 | 0.78 | 0.61 | 0.15 | | | | |

**Table 6.2 Model Item Fit Statistics for the 2006 MCAS Technology/Engineering Test**

| Item | Chi-Square | DF | Prob |
|---|---|---|---|
| 1 | 15.07 | 24 | 0.92 |
| 2 | 16.18 | 27 | 0.95 |
| 3 | 32.48 | 30 | 0.35 |
| 4 | 15.35 | 28 | 0.97 |
| 5 | 29.96 | 28 | 0.37 |
| 6 | 27.44 | 25 | 0.33 |
| 7 | 23.71 | 30 | 0.79 |
| 8 | 24.42 | 30 | 0.75 |
| 9 | 39.96 | 27 | 0.05 |
| 10 | 49.10 | 30 | 0.02 |
| 11 | 92.28 | 89 | 0.39 |
| 12 | 26.50 | 30 | 0.65 |
| 13 | 24.80 | 30 | 0.74 |
| 14 | 38.03 | 30 | 0.15 |
| 15 | 27.16 | 30 | 0.62 |
| 16 | 34.23 | 30 | 0.27 |
| 17 | 16.15 | 28 | 0.96 |
| 18 | 44.32 | 29 | 0.03 |
| 19 | 39.10 | 30 | 0.12 |
| 20 | 29.81 | 26 | 0.28 |
| 21 | 32.37 | 30 | 0.35 |
| 22 | 23.05 | 30 | 0.81 |
| 23 | 27.31 | 30 | 0.61 |
| 24 | 34.50 | 24 | 0.08 |
| 25 | 125.33 | 93 | 0.01 |
| 26 | 137.67 | 90 | 0.00 |
| 27 | 40.85 | 24 | 0.02 |
| 28 | 30.78 | 30 | 0.43 |
| 29 | 31.37 | 30 | 0.40 |
| 30 | 41.90 | 30 | 0.07 |
| 31 | 44.58 | 30 | 0.04 |
| 32 | 102.87 | 81 | 0.05 |
| 33 | 34.12 | 30 | 0.28 |
| 34 | 40.09 | 30 | 0.10 |
| 35 | 41.81 | 30 | 0.07 |
| 36 | 37.68 | 30 | 0.16 |
| 37 | 53.99 | 27 | 0.00 |

| | | | |
|---|---|---|---|
| 38 | 29.43 | 30 | 0.50 |
| 39 | 137.22 | 92 | 0.00 |
| 40 | 35.61 | 27 | 0.12 |
| 41 | 43.82 | 30 | 0.05 |
| 42 | 27.96 | 30 | 0.57 |
| 43 | 44.15 | 30 | 0.05 |
| 44 | 48.65 | 30 | 0.02 |
| 45 | 31.47 | 30 | 0.39 |

**Table 6.3 Summary Statistics of the IRT Item Parameter Estimates**

| Parameter | Mean | SD | N |
|---|---|---|---|
| **A** | 0.83 | 0.25 | 45 |
| **B** | 0.55 | 0.73 | 45 |
| **C** | 0.20 | 0.09 | 40 |
| **Proficiency Scores** | -0.01 | 0.93 | 2461 |

**Figure 6.1 Distribution of Standardized Residuals**

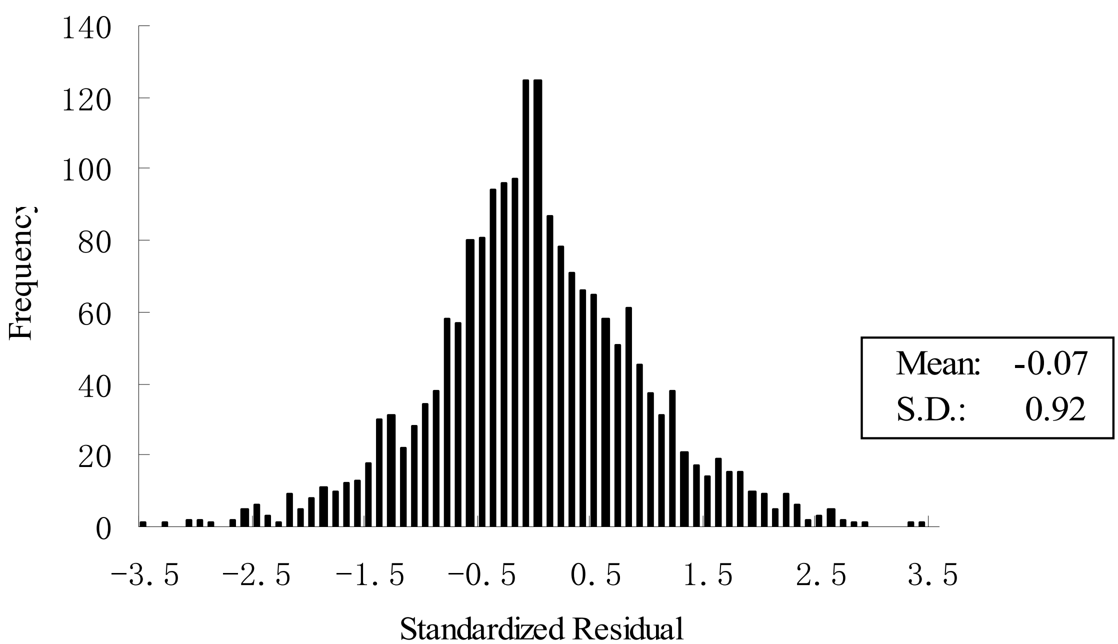


**Figure 6.2 Test Level Fit**

**(Observed versus Predicted Relative Frequency Distributions)**

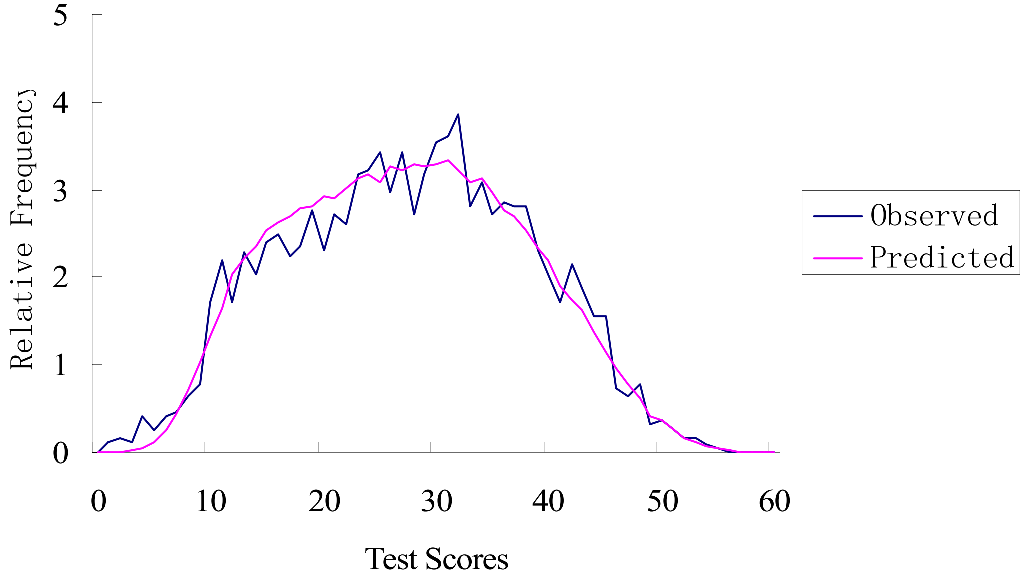
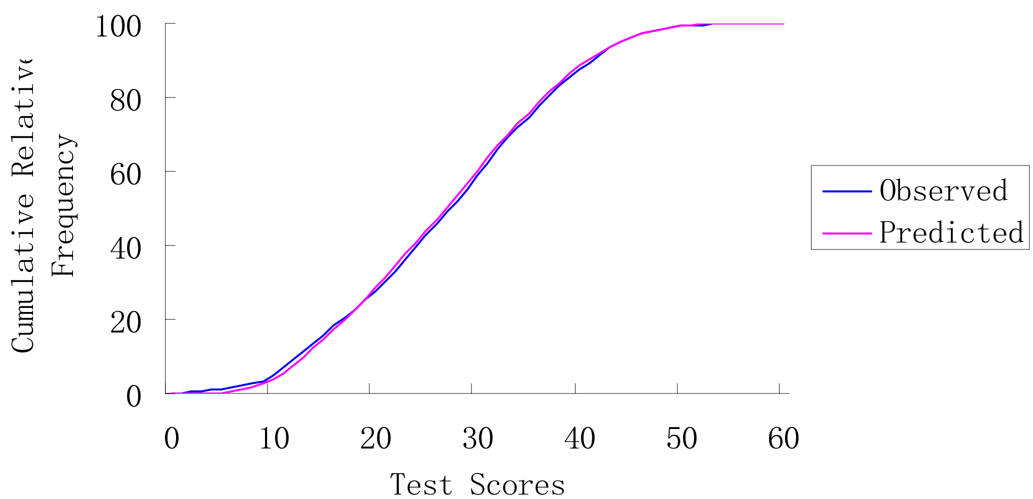



**Figure 6.3 Test Level Fit**

**Observed versus Predicted Relative Cumulative Frequency Distributions**



## 7. Test Information and Conditional Standard Errors

The test characteristic curve (TCC), test information function (TIF), and the standard error of measurement curves (SEM) are displayed in Figures 7.1, 7.2 and 7.3, respectively.

**Figure 7.1 Test Characteristic Curve**

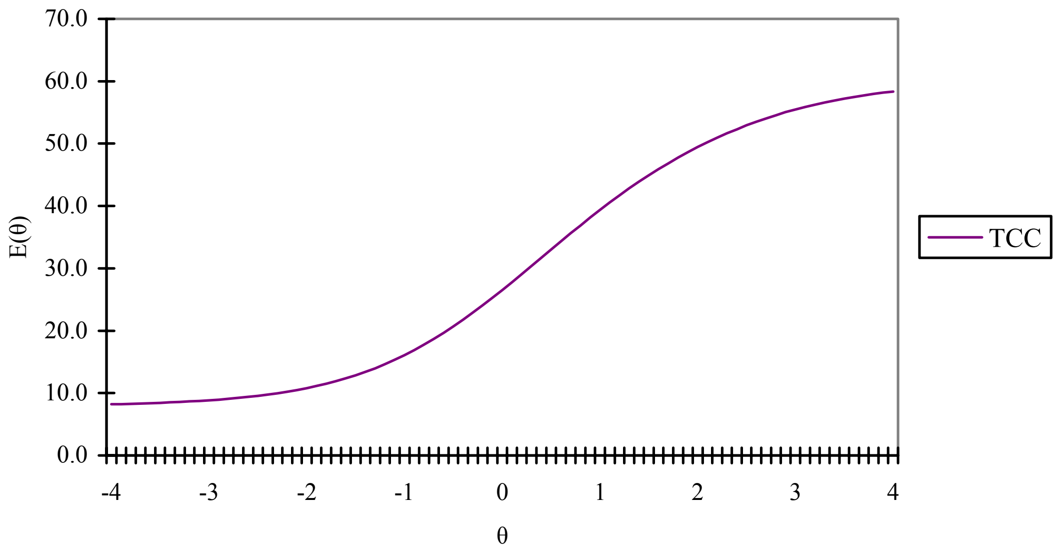



**Figure 7.2. Test Information Function**

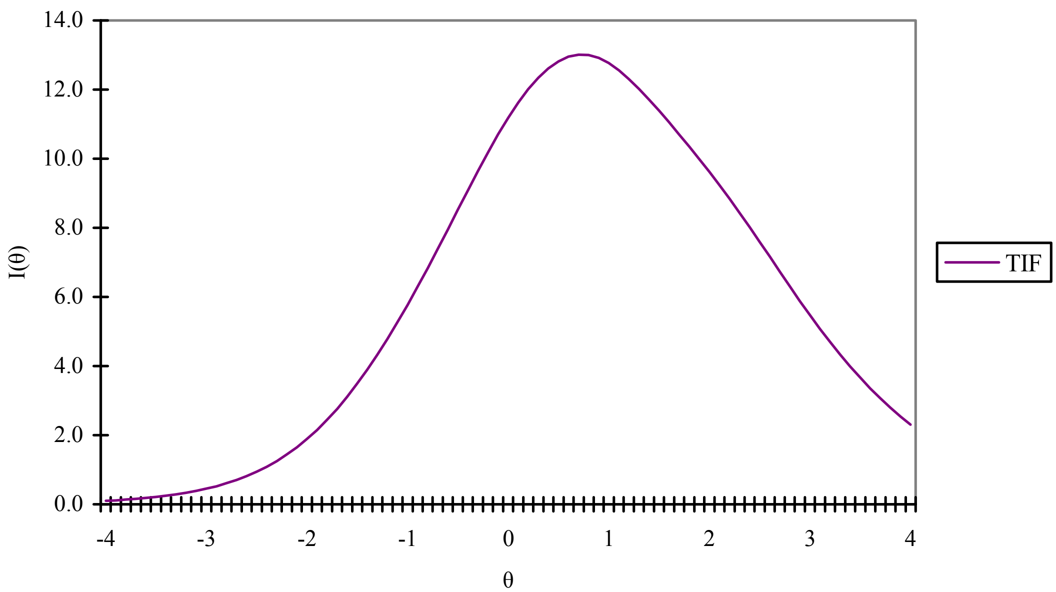


**Figure 7.3. Standard Error of Measurement**

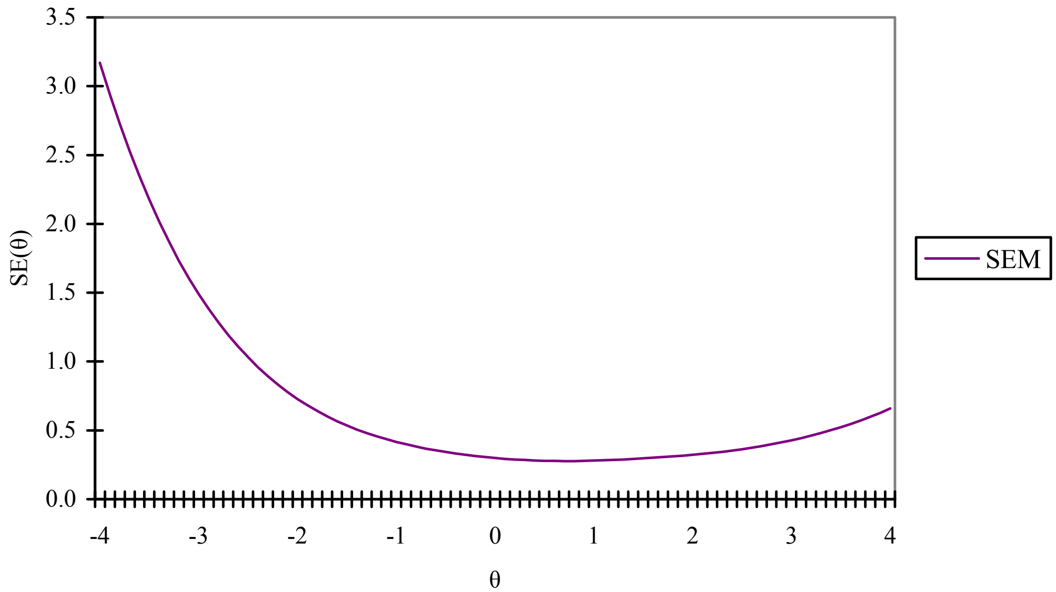



Figure 7.1 highlights again that the T/E test was generally difficult for students. The average student was not achieving a score of 50% on the test. Also, from Figures 7.2 and 7.3 it can be seen that the T/E test was providing a good level of measurement for students performing from about .5 SD below the mean to about two standard deviations above the mean. In future years, unless the anticipation is that there will be substantial student growth, the test might provide better measurement for more students were some of the more difficult items replaced with items providing good discrimination for students in the lower half of the test score distribution.

## 8. Identification of Differential Function Items

Considering the small sample sizes of ethnicity groups (see Tables 4.3 and 8.1), we approached the identification of ethnic DIF using a small sample approach. Only the

White-Hispanic comparison was investigated since sample sizes for the other ethnic groups were all less than 200. For the 40 dichotomously-scored items, Mantel-Haenszel statistics were computed and results are displayed in Table 8.2. Items 6 and 18 appear to be items worthy of a review, but they do not rise to the level of concern that is represented by C-type DIF items. For the five polytomously scored items, item mean differences conditioned on total test scores (1-15. 16-30, 31-45 and 46-60) were calculated for White and Hispanic groups. No DIF was found in the five items. See Figure 8.1 for a graphical presentation of these results.

**Table 8.1 Sample Sizes of the Ethnic Groups, in Four Test Score Groups**

| Ethnic Group | Test Score Group | | | | Total |
|---|---|---|---|---|---|
| | 1-15 | 16-30 | 31-45 | 46-60 | |
| Hispanic | 72 | 135 | 36 | 3 | 246 |
| White | 209 | 759 | 824 | 84 | 1876 |

**Table 8.2 Mantel-Haenszel Results for 40 Dichotomously Scored Items – White (N=1876) versus Hispanic (N=246)**

| Items | MH | DIF (MH>6.63 given $\alpha = 0.01$) | ETS Rule |
|---|---|---|---|
| Item 1 | 4.03 | OK | A |
| Item 2 | 1.14 | OK | A |
| Item 3 | 0.86 | OK | A |
| Item 4 | 2.03 | OK | A |
| Item 5 | 1.96 | OK | A |
| Item 6 | 7.40 | Flag | B |
| Item 7 | 1.57 | OK | A |
| Item 8 | 1.26 | OK | A |
| Item 9 | 0.51 | OK | A |
| Item 10 | 3.60 | OK | A |
| Item 12 | 1.00 | OK | A |
| Item 13 | 0.02 | OK | A |
| Item 14 | 4.91 | OK | A |
| Item 15 | 0.26 | OK | A |
| Item 16 | 0.58 | OK | A |

| | | | |
|---|---|---|---|
| Item 17 | 0.13 | OK | A |
| Item 18 | 15.15 | Flag | B |
| Item 19 | 0.41 | OK | A |
| Item 20 | 4.28 | OK | A |
| Item 21 | 0.00 | OK | A |
| Item 22 | 0.15 | OK | A |
| Item 23 | 0.01 | OK | A |
| Item 24 | 2.77 | OK | A |
| Item 27 | 3.72 | OK | A |
| Item 28 | 0.32 | OK | A |
| Item 29 | 0.62 | OK | A |
| Item 30 | 1.56 | OK | A |
| Item 31 | 1.23 | OK | A |
| Item 33 | 0.00 | OK | A |
| Item 34 | 0.01 | OK | A |
| Item 35 | 0.00 | OK | A |
| Item 36 | 0.13 | OK | A |
| Item 37 | 2.13 | OK | A |
| Item 38 | 0.75 | OK | A |
| Item 40 | 3.65 | OK | A |
| Item 41 | 1.98 | OK | A |
| Item 42 | 0.47 | OK | A |
| Item 43 | 2.53 | OK | A |
| Item 44 | 0.23 | OK | A |
| Item 45 | 0.09 | OK | A |

**Figure 8.1 Summary of Mantel-Haenszel Statistics for the 40 Dichotomously-Scored Items for White and Hispanic Groups**

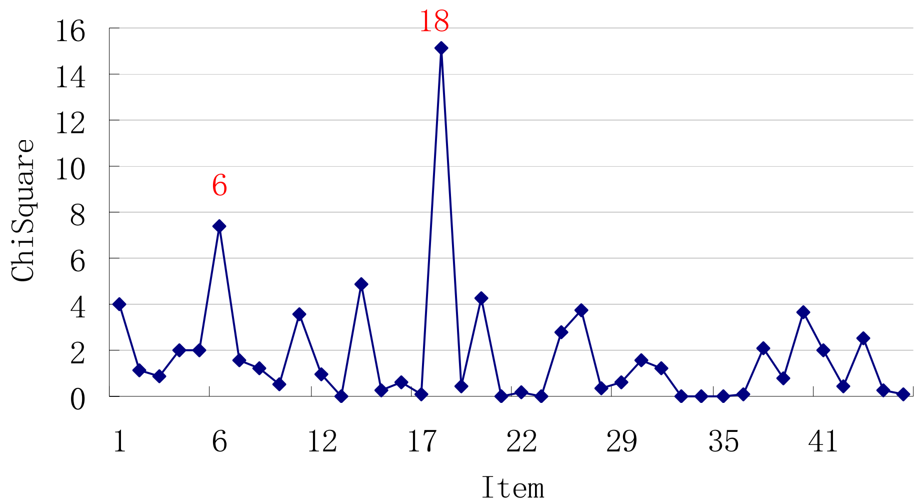
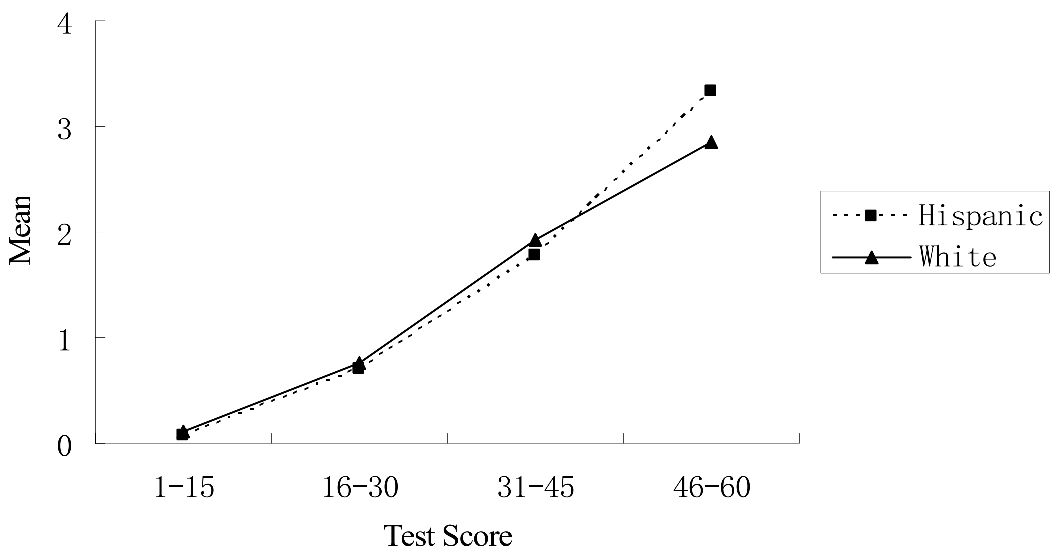
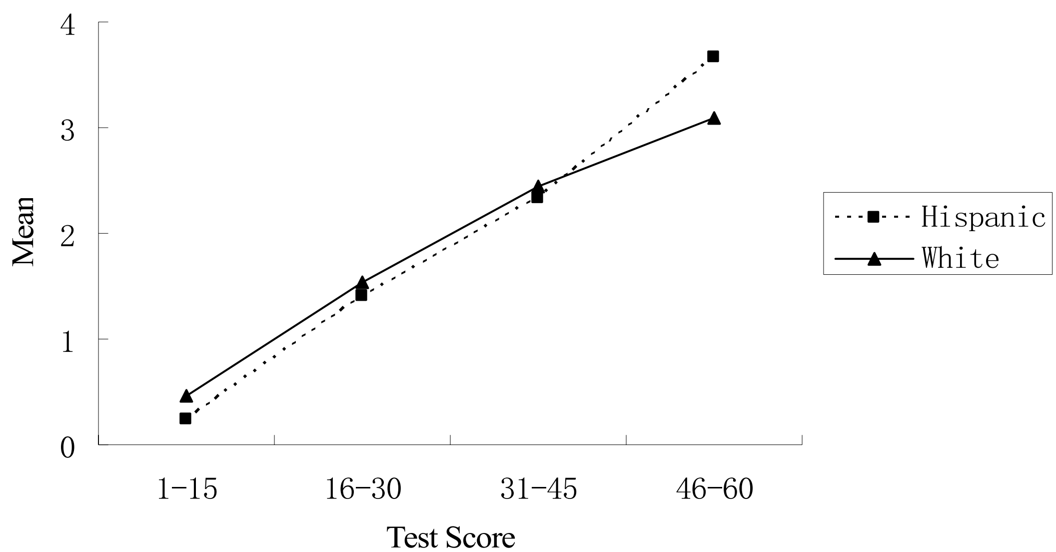



**Figure 8.2 Hispanic-White Group Differences on the Five Polytomously-Scored Items**

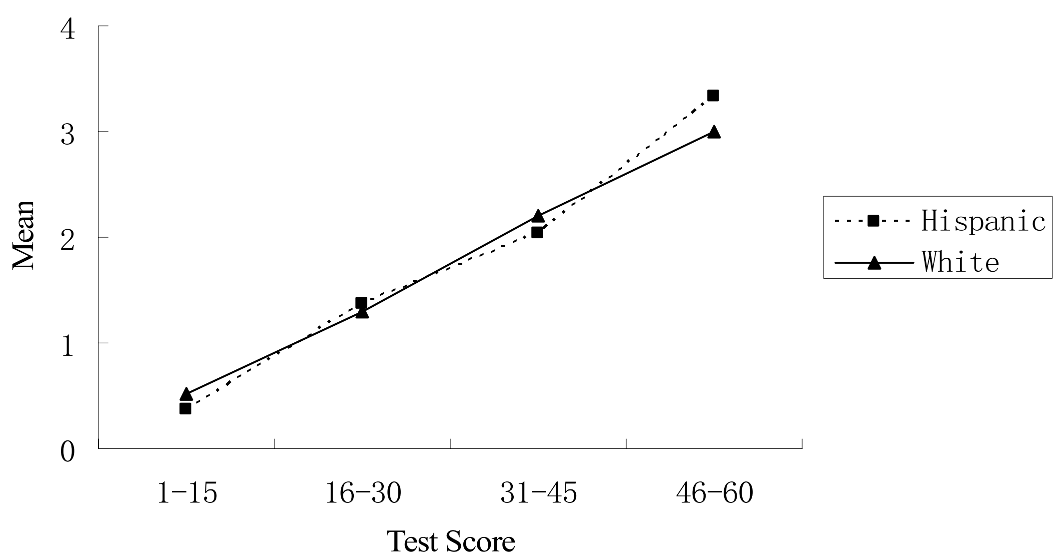
a. Item11
Mean Difference: 0.06
Absolute Mean Difference: 0.18



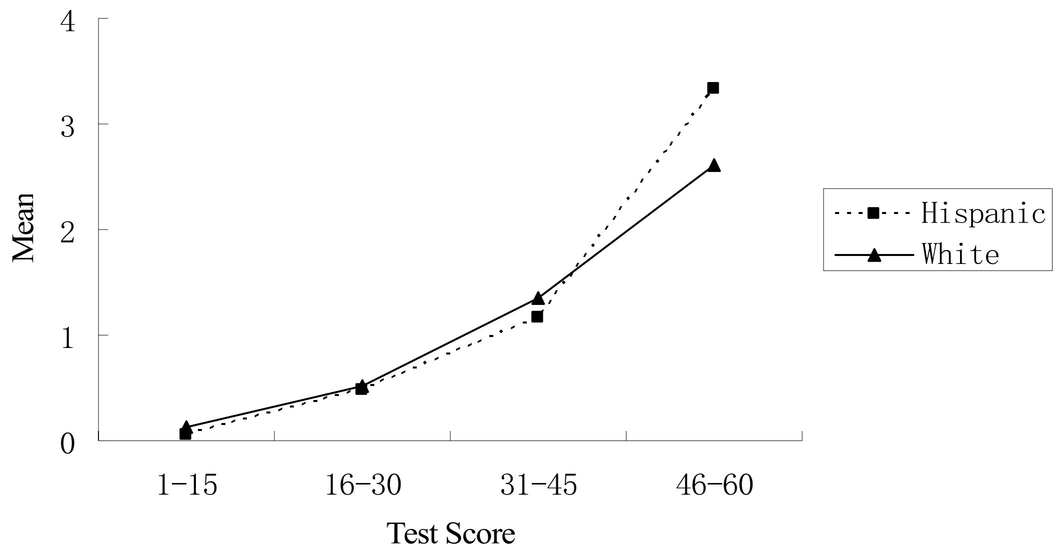
b. Item 25
Mean Difference: 0.03
Absolute Mean Difference: 0.26



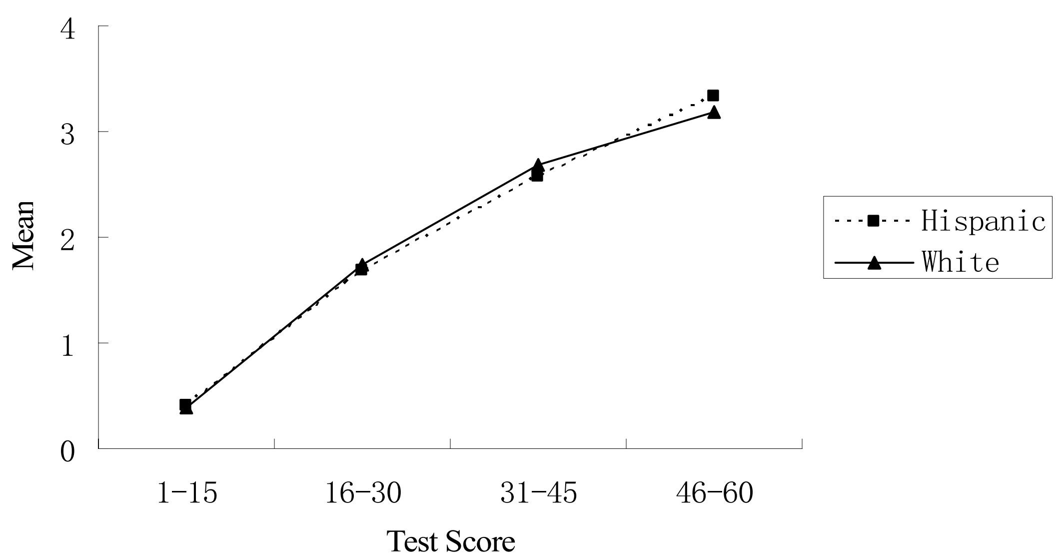

c. Item 26
Mean Difference: 0.02
Absolute Mean Difference: 0.18



d. Item 32
Mean Difference: 0.11
Absolute Mean Difference: 0.26




e. Item 39
Mean Difference: 0.003
Absolute Mean Difference: 0.08



For the gender groups, the test was examined using the computer program STDIF (Zenisky & Hambleton, 2007; Zenisky, Hambleton, & Robin, 2003, 2004). The program calculates the UDIF (unsigned DIF) statistics. This analysis was done in two stages. First, the program was run including all of the items when calculating the statistics. Then for the second stage, the items that showed DIF from the first stage were deleted from the conditioning variable to provide a bias-free matching variable. For the Male/Female comparison, males were the reference group. This analysis showed three DIF items (3, 18, and 21). Figure 8.3 displays the UDIF statistics but they are vey unstable because of the small samples (for this particular type of analysis). More interesting are the displays for the items 3, 18, and 21 shown in Figure 8.4. Though the graphs are unstable, there are clear and noticeable differences with the males outperforming the females. .

**Table 8.3 Sample Sizes of the Gender Groups**

| | Male | Female |
|---|---|---|
| Sample Size | 1627 | 737 |

**Figure 8.3 Summary of UDIF Statistics for Male-Female Comparisons**

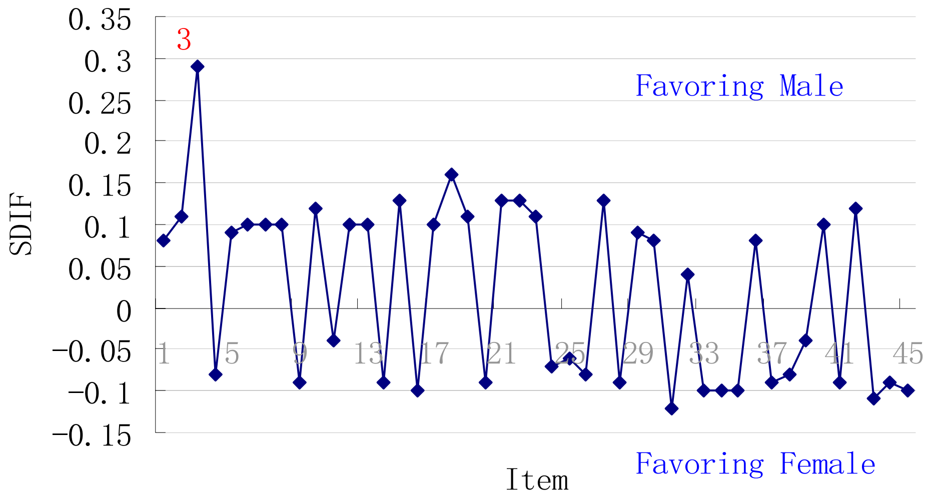
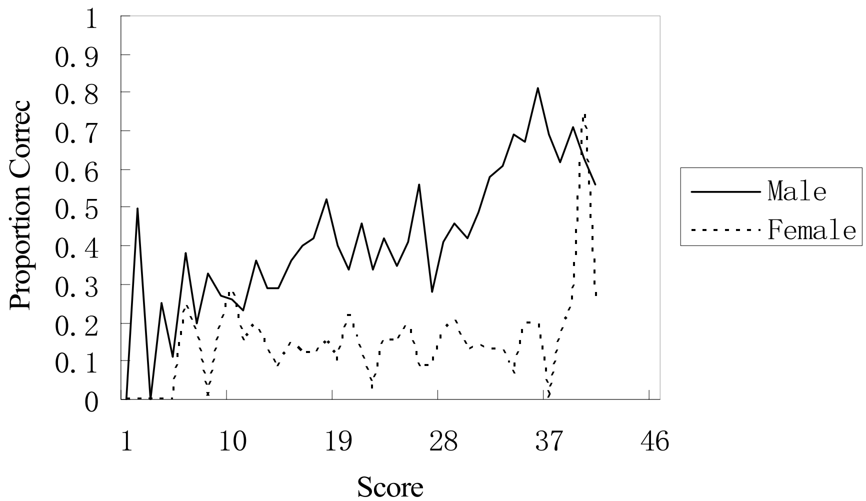
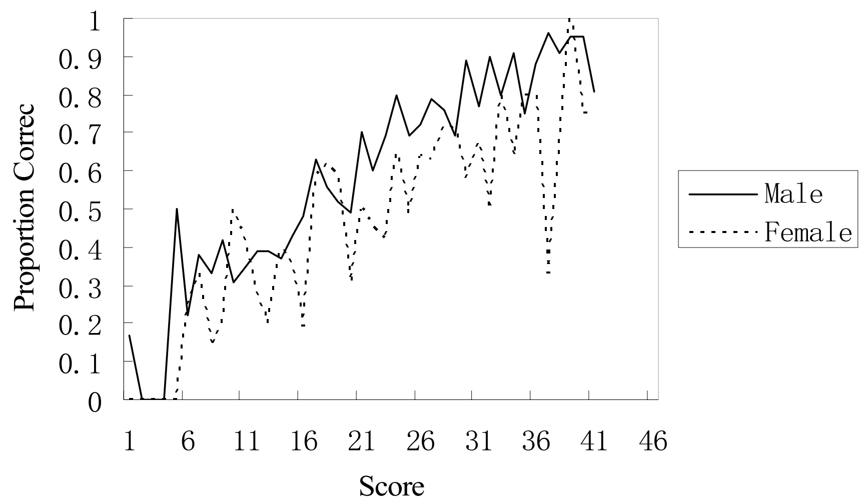



**Figure 8.4 DIF Plots for Males and Females**

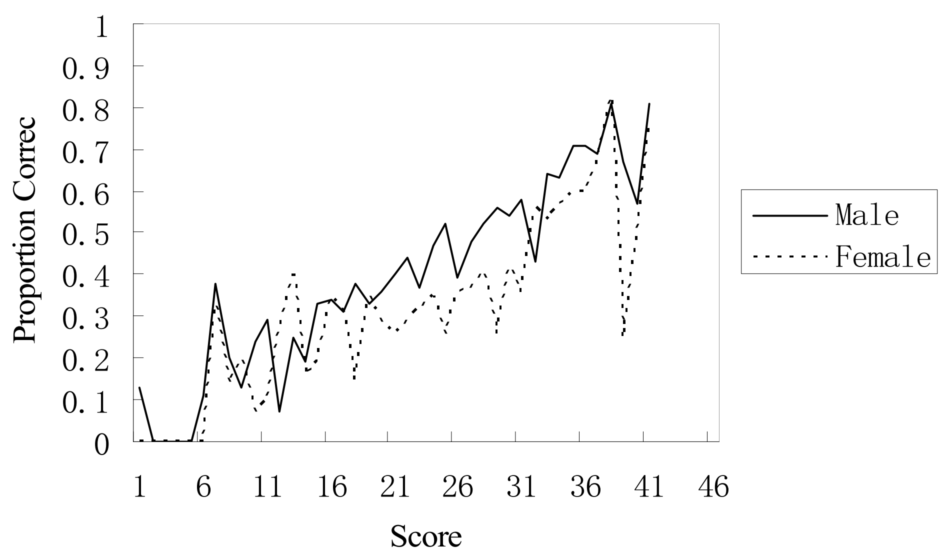
Item 3



Item 18




Item 21




## 9. Conclusions

Our psychometric analyses revealed that the 2006 MCAS grades 9/10 Technology/Engineering Test is of very high statistical quality. The item analysis work we did showed that the test items looked very good statistically though perhaps a bit on the difficult side for the students who took the test. Test reliability as estimated with Cronbach's coefficient alpha was .87 and this too is acceptable. Our study of test dimensionality revealed a strong first factor, with a small minor second factor, certainly strong enough evidence to support fitting a unidimensional IRT model or models to the data. Our fit of IRT models to the data revealed excellent fit at both the item and test level. A small number of items were not fit by the models. The information function we calculated showed good measurement precision across most parts of the reporting scale. More information at the lower end of the reporting scale would be important if the lower of the state's cut scores is placed in this region. Finally, our DIF analyses were limited but identified no items showing DIF against Hispanics. There was evidence of a small amount of DIF against females.

We did spot two areas in need of subsequent investigations. First, the information function for the test was not ideally placed for optimum measurement precision for a diverse group of students. With the likelihood of at least one of the cut scores (i.e., warning) being placed somewhat below the mean of the test score distribution, more test information in the lower portion of the test score distribution would be desirable. This might easily be accomplished in the future by substituting some of the hardest test items with test items capable of enhancing measurement precision for students scoring below the mean of the test score distribution. Secondly, there is some evidence of differential item functioning between males and females matched on Technology/Engineering test performance. This does not mean that the test items are flawed, but we do suggest that the test items be studied, to see what might be learned about the test items and the portions of the curriculum from which they came. Some insights about test items might be revealed, or areas of the curriculum where males may have an advantage or females a disadvantage because of backgrounds, culture, interests, etc. Whether the problems are due to backgrounds, culture, or curriculum, or a combination of factors, something valuable will be learned and can be attended to in the appropriate way in the future.

## References

Hambleton, R. K., Zhao, Y., Smith, Z., Lam, W., & Deng, N. (2008). *Psychometric analyses of the 2006 MCAS high school science tests* (Center for Educational Assessment Research Report No. 649). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Massachusetts Department of Education. (2006). *The Massachusetts Science and Technology/Engineering Curriculum Framework*. Malden, MA: Massachusetts Department of Education.

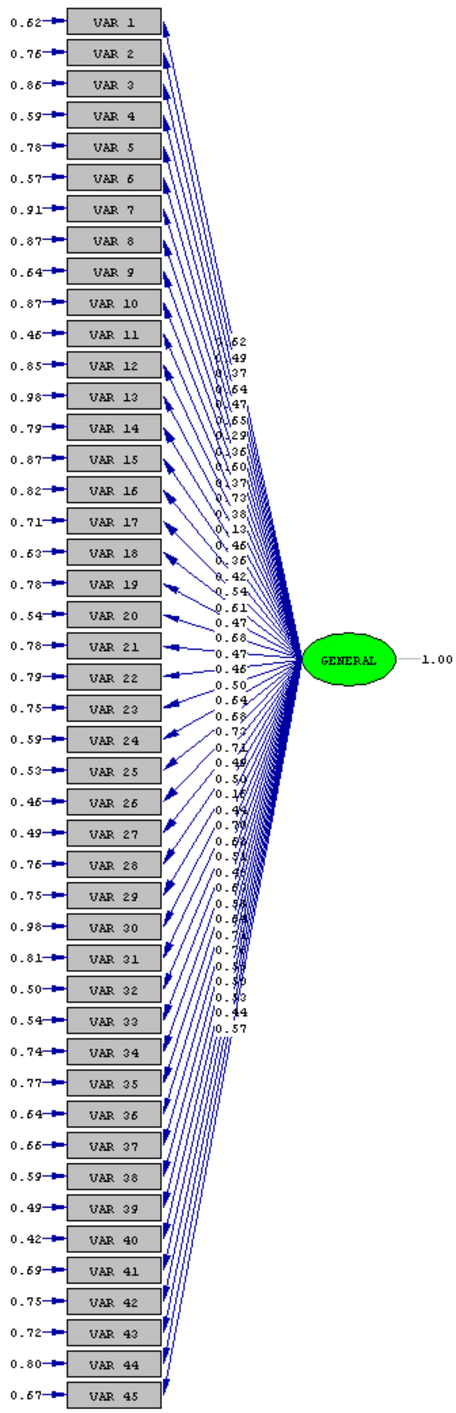
Zenisky, A. R., & Hambleton, R. K. (2007). *Differential item functioning analyses with STDIF: User's guide* (Unpublished report). Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Zenisky, A., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement, 63*(1), 51-64.

Zenisky, A., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment, 9*(1&2), 61-78.

## Appendix A. Graphical Display of the Technology/Engineering Factor Loadings

### Figure A.1 Factor loadings for a one factor model using LISREL software

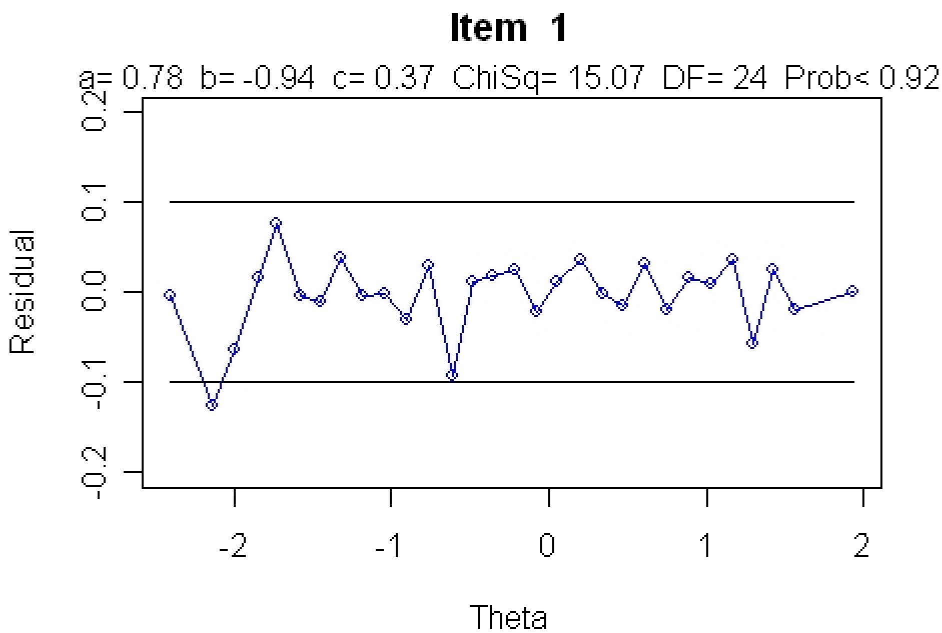


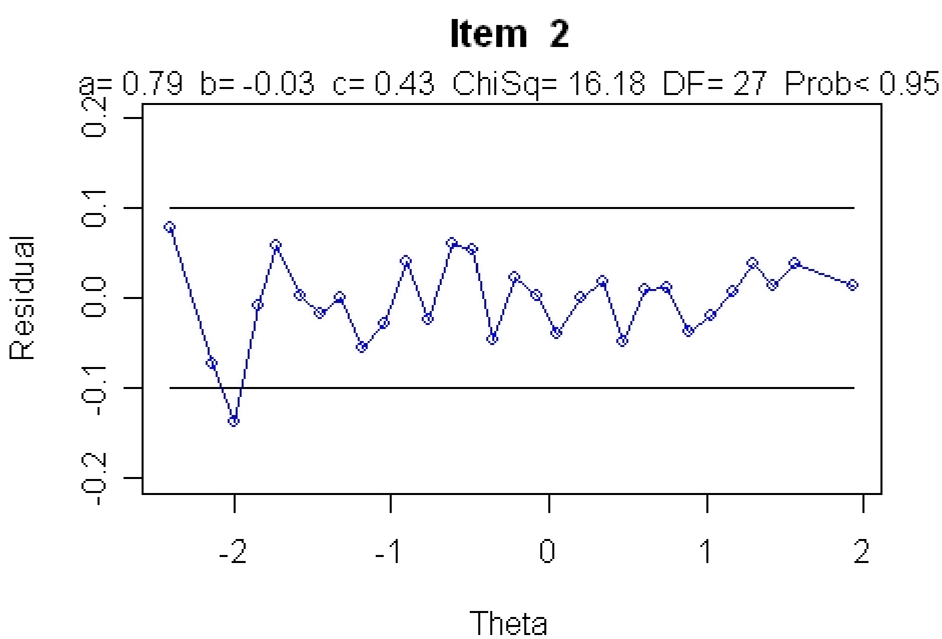

Chi-Square=2820.15, df=945, P-value=0.00000, RMSEA=0.028

## Appendix B. IRT Residual Plots

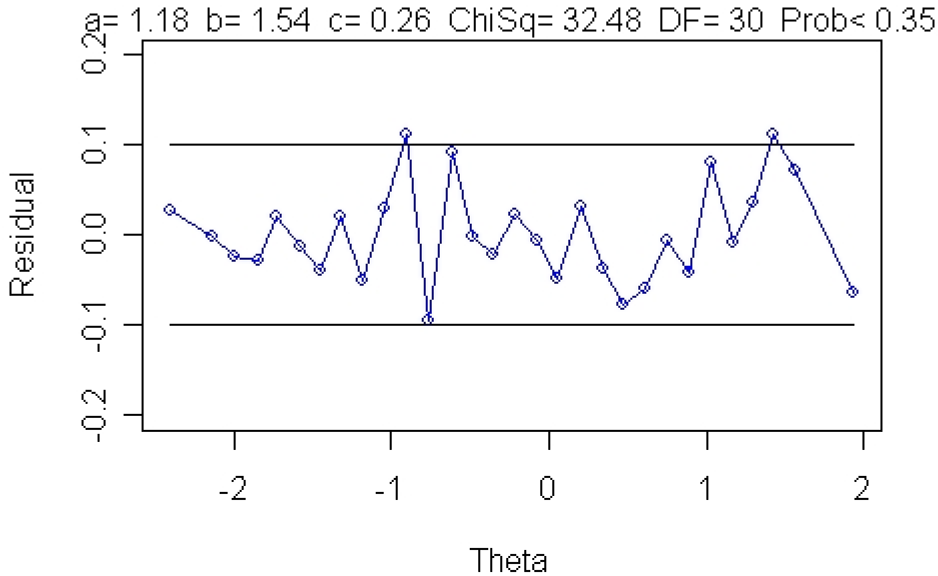
### Figure B.1 Raw residual plots for dichotomously-scored items (differences between the observed and expected item performances)

**Summary:**
From a review of the residual plots in Figures B.1 and B.2, it appeared that fits for items 9, 10, 27 and 37 may be problematic.

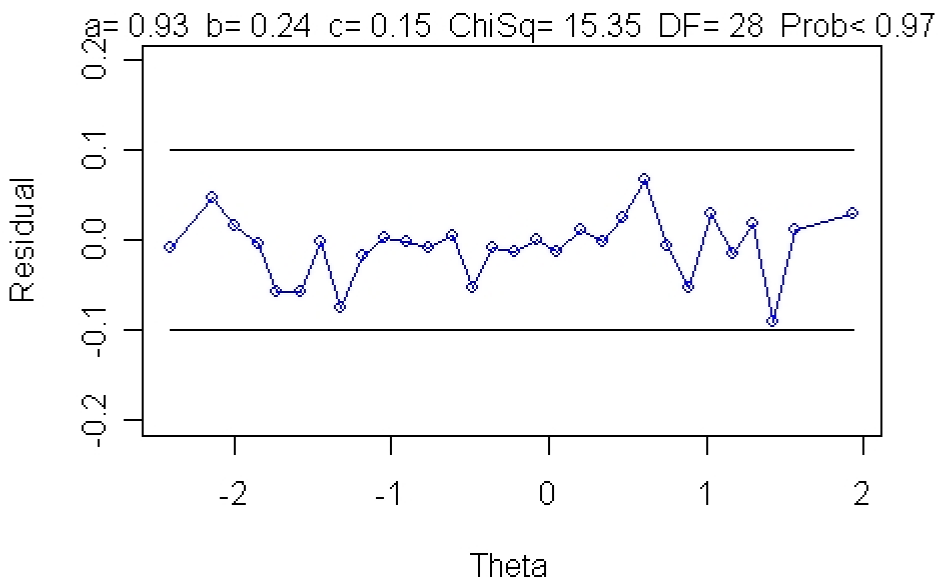

Item 1
a= 0.78 b= -0.94 c= 0.37 ChiSq= 15.07 DF= 24 Prob< 0.92
Residual
Theta
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2

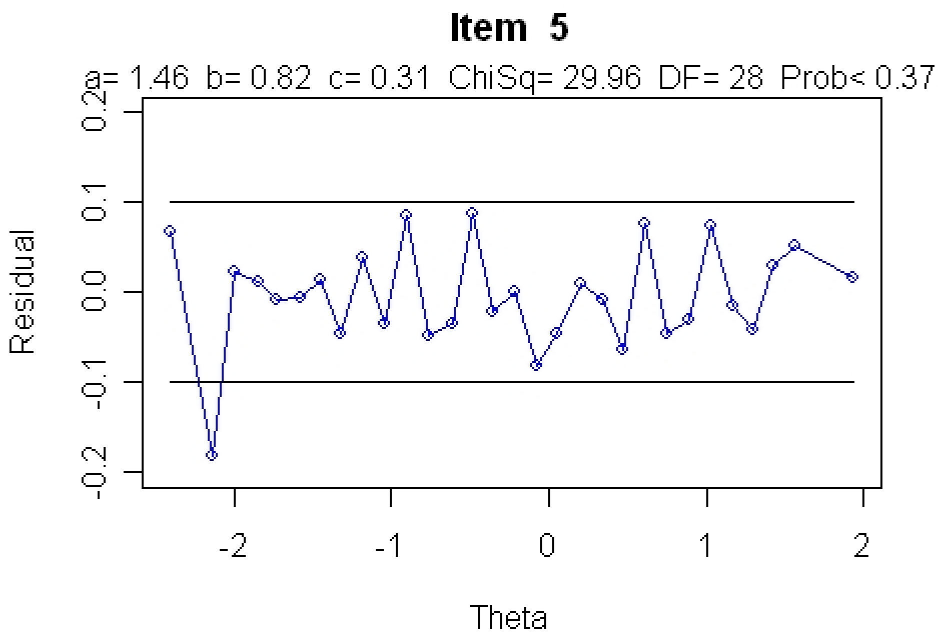

Item 2
a= 0.79 b= -0.03 c= 0.43 ChiSq= 16.18 DF= 27 Prob< 0.95
Residual
Theta
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2

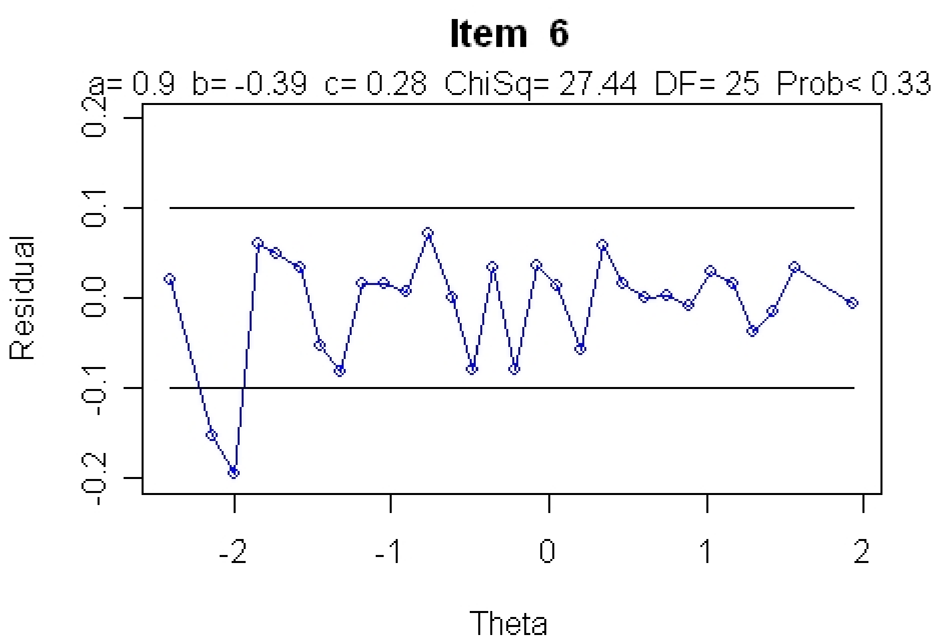
**Item 3**

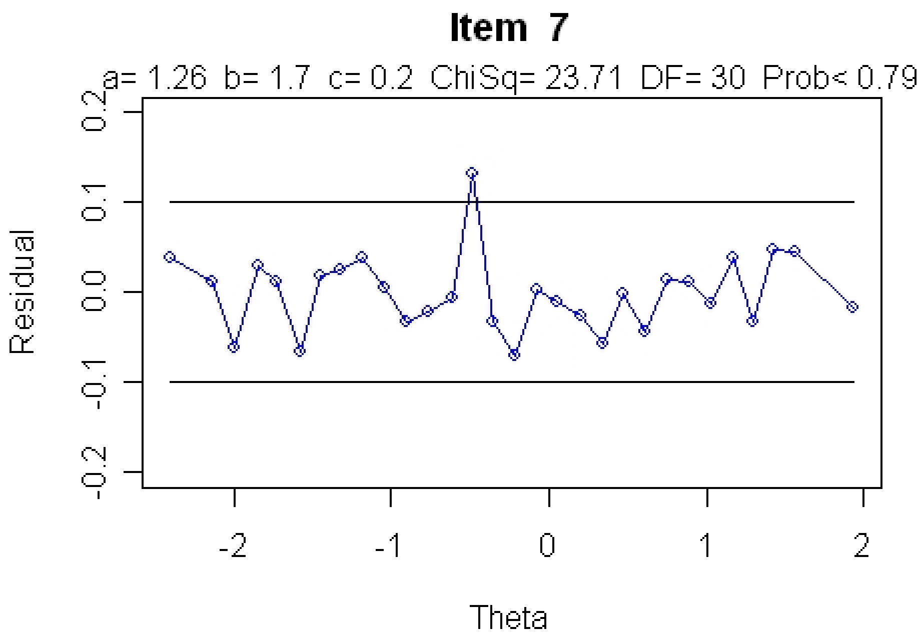

a= 1.18 b= 1.54 c= 0.26 ChiSq= 32.48 DF= 30 Prob< 0.35
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

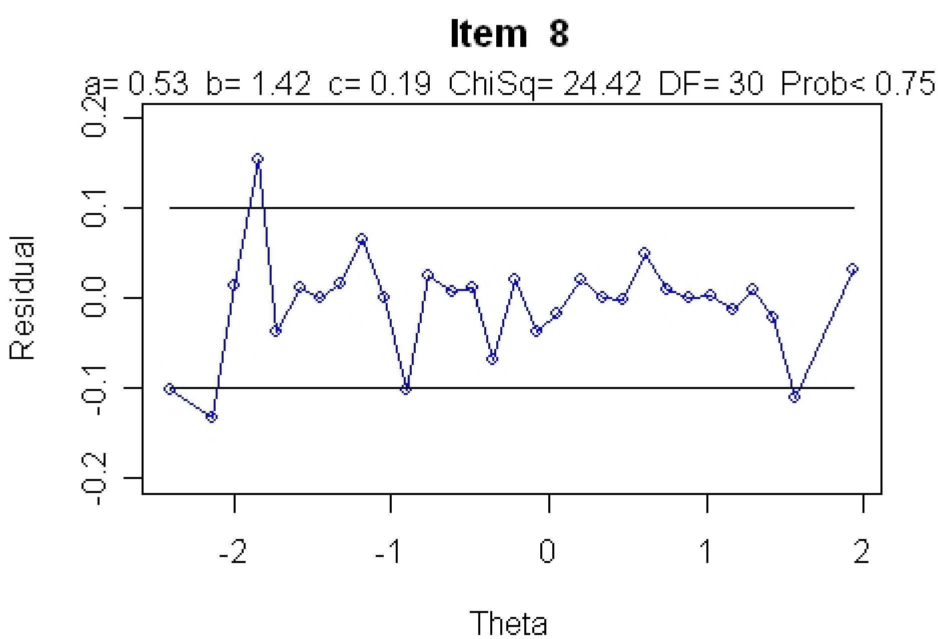
**Item 4**

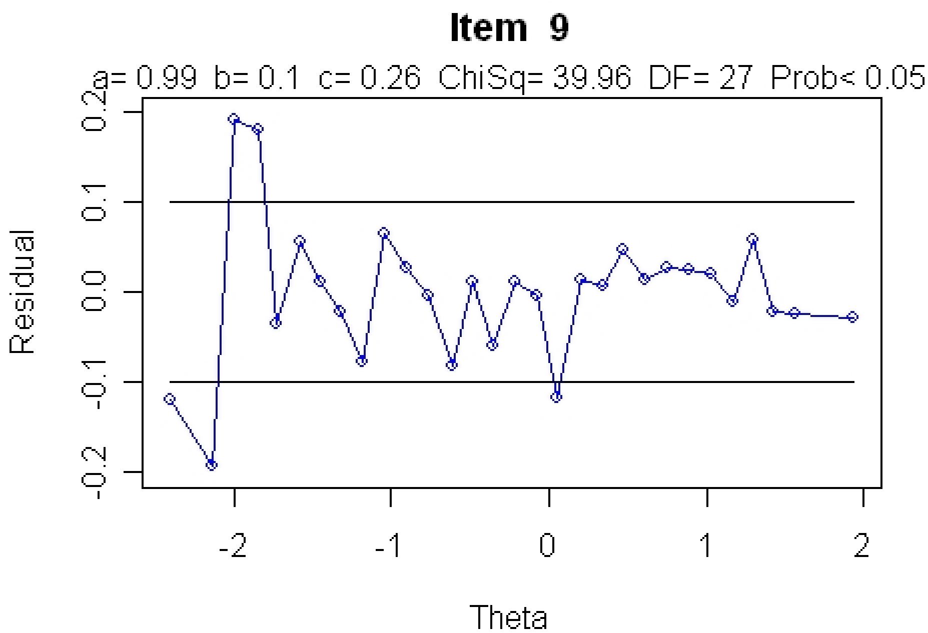

a= 0.93 b= 0.24 c= 0.15 ChiSq= 15.35 DF= 28 Prob< 0.97
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta


Item 5
a= 1.46 b= 0.82 c= 0.31 ChiSq= 29.96 DF= 28 Prob< 0.37
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

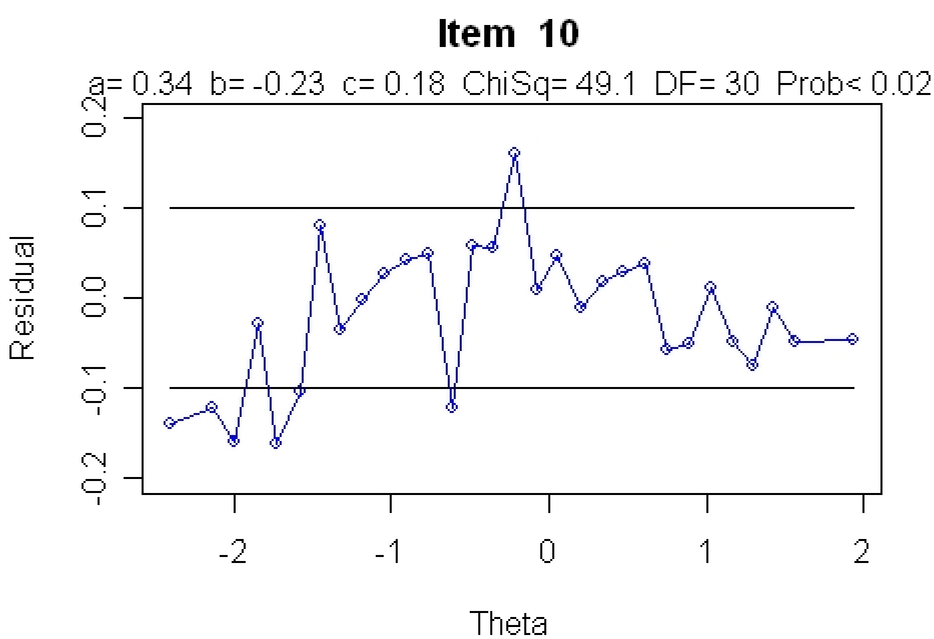

Item 6
a= 0.9 b= -0.39 c= 0.28 ChiSq= 27.44 DF= 25 Prob< 0.33
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta


Item 7
a= 1.26 b= 1.7 c= 0.2 ChiSq= 23.71 DF= 30 Prob< 0.79
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

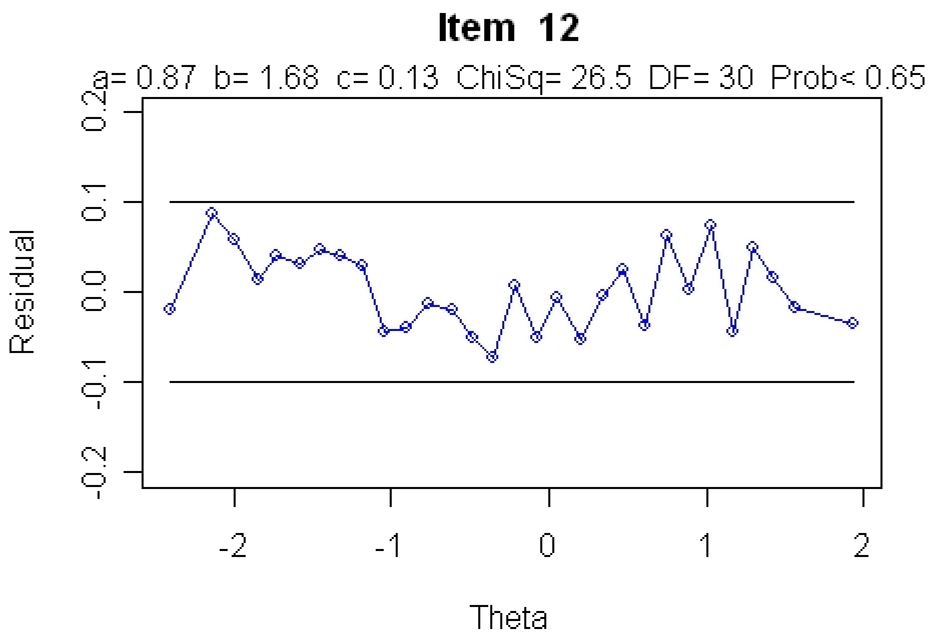

Item 8
a= 0.53 b= 1.42 c= 0.19 ChiSq= 24.42 DF= 30 Prob< 0.75
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

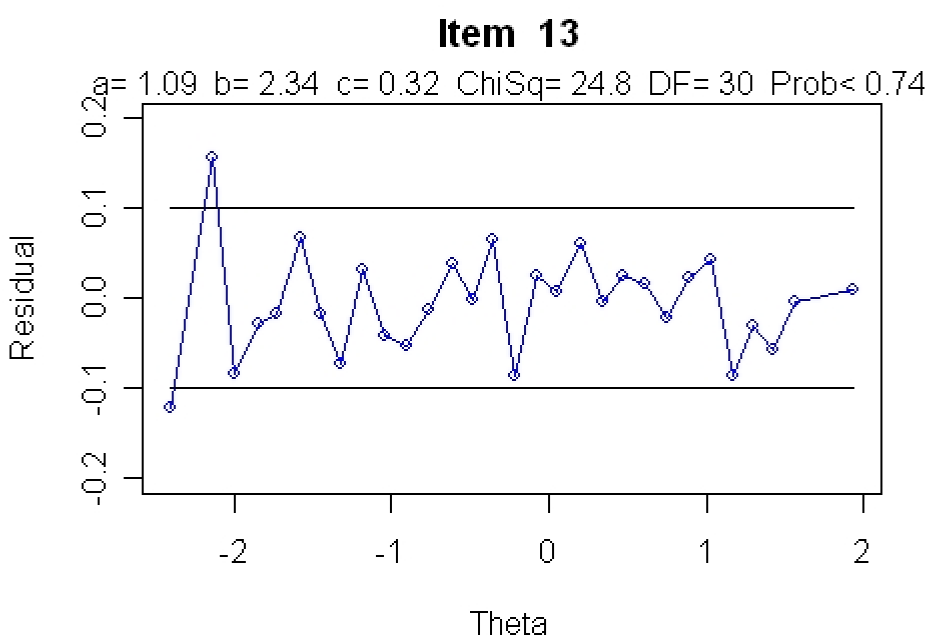


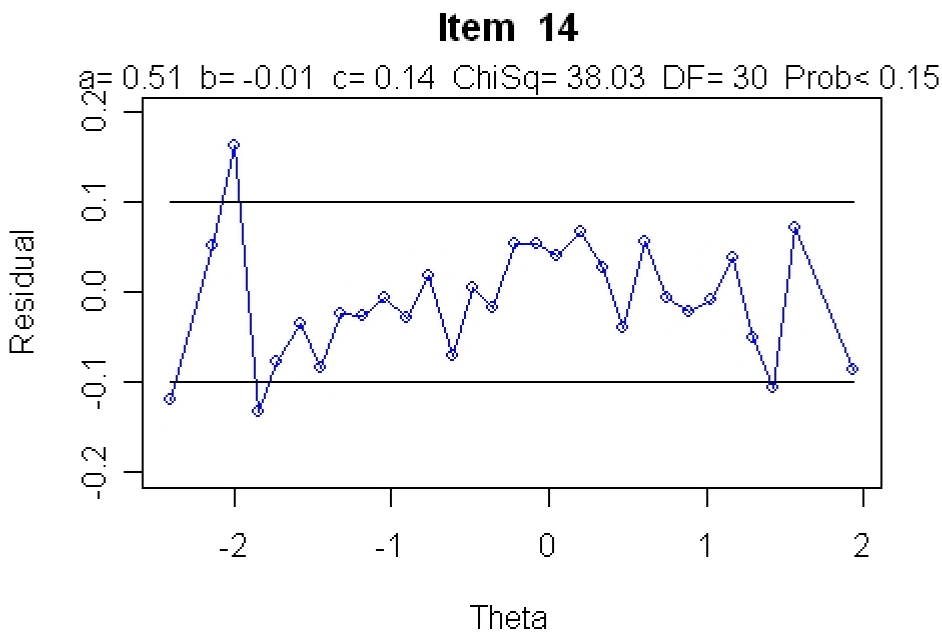
Item 9
a= 0.99 b= 0.1 c= 0.26 ChiSq= 39.96 DF= 27 Prob< 0.05
Residual
Theta

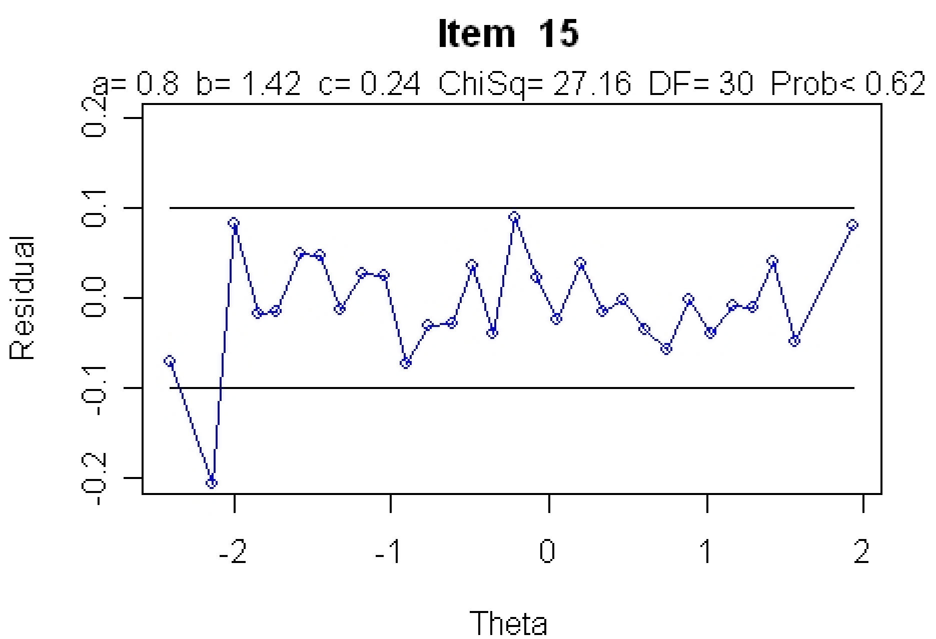



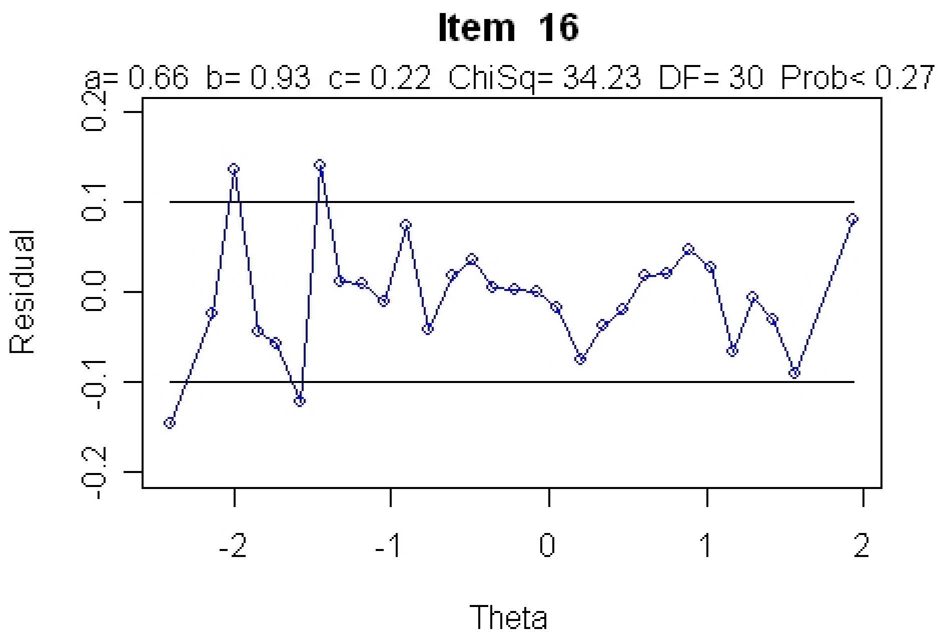
Item 10
a= 0.34 b= -0.23 c= 0.18 ChiSq= 49.1 DF= 30 Prob< 0.02
Residual
Theta

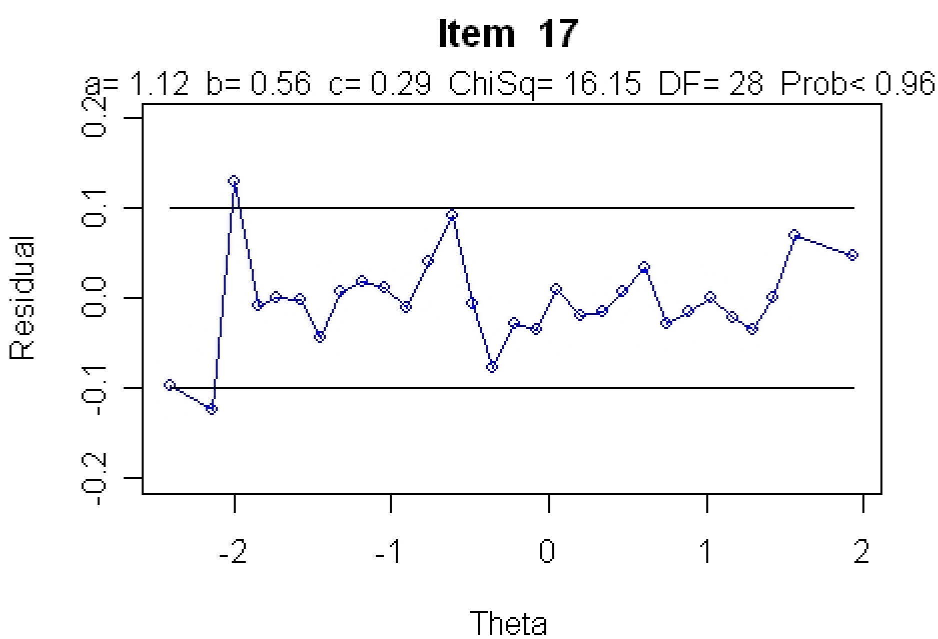

Item 12
a= 0.87 b= 1.68 c= 0.13 ChiSq= 26.5 DF= 30 Prob< 0.65
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

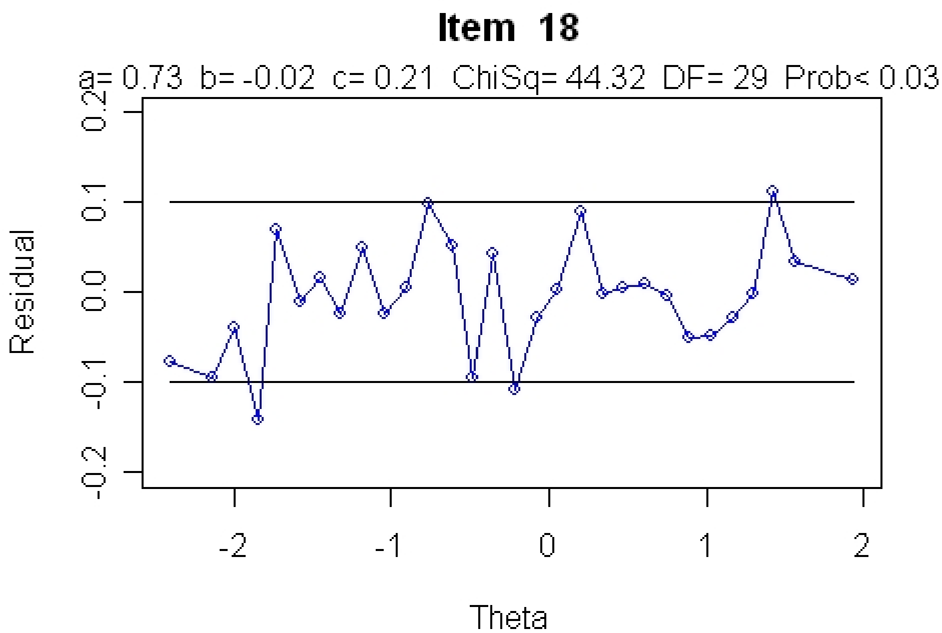

Item 13
a= 1.09 b= 2.34 c= 0.32 ChiSq= 24.8 DF= 30 Prob< 0.74
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

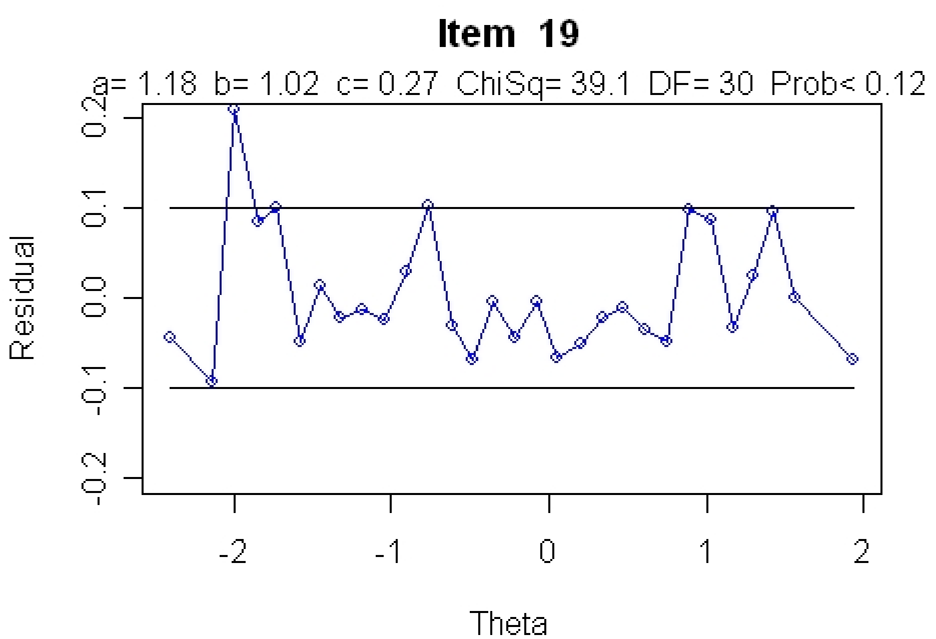

Item 14
a= 0.51 b= -0.01 c= 0.14 ChiSq= 38.03 DF= 30 Prob< 0.15
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

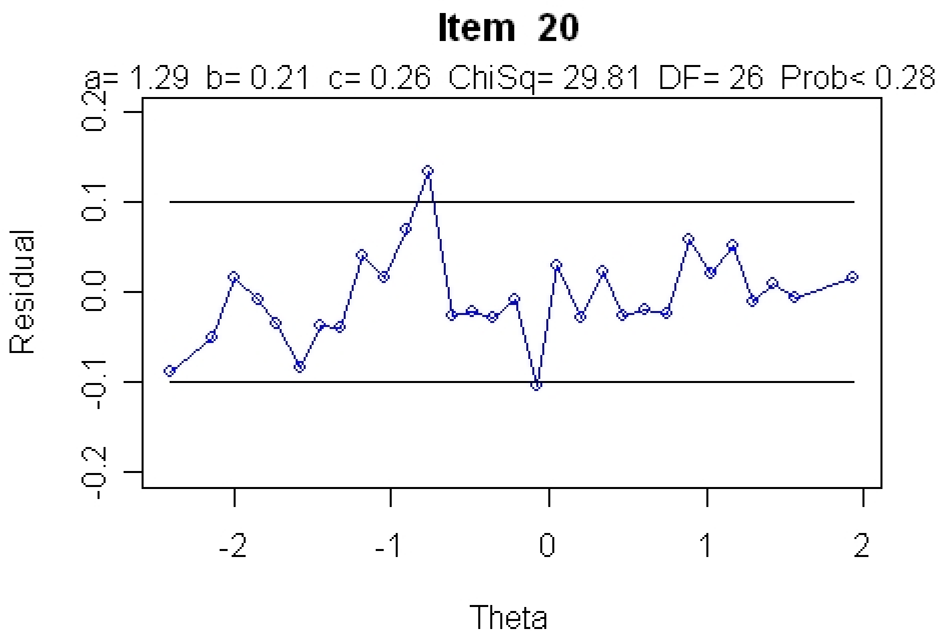

Item 15
a= 0.8 b= 1.42 c= 0.24 ChiSq= 27.16 DF= 30 Prob< 0.62
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

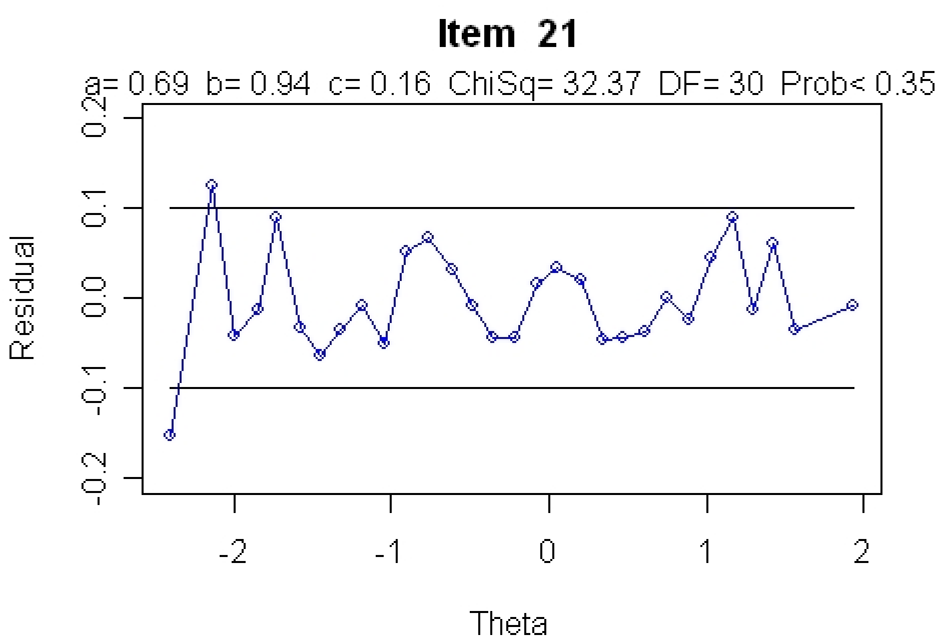

Item 16
a= 0.66 b= 0.93 c= 0.22 ChiSq= 34.23 DF= 30 Prob< 0.27
Residual
Theta
-0.2
-0.1
0.0
0.1
0.2
-2
-1
0
1
2

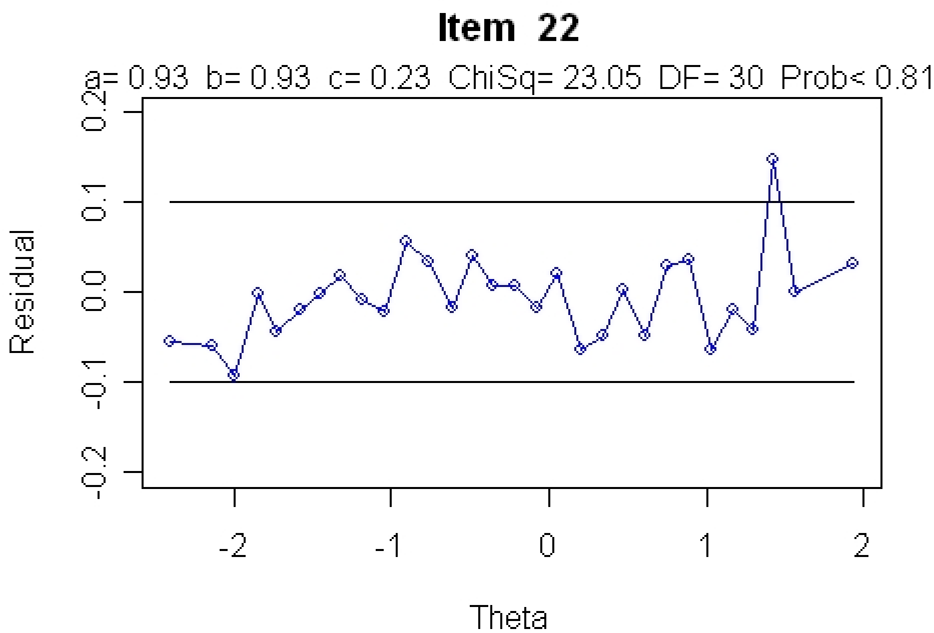

Item 17
a= 1.12 b= 0.56 c= 0.29 ChiSq= 16.15 DF= 28 Prob< 0.96
Residual
Theta
-0.2
-0.1
0.0
0.1
0.2
-2
-1
0
1
2

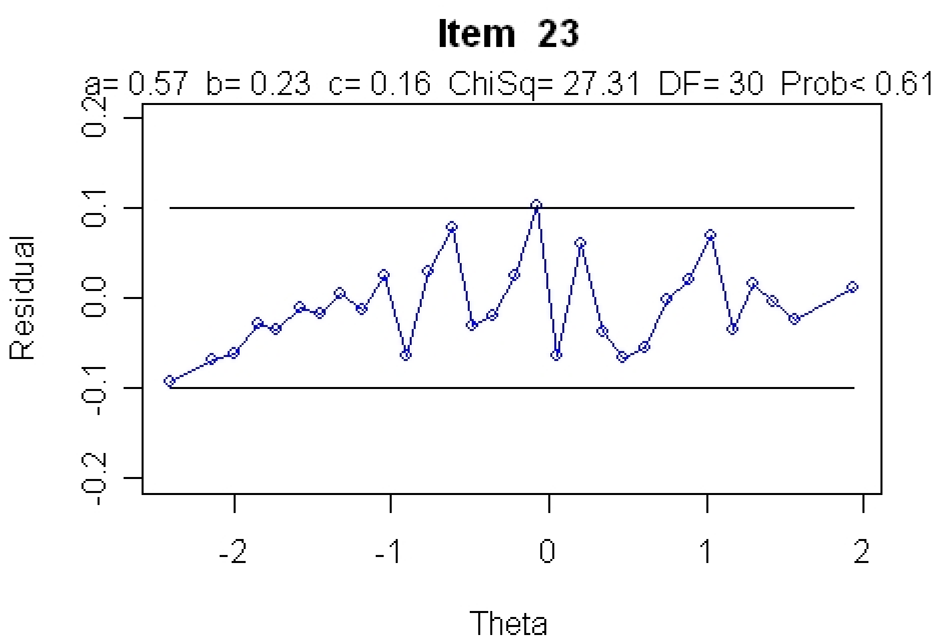

Item 18
a= 0.73 b= -0.02 c= 0.21 ChiSq= 44.32 DF= 29 Prob< 0.03
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

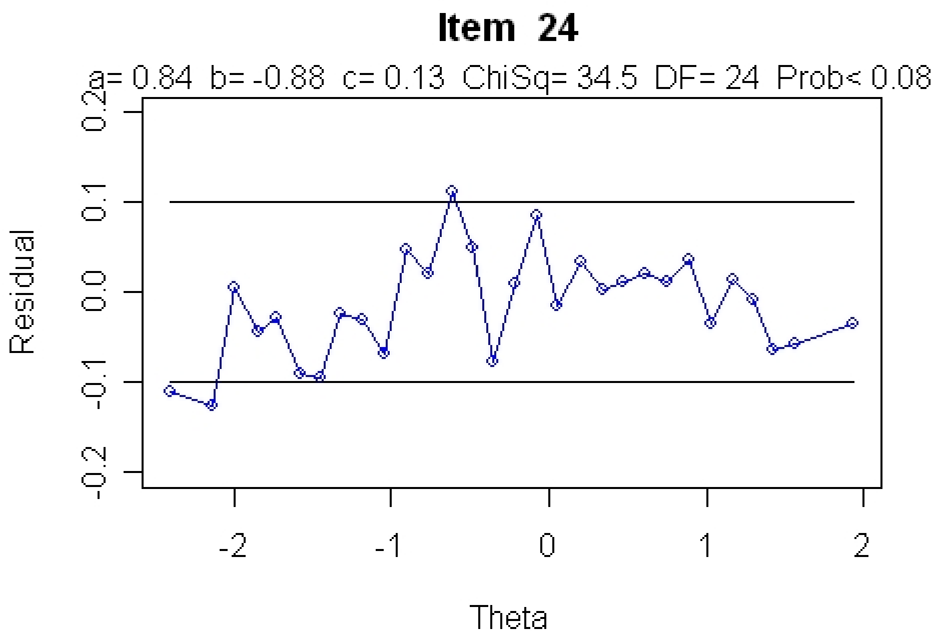

Item 19
a= 1.18 b= 1.02 c= 0.27 ChiSq= 39.1 DF= 30 Prob< 0.12
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

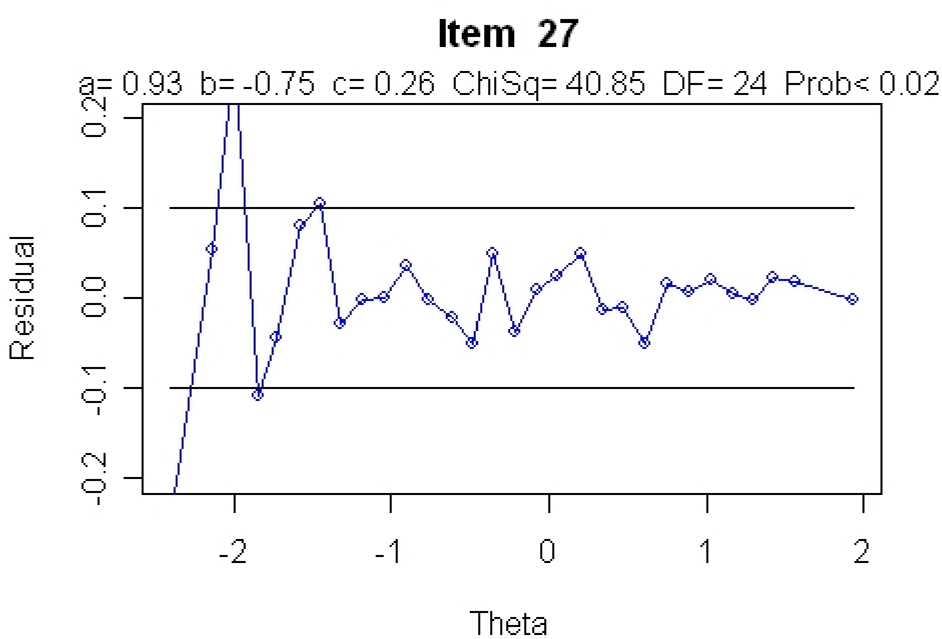

Item 20
a= 1.29 b= 0.21 c= 0.26 ChiSq= 29.81 DF= 26 Prob< 0.28
Residual
-0.2
-0.1
0.0
0.1
0.2
-2
-1
0
1
2
Theta

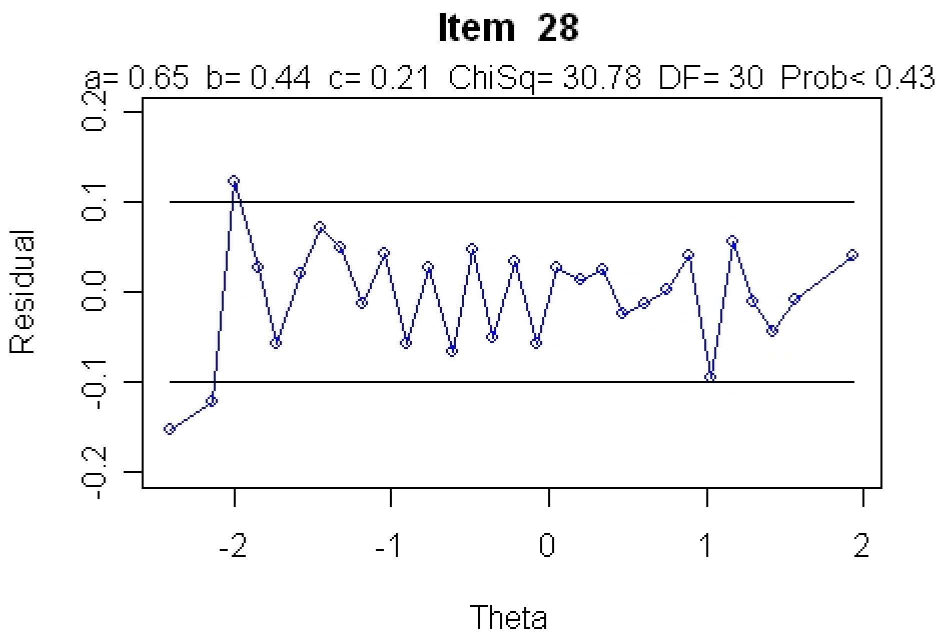

Item 21
a= 0.69 b= 0.94 c= 0.16 ChiSq= 32.37 DF= 30 Prob< 0.35
Residual
-0.2
-0.1
0.0
0.1
0.2
-2
-1
0
1
2
Theta

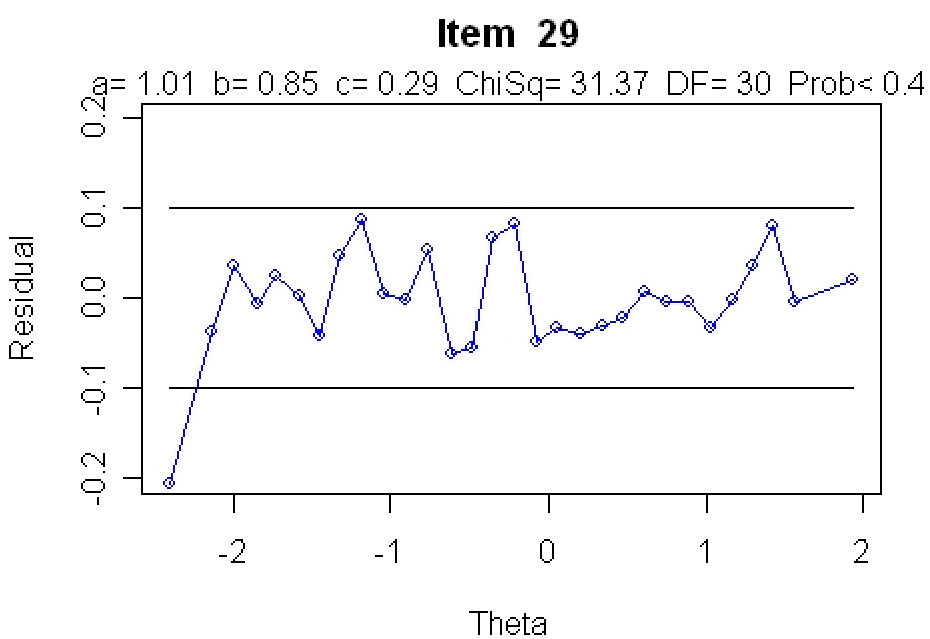

Item 22
a= 0.93 b= 0.93 c= 0.23 ChiSq= 23.05 DF= 30 Prob< 0.81
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

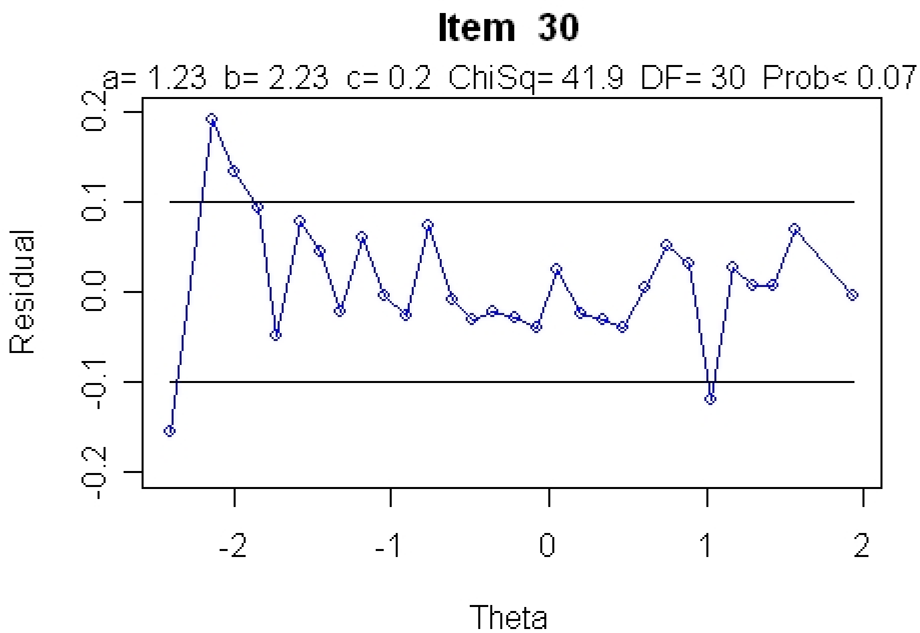

Item 23
a= 0.57 b= 0.23 c= 0.16 ChiSq= 27.31 DF= 30 Prob< 0.61
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta


Item 24
a= 0.84 b= -0.88 c= 0.13 ChiSq= 34.5 DF= 24 Prob< 0.08
Residual
Theta
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2

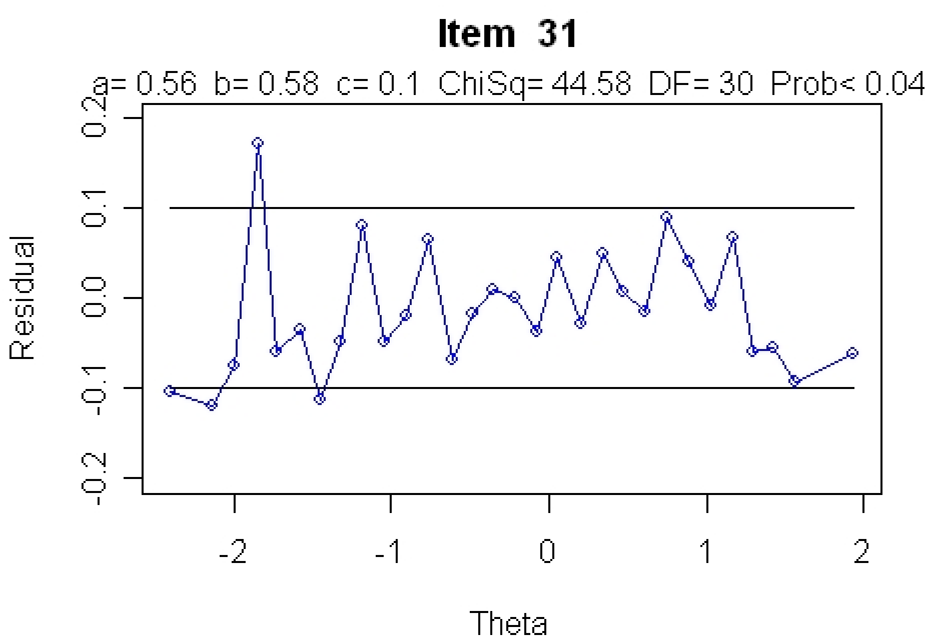

Item 27
a= 0.93 b= -0.75 c= 0.26 ChiSq= 40.85 DF= 24 Prob< 0.02
Residual
Theta
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2


Item 28
a= 0.65 b= 0.44 c= 0.21 ChiSq= 30.78 DF= 30 Prob< 0.43
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

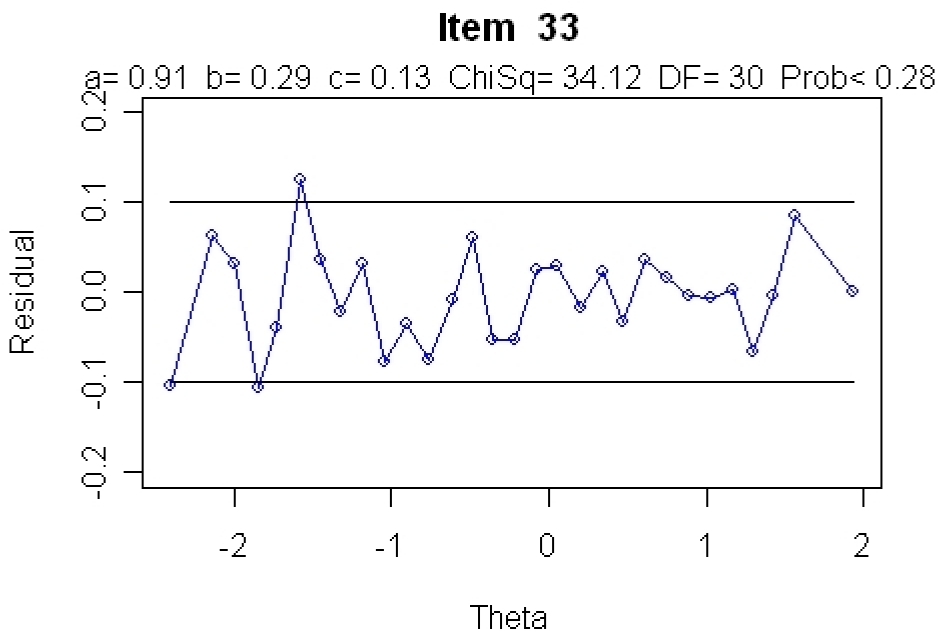

Item 29
a= 1.01 b= 0.85 c= 0.29 ChiSq= 31.37 DF= 30 Prob< 0.4
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

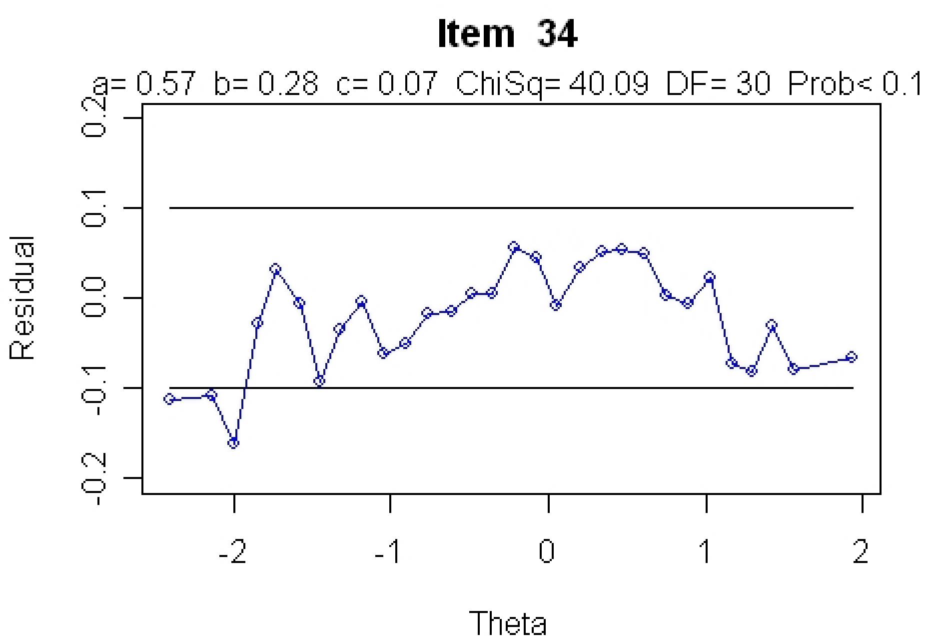


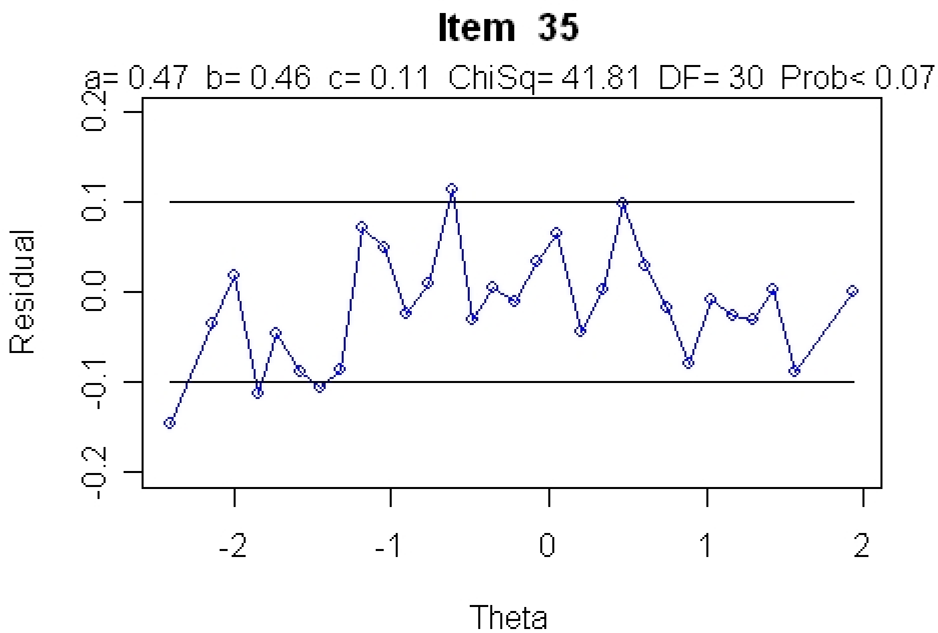
Item 30
a= 1.23 b= 2.23 c= 0.2 ChiSq= 41.9 DF= 30 Prob< 0.07
Residual
Theta

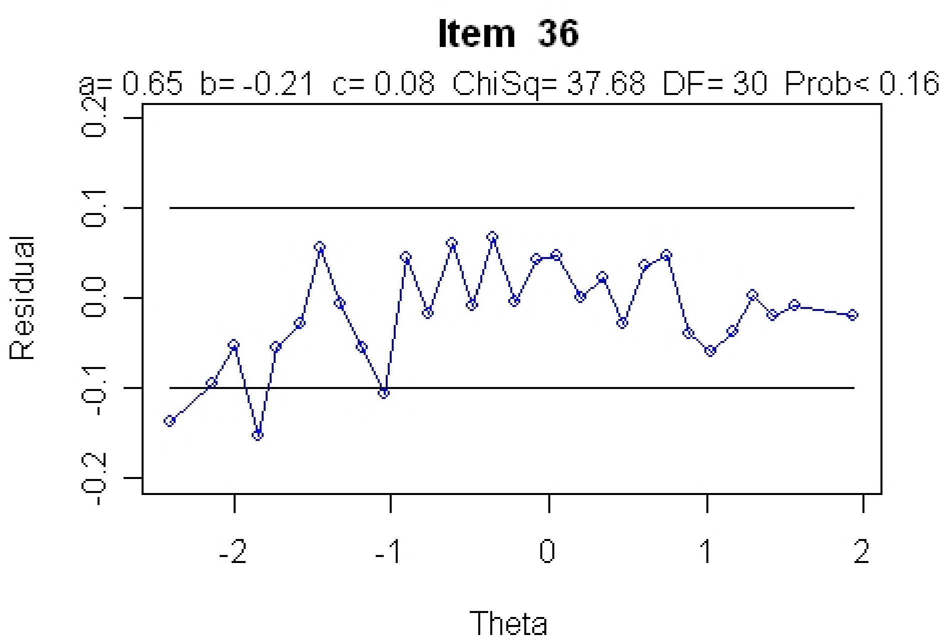



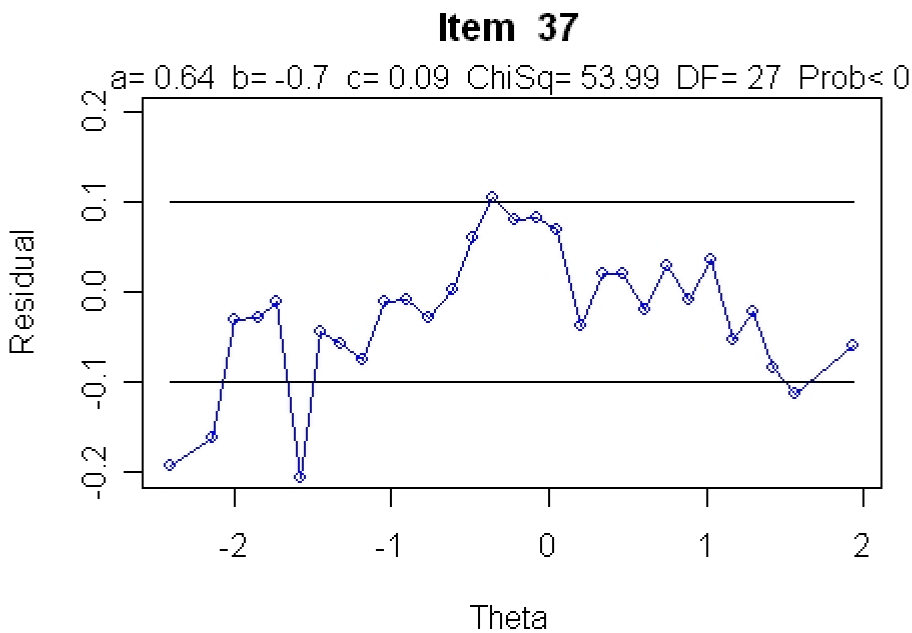
Item 31
a= 0.56 b= 0.58 c= 0.1 ChiSq= 44.58 DF= 30 Prob< 0.04
Residual
Theta

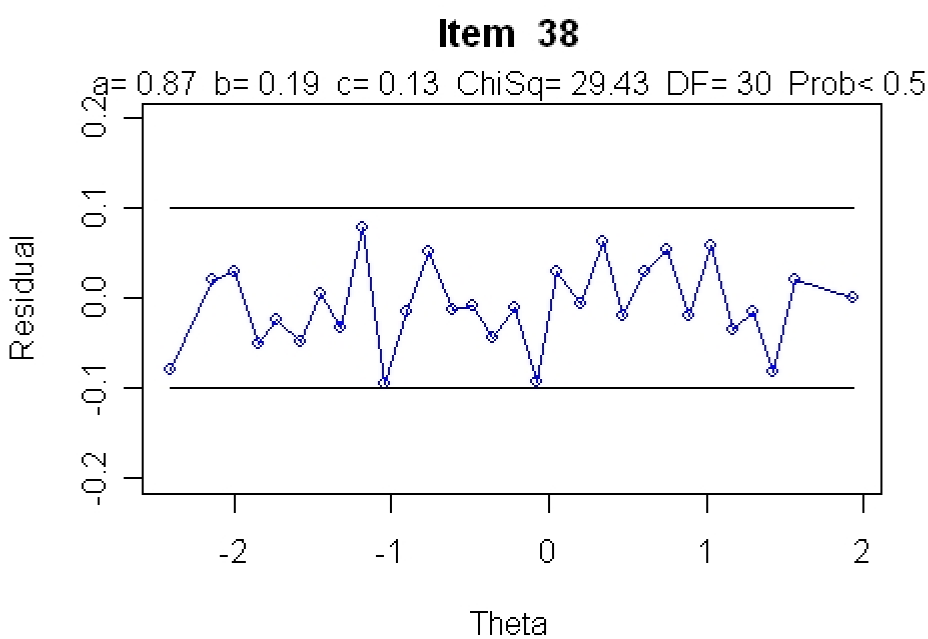

Item 33
a= 0.91 b= 0.29 c= 0.13 ChiSq= 34.12 DF= 30 Prob< 0.28
Residual
Theta
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2

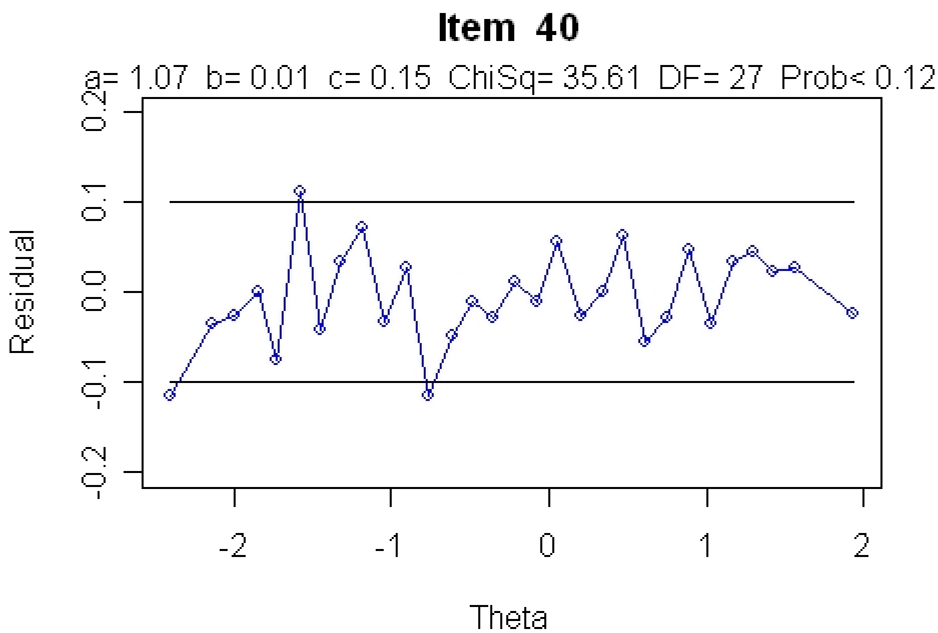

Item 34
a= 0.57 b= 0.28 c= 0.07 ChiSq= 40.09 DF= 30 Prob< 0.1
Residual
Theta
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2

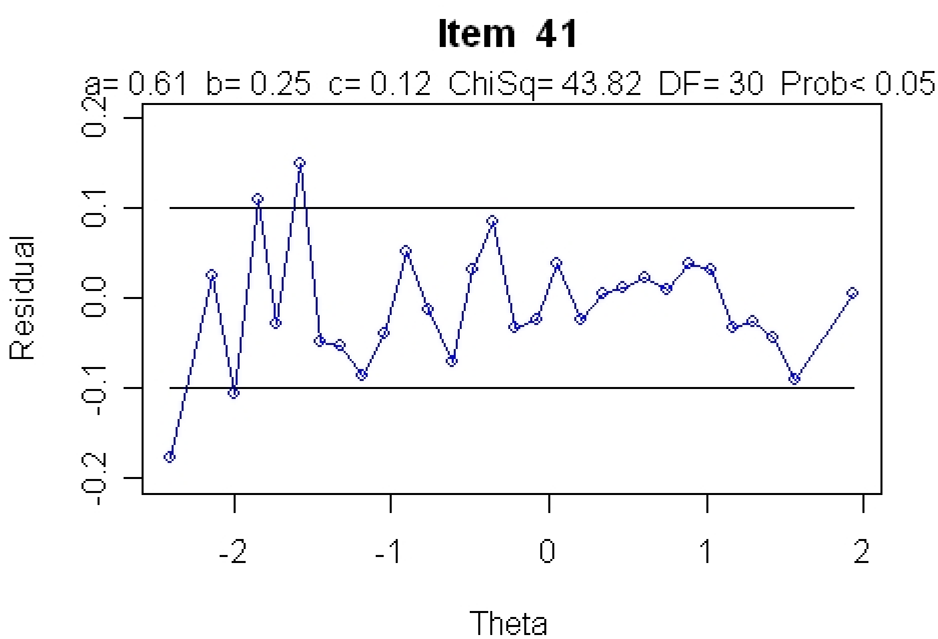

Item 35
a= 0.47 b= 0.46 c= 0.11 ChiSq= 41.81 DF= 30 Prob< 0.07
Residual
Theta
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2

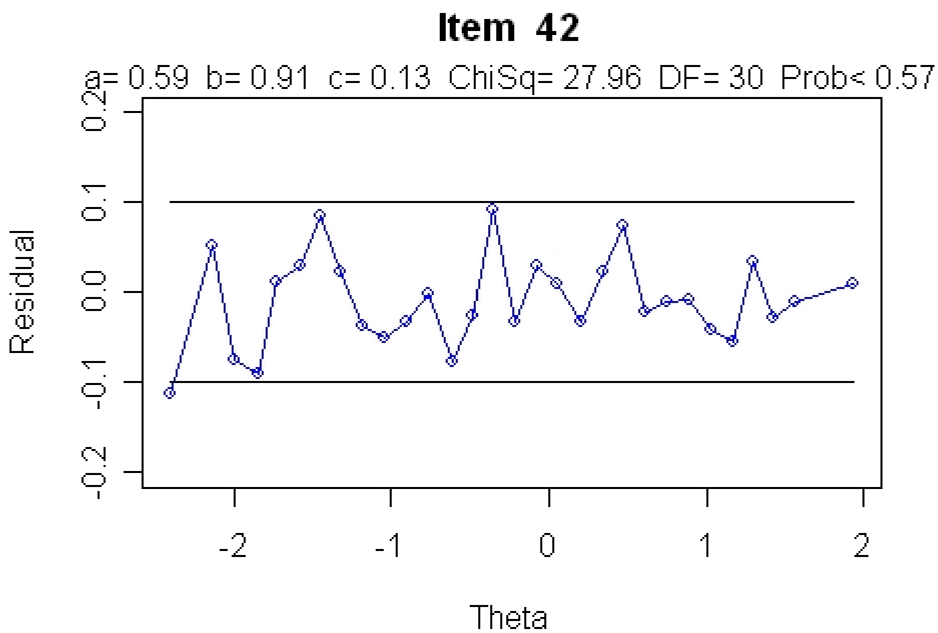

Item 36
a= 0.65 b= -0.21 c= 0.08 ChiSq= 37.68 DF= 30 Prob< 0.16
Residual
Theta
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2


Item 37
a= 0.64 b= -0.7 c= 0.09 ChiSq= 53.99 DF= 27 Prob< 0
Residual
-0.2
-0.1
0.0
0.1
0.2
-2
-1
0
1
2
Theta

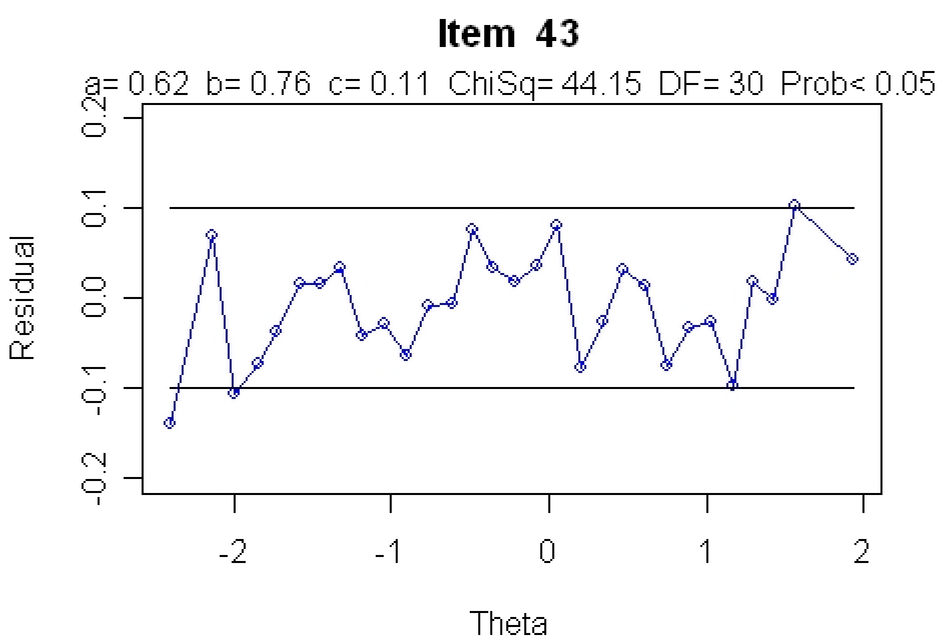

Item 38
a= 0.87 b= 0.19 c= 0.13 ChiSq= 29.43 DF= 30 Prob< 0.5
Residual
-0.2
-0.1
0.0
0.1
0.2
-2
-1
0
1
2
Theta


Item 40
a= 1.07 b= 0.01 c= 0.15 ChiSq= 35.61 DF= 27 Prob< 0.12
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta

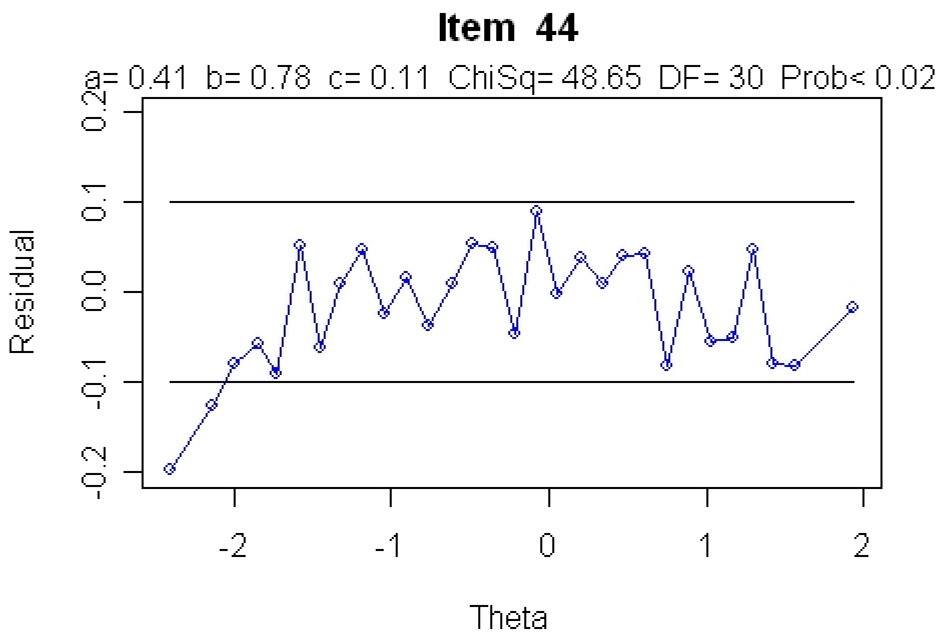

Item 41
a= 0.61 b= 0.25 c= 0.12 ChiSq= 43.82 DF= 30 Prob< 0.05
Residual
0.2
0.1
0.0
-0.1
-0.2
-2
-1
0
1
2
Theta



Item 42
a= 0.59 b= 0.91 c= 0.13 ChiSq= 27.96 DF= 30 Prob< 0.57
Residual
-0.2
-0.1
0.0
0.1
0.2
-2
-1
0
1
2
Theta

Item 43
a= 0.62 b= 0.76 c= 0.11 ChiSq= 44.15 DF= 30 Prob< 0.05
Residual
-0.2
-0.1
0.0
0.1
0.2
-2
-1
0
1
2
Theta


Item 44
a= 0.41 b= 0.78 c= 0.11 ChiSq= 48.65 DF= 30 Prob< 0.02
Residual
-0.2
-0.1
0.0
0.1
0.2
-2
-1
0
1
2
Theta

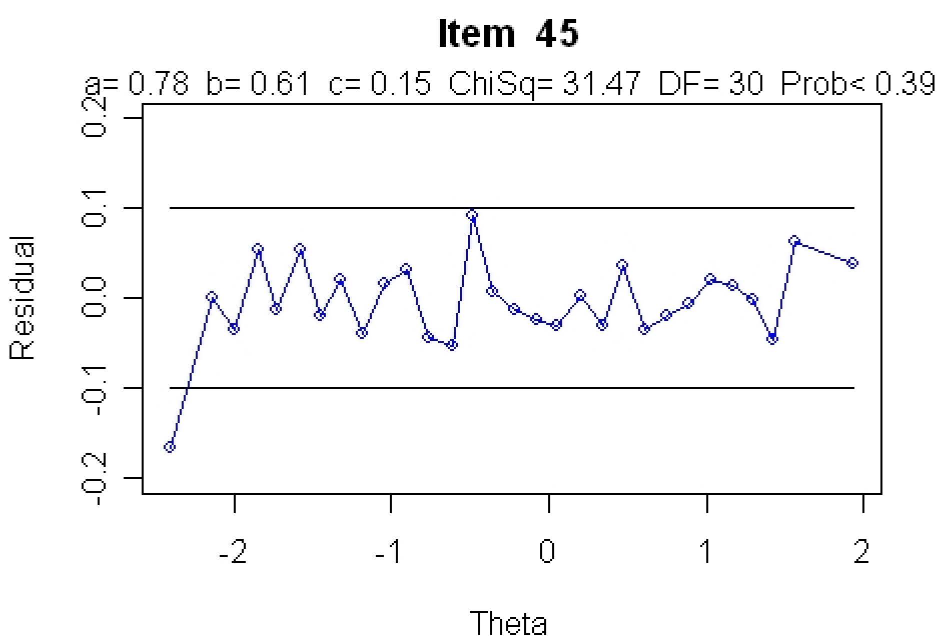
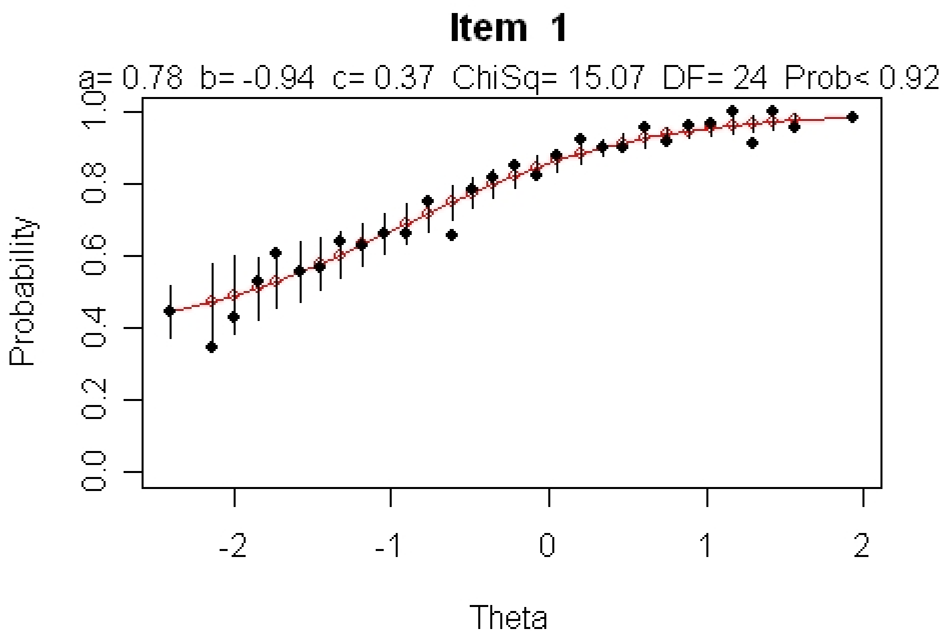
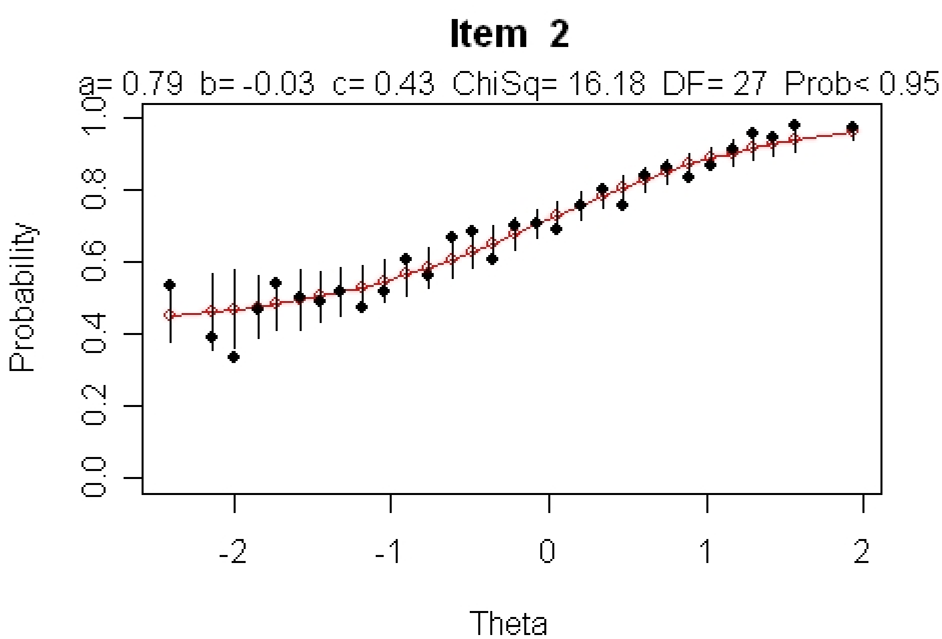
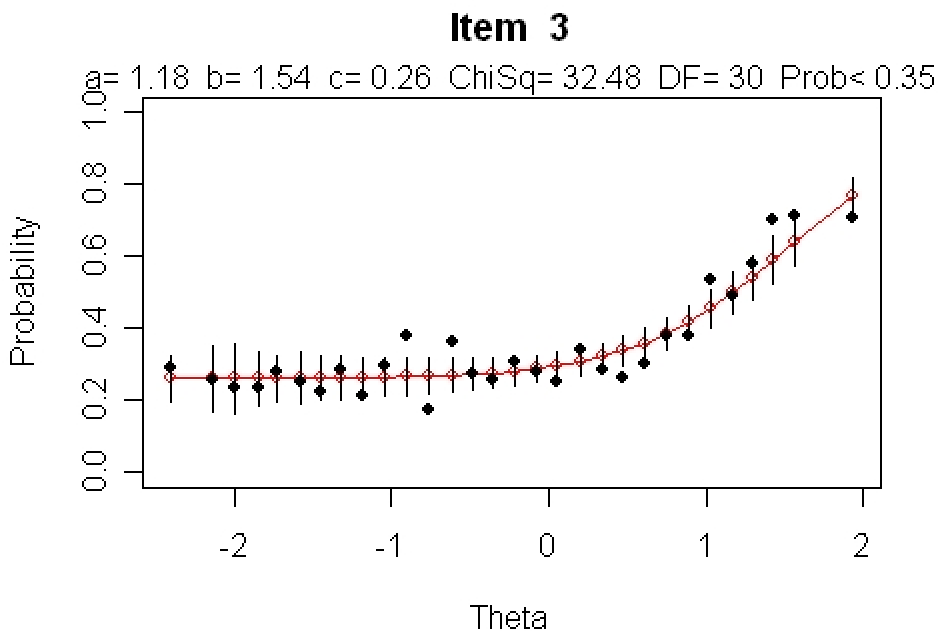

Item 45
a= 0.78 b= 0.61 c= 0.15 ChiSq= 31.47 DF= 30 Prob< 0.39
Residual
-0.2
-0.1
0.0
0.1
0.2
-2
-1
0
1
2
Theta

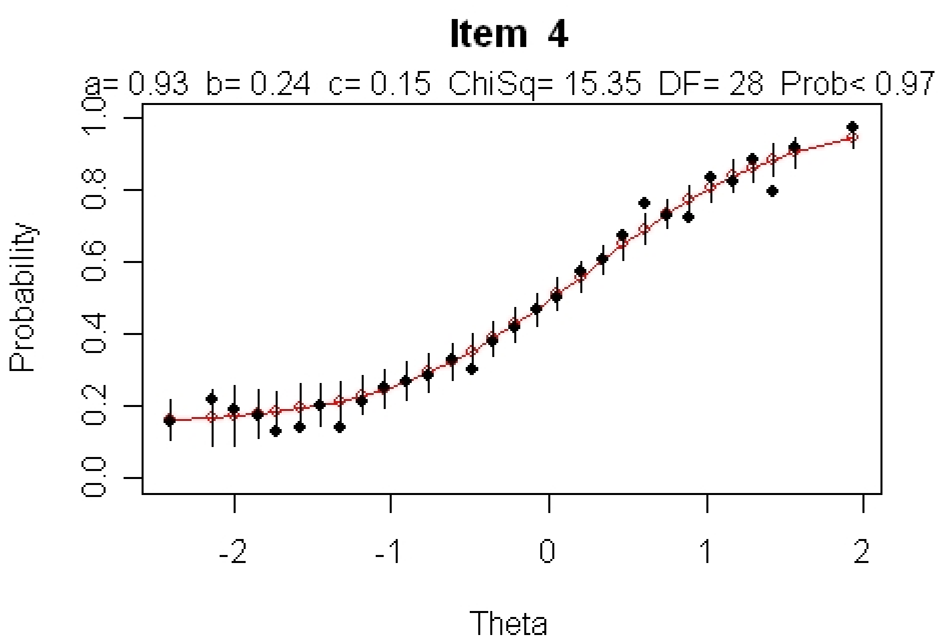
**Figure B.2 Probability plots for dichotomously-scored items highlighting the level of model misfit**



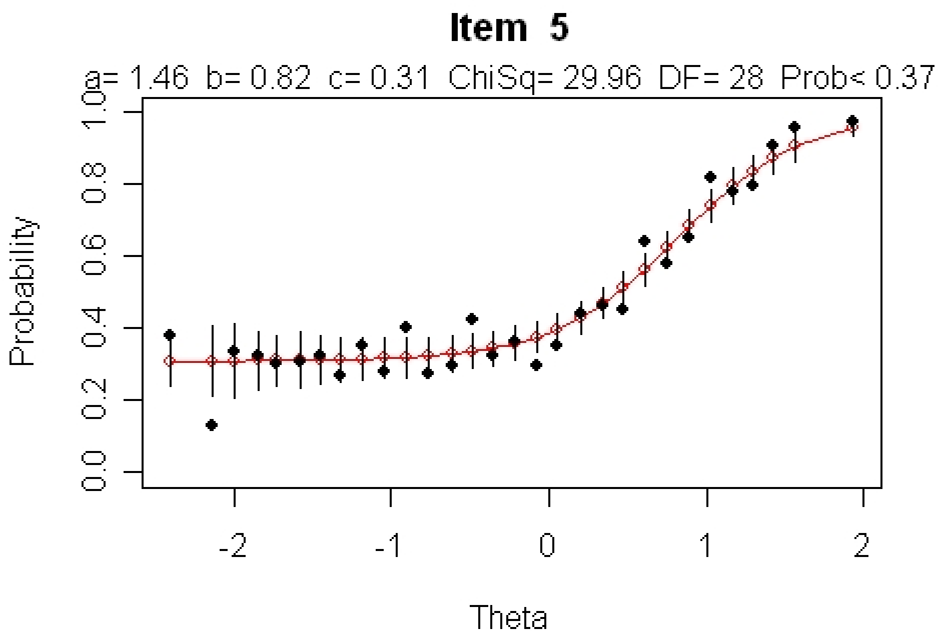



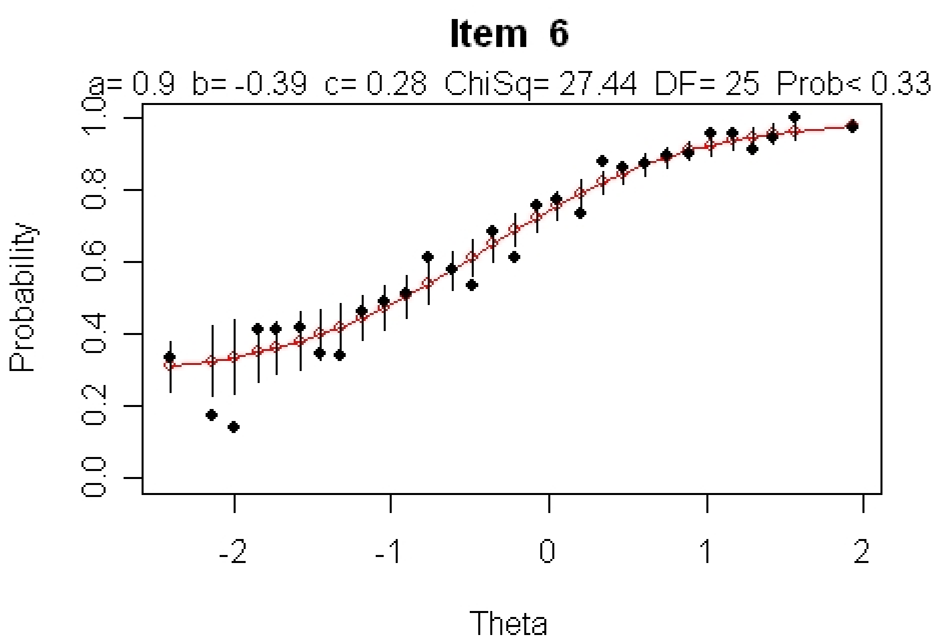


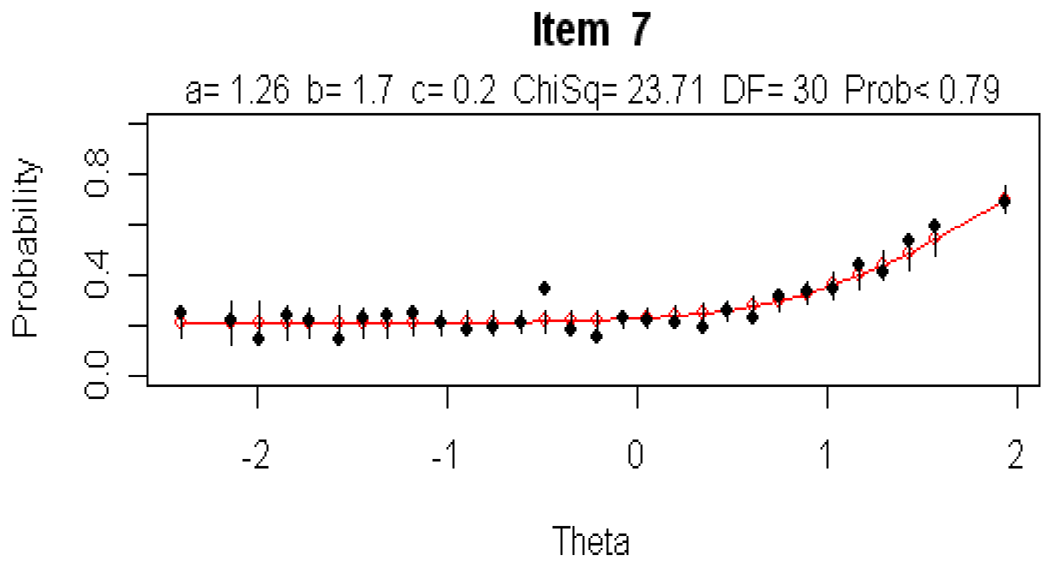
Item 3
a= 1.18 b= 1.54 c= 0.26 ChiSq= 32.48 DF= 30 Prob< 0.35
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

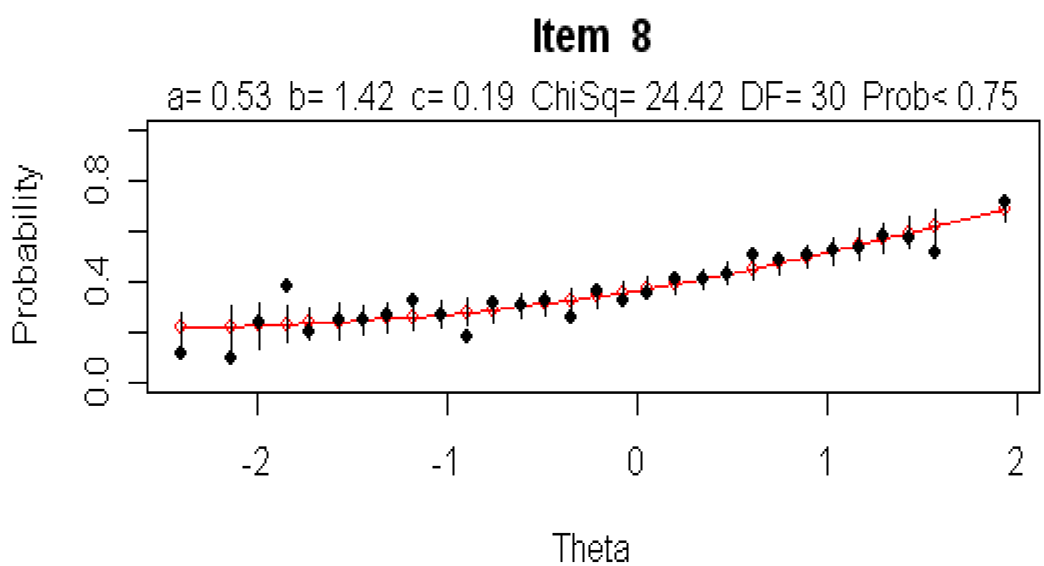



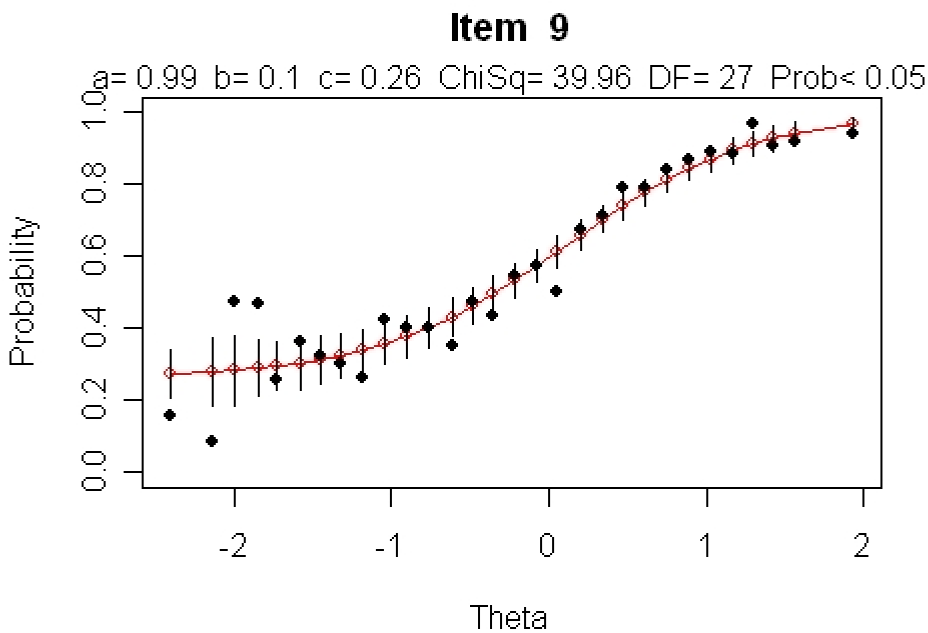
Item 4
a= 0.93 b= 0.24 c= 0.15 ChiSq= 15.35 DF= 28 Prob< 0.97
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

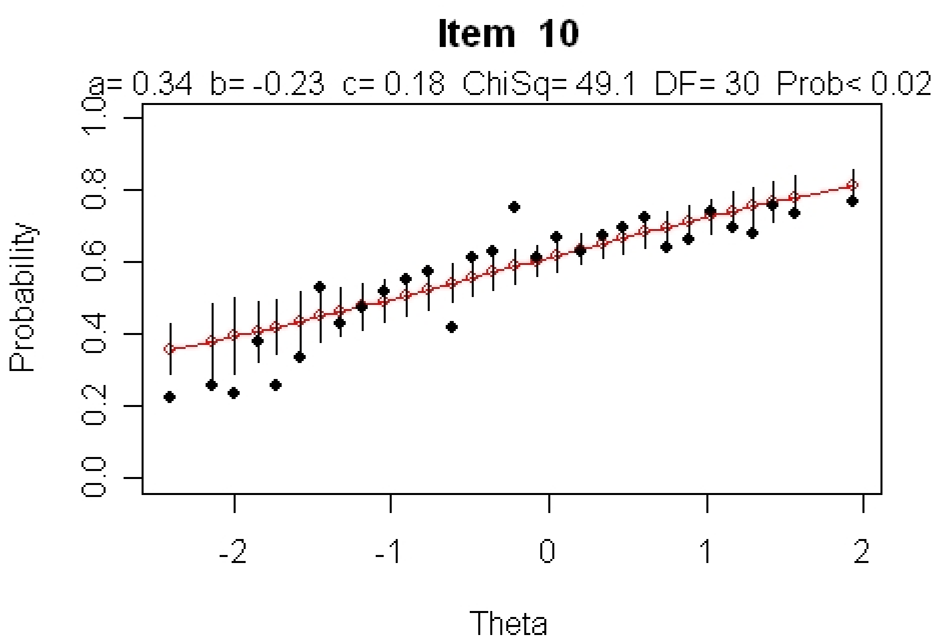

Item 5
a= 1.46 b= 0.82 c= 0.31 ChiSq= 29.96 DF= 28 Prob< 0.37
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

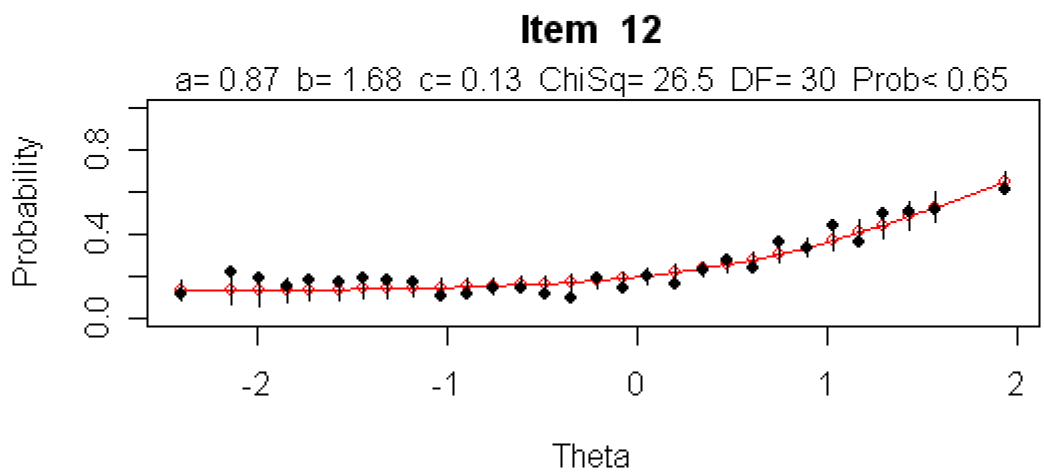

Item 6
a= 0.9 b= -0.39 c= 0.28 ChiSq= 27.44 DF= 25 Prob< 0.33
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

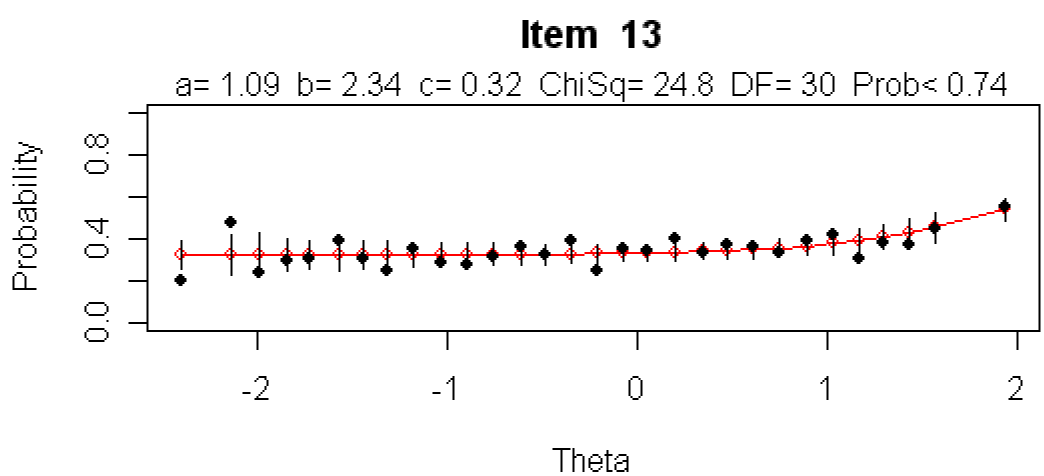

Item 7
a= 1.26 b= 1.7 c= 0.2 ChiSq= 23.71 DF= 30 Prob< 0.79
Probability
0.0
0.4
0.8
-2
-1
0
1
2
Theta

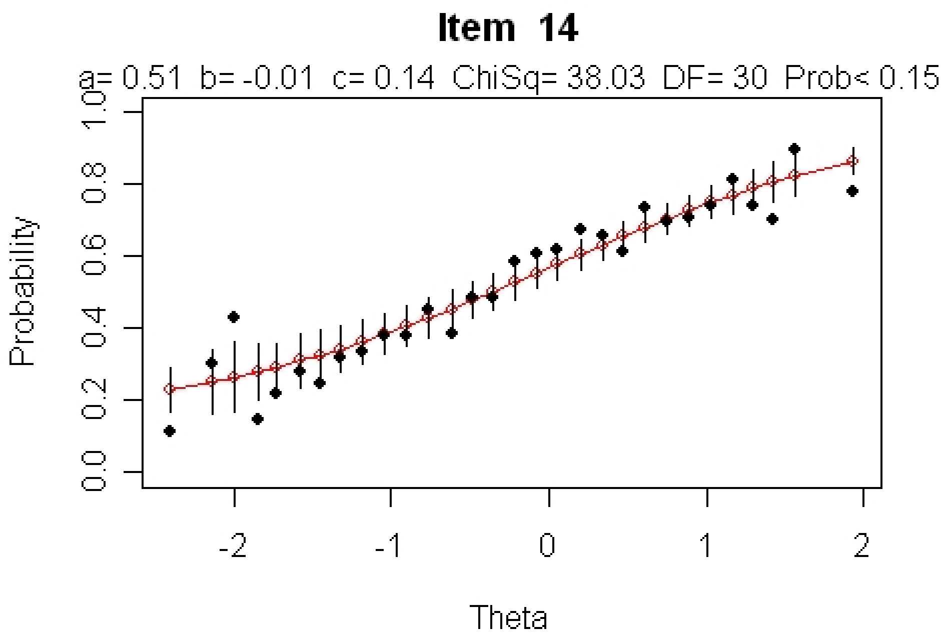

Item 8
a= 0.53 b= 1.42 c= 0.19 ChiSq= 24.42 DF= 30 Prob< 0.75
Probability
0.0
0.4
0.8
-2
-1
0
1
2
Theta

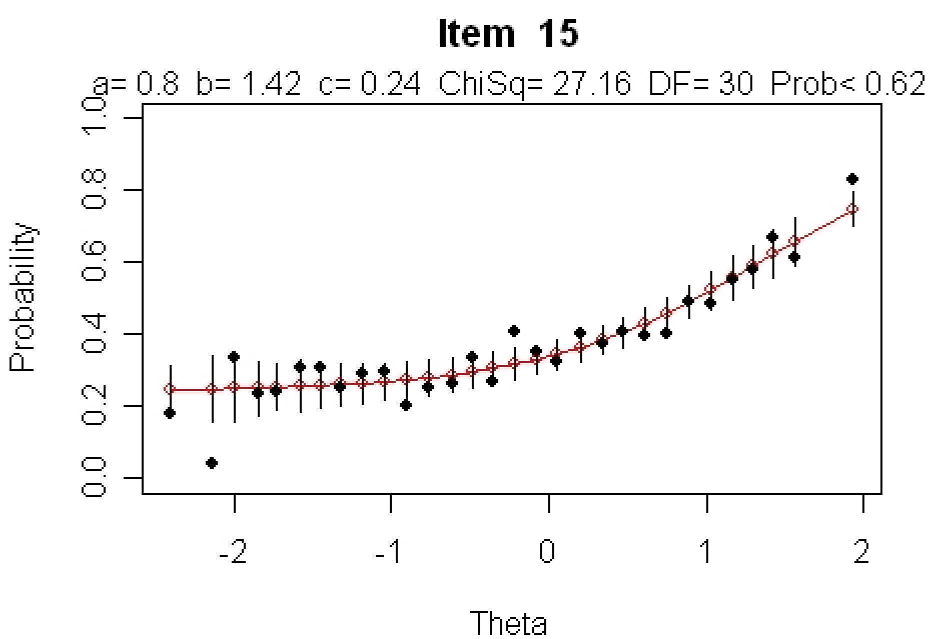

Item 9
a= 0.99 b= 0.1 c= 0.26 ChiSq= 39.96 DF= 27 Prob< 0.05
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

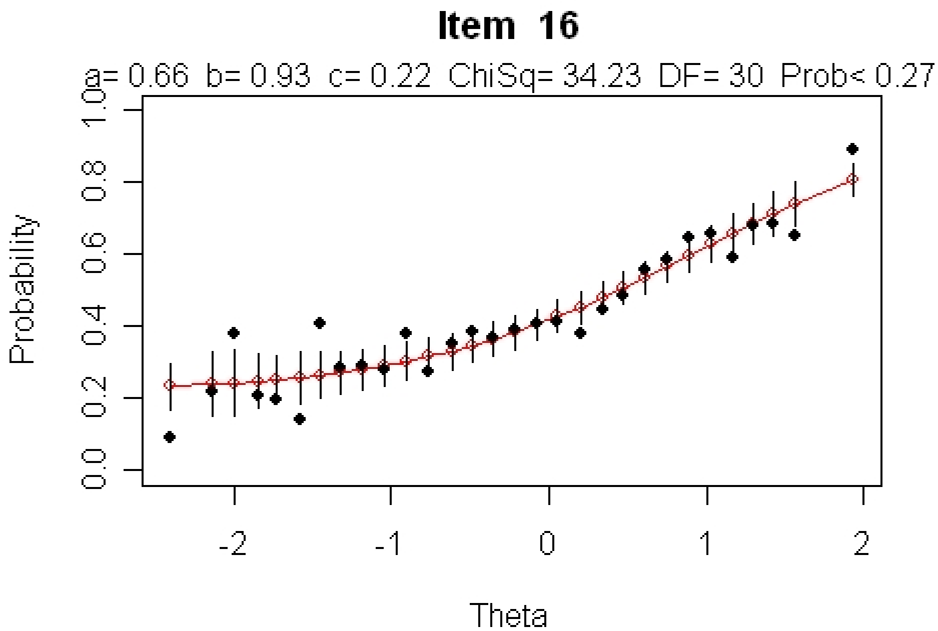

Item 10
a= 0.34 b= -0.23 c= 0.18 ChiSq= 49.1 DF= 30 Prob< 0.02
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

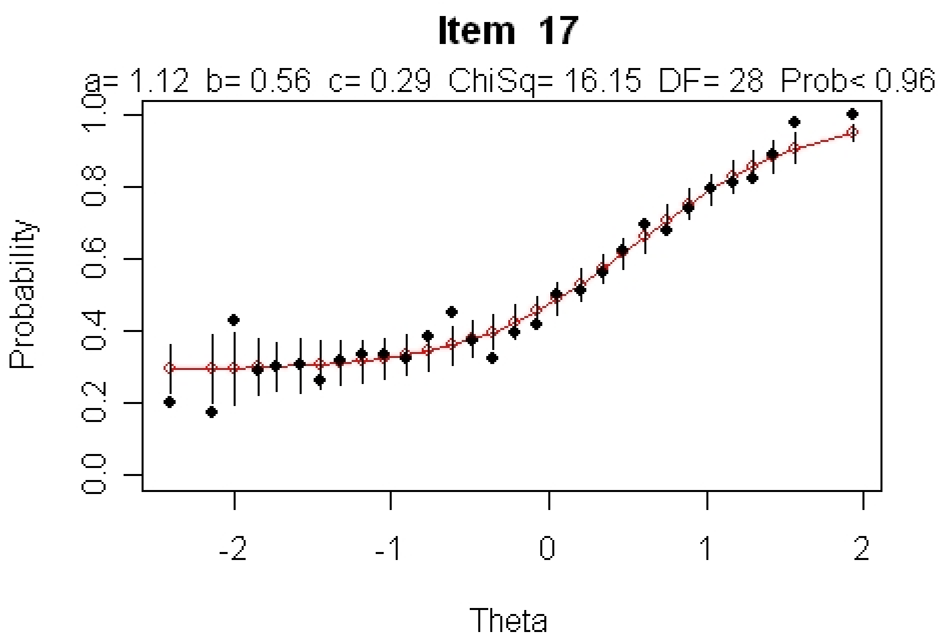

Item 12
a= 0.87 b= 1.68 c= 0.13 ChiSq= 26.5 DF= 30 Prob< 0.65
Probability
0.0
0.4
0.8
-2
-1
0
1
2
Theta

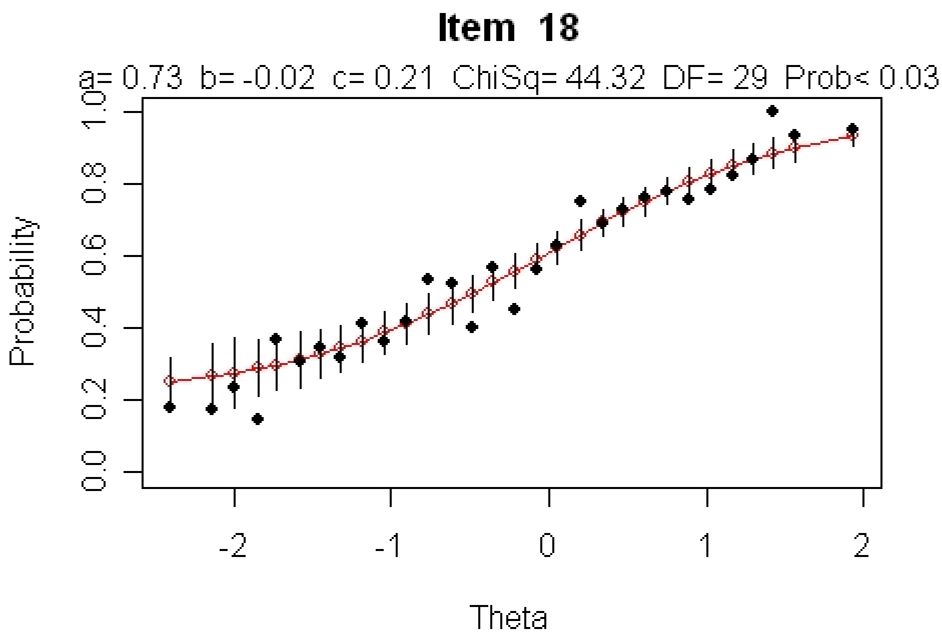

Item 13
a= 1.09 b= 2.34 c= 0.32 ChiSq= 24.8 DF= 30 Prob< 0.74
Probability
0.0
0.4
0.8
-2
-1
0
1
2
Theta

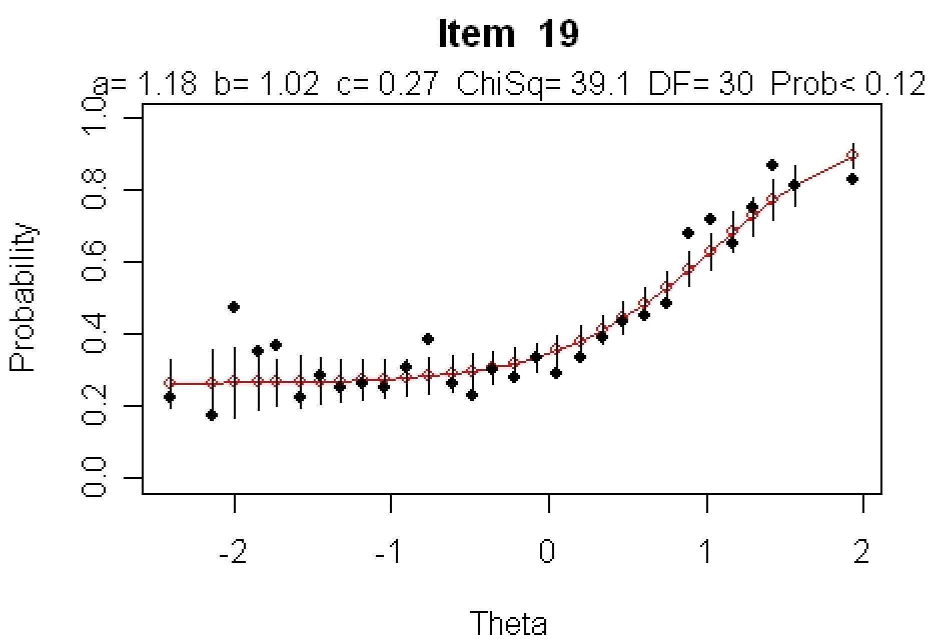

Item 14
a= 0.51 b= -0.01 c= 0.14 ChiSq= 38.03 DF= 30 Prob< 0.15
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

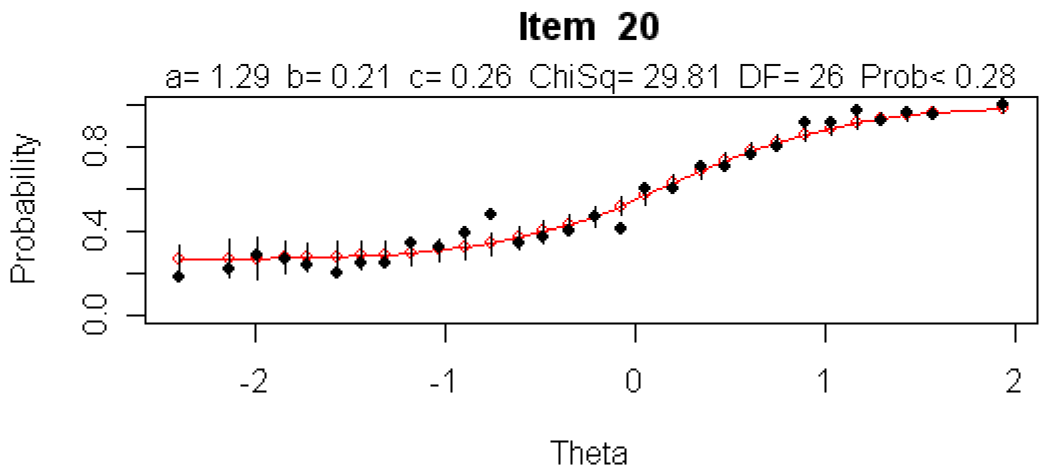

Item 15
a= 0.8 b= 1.42 c= 0.24 ChiSq= 27.16 DF= 30 Prob< 0.62
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

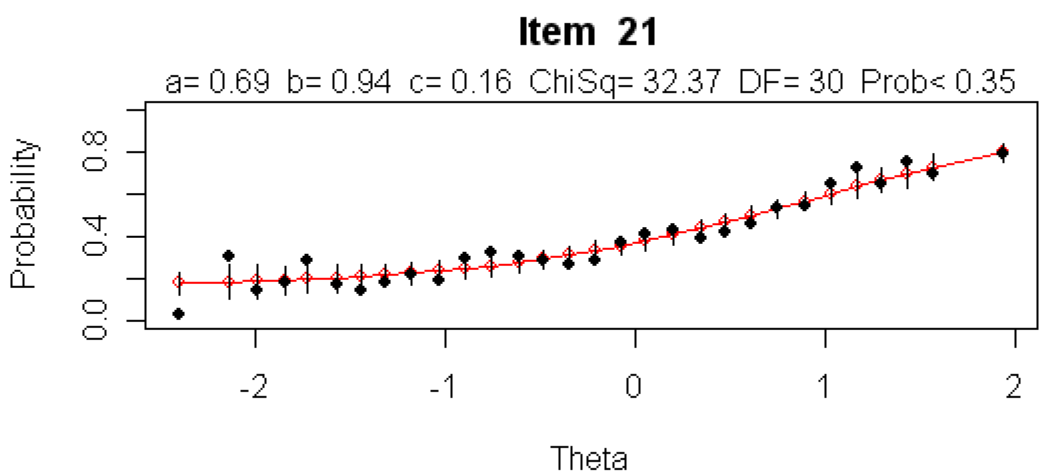

Item 16
a= 0.66 b= 0.93 c= 0.22 ChiSq= 34.23 DF= 30 Prob< 0.27
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

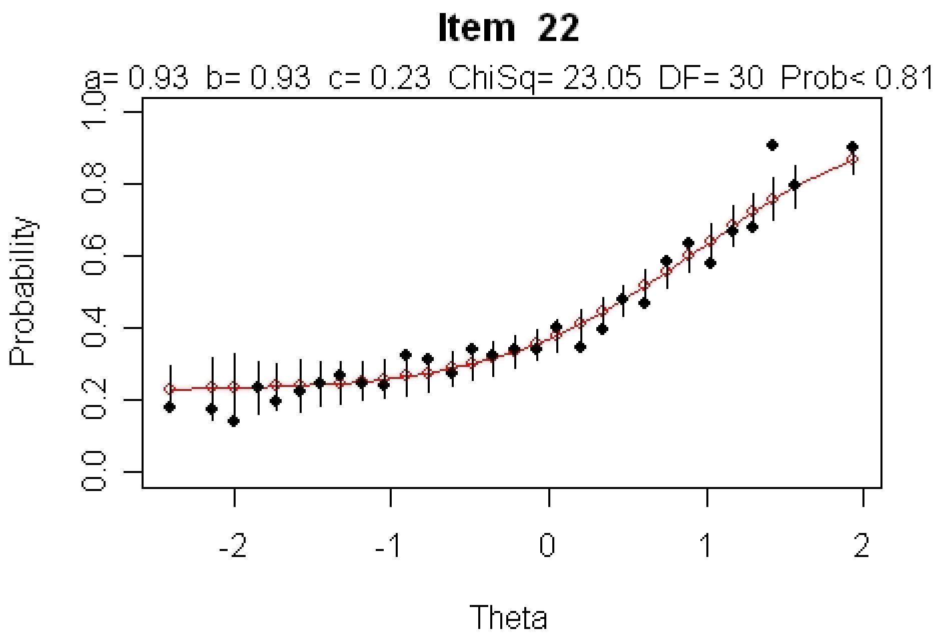

Item 17
a= 1.12 b= 0.56 c= 0.29 ChiSq= 16.15 DF= 28 Prob< 0.96
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

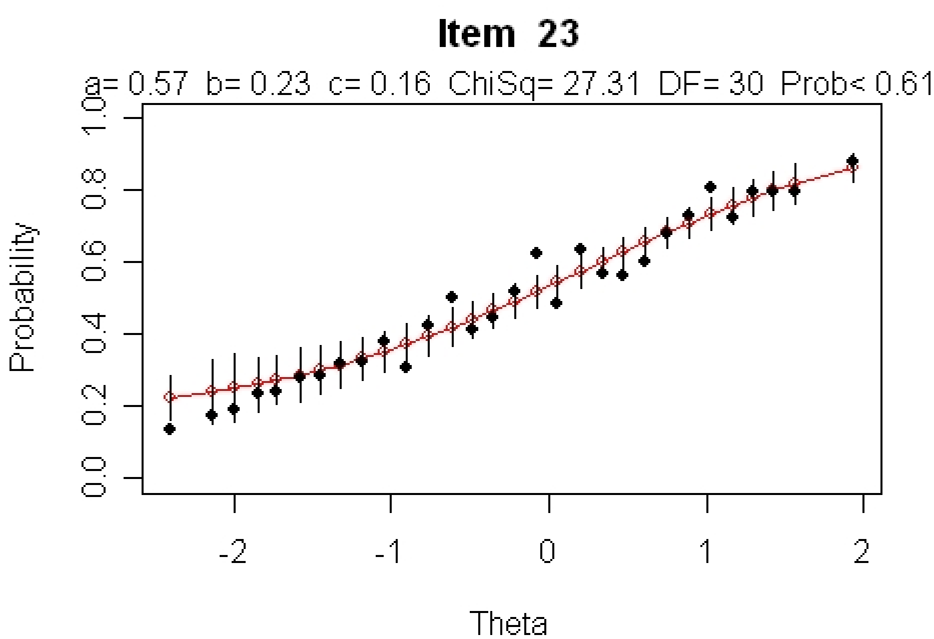

Item 18
a= 0.73 b= -0.02 c= 0.21 ChiSq= 44.32 DF= 29 Prob< 0.03
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

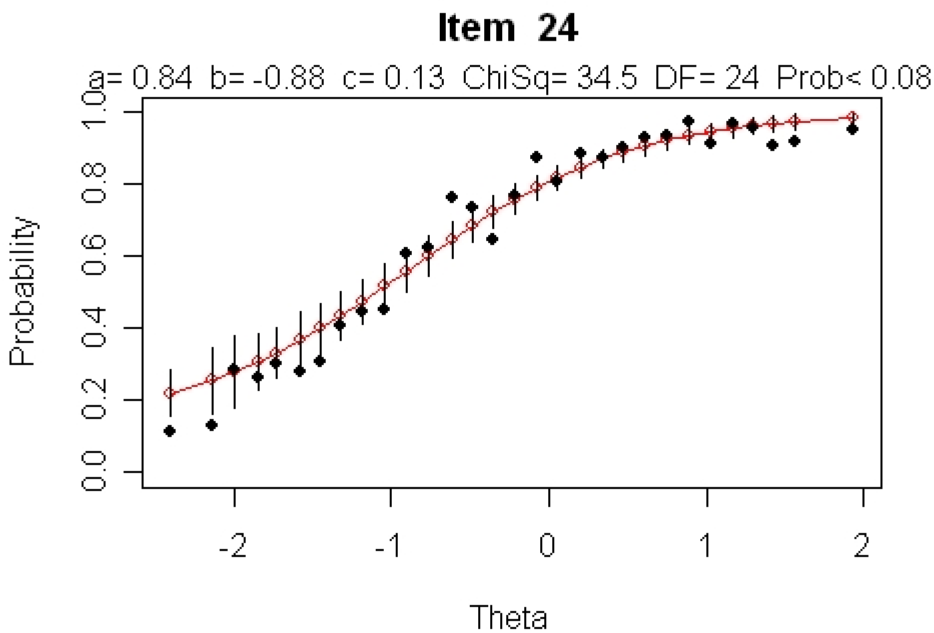

Item 19
a= 1.18 b= 1.02 c= 0.27 ChiSq= 39.1 DF= 30 Prob< 0.12
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

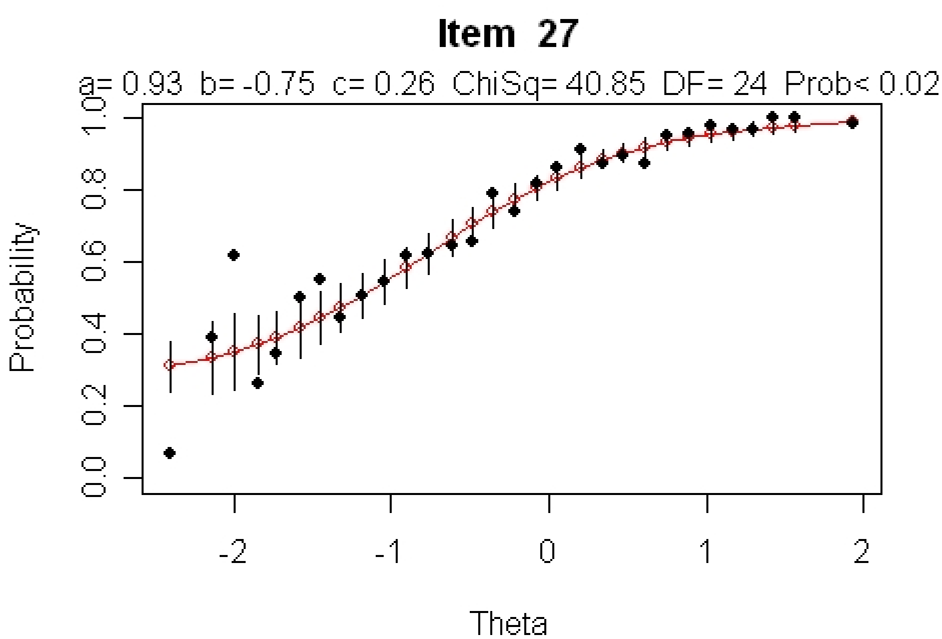

Item 20
a= 1.29 b= 0.21 c= 0.26 ChiSq= 29.81 DF= 26 Prob< 0.28
Probability
0.0
0.4
0.8
-2
-1
0
1
2
Theta

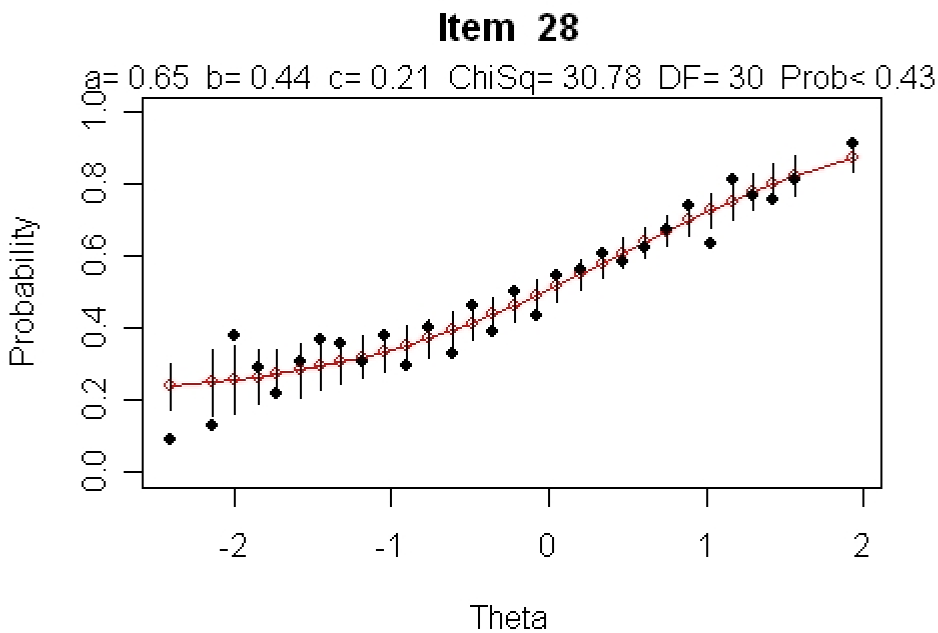

Item 21
a= 0.69 b= 0.94 c= 0.16 ChiSq= 32.37 DF= 30 Prob< 0.35
Probability
0.0
0.4
0.8
-2
-1
0
1
2
Theta

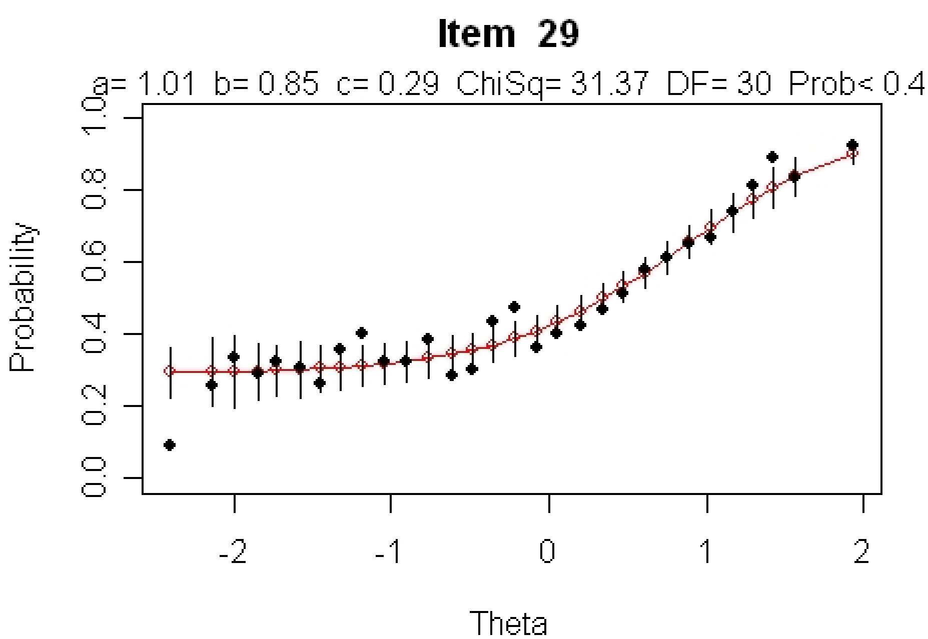

Item 22
a= 0.93 b= 0.93 c= 0.23 ChiSq= 23.05 DF= 30 Prob< 0.81
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

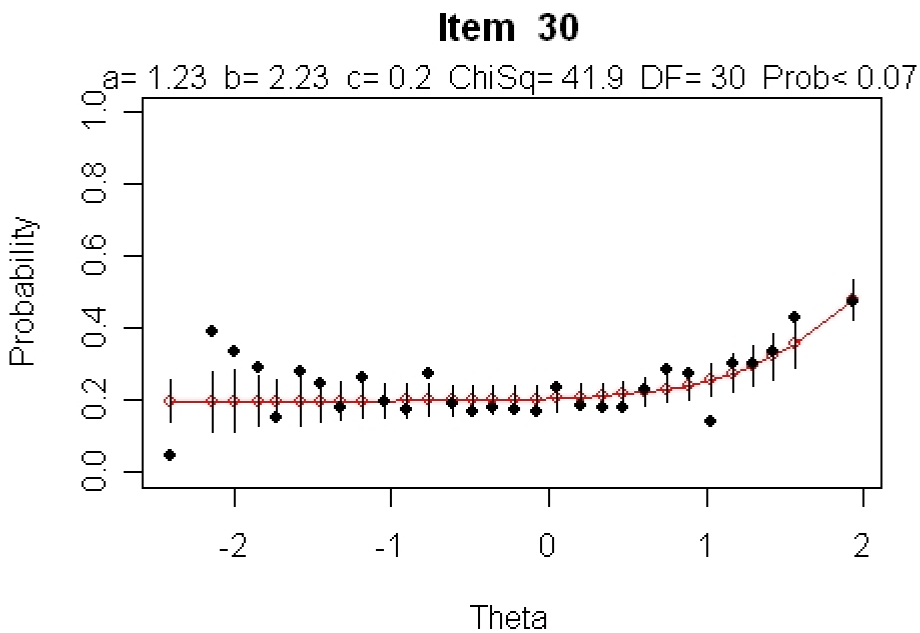

Item 23
a= 0.57 b= 0.23 c= 0.16 ChiSq= 27.31 DF= 30 Prob< 0.61
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

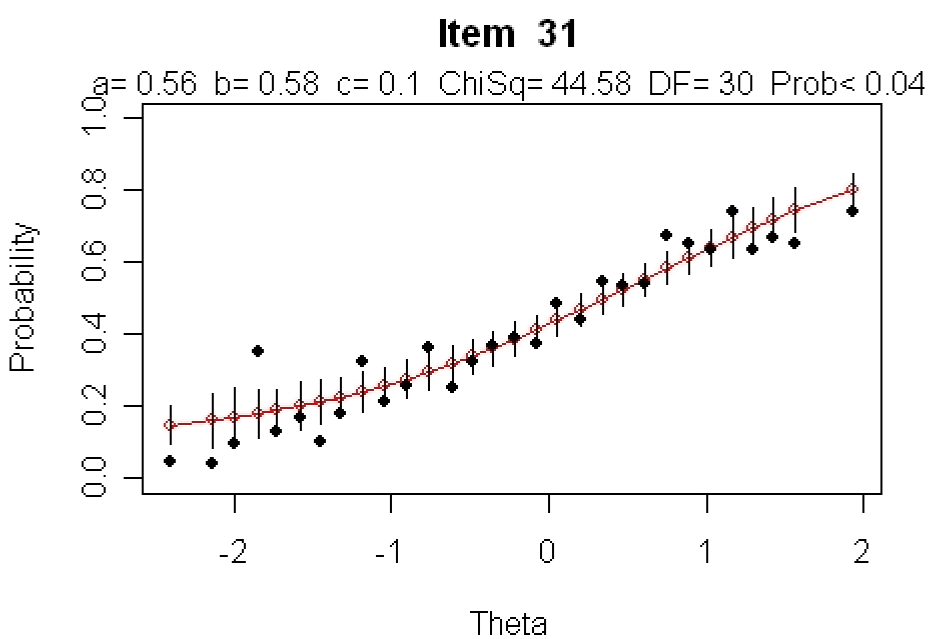

Item 24
a= 0.84 b= -0.88 c= 0.13 ChiSq= 34.5 DF= 24 Prob< 0.08
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

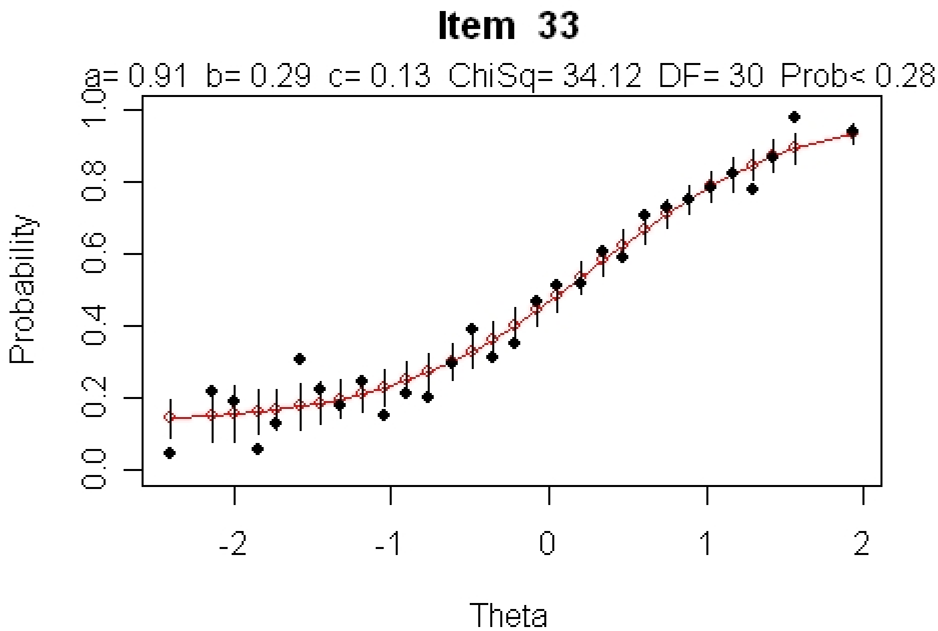

Item 27
a= 0.93 b= -0.75 c= 0.26 ChiSq= 40.85 DF= 24 Prob< 0.02
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

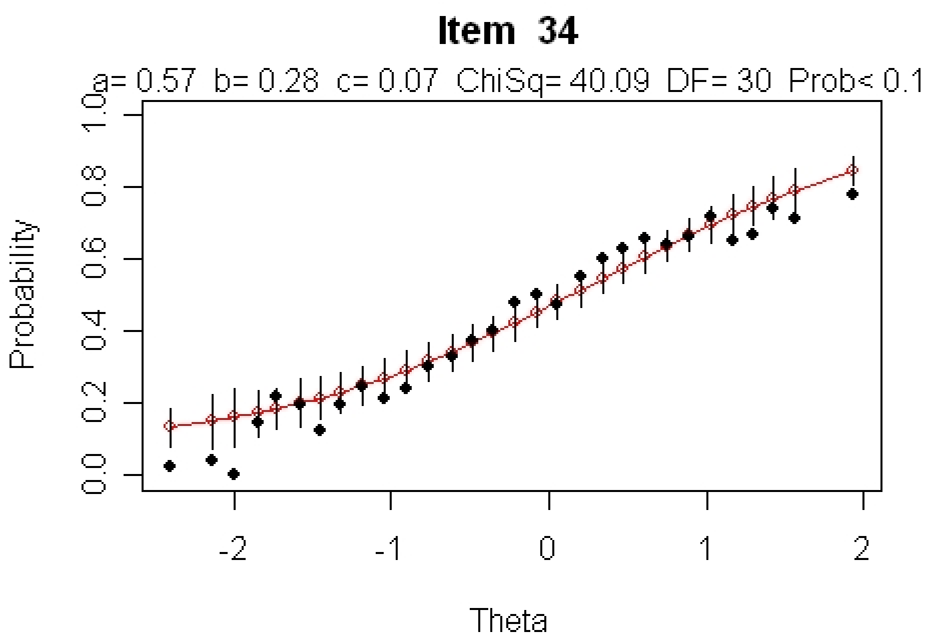

Item 28
a= 0.65 b= 0.44 c= 0.21 ChiSq= 30.78 DF= 30 Prob< 0.43
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

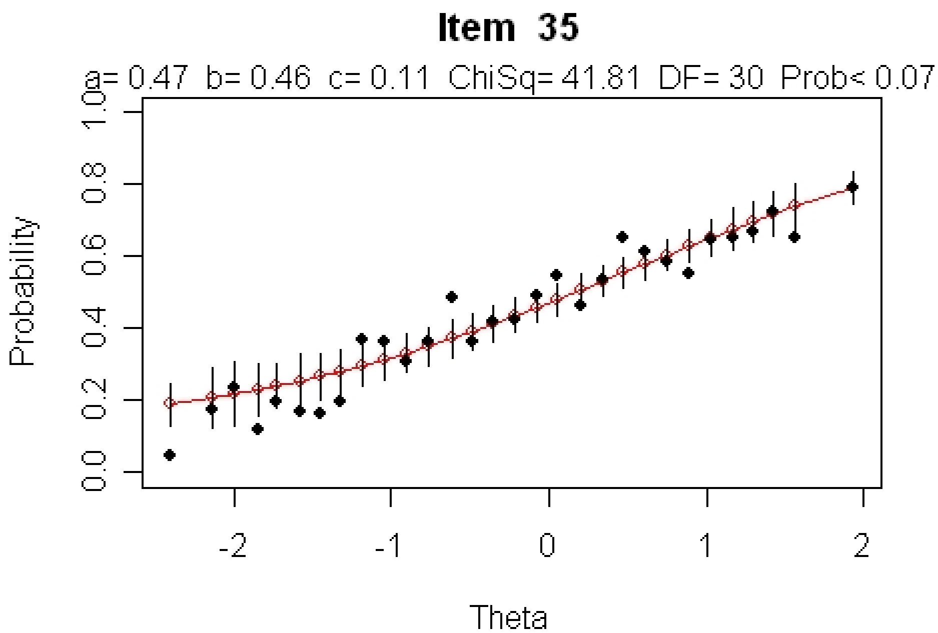

Item 29
a= 1.01 b= 0.85 c= 0.29 ChiSq= 31.37 DF= 30 Prob< 0.4
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

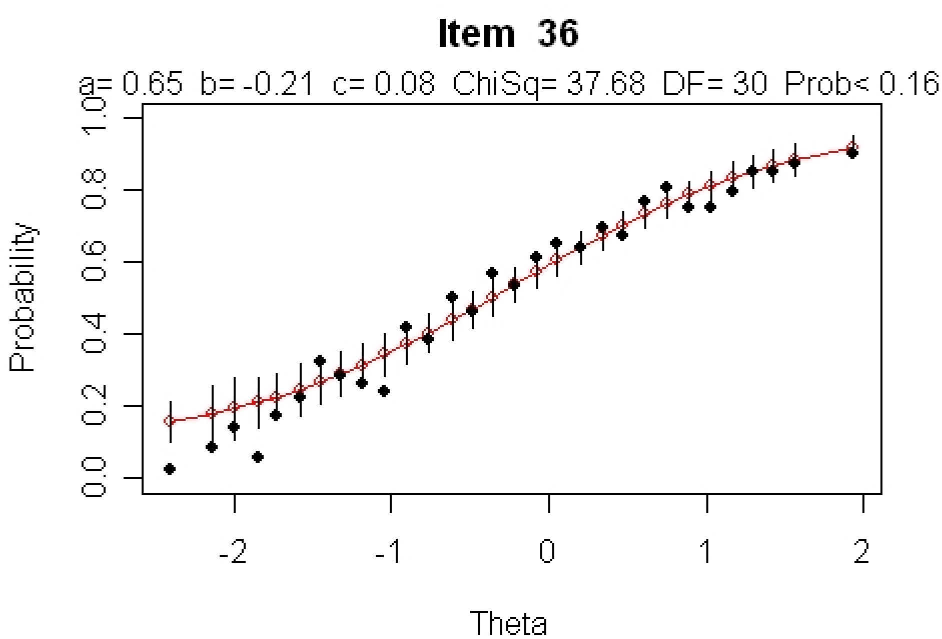

Item 30
a= 1.23 b= 2.23 c= 0.2 ChiSq= 41.9 DF= 30 Prob< 0.07
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

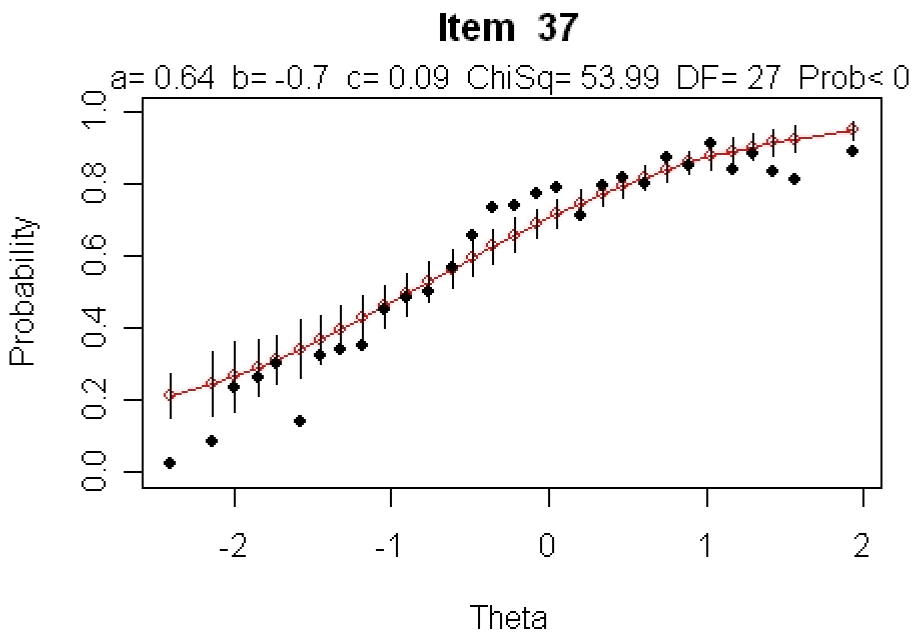

Item 31
a= 0.56 b= 0.58 c= 0.1 ChiSq= 44.58 DF= 30 Prob< 0.04
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

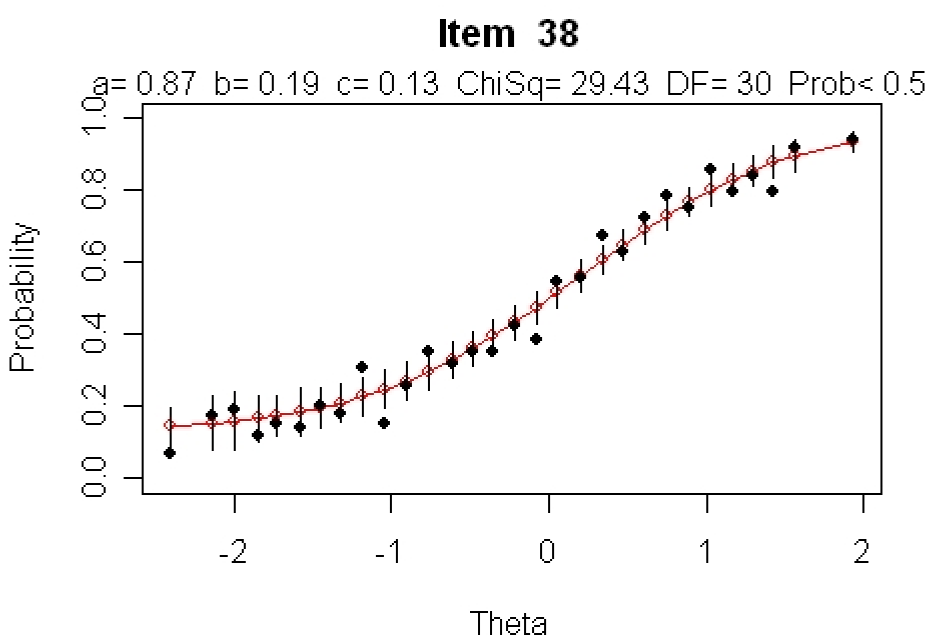

Item 33
a= 0.91 b= 0.29 c= 0.13 ChiSq= 34.12 DF= 30 Prob< 0.28
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

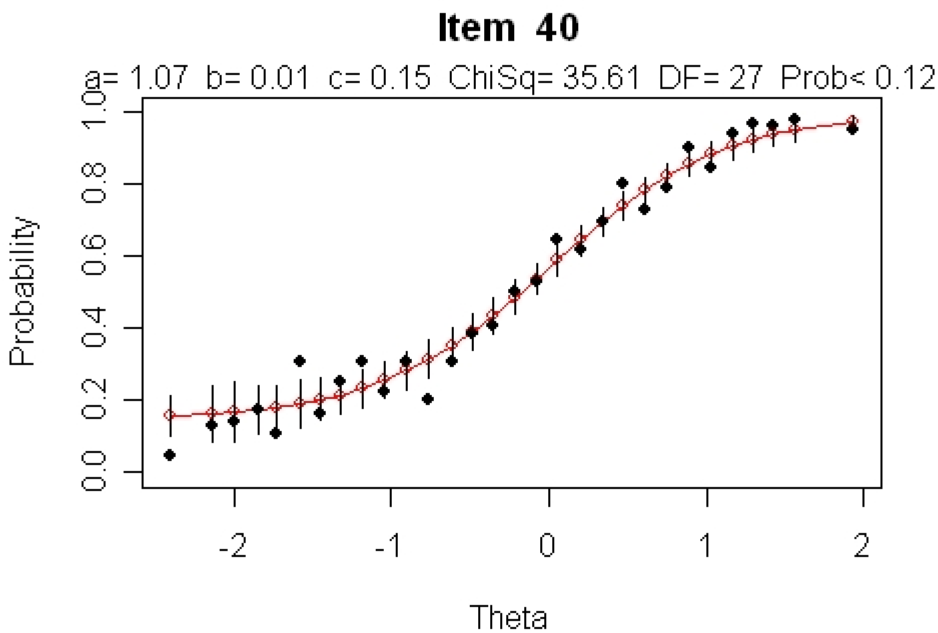

Item 34
a= 0.57 b= 0.28 c= 0.07 ChiSq= 40.09 DF= 30 Prob< 0.1
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

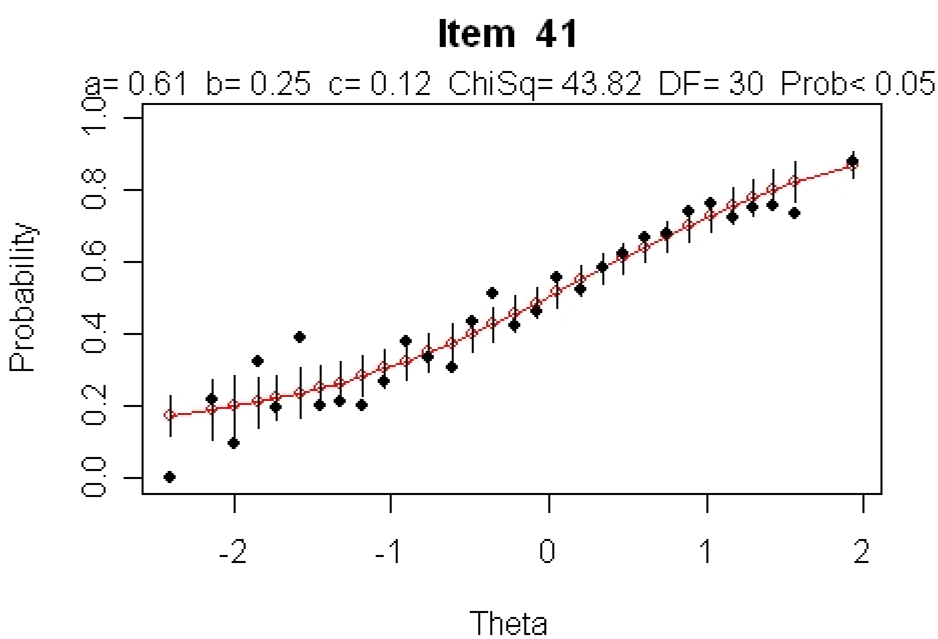

Item 35
a= 0.47 b= 0.46 c= 0.11 ChiSq= 41.81 DF= 30 Prob< 0.07
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

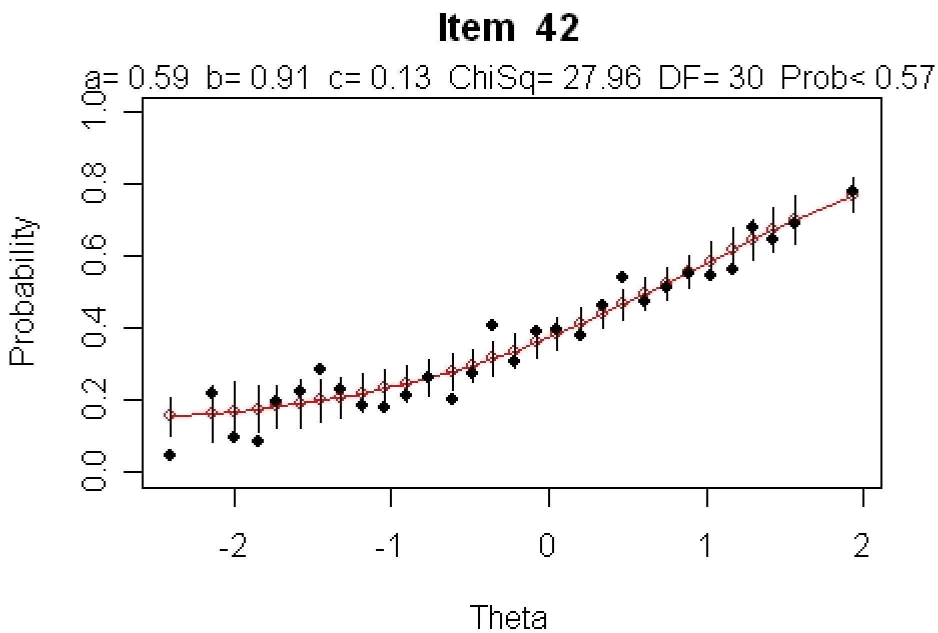

Item 36
a= 0.65 b= -0.21 c= 0.08 ChiSq= 37.68 DF= 30 Prob< 0.16
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

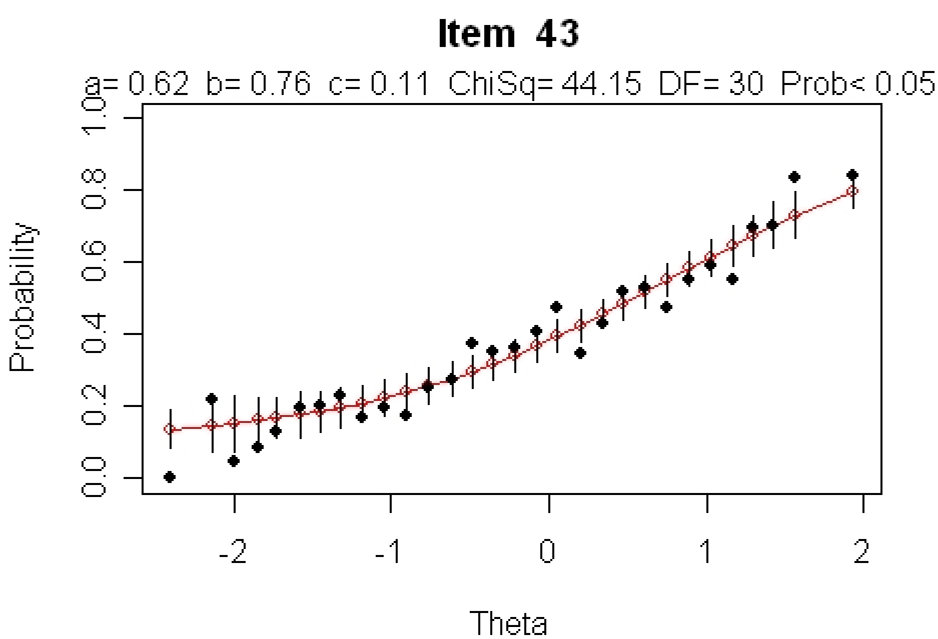

Item 37
a= 0.64 b= -0.7 c= 0.09 ChiSq= 53.99 DF= 27 Prob< 0
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

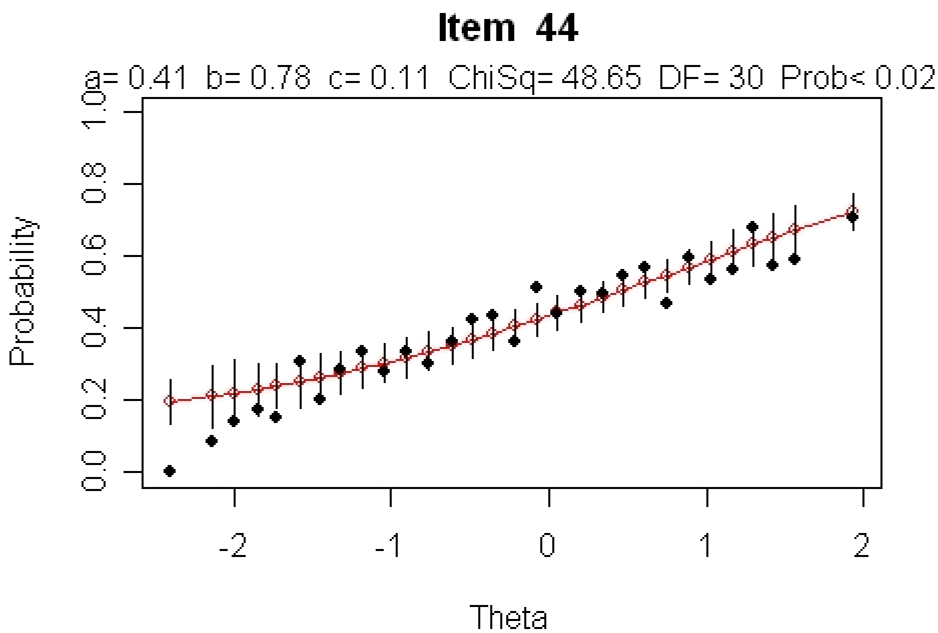

Item 38
a= 0.87 b= 0.19 c= 0.13 ChiSq= 29.43 DF= 30 Prob< 0.5
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta


Item 40
a= 1.07 b= 0.01 c= 0.15 ChiSq= 35.61 DF= 27 Prob< 0.12
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta


Item 41
a= 0.61 b= 0.25 c= 0.12 ChiSq= 43.82 DF= 30 Prob< 0.05
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta


Item 42
a= 0.59 b= 0.91 c= 0.13 ChiSq= 27.96 DF= 30 Prob< 0.57
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta


Item 43
a= 0.62 b= 0.76 c= 0.11 ChiSq= 44.15 DF= 30 Prob< 0.05
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta


Item 44
a= 0.41 b= 0.78 c= 0.11 ChiSq= 48.65 DF= 30 Prob< 0.02
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

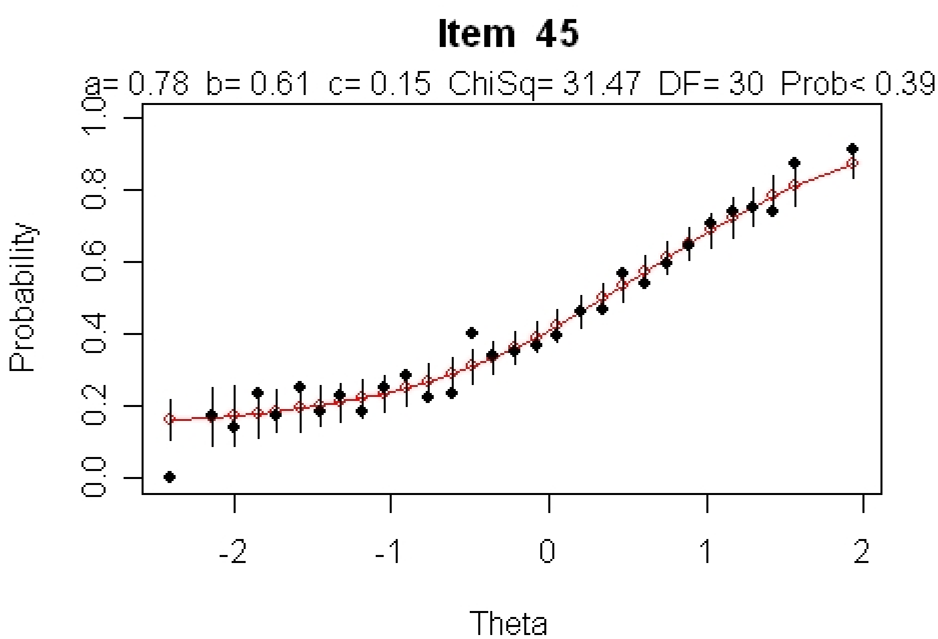

Item 45
a= 0.78 b= 0.61 c= 0.15 ChiSq= 31.47 DF= 30 Prob< 0.39
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

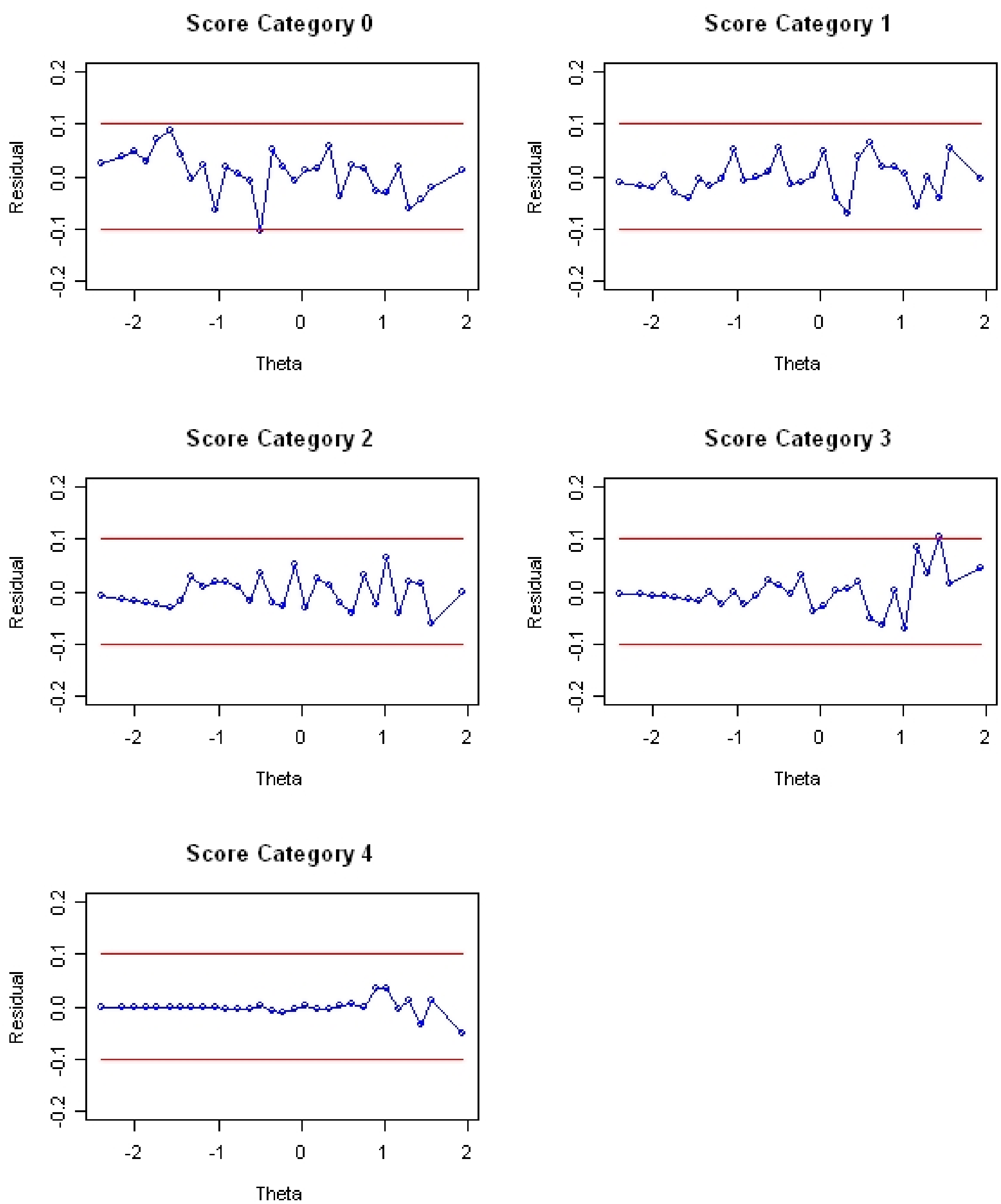
**Figure B.3 Raw residual plots for polytomously-scored items (differences between the observed and expected item performances)**

**Note:** Items 25 and 39 showed a level of model misfit that warrants further investigation.

**a. Item 11**

a = 0.89 b1 = -0.02 b2 = 0.42 b3 = 1.17 b4 = 2.81

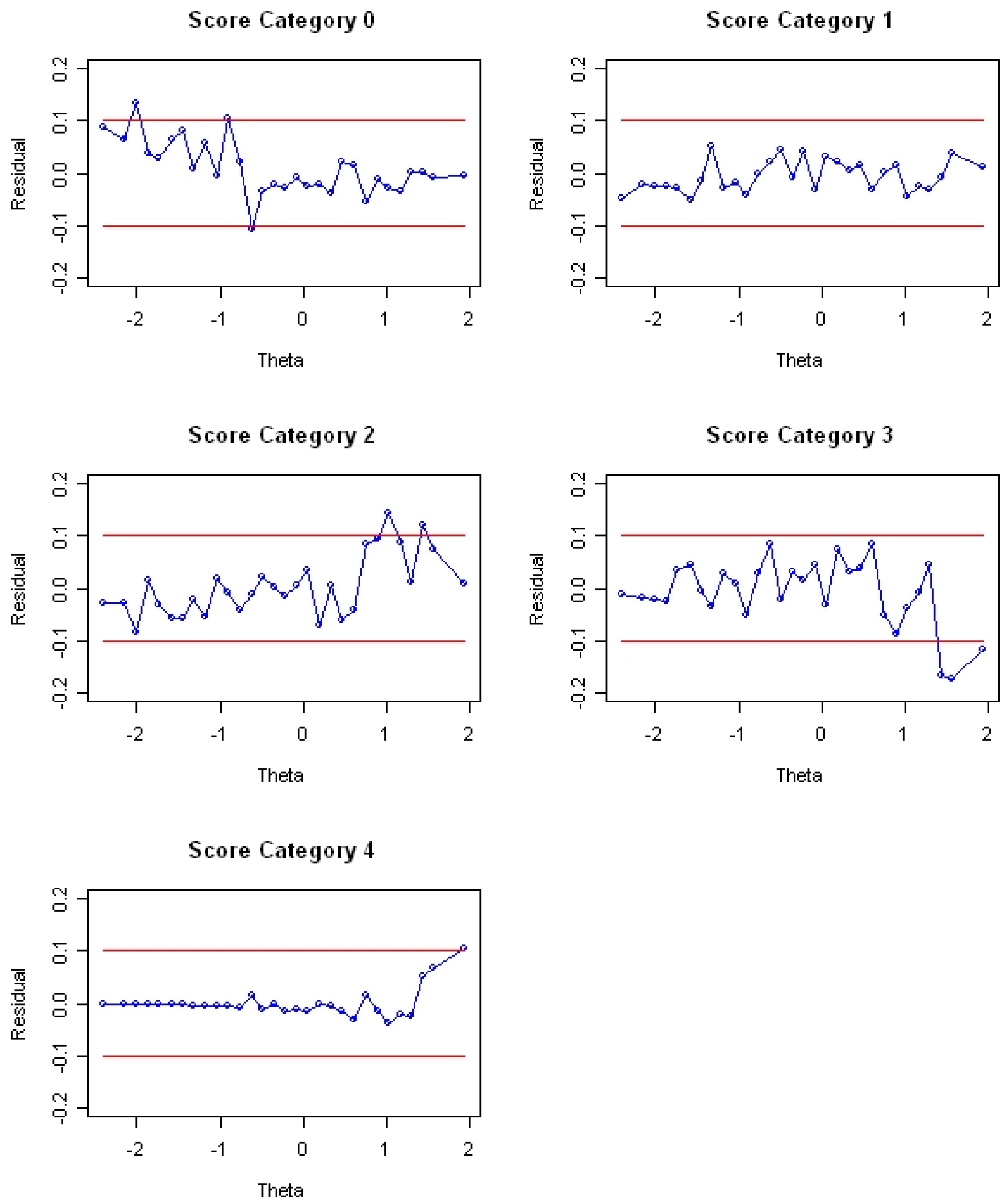
Chisq = 92.28 DF =89 Prob < 0.39



## b. Item 25

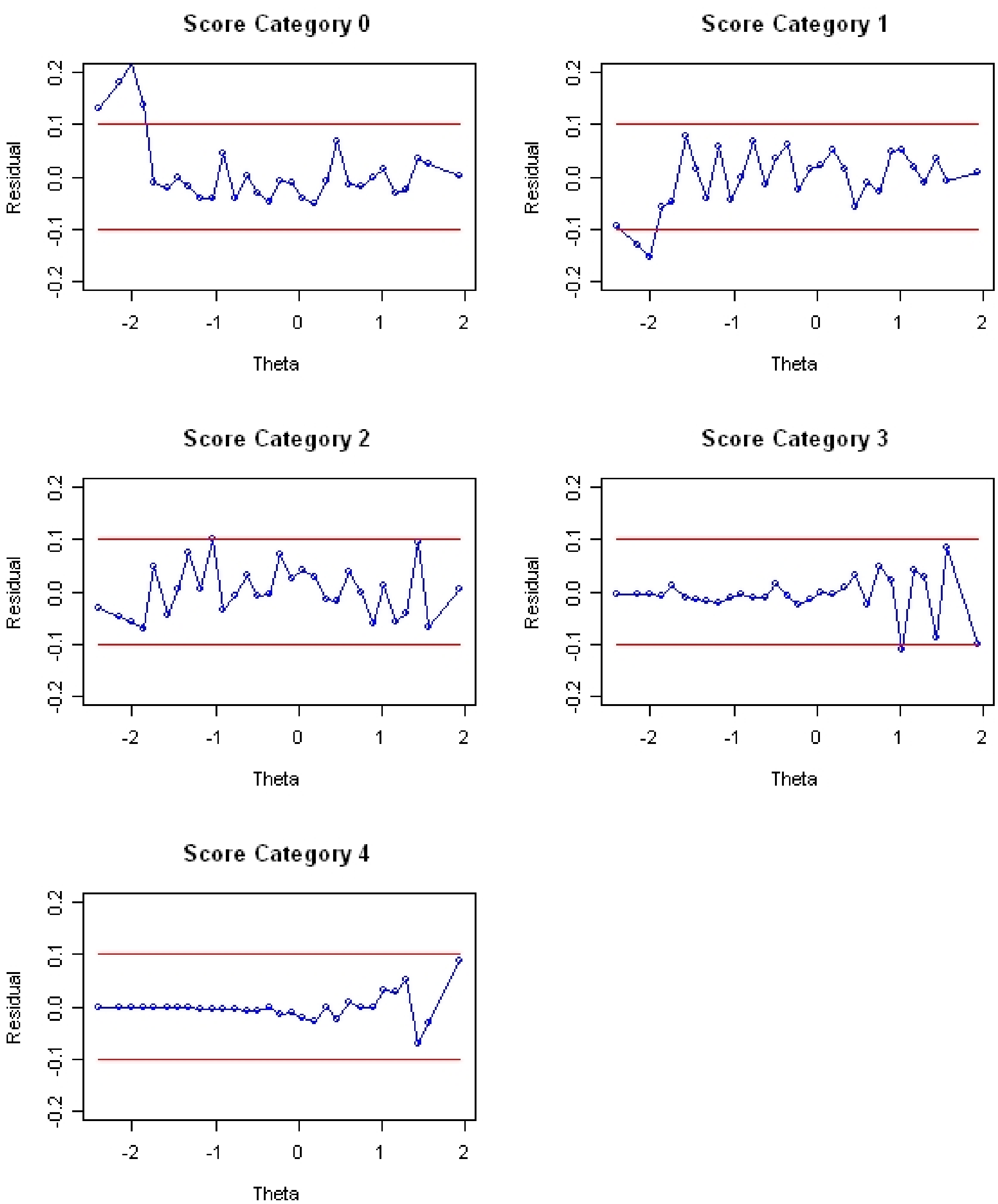
a = 0.83  b1 = -0.93  b2 = -0.50  b3 = 0.67  b4 = 2.76

Chisq = 125.33  DF =93  Prob < 0.01



### c. Item 26

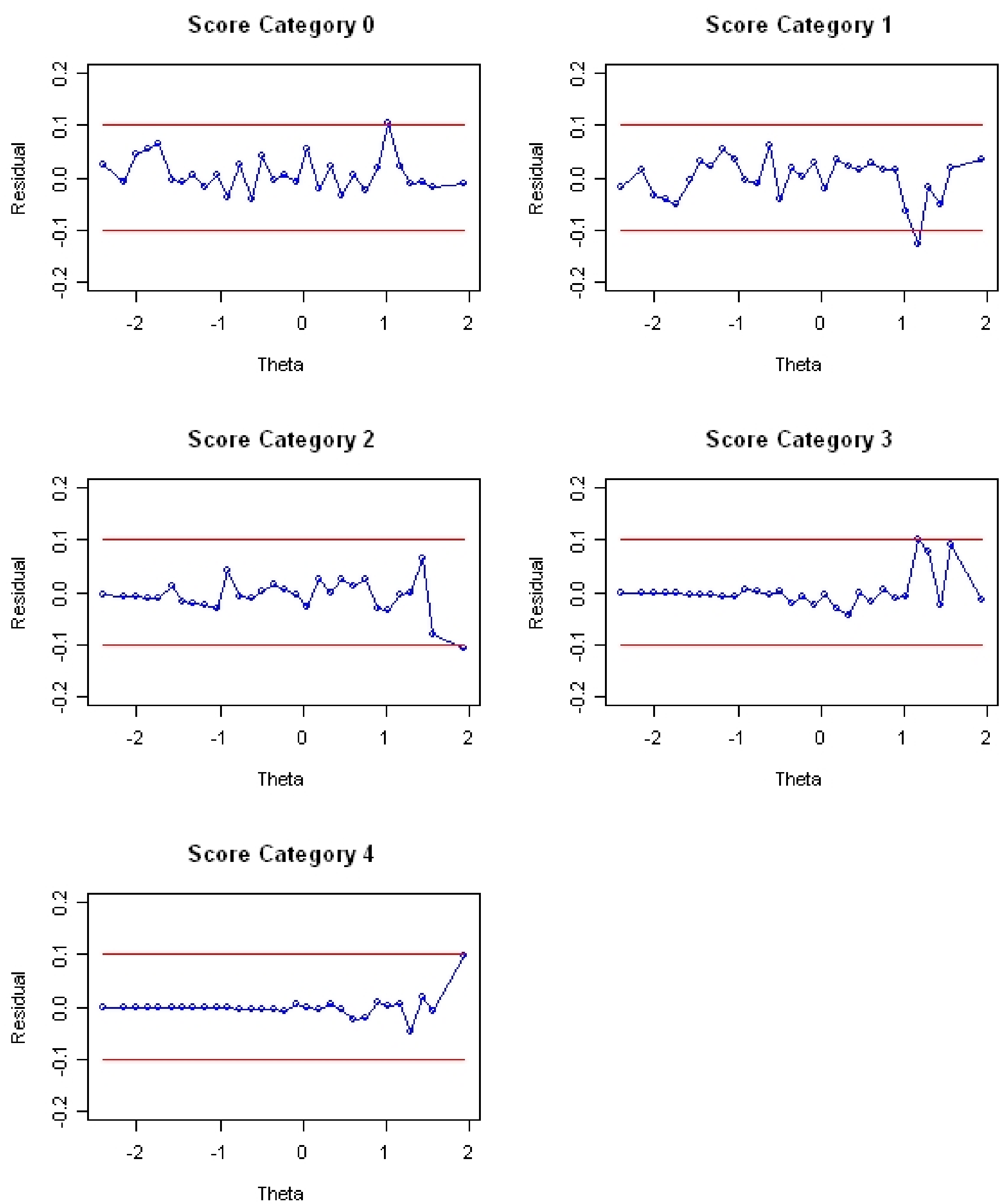
a = 0.88 b1 = -1.14 b2 = -0.21 b3 =1.30 b4 = 2.61
Chisq = 137.67 DF =90 Prob < 0.00



## d. Item 32

a = 0.87  b1 = 0.07  b2 = 1.08  b3 = 2.06  b4 = 3.13

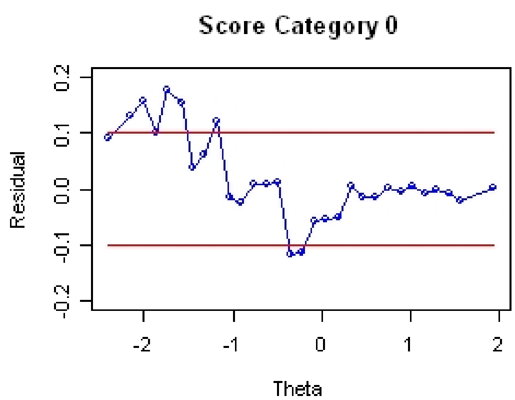
Chisq = 102.87  DF =81  Prob < 0.05

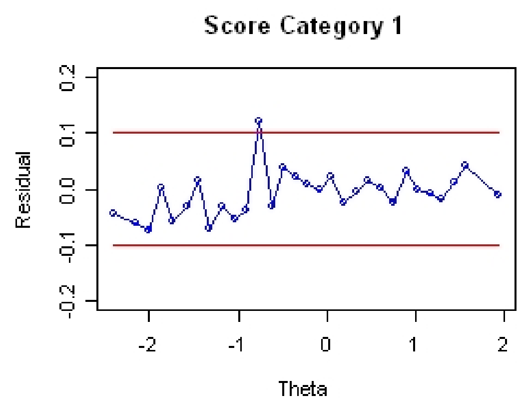


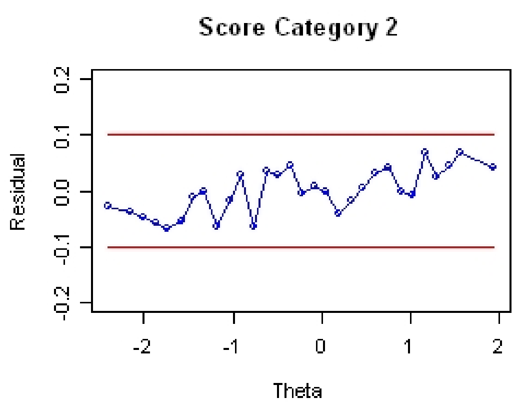
### e. Item 39

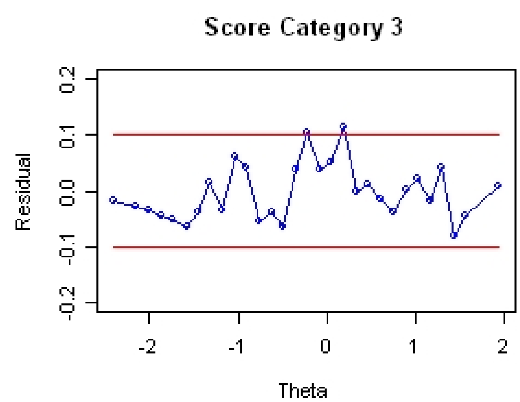
a = 0.91 b1 = -0.91 b2 = -0.45 b3 = 0.12 b4 = 2.29

Chisq = 137.22 DF =92 Prob < 0.00



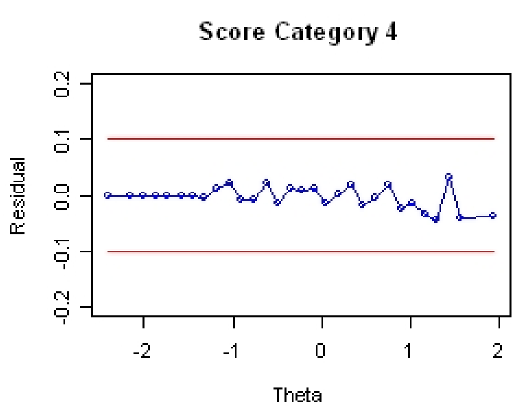
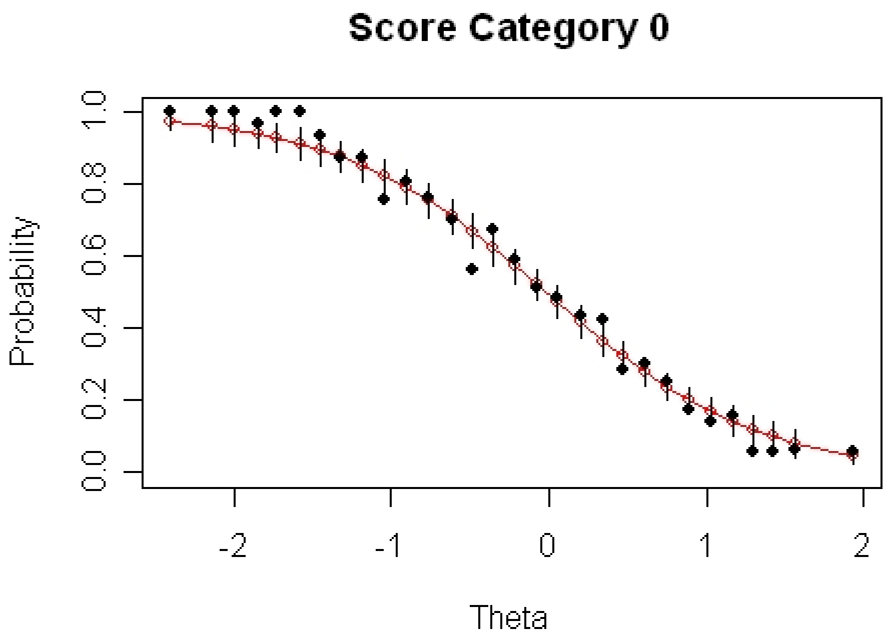



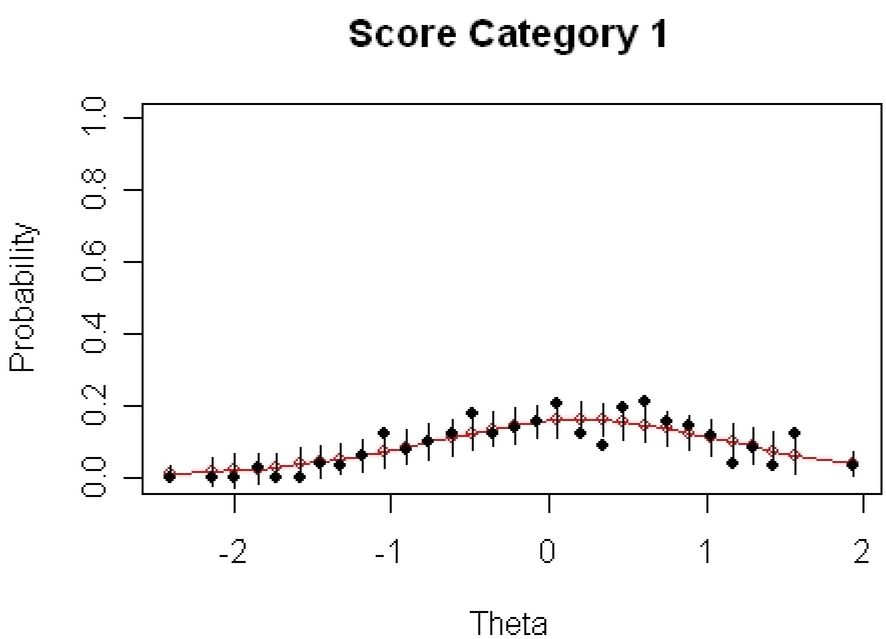
**Figure B.4 Probability plots for polytomously-scored items highlighting the level of model misfit**

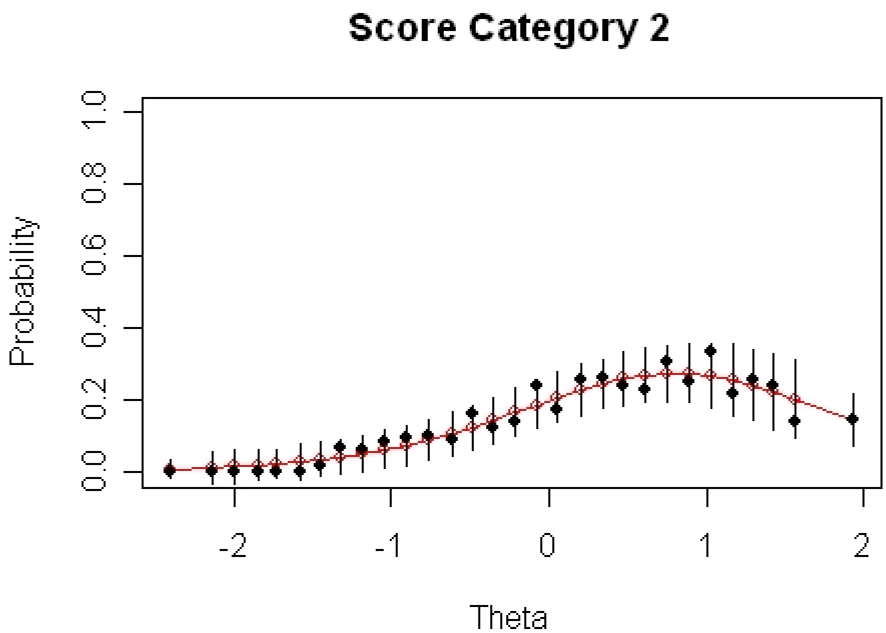
**a. Item 11**

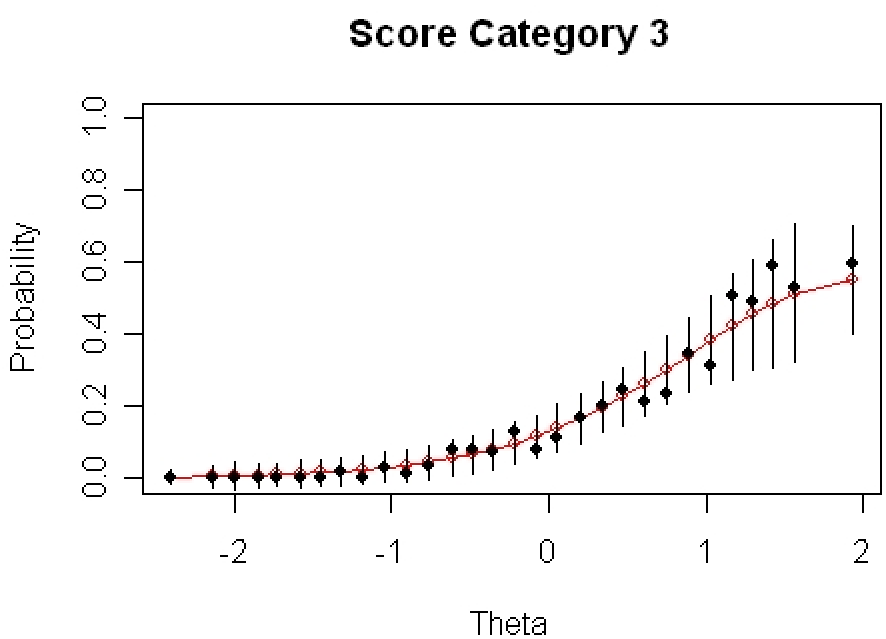
a = 0.89 b1 = -0.02 b2 = 0.42 b3 = 1.17 b4 = 2.81

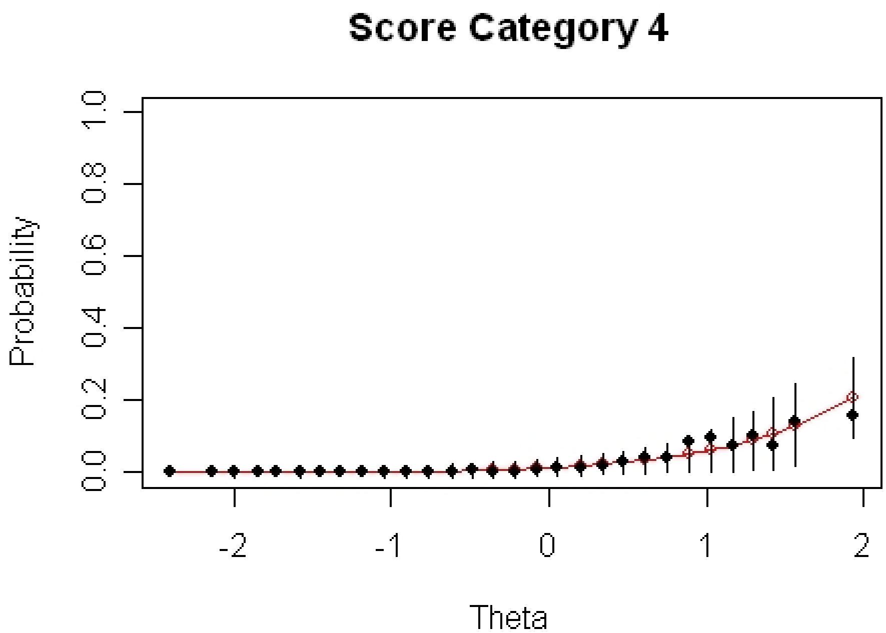
Chisq = 92.28 DF =89 Prob < 0.39


Score Category 0
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta


Score Category 1
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta


Score Category 2
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

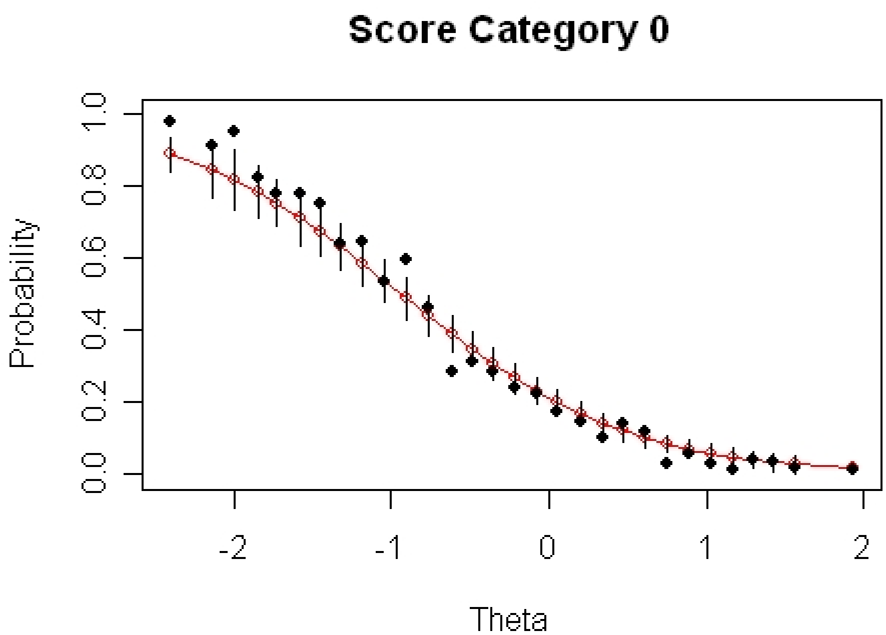

Score Category 3
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

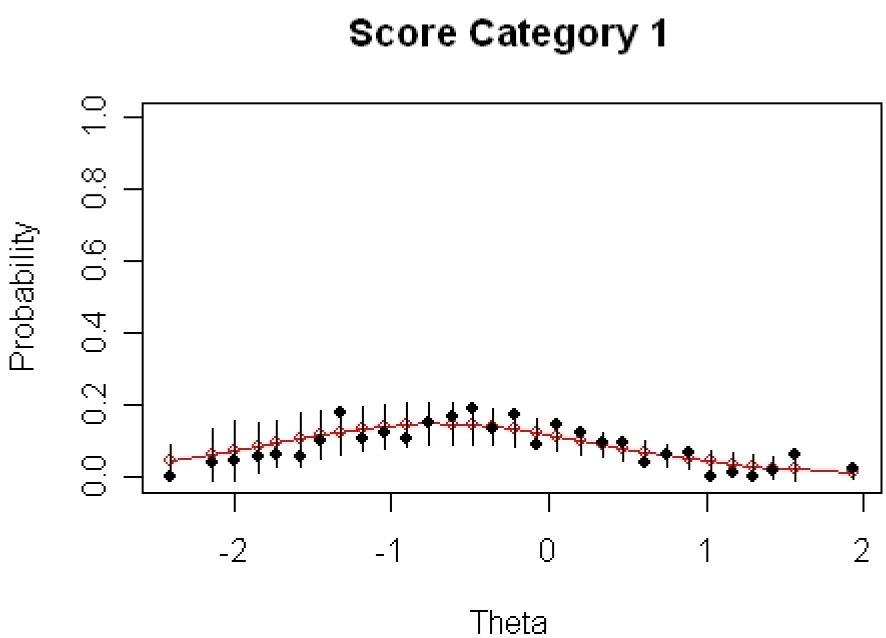


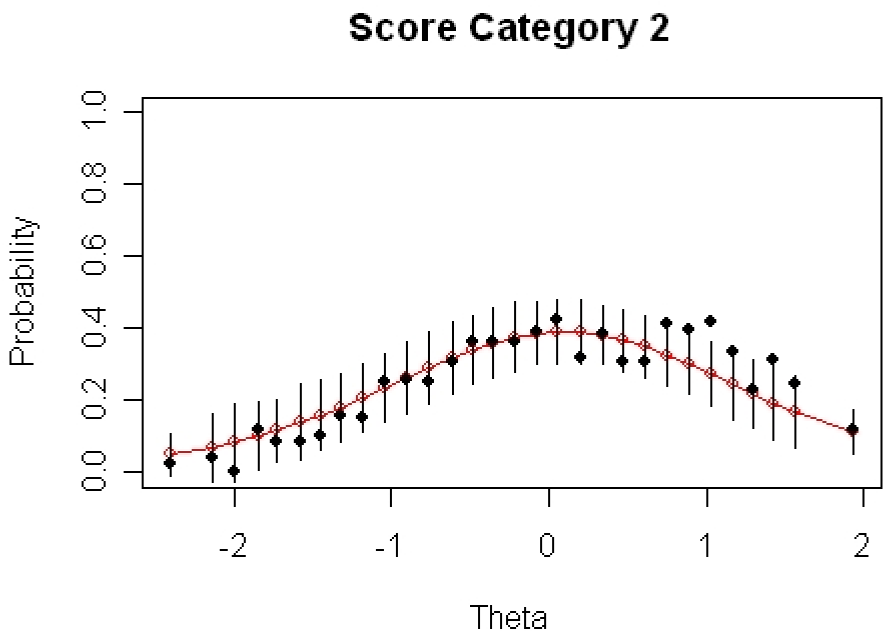
Score Category 4
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

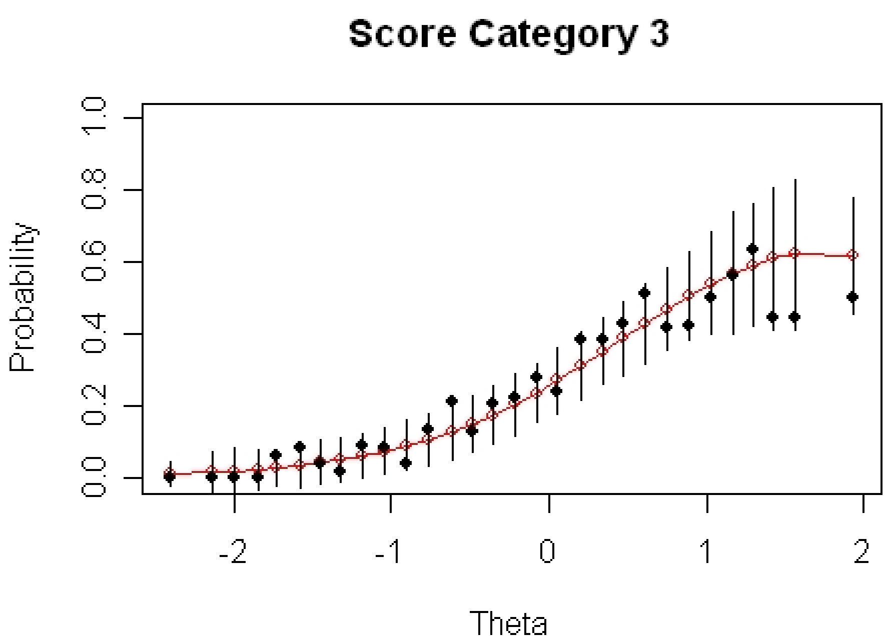
**b. Item 25**

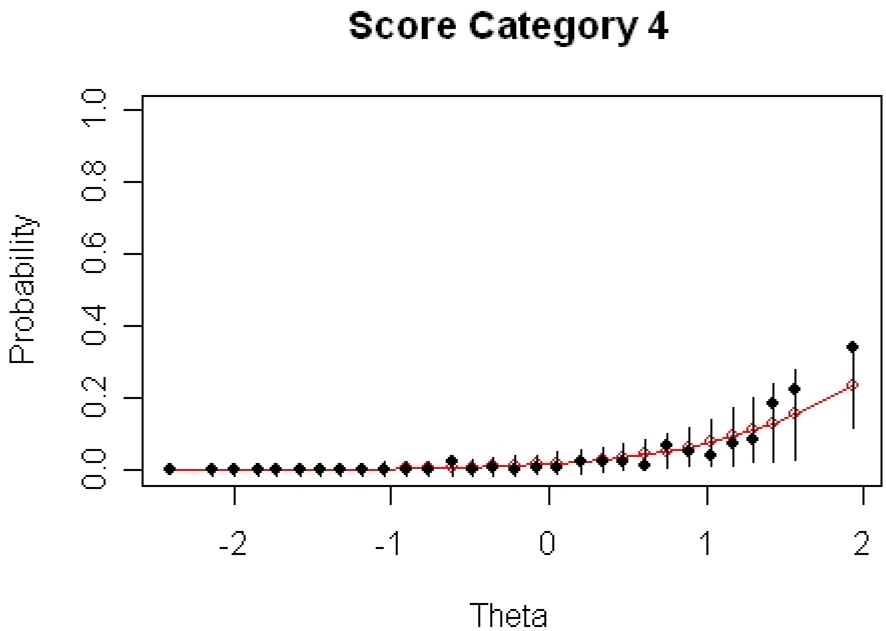
a = 0.83 b1 = -0.93 b2 = -0.50 b3 = 0.67 b4 = 2.76
Chisq = 125.33 DF =93 Prob < 0.01




Score Category 2
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

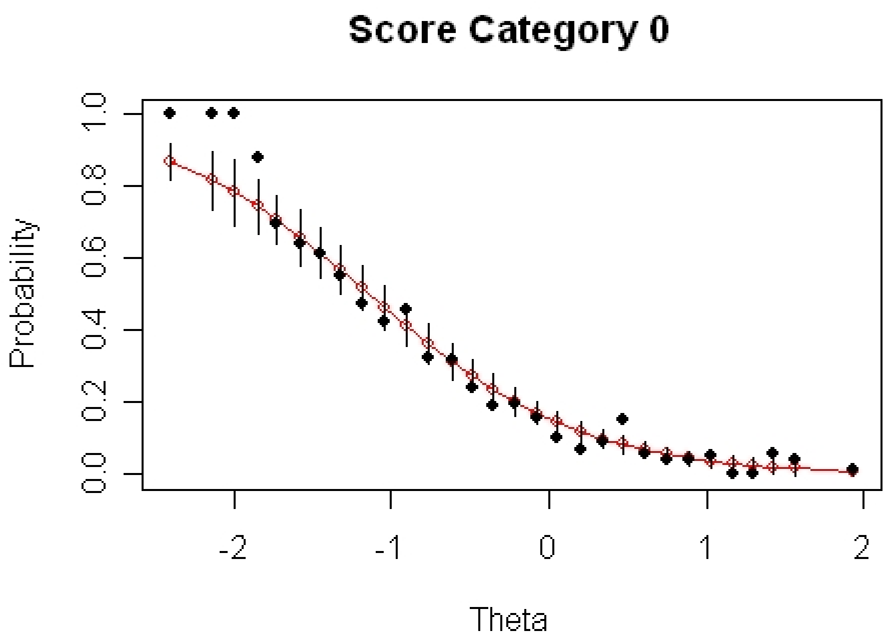

Score Category 3
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

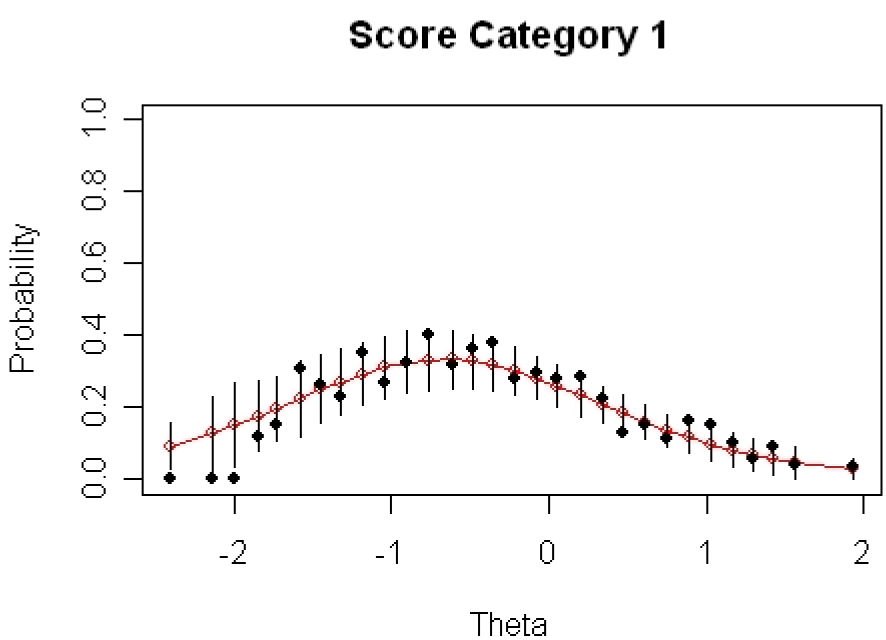


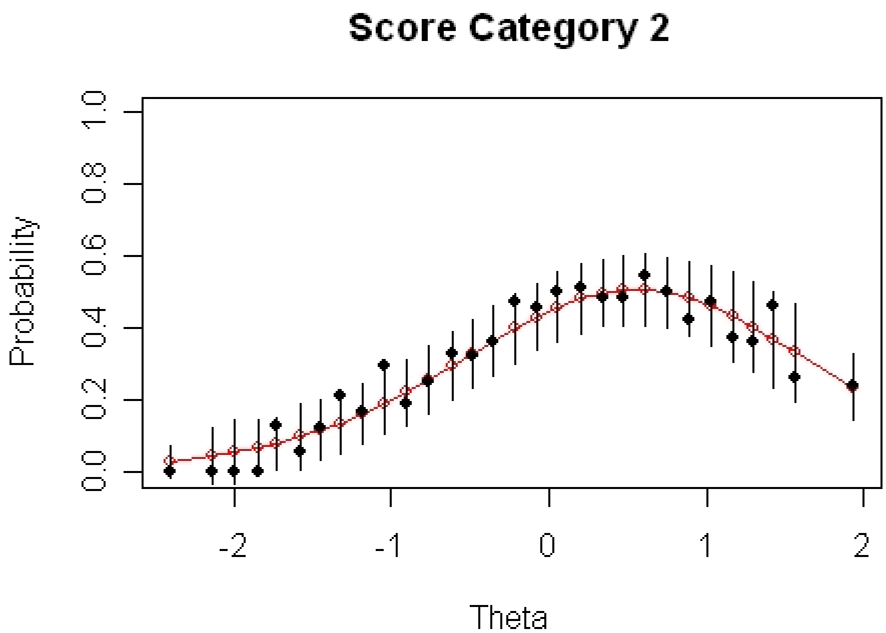
Score Category 4
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

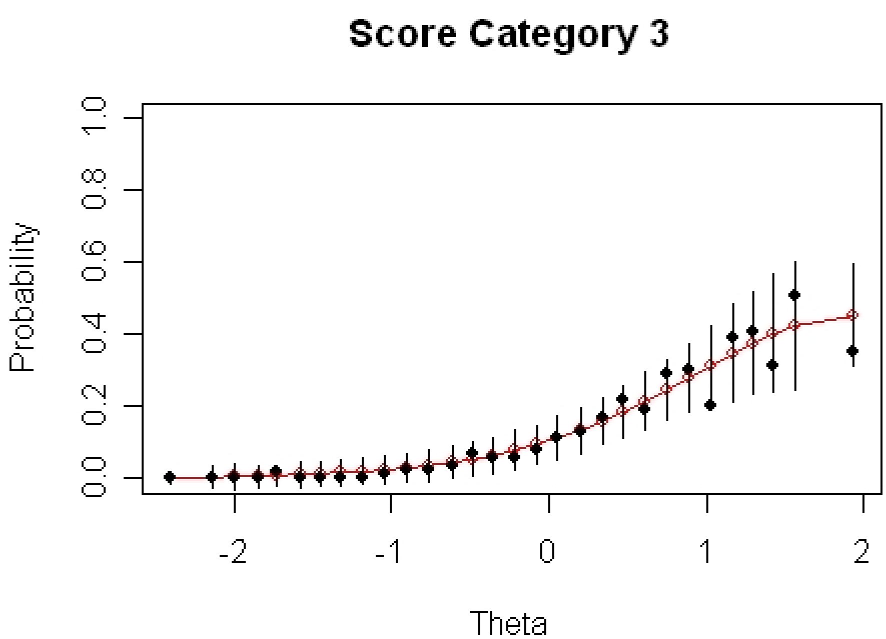
**c. Item 26**

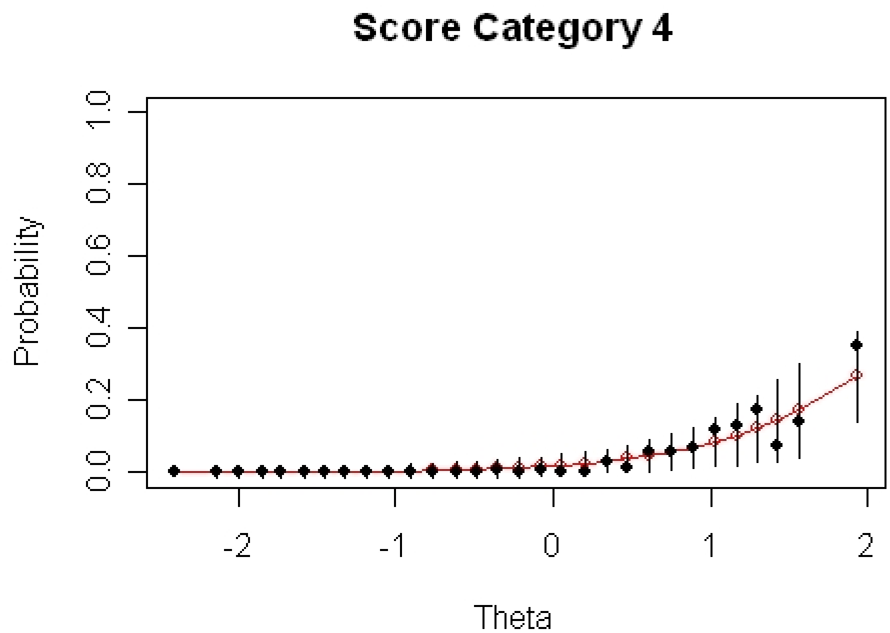
a = 0.88  b1 = -1.14  b2 = -0.21  b3 =1.30  b4 = 2.61
Chisq = 137.67  DF =90  Prob < 0.00



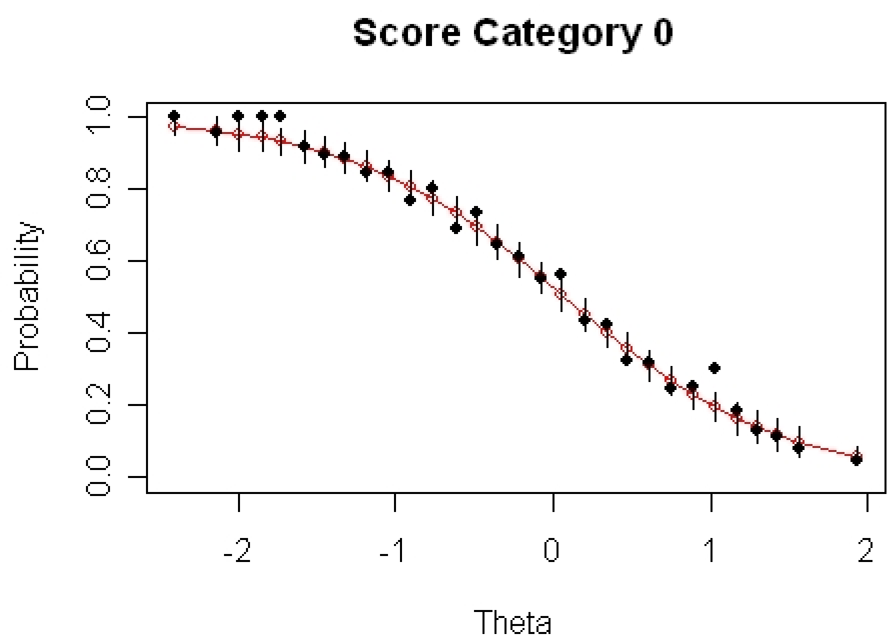

Score Category 2
Probability
1.0
0.8
0.6
0.4
0.2
0.0
-2
-1
0
1
2
Theta

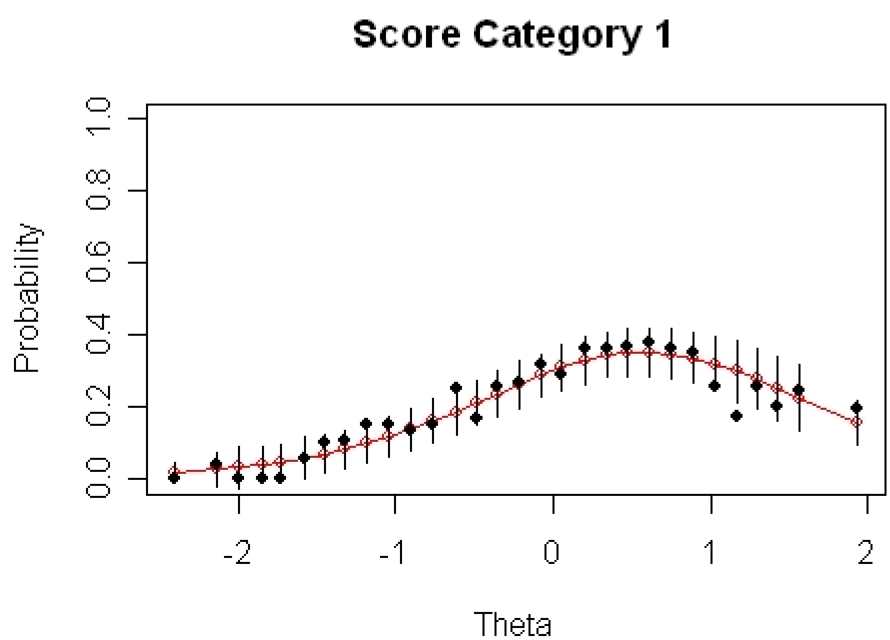

Score Category 3
Probability
1.0
0.8
0.6
0.4
0.2
0.0
-2
-1
0
1
2
Theta

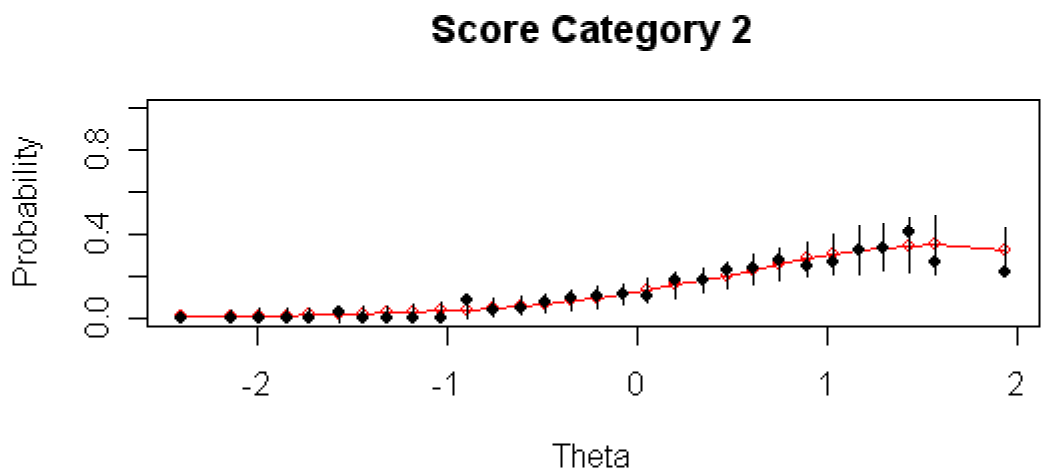

Score Category 4
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

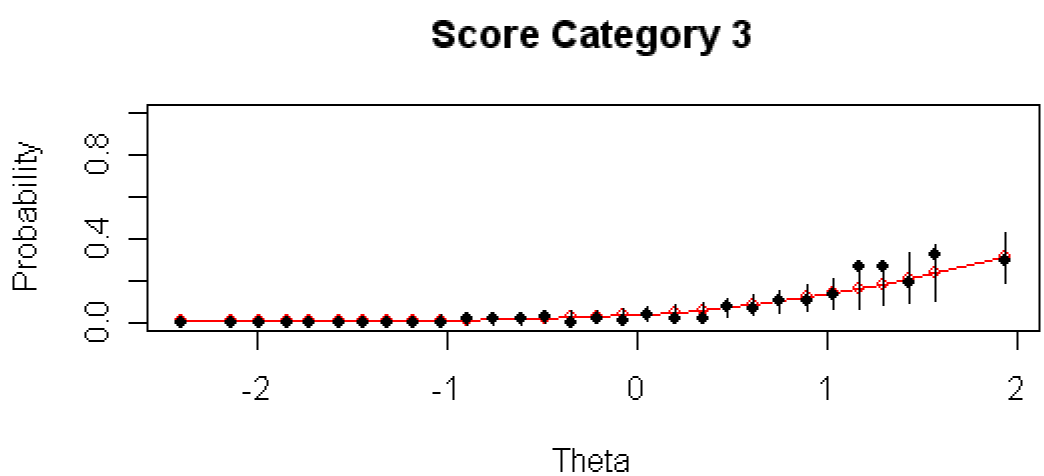
### d. Item 32

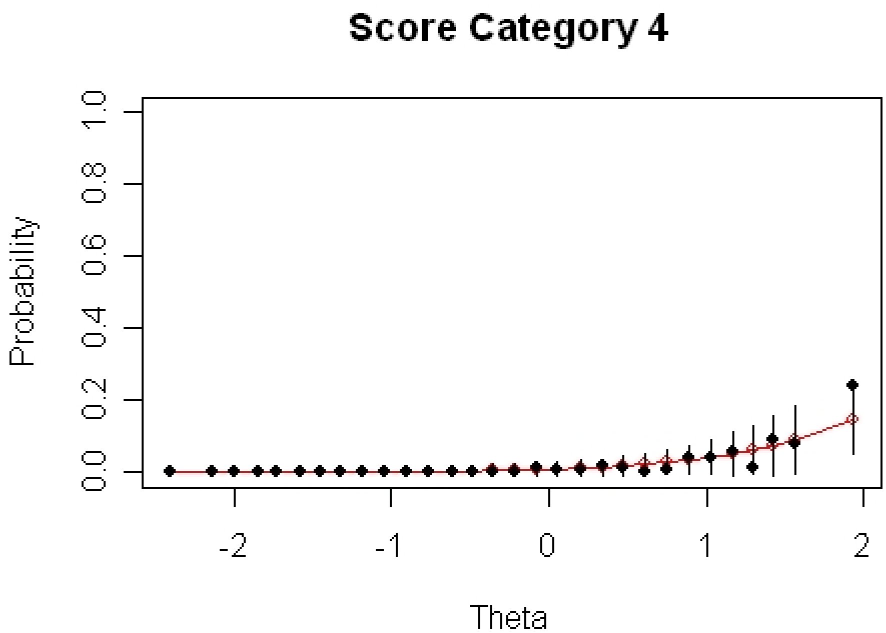
a = 0.87 b1 = 0.07 b2 = 1.08 b3 = 2.06 b4 = 3.13
Chisq = 102.87 DF =81 Prob < 0.05



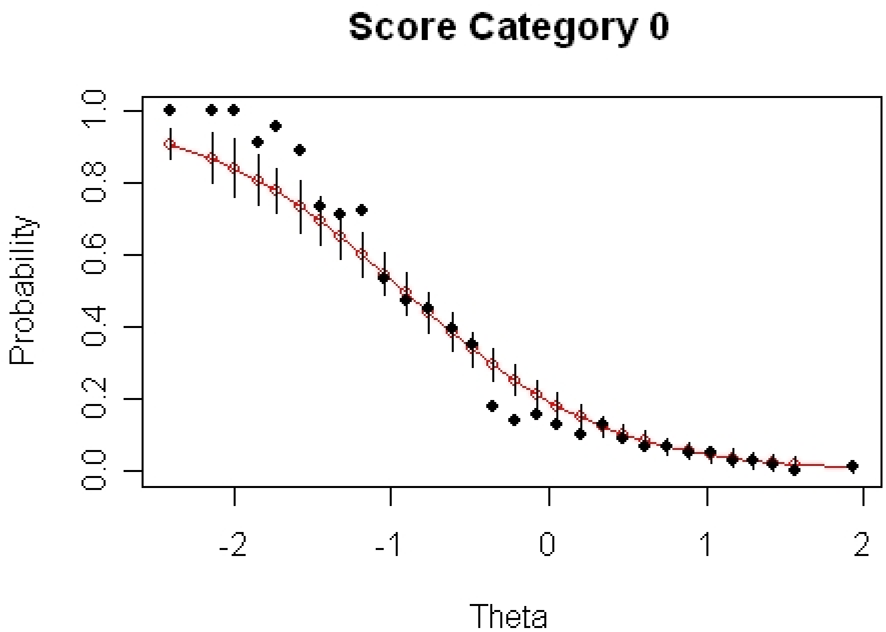

Score Category 2
Probability
0.0
0.4
0.8
-2
-1
0
1
2
Theta

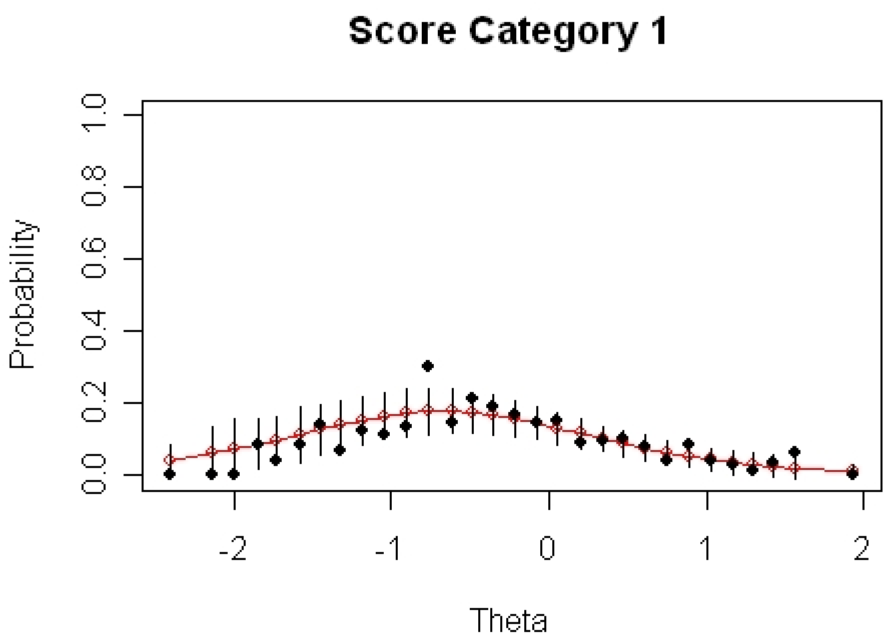

Score Category 3
Probability
0.0
0.4
0.8
-2
-1
0
1
2
Theta

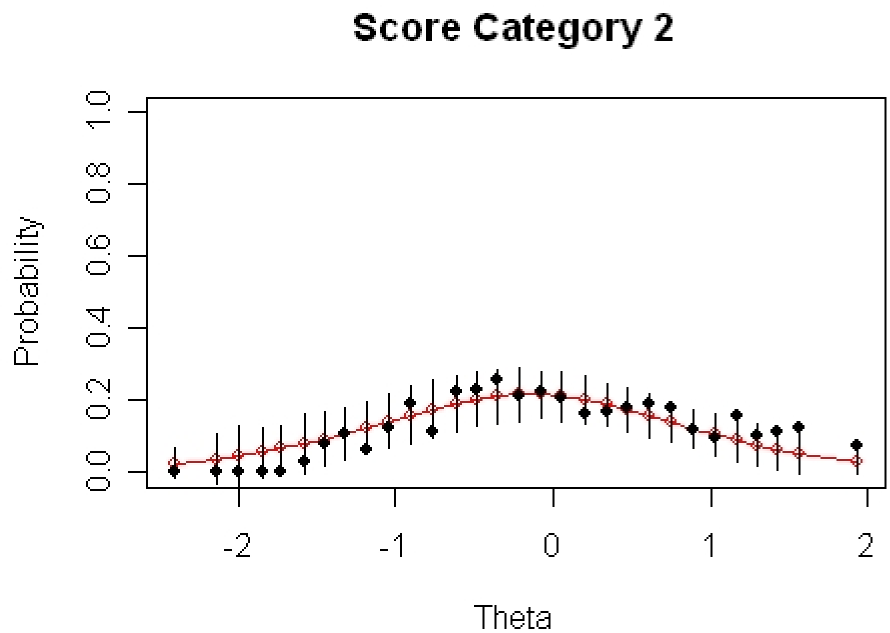

Score Category 4
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta

**e. Item 39**

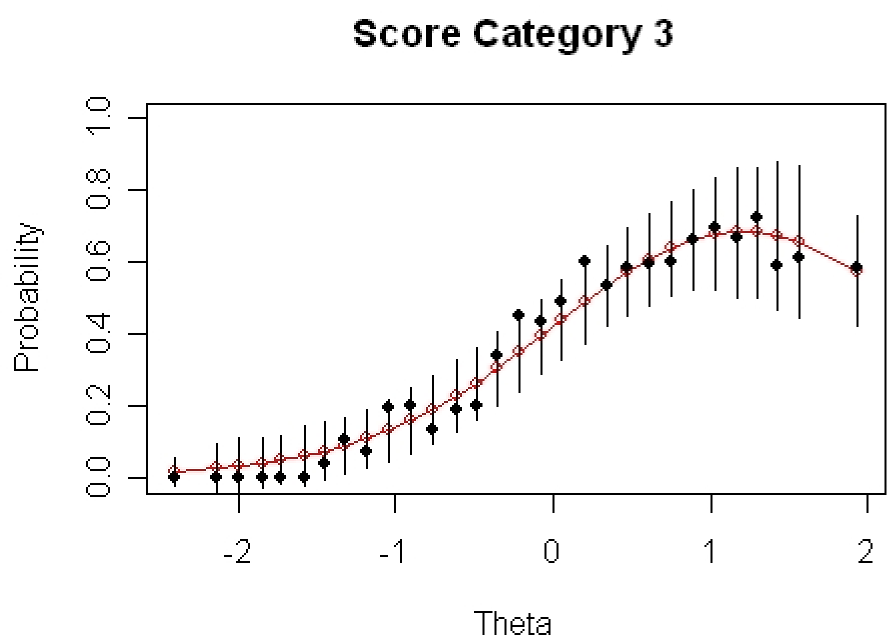
a = 0.91 b1 = -0.91 b2 = -0.45 b3 = 0.12 b4 = 2.29
Chisq = 137.22 DF =92 Prob < 0.00



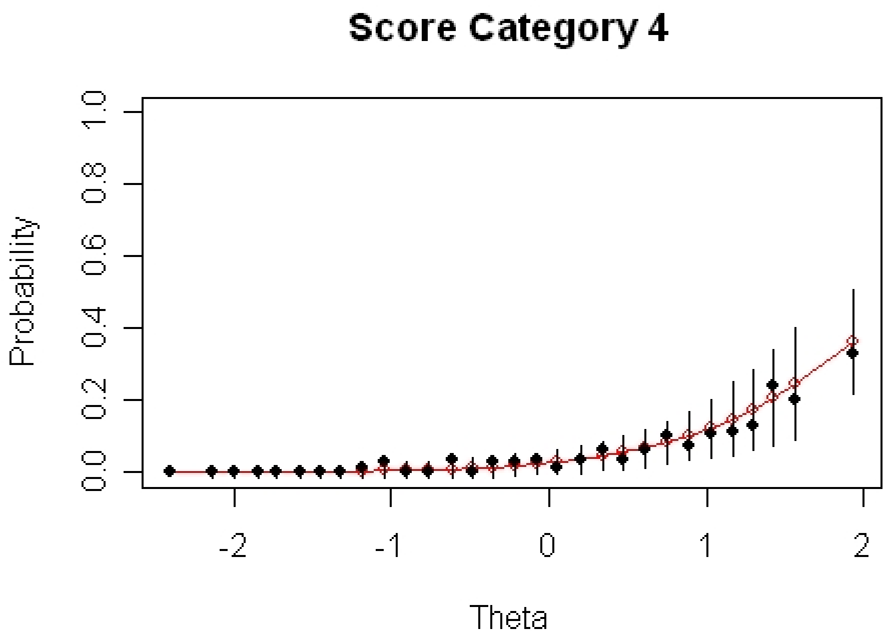





Score Category 2
Probability
1.0
0.8
0.6
0.4
0.2
0.0
-2
-1
0
1
2
Theta

Score Category 3
Probability
1.0
0.8
0.6
0.4
0.2
0.0
-2
-1
0
1
2
Theta


Score Category 4
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
Theta