Listening. Learning. Leading.®

Research Report
ETS RR–11-29

Fit of Item Response Theory Models: A Survey of Data from Several Operational Tests

Sandip Sinharay

Shelby J. Haberman

Helena Jia

July 2011

Fit of Item Response Theory Models: A Survey of Data From Several Operational Tests

Sandip Sinharay, Shelby J. Haberman, and Helena Jia
ETS, Princeton, New Jersey

July 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Matthias von Davier

Technical Reviewers: Hongwen Guo and Frederic Robin

Abstract

Standard 3.9 of the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council for Measurement in Education, 1999) demands evidence of model fit when an item response theory (IRT) model is used to make inferences from a data set. We applied two recently suggested methods for assessing goodness of fit of IRT models—generalized residual analysis (Haberman, 2009) and residual analysis for assessing item fit (Bock & Haberman, 2009)—to several operational data sets. We assessed the practical significance of misfit whenever possible. This report summarizes our findings. Though evidence of misfit of the IRT model was found for all the data sets, the misfit was not always practically significant.

Key words: generalized residual, item fit, residual analysis, two-parameter logistic model, three-parameter logistic model

Acknowledgments

We are grateful to Neil Dorans, Andreas Oranje, and Matthias von Davier for their helpful advice, to Jill Carey, Behroz Maneckshana, Anthony Giunta, Rui Gao, Kevin Larkin, Aleta Sclan, and Yi-Hsuan Lee for their help with the data, and to Ruth Greenwood for her help with copy editing.

Table of Contents

Standard 3.9 of the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) demands evidence of model fit when an item response theory (IRT) model is used to make inferences from a data set. van der Linden and Hambleton (1997) recommended the collection of a wide variety of evidence about model fit of common IRT models. However, a recent survey of operational practice of several tests that employ IRT models (Sinharay, 2008) showed that almost all of these testing programs that were surveyed use only the item fit plots and item fit statistics that are available from commercial software packages such as PARSCALE (Muraki & Bock, 1991) to evaluate model fit. To make matters worse, the item fit statistics available in commercial software packages are assumed to follow the χ^2 distribution, whereas the assumption is not based on any theoretical proof and often is incorrect. See, for example, Chon, Lee, and Dunbar (2010) and DeMars (2005), who often found high Type I error rate of the PARSCALE G^2 model fit statistic in detailed simulation studies. The increased Type I error rate for the item fit statistics available in commercial software packages results in some items being wrongly labeled as misfitting (and hence thrown out of the item pool). In addition, the judgment on model fit is subjective and depends heavily on the investigator examining the plots. Given the substantial amount of work on IRT model fit assessment in recent years (for example, see Bock & Haberman, 2009; Haberman, 2009; Hambleton & Han, 2005; Maydeu-Olivares & Joe, 2005; Sinharay, 2005; Sinharay, Johnson, & Stern, 2006; Swaminathan, Hambleton, & Rogers, 2006), to ensure continual improvement in the operational practice, it is important to apply more model fit assessment methods, especially those that were recently suggested, to operational testing.

In this paper, we consider two types of recently suggested methodologies for assessing fit of IRT models: generalized residual analysis (Haberman, 2009) and residual analysis to assess item fit (Bock & Haberman, 2009). We apply these methodologies to data from the following operational test forms:

- One form of a test of basic skills,
- Two forms each of the reading, writing, and mathematics tests belonging to a series of tests of basic skills,
- Two forms each from the verbal and quantitative parts of pretest sections of a graduate admissions test,

- One form of a computer-based science test,
- Two forms each of two sections of an test of English proficiency,
- One form each of two subjects from a battery of examinations,
- Two forms of one subject of a state test.

Methodologies

Here, we describe the two types of analysis of fit of IRT models employed in this paper: generalized residual analysis (Haberman, 2009) and residual analysis to assess item fit (Bock & Haberman, 2009).

Generalized Residual Analysis

Haberman (2009) suggested the use of generalized residuals to assess the fit of IRT models. Let x denote a possible response pattern in the data set to which an IRT model has been fitted. For example, x could be (0,1,1,...1) for a test with dichotomous items.

Suppose

$$p(x) = \text{probability of observing the response } x,$$

and

$$n(x) = \text{the number of examinees in the sample whose response pattern is } x.$$

An empirical estimate of $p(x)$ is given by

$$\hat{p}(x) = \frac{n(x)}{n},$$

where n is the sample size.

Let τ denote the set of possible values of x. Let $d(x)$ be a real-valued function. A test statistic O is computed as

$$O = \sum_{x \in \tau} d(x)\hat{p}(x)\cdot \tag{1}$$

Then its estimated mean $\hat{E}(O)$ and estimated standard deviation $\hat{s}(O)$ is computed under the assumption that the IRT model fitted to the data set is the true model. The next step is to

compute a generalized residual

$$g = \frac{O - \hat{E}(O)}{\hat{s}(O)}. \tag{2}$$

If the fitted IRT model provides a good fit to the data and the sample is large, the distribution of g is well approximated by the standard normal distribution (Haberman, 2009). Thus, a statistically significant value of g indicates that the IRT model does not adequately predict the statistic O and hence shows evidence of misfit of the IRT model to the data. The method is quite flexible. Any common data summary such as the item proportion correct, proportion simultaneously correct for a pair of items, and observed score distribution can be expressed as the statistic O by defining $d(x)$ appropriately. For example, assume that the first item of a test is dichotomous. If one defines

$$d(x) = \begin{cases} 1 \text{ whenever there is a 1 in the first component of } x \\ 0 \text{ otherwise} \end{cases},$$

then O of Equation 1 becomes the proportion correct for Item 1. Haberman (2009) used generalized residuals to assess the extent to which an IRT model predicted the observed score distribution of an operational data set.

We applied the generalized residual methodology of Haberman (2009) to assess the fit of unidimensional IRT models to several operational data sets. To do this, we fitted the IRT model used by each test (and other models that might fit the data better) and then defined O appropriately to assess the extent to which the IRT model predicts several simple data summaries such as item proportion correct, proportion simultaneously correct for pairs of items, and observed score distribution. If the IRT model fits the data sets, then we should see relatively few statistically significant residuals in our analysis of generalized residuals. Strong evidence of misfit, for example, many more than 5% statistically significant generalized residuals at the 5% level, will raise questions about the appropriateness of the IRT model. The software developed by Haberman (2009) was used to perform the computations for the generalized residuals. Missing or omitted responses were treated as wrong answers. We believe that this assumption does not affect the conclusions on model misfit because of a small proportion of missing and omitted responses (less than 1%) for the data analyzed.

Residual Analysis for Assessing Item Fit

Bock and Haberman (2009) applied a form of residual analysis to assess item fit. This residual

analysis is based on a comparison of two approaches to estimation of the item-response function. The approach leads to residuals that may be standardized to have approximate standard normal distributions in large samples. Suppose $\hat{I}_j(\theta)$ is the estimated item characteristic curve of an item. For example, for the two-parameter logistic (2PL) model,

$$\hat{I}_j(\theta) = \frac{\exp[\hat{a}_j(\theta - \hat{b}_j)]}{1 + \exp[\hat{a}_j(\theta - \hat{b}_j)]},$$

where $\hat{a}_j$ and $\hat{b}_j$ are the estimated discrimination and difficulty parameters, respectively.

The assessment of item fit of Bock and Haberman (2009) involves the value of

$$t_j(\theta) = \frac{\bar{I}_j(\theta) - \hat{I}_j(\theta)}{s_j(\theta)}, \tag{3}$$

where $t_j(\theta)$ is the residual at an examinee ability level θ, $\bar{I}_j(\theta)$ is an unconditional estimate of the probability of a correct response by an individual with latent ability θ and is a weighted average of the responses of the examinees to Item j, with the weights being proportional to the estimated posterior distribution of the examinee ability, and $s_j(\theta)$ is the estimated standard deviation of $[\hat{I}_j(\theta) - \bar{I}_j(\theta)]$. If the model is a good fit to the data, then $t_j(\theta)$ follows an approximate standard normal distribution conditional on θ. If the model does not fit the data and the sample is large, then many residuals will be significantly larger or smaller than can be reasonably expected based on the standard normal distribution.

To apply residual analysis to assess item fit (Bock & Haberman, 2009) in the operational data sets, we fitted the IRT model used in the test (and other IRT models that might fit the data better) and then assessed item fit for each item in the data by plotting the above-mentioned residuals for 31 equally-spaced values of θ between -3 and 3. We also created plots of item fit that show the values of $\hat{I}_j(\theta)$ from Equation 3 for an item as a dotted line and the values of $\bar{I}_j(\theta) - 2s_j(\theta)$ and $\bar{I}_j(\theta) + 2s_j(\theta)$ as solid lines. A dot outside the band formed by the two solid lines indicates a statistically significant residual. If the IRT model fits the data set, then we should not see much evidence of item misfit in our analysis of item fit. Substantial item misfit in the form of too many significant residuals raises questions about the appropriateness of the IRT model. We judged that there is a substantial misfit for an item when (a) the number of statistically significant residuals between $\theta = -2$ and $\theta = 2$ is five or more, and (b) at least five of those statistically significant residuals are associated with a difference of at least 0.01 between $\hat{I}_j(\theta)$ and $\bar{I}_j(\theta)$. [1] In our item fit analyses, the missing or omitted responses were assumed to be wrong answers. We

believe that this assumption does not affect the conclusions on model misfit because of the small proportion of missing and omitted responses for the data analyzed.

Why We Chose These Methods

We used the techniques suggested by Haberman (2009) and Bock and Haberman (2009) ahead of the several IRT model fit techniques in the literature (for a review of them, see, for example, Swaminathan et al., 2006). There are several reasons for this choice:

- There is the intuitive appeal that both of these techniques are forms of residual analysis.
- These techniques have a solid theoretical basis. Each of the residuals g and $t_j(\theta)$ has a distribution with an approximate standard normal distribution if the IRT model fits the data, where the approximation becomes arbitrarily accurate as the sample size becomes increasingly large.
- The computation involved in these techniques does not involve any recursion and hence takes little time once the item parameters have been estimated.
- Use of generalized residuals provides a framework to assess several aspects of misfit of the IRT model.

We did not use the statistics for fit of IRT models that are available in popular software packages. Hambleton and Han (2005) commented that many χ^2 statistics for item fit commonly employed in many software packages have well-known shortcomings such as uncertain sampling distribution. In addition, researchers such as Chon et al. (2010) demonstrated the limitations of such statistics.

We performed some limited simulation studies to study the properties of the model fit techniques suggested by Haberman (2009) and Bock and Haberman (2009) that we would use later. We found that both these set of techniques have satisfactory Type I error rate (of around 5% at the 5% level) and power.

Evaluating Practical Consequences of Misfit

The famous statistician George E. P. Box said that all models are wrong but some are useful. If we find an item response data set that has a sufficient number of examinees, no IRT model

will fit the data. However, an IRT model that shows misfit can still be useful. Therefore, several researchers such as Hambleton and Han (2005) and Sinharay (2005) recommended the evaluation of whether the IRT model misfit found is of practical significance. However, other than Sinharay (2005) and Lu and Smith (2007), there are few examples of evaluation of practical significance of IRT model misfit. If the misfit is not practically significant, then the IRT model is useful despite its limitations. In this paper, we evaluated the practical significance of misfit whenever possible.

A Test of Basic Skills

We had the responses of 8,686 examinees to a separately timed section with 45 five-option multiple-choice (MC) items in the writing assessment part of a basic skills test. The first 25 items are on finding errors while the last 20 are on correcting errors. Experts believed that the test was speeded. The test of basic skills used both paper-and-pencil and computerized forms; the computerized forms, which were very similar to the paper-and-pencil forms, used the 3PL model to equate the scores. This particular form is a paper-and-pencil form so that no IRT model was operationally used on this. Sinharay (2005) applied a Bayesian model checking method to find some evidence of misfit of the three-parameter logistic (3PL) model.

We fitted the 2PL model to the data and then used the framework of generalized residuals to define $d(x)$ in such a way that O of Equation 1 is equal to p_{11}, the second-order marginal, that is, the proportion of examinees who answer a pair of items correctly. If, for example, the data are multidimensional (due to, for example, the items on finding errors measuring a different skill from those on correcting errors), then the generalized residuals will be positive for pairs of items that measure the same trait (because of high correlation between such item pairs) and negative for pairs of items that measure different traits.

As an example, the value of p_{11} for the item pair $\{1, 2\}$ is 0.136, the corresponding estimated expected value is 0.149, the corresponding standard deviation of the difference is 0.0025—that results in a generalized residual of -5.24. Figure 1 summarizes the generalized residuals for the statistic p_{11} for all item pairs.

For any pair of items, a plus sign in the plot indicates that the generalized residual corresponding to p_{11} for the item pair is positive and statistically significant at the 5% level, so that the number of examinees who answered the pair correctly, or, the association for the pair, is more than that predicted by the 2PL model. A minus sign denotes that the corresponding residual

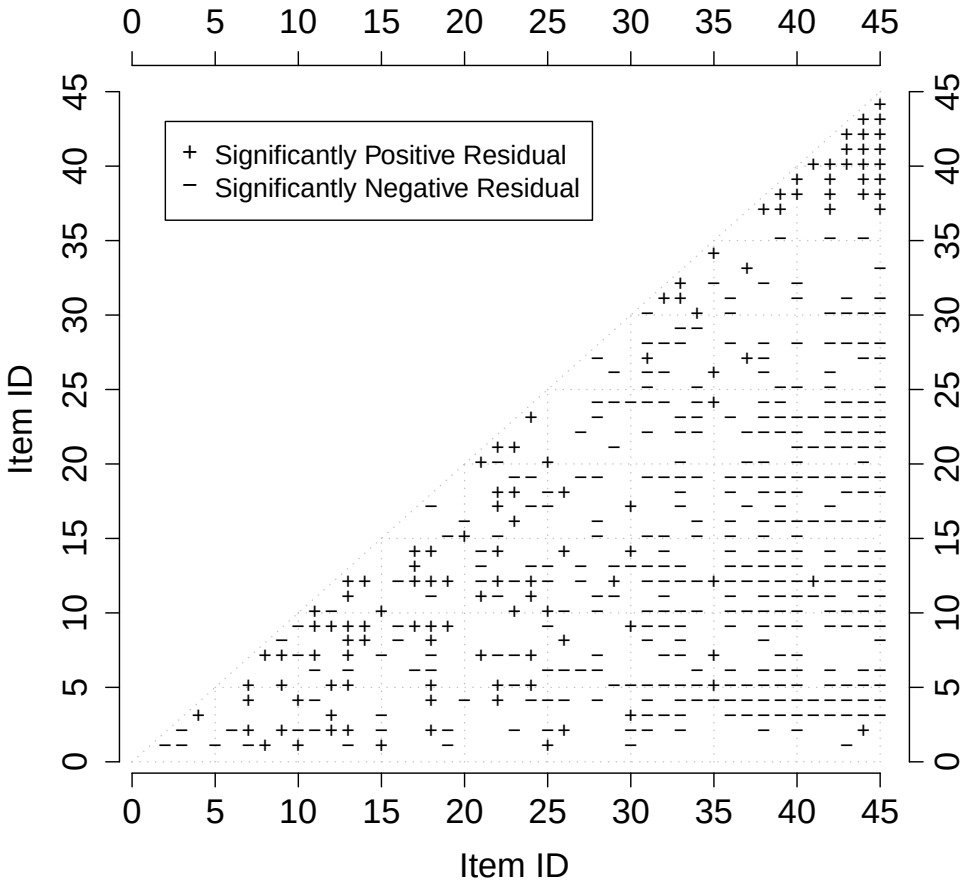

***Figure* 1. The fit of the IRT model to the second-order marginal totals for the test of basic skills.**

is negative and statistically significant so that the number of examinees who answered the pair correctly is less than that predicted by the 2PL model. No symbol for a pair of items denotes a nonsignificant p-value, which indicates that the 2PL model adequately predicts the association for the pair. The plot shows that overall, the 2PL model cannot adequately predict the association among the items. The second-order marginal totals are higher than the model predicts for a large number of pairs involving the last nine items (37–45) of the test. The generalized residuals for the last few items are very large. For example, for the item pair {44,45}, the value of p_{11} is 0.28, the corresponding expected proportion is 0.23, and the generalized residual g is 22.6. In addition, clearly visible are the large number of negative generalized residuals for pairs involving one item from the group numbered from 38 to 45 and another from the group numbered from 1 to 25. Thus, Items 38 to 45 seem to load on a dimension different (in the words of, e.g., Stout, 1987) from that measured by Items 1 to 37. It is highly likely that this phenomenon is caused by speededness in the test, especially because Items 38 to 45 are discrete/stand-alone items on correcting errors, just like Items 26 to 37, and a study of their content does not suggest any unexpected connection among them.

Evaluating Practical Consequences of Misfit

Though Figure 1 shows statistically significant model violation that may be explained by speededness, it does not indicate whether the amount of speededness is practically significant; hence, there is a need for further analysis.

Boughton, Larkin, and Yamamoto (2004, April) thoroughly examined the existence of speededness in a number of tests, including this particular test (the test considered in this paper is referred to as PPW1). They applied the HYBRID model (Yamamoto, 1989), which assumes that subsets of examinee response patterns are described by a discrete latent class model, while the remaining responses satisfy an IRT model. Boughton et al. (2004, April) found that, for the test considered here, more than 20% of the examinees switched to a random response pattern by Item 38, which, according to their opinion, was proof of substantial speededness. In the end, a decision was made to reduce the length of the test by seven items in future administrations.

In addition, we computed estimated abilities (posterior means) using the 2PL model twice, once from all the 45 items and once more from the first 38 items. Figure 2 compares these two sets of estimates. The top panel shows a bivariate scatter-plot of the ability estimate based on the first

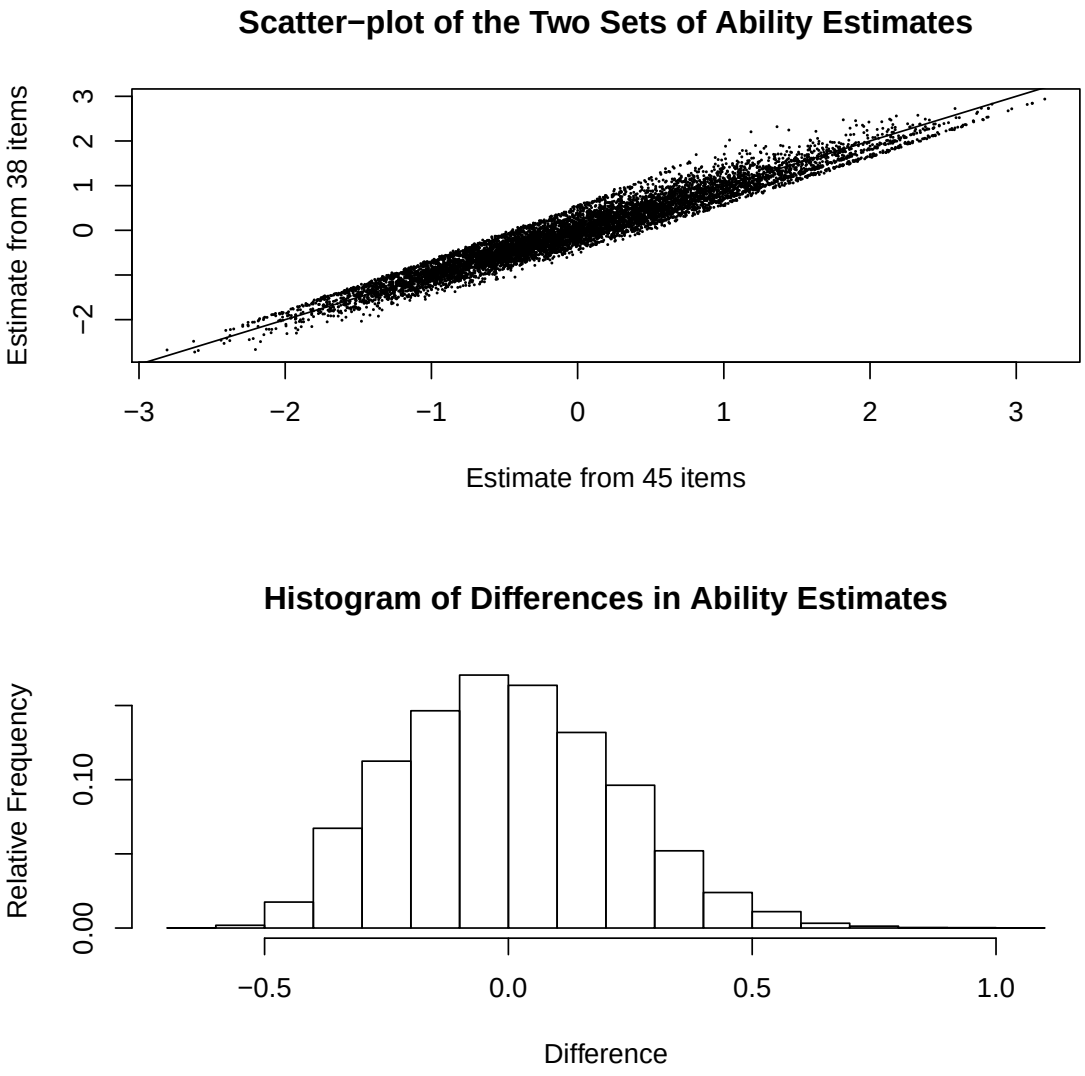

***Figure* 2. A comparison of the ability estimates based on the first 38 items and those based on all 45 items.**

38 items versus the ability estimate based on all the 45 items. The panel also shows a 45-degree line for convenience; if there is no speededness, we would expect the points in the scatter-plot to fall evenly around this line. The bottom panel plots the differences of these ability estimates. If there is no speededness, the differences will be distributed evenly around zero. However, there are more points above the 45-degree line than below it in the top panel and the right tail of the histogram is longer than the left tail in the bottom panel—these support the finding of Boughton et al. (2004, April) that there may be speededness in the data.

Therefore, the model misfit suggested by the generalized residual analysis in Figure 1 was of substantial practical consequence.

More Tests of Basic Skills

We had data from two forms each of the reading, writing, and mathematics tests belonging to a series of tests of basic skills. The reading and mathematics tests have 40 multiple choice items each. The writing test has 38 multiple choice items; this test also has an essay, but no IRT model is applied to the essay score. These forms are computer-delivered but not adaptive. An examinee can take the test on any day he/she wants and is randomly administered one of the many operational forms. The sample sizes for these forms were between 2,500 to 3,000. Operationally, the 3PL model is used to perform an equating of a new form to a base form. However, there are no common items between the new form and the base form. There is a large pool of items that are all calibrated on the same scale using the Stocking-Lord algorithm (Kolen & Brennan, 2004). A new form is equated to the base form using IRT true score equating.

We fitted the 3PL model to each of these data sets. However, the parameters of the model were not well-identified. This was revealed from a comparison of the standard error estimates from the stabilized Newton-Raphson algorithm (Haberman, 1988) used to fit the model and standard error estimates computed using the formula of Louis (1982). Large differences between the two estimates normally reflect near-singularity of the Hessian matrix used by the stabilized Newton-Raphson algorithm. This issue normally suggests poor identification of model parameters. In such a case, the normal approximations required in residual analysis cannot be regarded as trustworthy.

Therefore, we fitted a restricted version of the 3PL model where we forced all the guessing parameters to be the same—this common guessing parameter was estimated from the data.

The identifiability problems we observed for the unrestricted 3PL model were not seen for the restricted 3PL model—so we report results for this model in the remainder of this section.

Figure 3 shows the histograms for the generalized residuals when O of Equation 1 is the proportion correct for an item. In the figure, two vertical dashed lines constitute a 95% confidence interval under perfect model fit—any generalized residual outside this interval is statistically significant at the 5% level. The residuals are mostly quite large. Except the second reading form, all the forms had more than 50% of the generalized residuals statistically significant and negative. However, this result reflects the very small standard error of the difference between the observed and expected proportions correct. Given this very small standard error, the large magnitude of the generalized residual reflects small errors in approximation of maximum-likelihood estimates by an iterative algorithm that must stop at some point. Histograms for the differences between the observed and expected proportions correct for the six forms are shown in Figure 4. The histograms show that the differences between observed and expected proportion corrects are mostly very small.

Large generalized residuals for the proportions correct were observed for all the remaining data sets in the paper, irrespective of the IRT model used, whereas the differences between observed and expected proportions correct were always very small—so results for this statistic are not discussed further in this paper.

Figure 5 shows histograms for the generalized residuals when O of Equation 1 is the number of examinees who obtained a raw score of s, $s = 0, 1, \ldots I$, where I is the total number of items in a form. These values of O constitute the marginal score distribution. In each panel, the raw score is plotted along the X-axis and the generalized residual for each raw score is plotted along the Y-axis. For convenience, the range of the Y-axis is the same in all the panels of the figure and there are horizontal lines at the 2.5th and 97.5th percentiles of the standard normal distribution. The figure does not show much evidence of misfit of the IRT model to the marginal score distribution—few of the generalized residuals lie beyond the 2.5th or 97.5th percentiles. For the six forms together, the residual is between -49 and -55 for a total of 24 score-points—these residuals are not shown in Figure 5 in which the range of Y-axis is chosen to be between -3.2 and 3. However, these score-points are at the bottom end of the score scale (0 to 9) and the observed and expected values both are very close and close to 0 (the maximum observed count is 7) for these score-points. So these large residuals should not cause concern because very few examinees

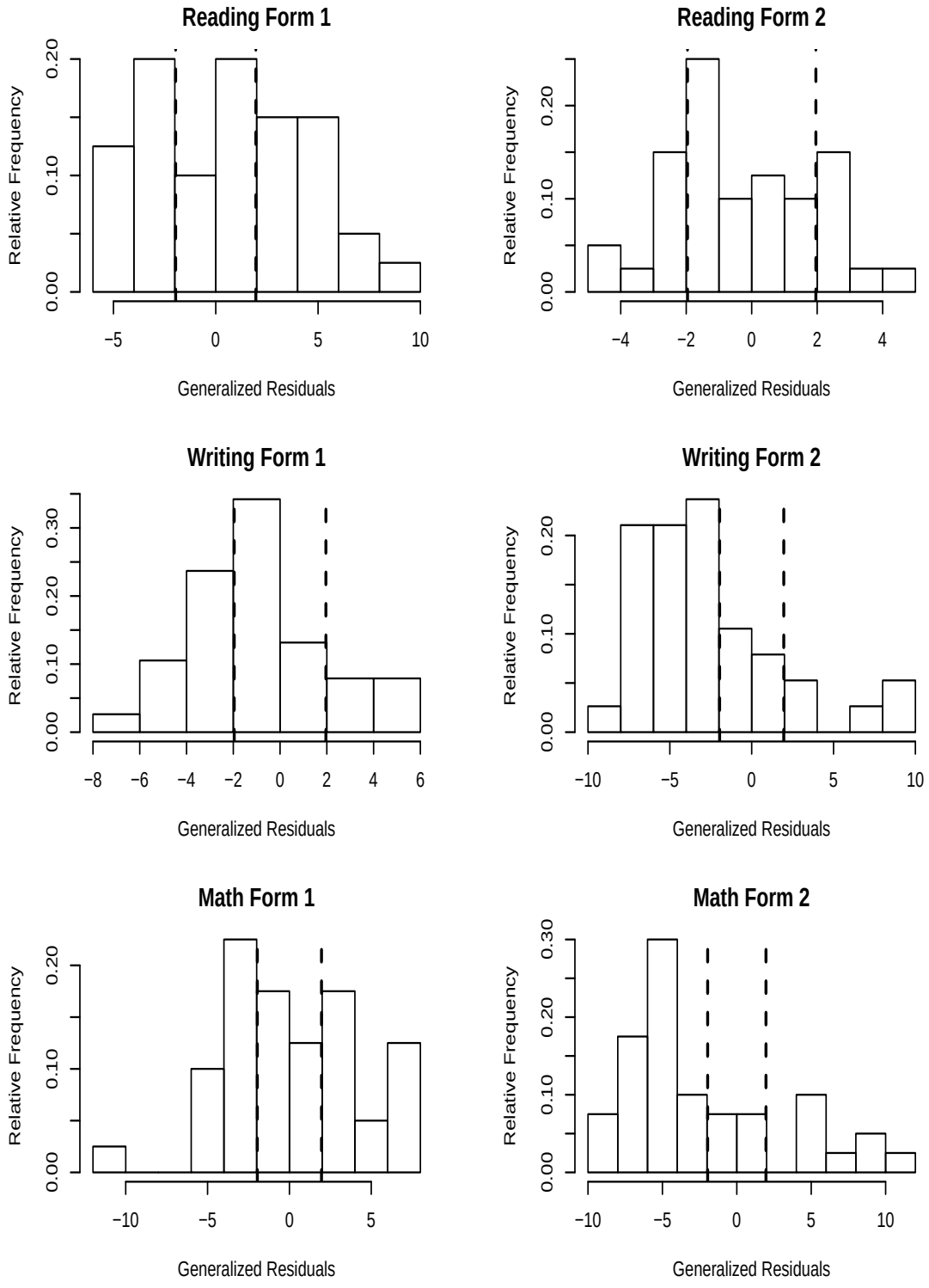

***Figure* 3. The generalized residuals for proportions correct for the six forms of the test of basic skills.**

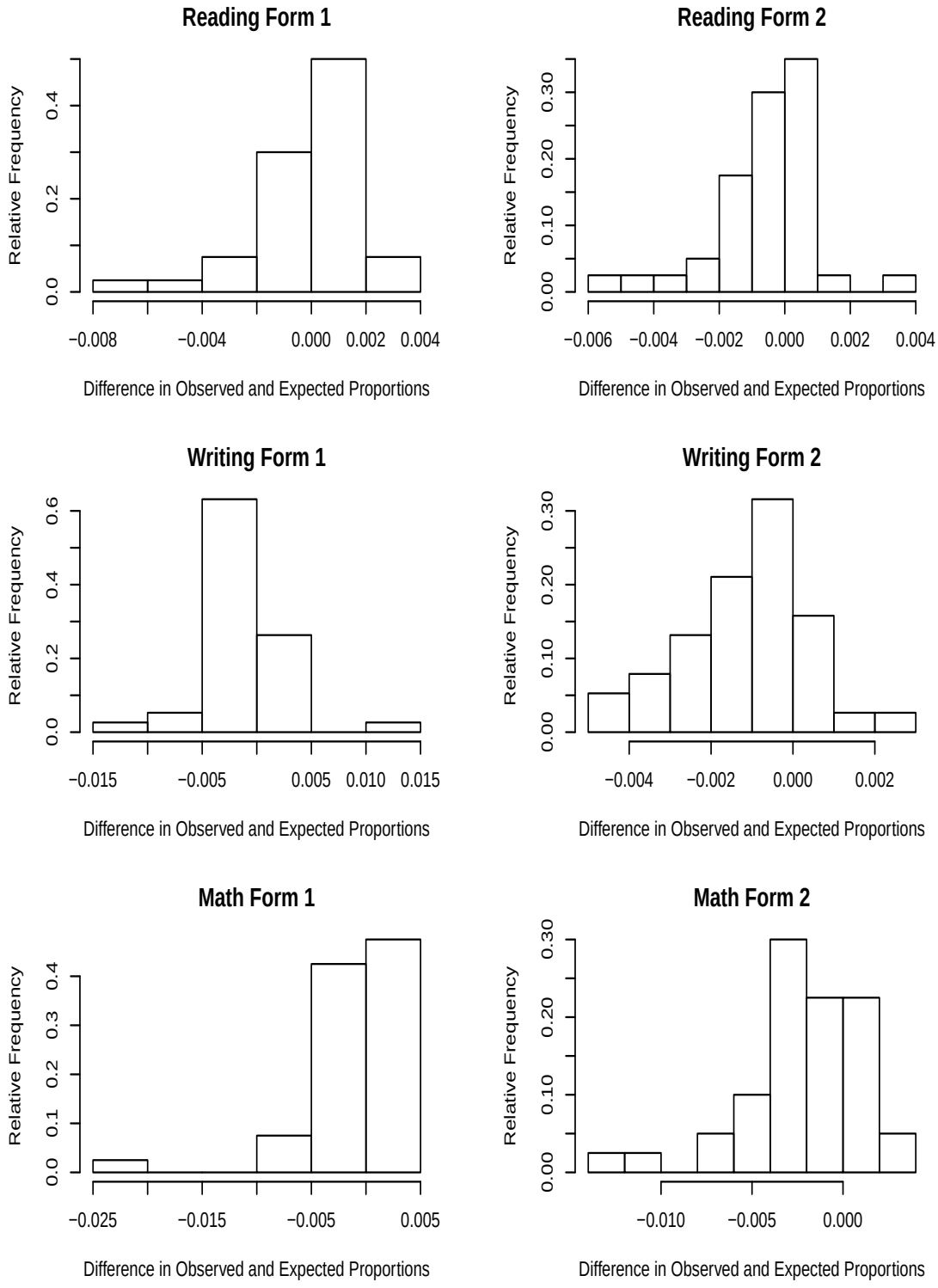

***Figure* 4. The differences between observed and expected proportions correct for the the six forms of the test of basic skills.**

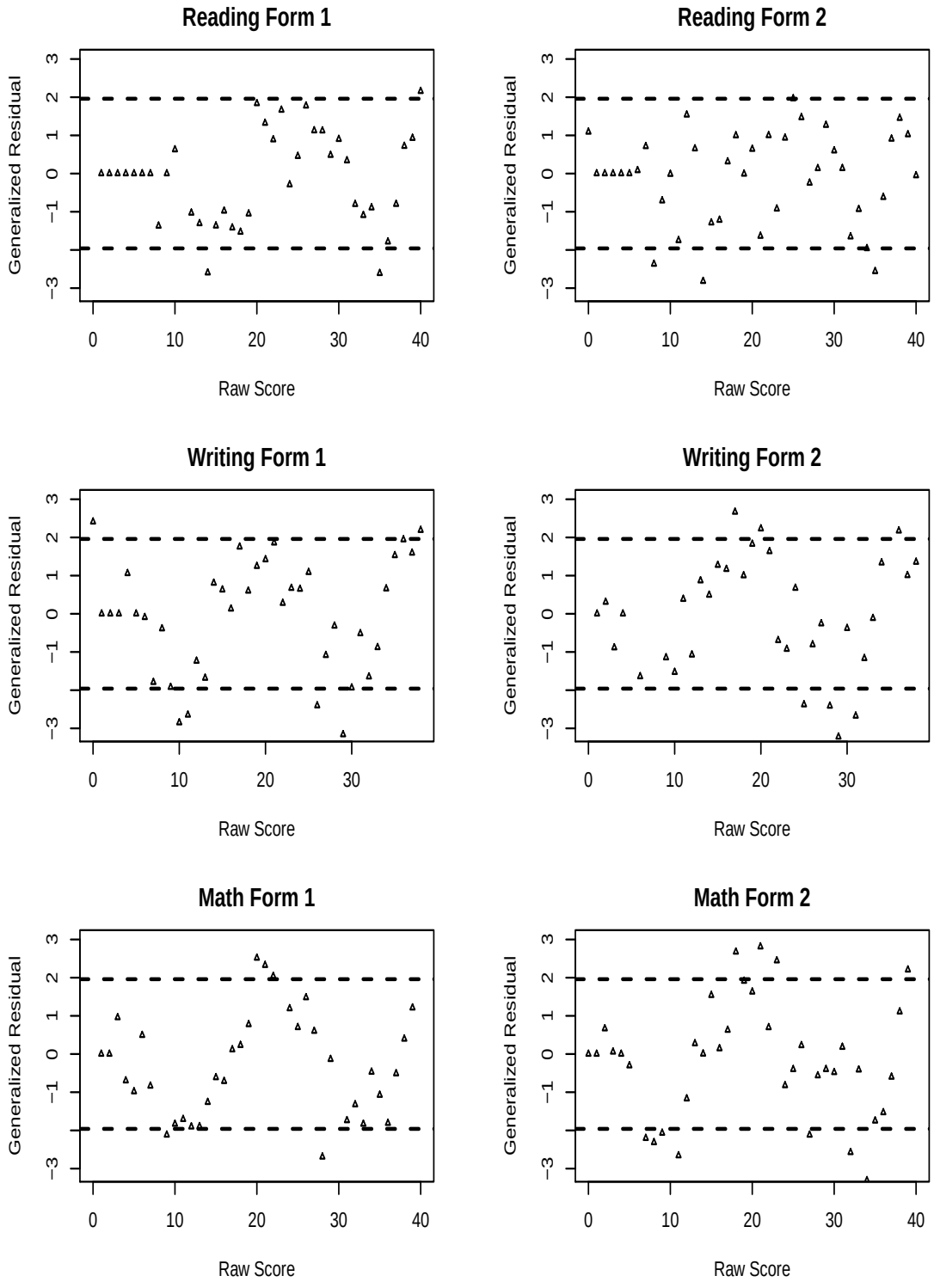

***Figure* 5. The generalized residuals for marginal score distribution for the six forms for the test of basic skills.**

score in that region.

Similar to Figure 1, each panel of Figure 6 shows the significant generalized residuals for the statistic p_{11} for all pairs of items for all six forms. The figure shows some evidence of misfit of the restricted 3PL model to the second-order marginal totals, though the evidence is not as strong as in Figure 1. There is some evidence of clustering of items into two for reading Form 1 and mathematics Form 1—there are a large number of positive and statistically significant residuals for pairs involving the first few items. The percentage of p-values that are significant at the 5% level are 18, 11, 20, 21, 29, and 32, respectively, for the six forms.

Figure 7 shows the residuals for item fit given by Equation 3 for all the items in the first reading form.

In the plot for any item, the examinee ability θ is plotted along the X-axis and the residual given by Equation 3 is plotted along the Y-axis. For convenience, two horizontal lines are shows at -1.96 and 1.96—a residual beyond these lines is statistically significant.

Consider for example the first item in Figure 7. The residual for $\theta = -4.00$ is -0.52, but the residual decreases as θ increases till the residual is -3.58 for $\theta = -2.90$. Then the residual increases as θ increases till $\theta = -1.52$ and then decreases again. Finally, after a couple of more oscillations, the residual increases from -2.24 to 0.63 as θ increases from $\theta = 0.97$ to $\theta = 4.00$.

The figure shows several large residuals for item fit for values of θ with absolute value greater than 2. For example, for Item 2 of reading Form 1, the residuals are quite large and as large as 16 for θ between {2,4}. However, there are usually very few individuals outside {-2,2}. In addition, several of these large residuals (for example, for Item 2 on reading Form 1) are associated with a small standard error in Equation 3 and with very similar values of $\hat{I}_j(\theta)$ and $\bar{I}_j(\theta)$. [2] Hence the large residuals for values of θ outside {-2,2} are not practically significant in most cases. For values of θ in {-2,2}, the residuals are rarely larger than 1.96 in absolute value. Even if they are, sometimes it is because the item is very easy; for example, for Item 2 on reading Form 1, for θ=0.69, $\hat{I}_j(\theta)$=0.99 and $\bar{I}_j(\theta)$=0.986, so that their difference is quite small; however, the residual is 2.15.

Figures 8 to 13 show the plots of item fit for only the items for which substantial misfit was found.

Overall, the analysis of item fit suggests some amount of misfit of the IRT model to these data.

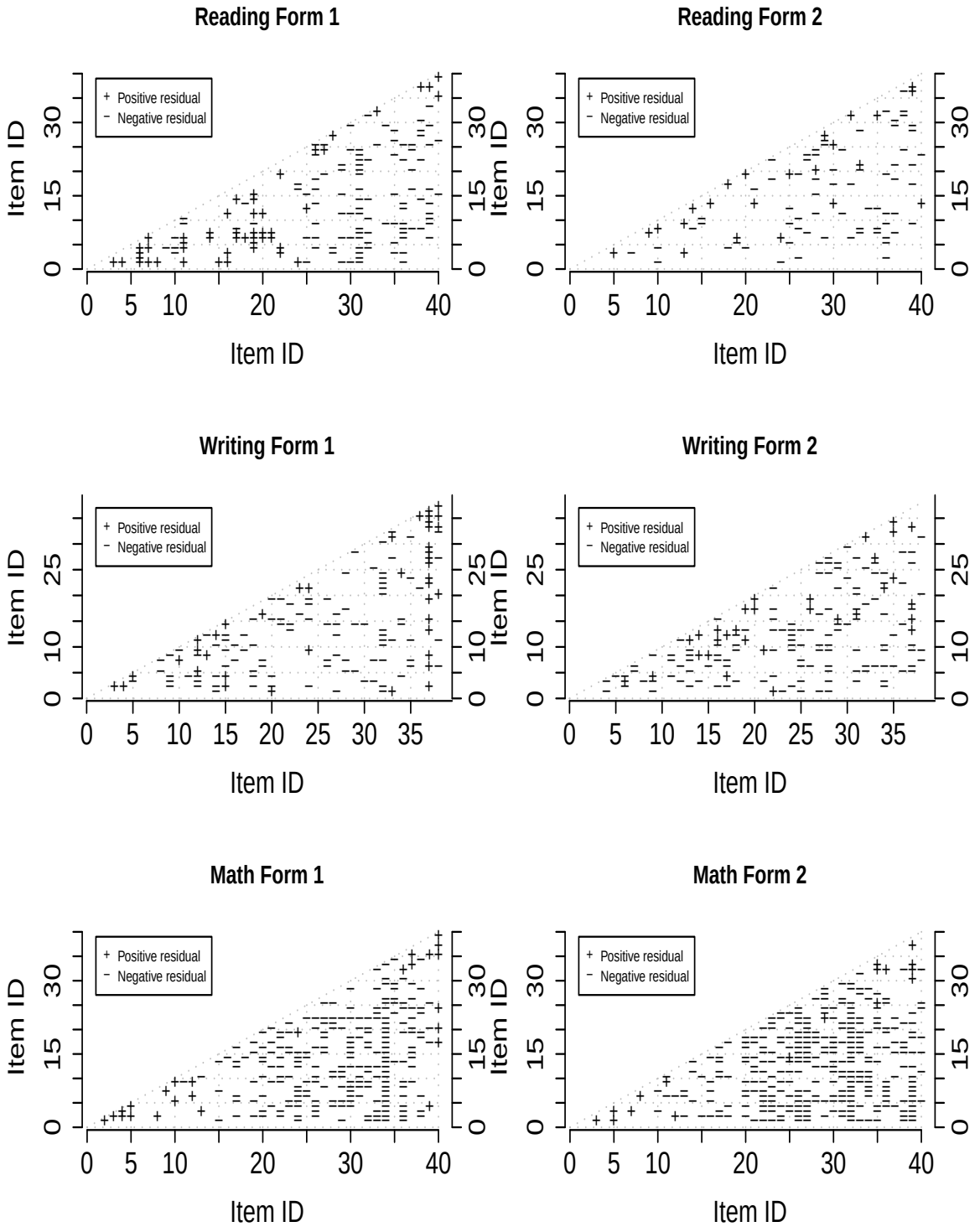

***Figure* 6. The fit of the IRT model to the second-order marginal totals for the six forms of the test of basic skills.**

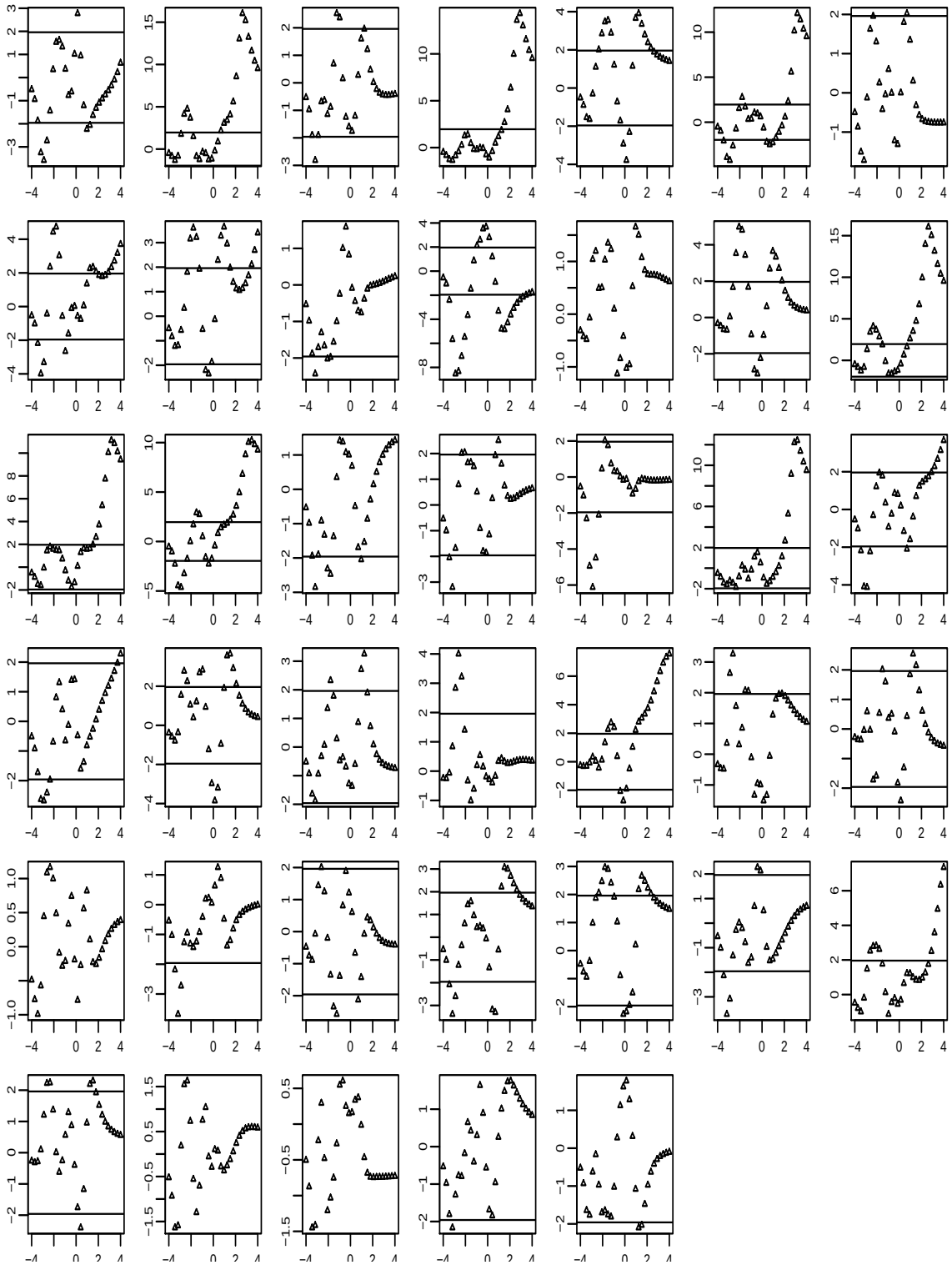

***Figure* 7. The residuals for item Fit for the test of basic skills—First reading form.**

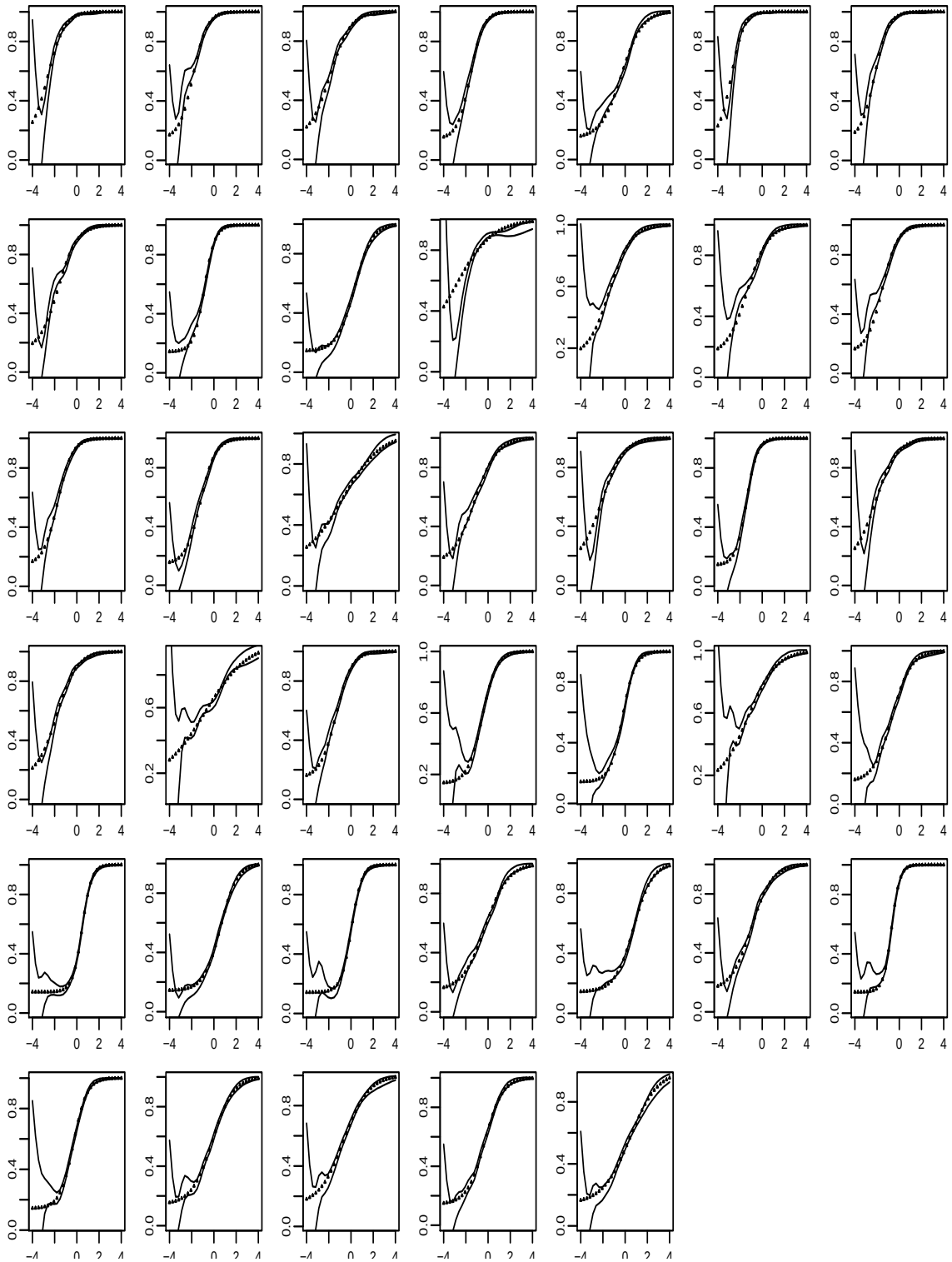

***Figure* 8. Plots of item fit for the test of basic skills—First reading form.**

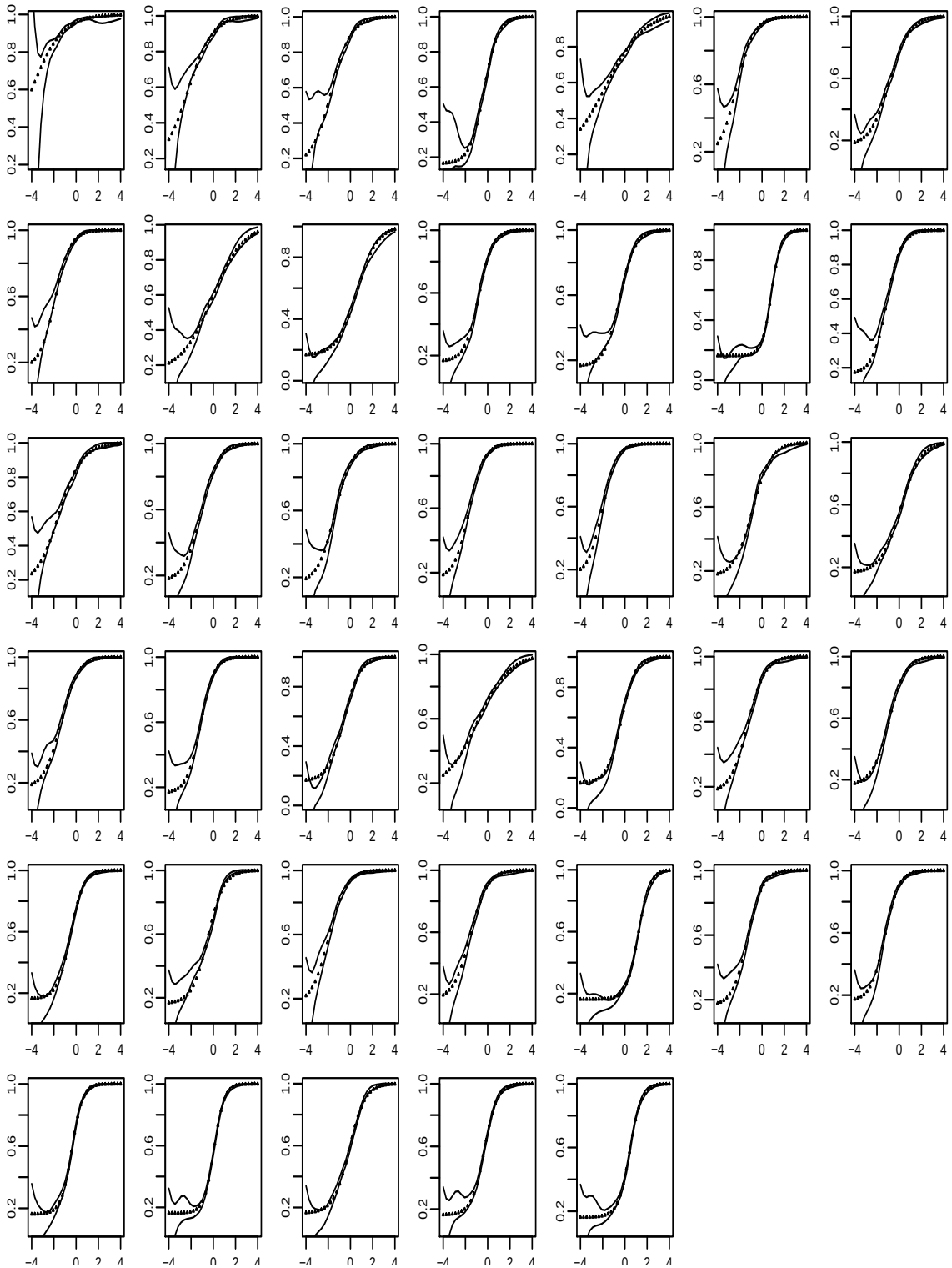

***Figure* 9. Plots of item fit for the test of basic skills—Second reading form.**

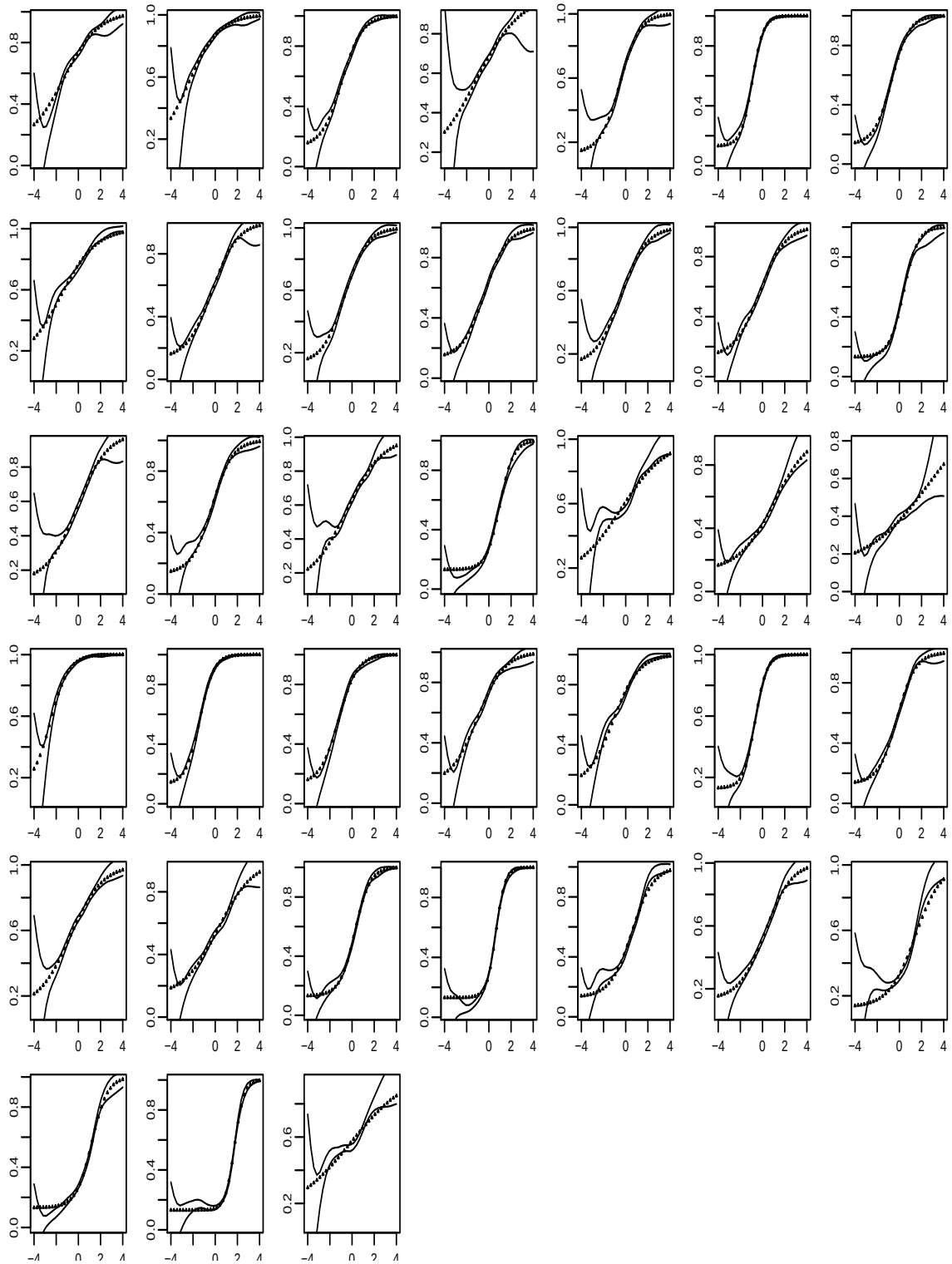

***Figure* 10. Plots of item fit for the test of basic skills—First writing form.**

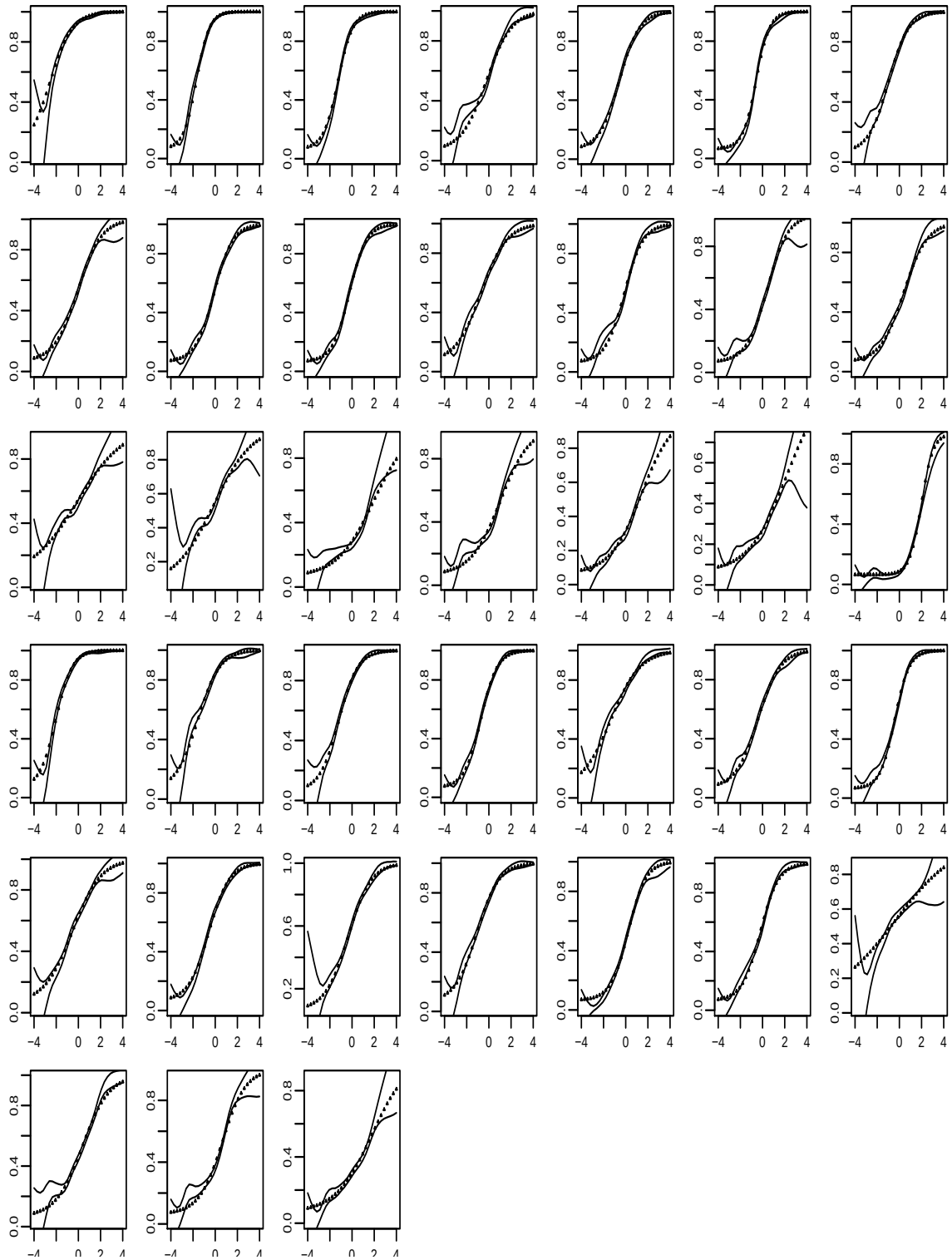

***Figure* 11. Plots of item fit for the test of basic skills—Second writing form.**

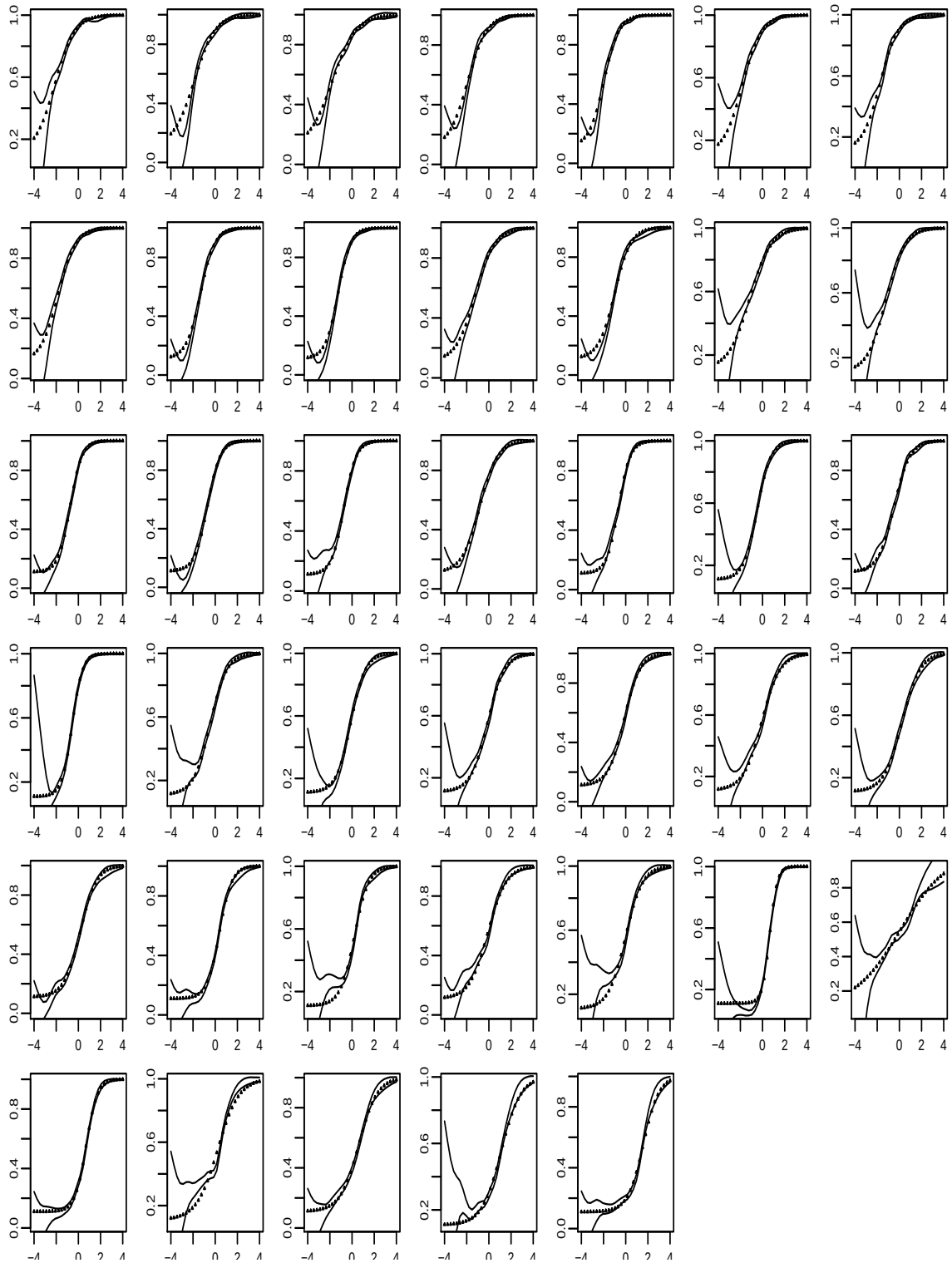

***Figure* 12. Plots of item fit for the test of basic skills—First mathematics form.**

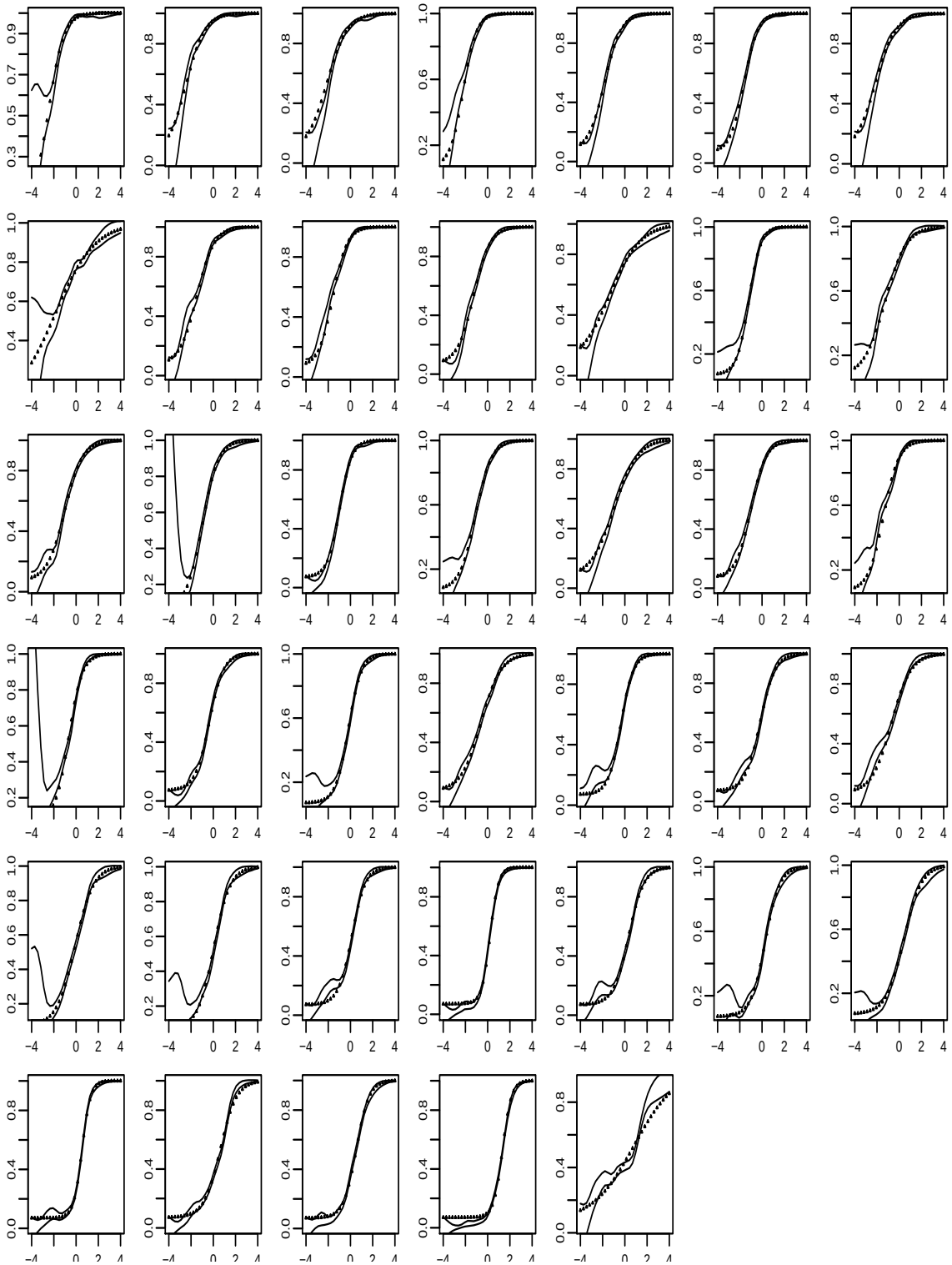

***Figure* 13. Plots of item fit for the test of basic skills—Second mathematics form.**

Evaluating Practical Consequences of Misfit

Mathematics. To evaluate practical consequences of misfit for the mathematics test of the series of tests of basic skills, we performed IRT true score equating for the first mathematics form three times to equate the raw scores on the form to those on the base form:

- Equating using all the 40 items.
- Equating after removing the 5 items that show most significant amount of misfit in Figure 12—Items 4, 32, 35, 37, and 40.
- Equating after removing 5 items that were randomly chosen from the 40 items of the form—Items 8, 19, 26, 35, and 38.

The item parameters were re-estimated in each of the above three equatings. We then calculated (a) E_1, the vector of equated scores of all the examinees from the first of the three equatings above, (b) E_2, the vector of equated scores of all the examinees from the second of the three equatings above, and E_3, the vector of equated scores of all the examinees from the third of the three equatings above. We then plotted E_1 versus E_2 on a graph (the top panel of Figure 14). We computed the correlation coefficient between E_1 and E_2. We also computed the square root of the sum of the squares of the differences between E_1 and E_2. We then plotted E_1 versus E_3 on a graph (the bottom panel of Figure 14). We computed the correlation coefficient between E_1 and E_3. We also computed the square root of the sum of the squares of the differences between E_1 and E_3. A larger association between E_1 and E_3 than between E_1 and E_2 would indicate that the misfit observed in Figure 12 has a practical significance. However, Figure 14 shows that the association between E_1 and E_3 is very similar to that between E_1 and E_2. In addition, the correlation (and rank correlation) is 0.99 between E_1 and E_3 and also between E_1 and E_2. The square root of the sum of the squares of the differences is 1.01 for E_1 and E_2 and 0.99 for E_1 and E_3. Thus the misfit of the items of Form 1 of the mathematics test of the series of tests of basic skills does not seem to have much practical significance.

Writing. To evaluate practical consequences of misfit for the writing test of the series of tests of basic skills, we performed IRT true score equating for the first writing form three times to equate the raw scores on the form to those on the base form:

- Equating using all the 38 items in the form.

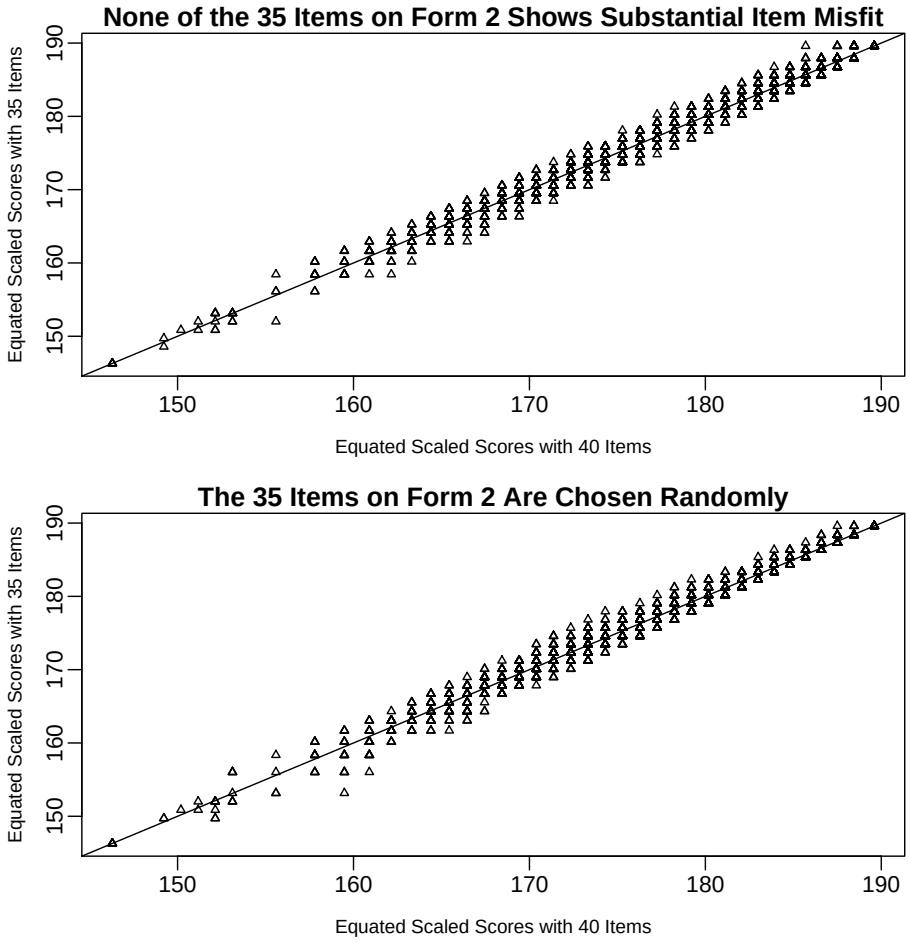

***Figure* 14. Evaluating practical consequences of misfit for Form 1 of the mathematics test from the series of tests of basic skills.**

- Equating after removing the 5 items that show most significant amount of misfit in Figure 10—Items 8, 19, 26, 35, and 38.

- Equating after removing 5 items that were randomly chosen from the 40 items of the form—Items 4, 10, 21, 32, and 38.

The item parameters were re-estimated in each of the three above equatings. We then calculate (a) E_1, the vector of equated scores of all the examinees from the first of the three equatings above, (b) E_2, the vector of equated scores of all the examinees from the second of the three equatings above, and E_3, the vector of equated scores of all the examinees from the third of the three equatings above. We then plotted E_1 versus E_2 on a graph (the top panel of Figure 15). Note that the writing test has a constructed response (CR) component as well, but the equating is performed through the MC items only. The outcome is an equating transformation of the composite of the MC and CR score on a form to the operational score scale. That is why the X-axis of the above figure ranges between 0 and 76 (even though there are only 38 MC items on the writing test).

We computed the correlation coefficient between E_1 and E_2. We also computed the square root of the sum of the squares of the differences between E_1 and E_2. We then plotted E_1 versus E_3 on a graph (the bottom panel of Figure 15). We computed the correlation coefficient between E_1 and E_3. We also computed the square root of the sum of the squares of the differences between E_1 and E_3. A larger association between E_1 and E_3 than between E_1 and E_2 would indicate that the misfit observed in Figure 10 has a practical significance. However, Figure 15 shows that the association between E_1 and E_3 is very similar to that between E_1 and E_2. In addition, the correlation (and rank correlation) is 0.99 between E_1 and E_3 and also between E_1 and E_2. The square root of the sum of the squares of the differences is 0.54 for E_1 and E_2 and 0.56 for E_1 and E_3. Thus the misfit of the items of Form 1 of the writing test of the series of tests of basic skills does not seem to have much practical significance.

A Graduate Admissions Test

We had data from the pretest sections of a graduate admissions test—two data sets from the verbal part and two from the quantitative part of the test. The number of items for the quantitative part is 30 and that for the verbal part is 28. The sample size was close to 2,000 for all

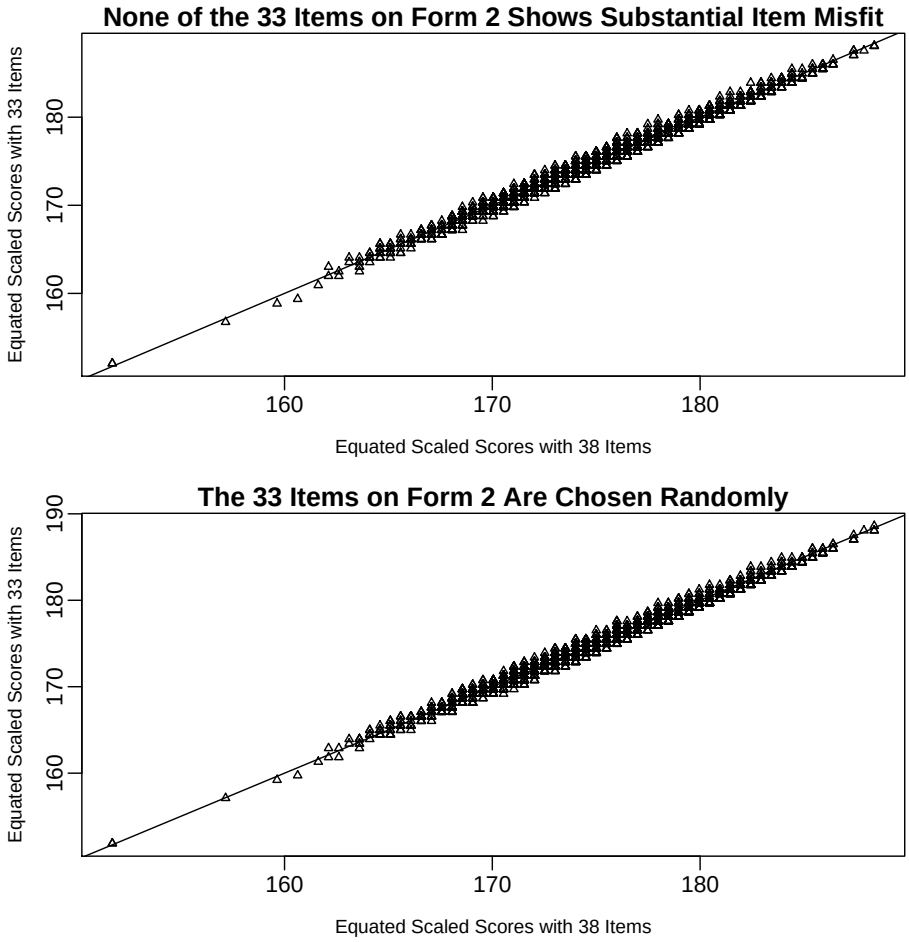

***Figure* 15. Evaluating practical consequences of misfit for Form 1 of the writing test of the series of tests of basic skills.**

the four data sets. The 3PL model is the operationally used model. However, there were similar problems of parameter identifiability as with the basic skills data when we fitted the 3PL model to these data. Therefore, we fitted a restricted version of the 3PL model where we forced all the guessing parameters to be the same.

Figure 16 shows histograms for the generalized residuals when O is the marginal score distribution. The figure does not show any evidence of misfit of the restricted 3PL model to the marginal score distribution—few of the generalized residuals lie beyond the 2.5th or 97.5th percentiles. For the four forms together, the residual is between -45 and -44 for five score-points (not shown in the plot). However, these are all very low scores (0, 1, or 2) and the observed and expected values both are very close and close to 0. Because very few examinees score so low, these large residuals should not cause much concern.

Figure 17, which is similar to Figure 1, shows the fit of the model to the statistic p_{11} for all pairs of items. The figure shows some evidence of misfit of the restricted 3PL model to the second-order marginal totals, though the evidence is not as strong as in Figure 1. The percentage of p-values that are significant at the 5% level are 30, 17, 17, and 12 for the four data sets.

Figures 18 to 21 show the plots of item fit for the four forms only for the items for which substantial misfit was found.

Overall, the analysis of item fit shows that there is some evidence of item misfit for these data.

A Computer-Based Science Test

In this study, a sample of about 2,000 examinees responded to three separately timed computer-based interactive tasks in a science test. The goal of the test was to provide in-depth information about student performance that cannot be directly accessed in paper-and-pencil tests by capturing students' behavioral actions by use of a computer. Several items, both MC and CR items, followed each task.

For one of the tasks, students' problem-solving behaviors were recorded and scored along with the CR and MC items answered by students. Table 1 lists the number of items by item type (MC, CR, and behavior indictor or BI) and maximum block scores for each task.

The score range varies over the 19 CR items. Two of them have a score range of 0 to 1, nine have score range of 0 to 2, seven have range of 0 to 3, and one has score range of 0 to 4. In

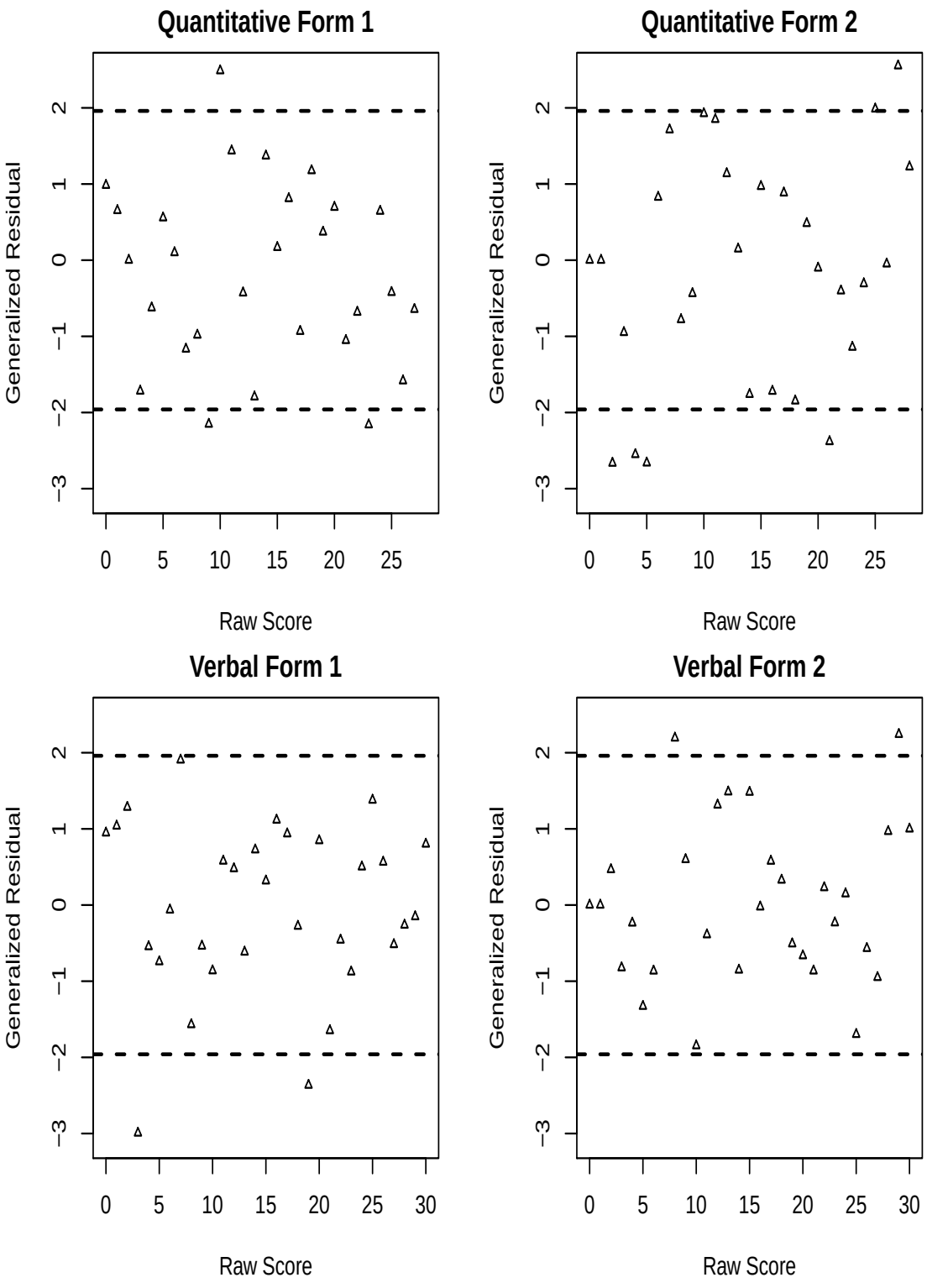

***Figure* 16. The generalized residuals for marginal score distribution for the admissions tests.**

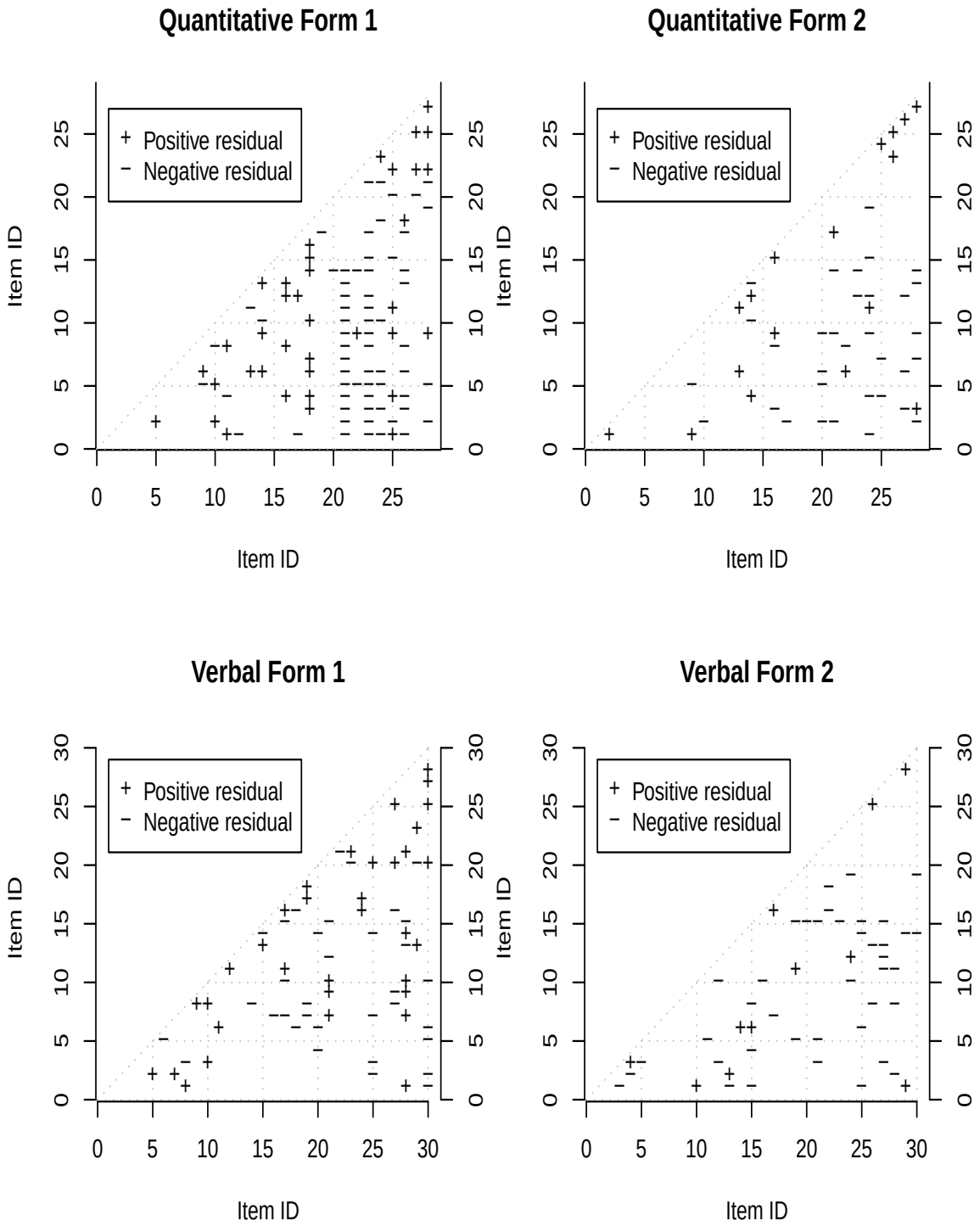

***Figure* 17. The fit of the IRT model to the second-order marginal totals for the admissions tests.**

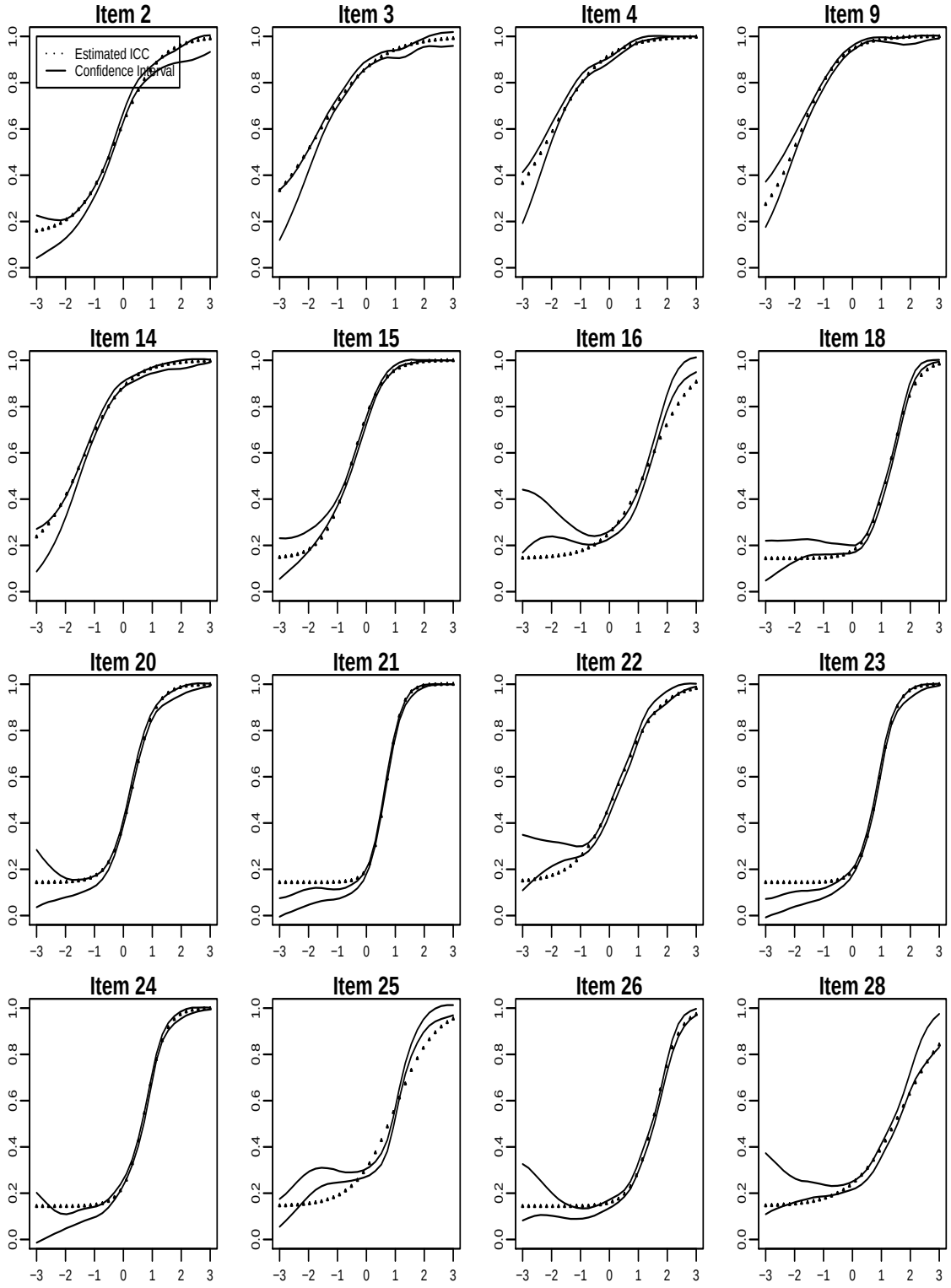

***Figure* 18. Plots of item fit for the admissions tests—First quantitative form.**

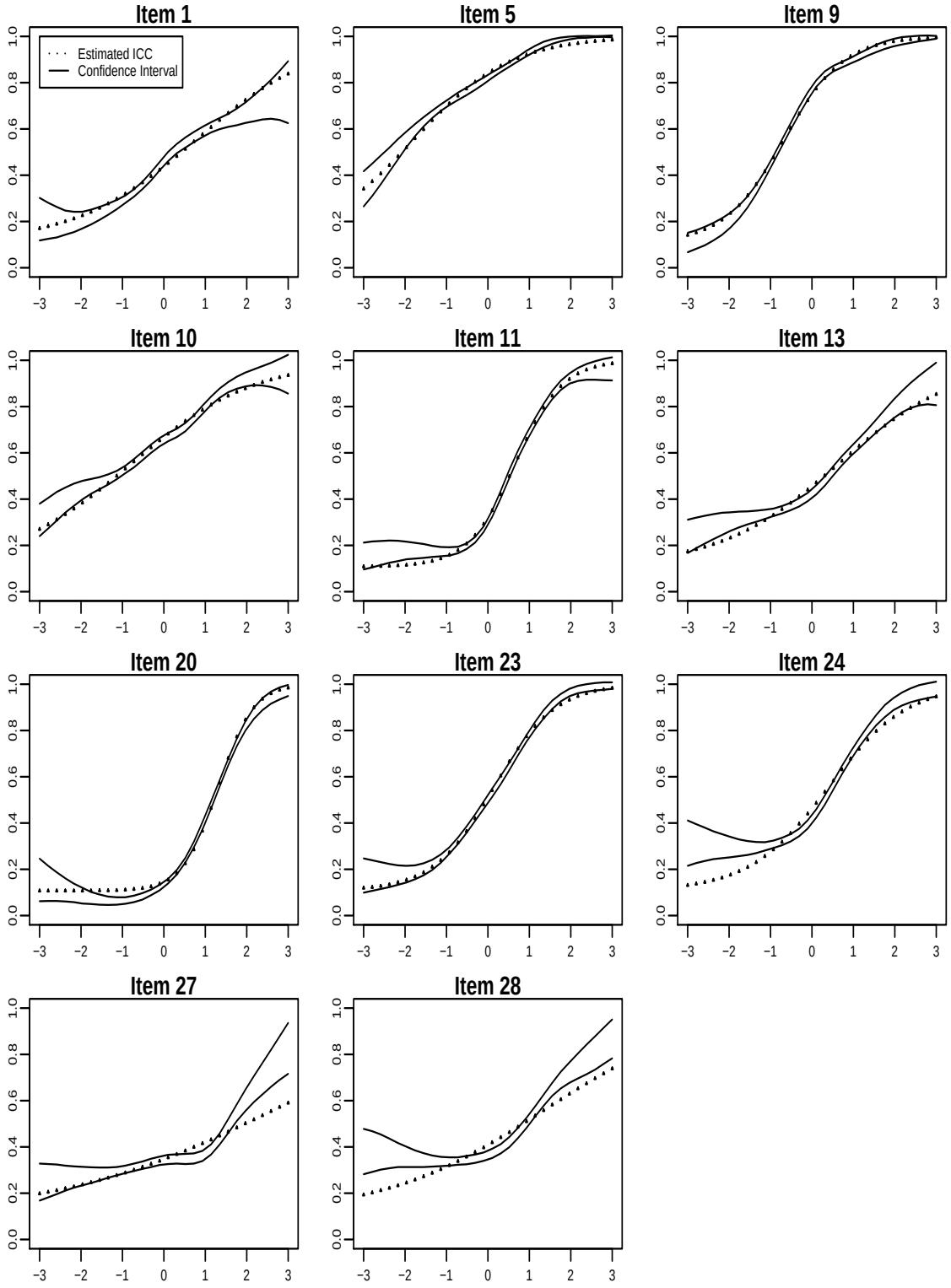

***Figure* 19. Plots of item fit for the admissions tests—Second quantitative form.**

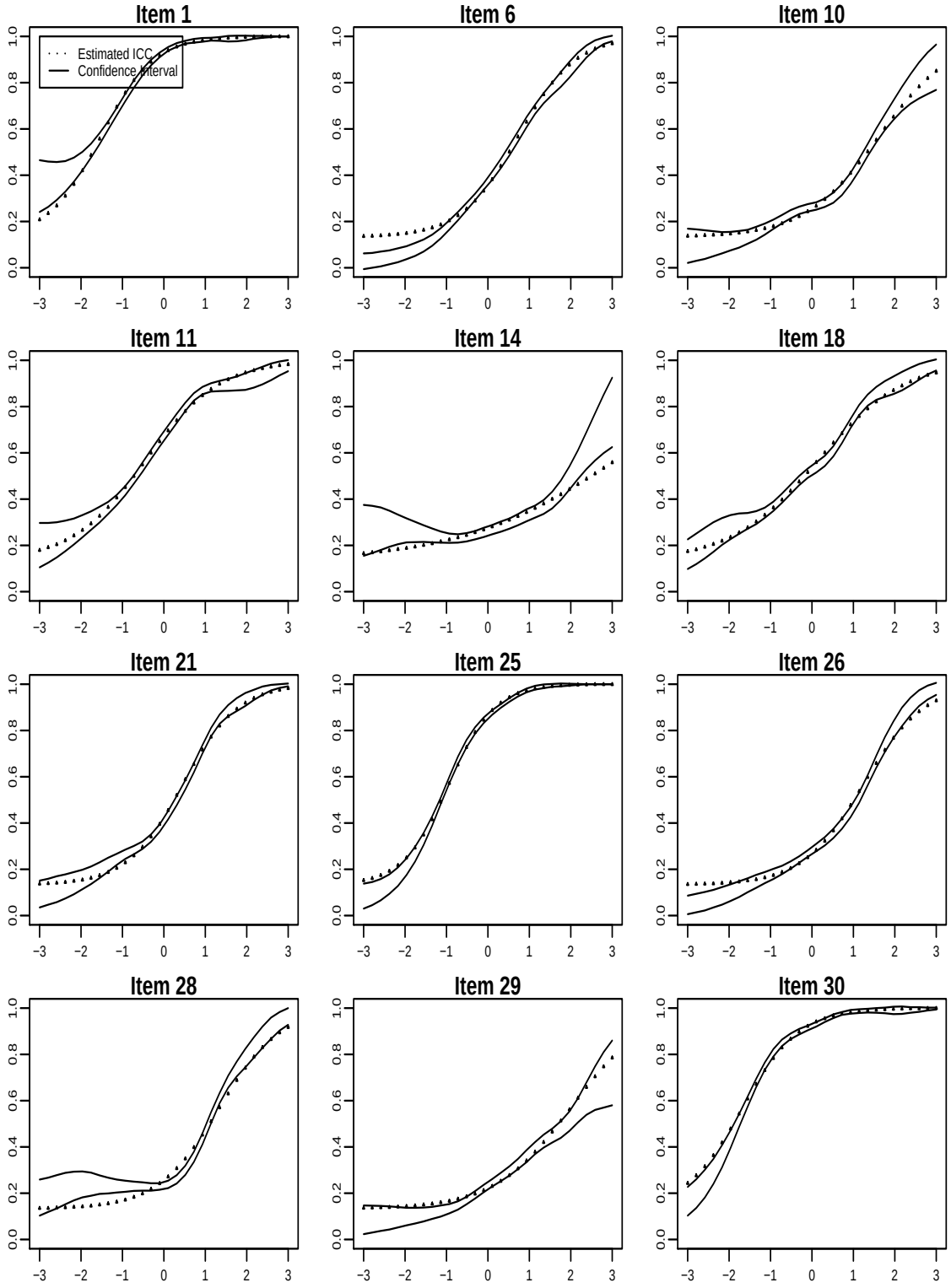

***Figure* 20. Plots of item fit for the admissions tests—First verbal form.**

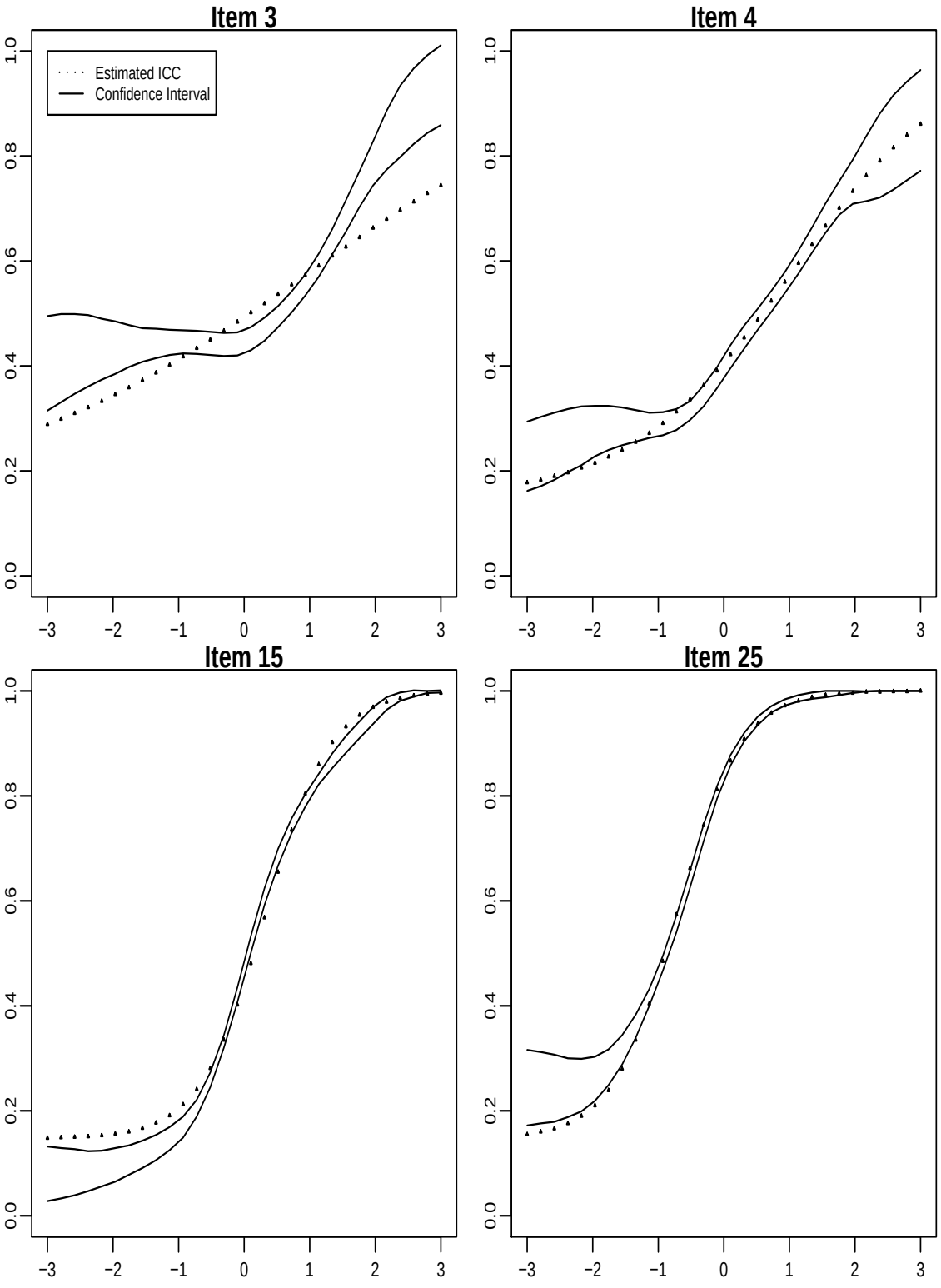

***Figure* 21. Plots of item fit for the admissions tests—Second verbal form.**

Table 1

Description of the Science Test

Task	Number of MC items	Number of CR items	Number of BI items	Maximum block score
1	1	6	4	25
2	0	9	0	20
3	2	4	0	14
Total	3	19	4	59

Note. BI = behavior indicator, CR = constructed response, MC = multiple choice.

addition, the score range is 0 to 2 for one of the four behavior indicators, and 0 to 3 for the rest of three behavior indicators. Both principal component analysis and confirmatory factor analysis suggested that the data do not strongly support multi-dimensionality. However, there is concern about too much association among items that belong to the same task that would lead to the violation of the local independence assumption that an IRT model makes. The goal here of the analysis of model fit is to examine if a unidimensional IRT model can provide an adequate fit to the data. It is suspected that the correlation between the items within each task may be too high to be explained by a unidimensional IRT model. We fitted the 2PL model to the MC items and the generalized partial credit model to the CR and behavior indicator items.

Because several items are polytomous, there are several residuals for each pair of items. For example, for Items 1 and 2, because Item 1 has 3 score categories and Item 2 has 4, there are 12 residuals, one for each pair of score categories, where an example of a pair of score categories is {2,3}, where 2 denotes the score on Item 1 and 3 the score on Item 2. Of the 3,464 such residuals for pairs of items, about 42% were statistically significant at the 5% level. Figure 22 is an attempt to show the pairs of items with statistically significant association. If a pair of items is highly correlated, then we would expect that the number of examinees who obtained a high score on both the items would be higher than predicted by a unidimensional IRT model and the residual will be high for those pairs of high score categories. Therefore, for each pair of items, we computed n_+, the number of the generalized residuals among the pairs of high score categories that are larger than 1.96 and n_-, the number of the generalized residuals among the pairs of high

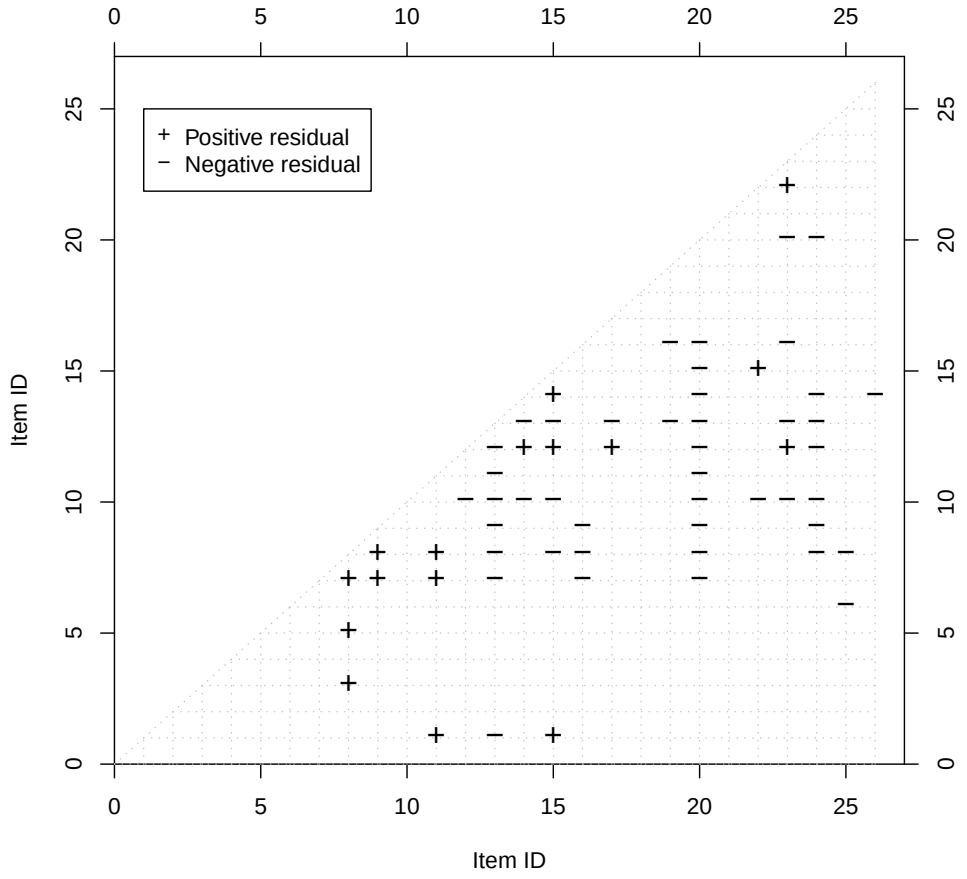

***Figure* 22. A plot showing the association among the pairs of items for the computer-based science test.**

score categories that are smaller than -1.96. If $n_+ - n_-$ is larger than half the number of pairs of high score categories for an item pair, then we judge the item pair to have a significantly positive association. If $n_- - n_+$ is larger than half the number of pairs of high score categories for an item pair, then we judge the item pair to have a significantly negative association. For items with 2, 3, 4, and 5 score categories, we defined 1, 2, 2 and 3, and 3 and 4, respectively, as the high score categories. Thus, for example, for the item pair consisting of Items 16 and 19, where both items have 4 score categories, $n_+ = 0$, $n_- = 3$, and the number of pairs of high score categories is 4 (for both these items, the high score categories are 2 and 3). Thus the item pair was judged to have a significantly negative association. [3] Figure 22 shows the item pairs that were found to have significantly positive association (a plus sign) or significantly negative association (a minus sign). No symbol for an item pair denotes that the pair had neither a significantly positive association nor a significantly negative association.

Figure 22 shows the existence of statistically significant residuals for several pairs of items. For example, the three pairs involving Items 7, 8, and 9 have significantly positive association. Further study of the content of the assessment suggests that these items were related to one test scenario where the examinees conduct an experiment and are asked the behavior-indicator Items 8 and 9 and asked to draw conclusions based on their experimental result in Item 7. Thus it is no surprise that they are more correlated than what the IRT model can predict. Several pairs of items that involve Items 6, 10, 13, 16, 20, 24, and 25 have negative and significant residuals. Most of these items are difficult items, although they are not the seven most difficult items; however, Figure 22 does not indicate any clustering of items such as clustering into the three tasks, except for clustering of Items 7 to 9.

It is impossible to assess the practical significance of misfit for this test because little additional information was available and this test was only used once. However, too much association between some item pairs (as shown in Figure 22), along with other considerations, such as the limited number of test items, led to the decision not to use any IRT model for reporting results of the computer-based science test.

An English Proficiency Test

We had data from two forms each of two (out of four) parts of a special administrations of an English proficiency test. Operationally, for each part, the raw score on a form is equated to that

on a reference form using IRT true score equating. The tests consist of mostly binary items. In typical administrations, there are 34 dichotomous items in Part 1 and three 3-category polytomous items in addition to 39 dichotomous items in Part 2; however, minor variations on this format can arise. The 2PL model is currently used for the binary items; however, at the time of the special administration, the 3PL model was used. The generalized partial credit model (GPCM) is used for polytomous items. In these special administrations, scores on Form 2 of Part 1 could be equated to scores on Form 1 of Part 1 using 17 external anchor items and scores on Form 2 of Part 2 could be equated to scores on Form 1 of Part 2 using 28 external anchor items. In usual operational administrations, it is very unusual for two forms to have so many common items for the same part. The sample size was about 1,500 for both of these forms. For these data, the goal of our analysis of model fit is to examine if the operationally used IRT model provides an adequate fit to the data. We would also like to examine, if misfit is found, the consequences of the misfit.

We combined the operational items with the anchor items and fitted the IRT models to the combined data. There were the previously-mentioned identifiability problems with the 3PL model. Therefore, we fitted a restricted 3PL model where we forced all the guessing parameters to be the same for the binary items. Because the fit of the 2PL model is very similar to that of this restricted 3PL model, results for the 2PL model are not reported. We fitted the generalized partial credit model (GPCM) to the 3-category items.

Figure 23, which is similar to Figure 1, shows all the item pairs for which a significantly positive or negative association was found. [4] The figure shows strong evidence of misfit of the restricted 3PL model to the second-order marginal totals for the two forms for Part 2 and some evidence of misfit for the two forms for Part 1. There are more significantly negative generalized residuals than significantly positive generalized residuals for the two forms for Part 2. We do not have a fully satisfactory explanation for the observed pattern of generalized residuals, although it should be noted that Part 1 includes nine item sets and Part 2 includes five item sets. Our difficulty is that the relationship of sign to membership in the same set of items does not appear to be clear. For Part 1, Form 2, the generalized residual is significant and negative for almost all pairs of items involving Item 10, which was found to be the most difficult item in the data set. The item was answered correctly by only 12% of examinees, the estimated difficulty parameter was 5.14, and the estimated slope parameter was 2.42. The percentage of p-values that are significant at the 5% level is 11, 16, 31, and 40 for the four data sets.

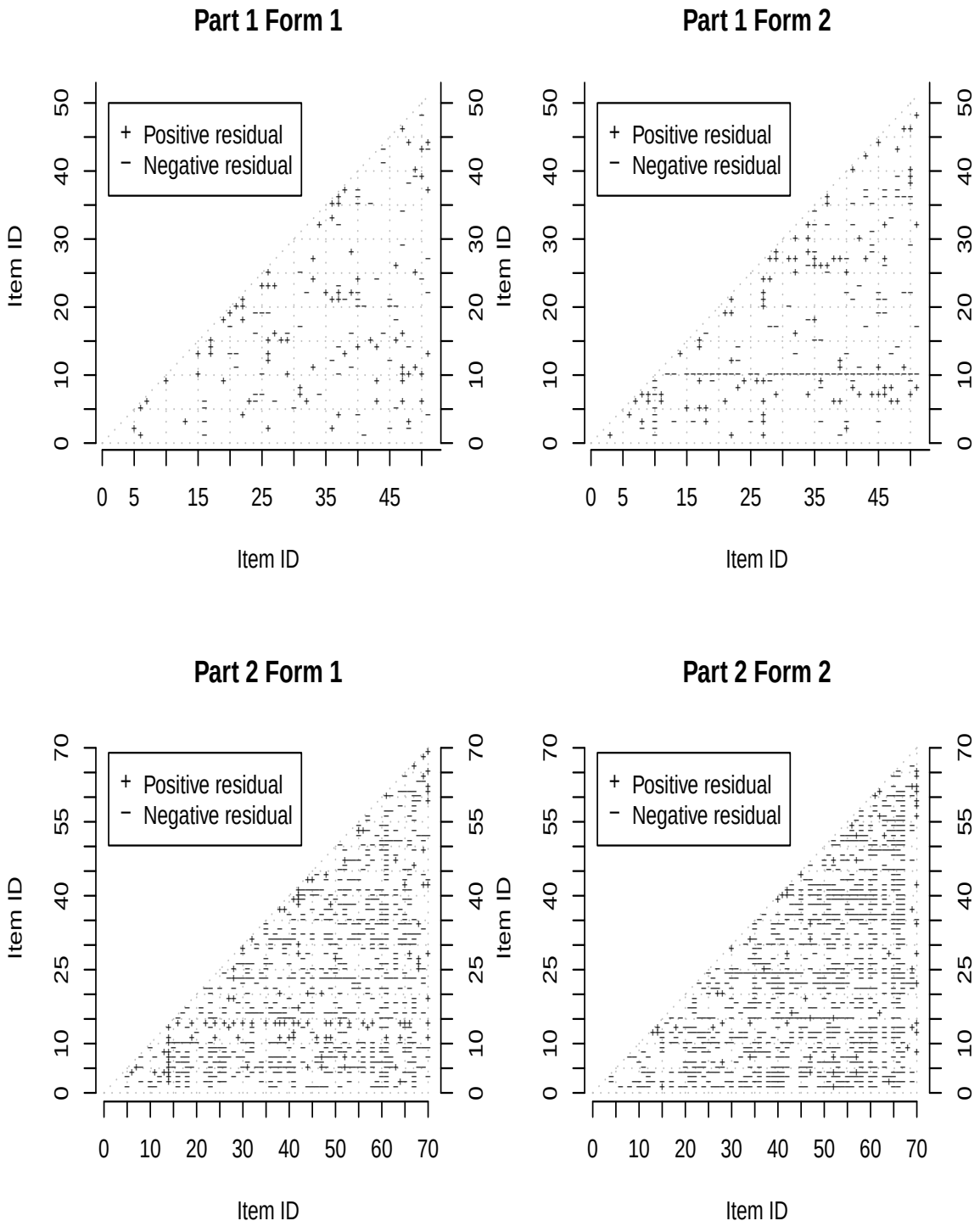

***Figure* 23. The fit of the IRT model to the second-order marginal totals for the English proficiency test.**

Figures 24 to 27 show the plots of item fit for only the items for which substantial misfit was found in the four forms. In these plots, the anchor items are treated as Items 35 to 51 for the two forms for Part 1 and Items 43 to 70 for the two forms from Part 2 forms in these plots. Some amount of item misfit is evident for the two parts of the test. More misfit is found for Part 2.

The Practical Significance of Misfit

Because an IRT model is used to perform a true score equating of the raw scores for this test, one way to assess the practical significance of misfit is to examine if the omission of the misfitting items from the anchor test leads to a difference in true score equating of the raw scores. The two forms of Part 2 can be equated through 28 common items. Figures 26 and 27 show that there is considerable misfit for Items 44, 49, 56, 61, and 70 in both forms for Part 2 forms. [5] Figure 28 shows the impact of omitting these five items from the anchor set. The top panel of the figure shows the equating conversions when (a) equating was performed using all anchor items and (b) equating was performed after omitting the above-mentioned five items from the anchor, that is, using the remaining 23 anchor items. The equating method used was the IRT true score equating using the Stocking-Lord algorithm (Kolen & Brennan, 2004) that is used operationally. The bottom panel of Figure 28 shows the differences between the equating conversions. In interpreting these plots, we will use the difference that matters or DTM criterion suggested by Dorans and Feigenbaum (1994) as a difference in equating conversions that is practically large. For raw score conversions, the DTM is 0.5 according to Dorans and Feigenbaum (1994). Figure 28 shows that the differences between the equating conversions here are less than the DTM except for the very low score points. Thus, the lack of item fit observed in Figures 26 and 27 does not seem to matter practically, at least when only equating of two forms is a concern.

We performed another set of analysis to assess the practical significance of the misfit for Part 2. The steps of the analysis are the following:

- Remove from Form 2 all the 16 misfitting items featured in Figure 27. The form has 32 unique items (items that are not common with Form 1) after this and the anchor test has 22 items after this.

- Fit the IRT model to this cleaned data file and perform IRT true score equating of 32-item version of Form 2 to the 42-item version of Form 1 using the 22 anchor items.

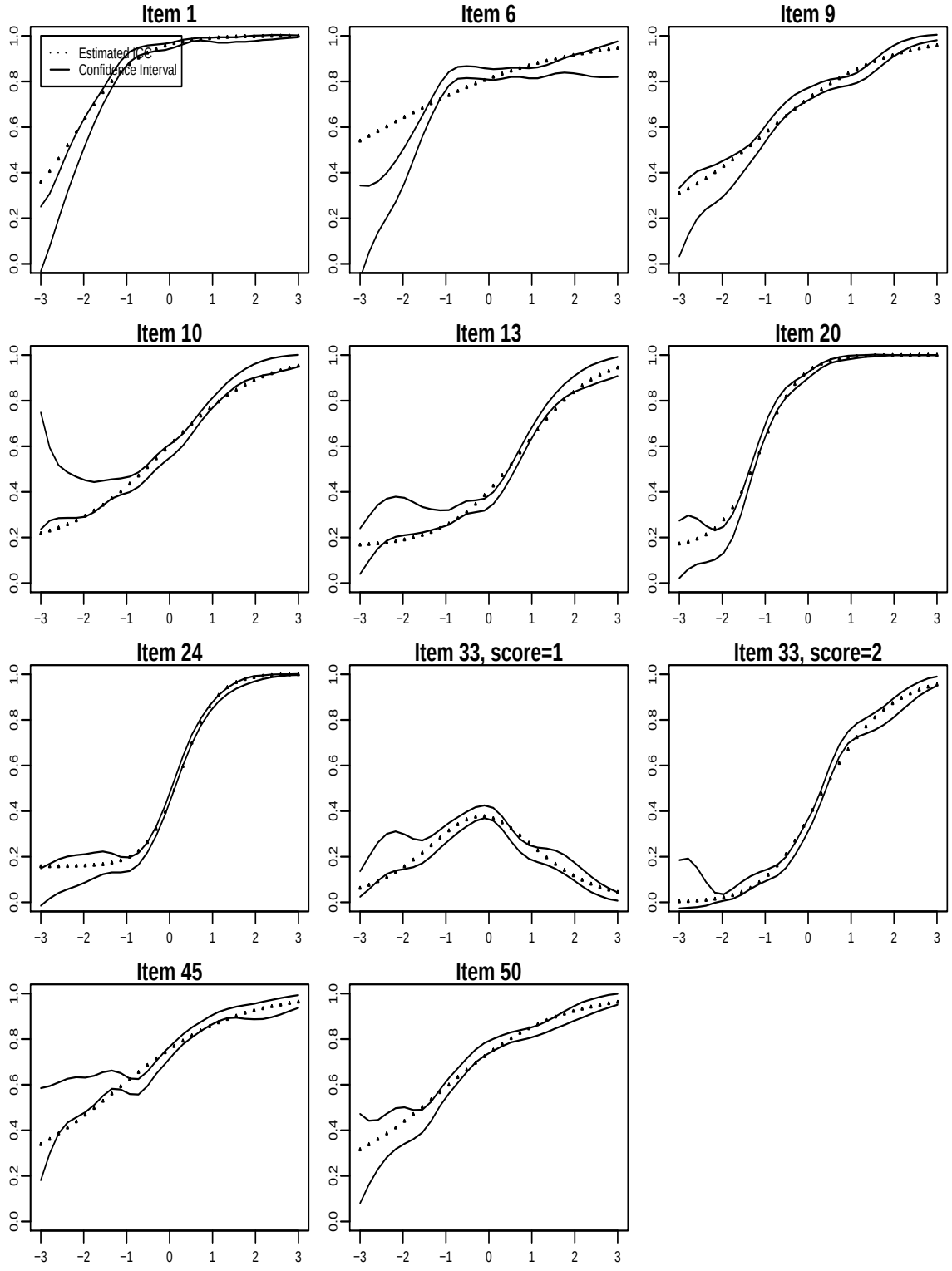

***Figure* 24. Plots of item fit for the test of English proficiency—First form of Part 1.**

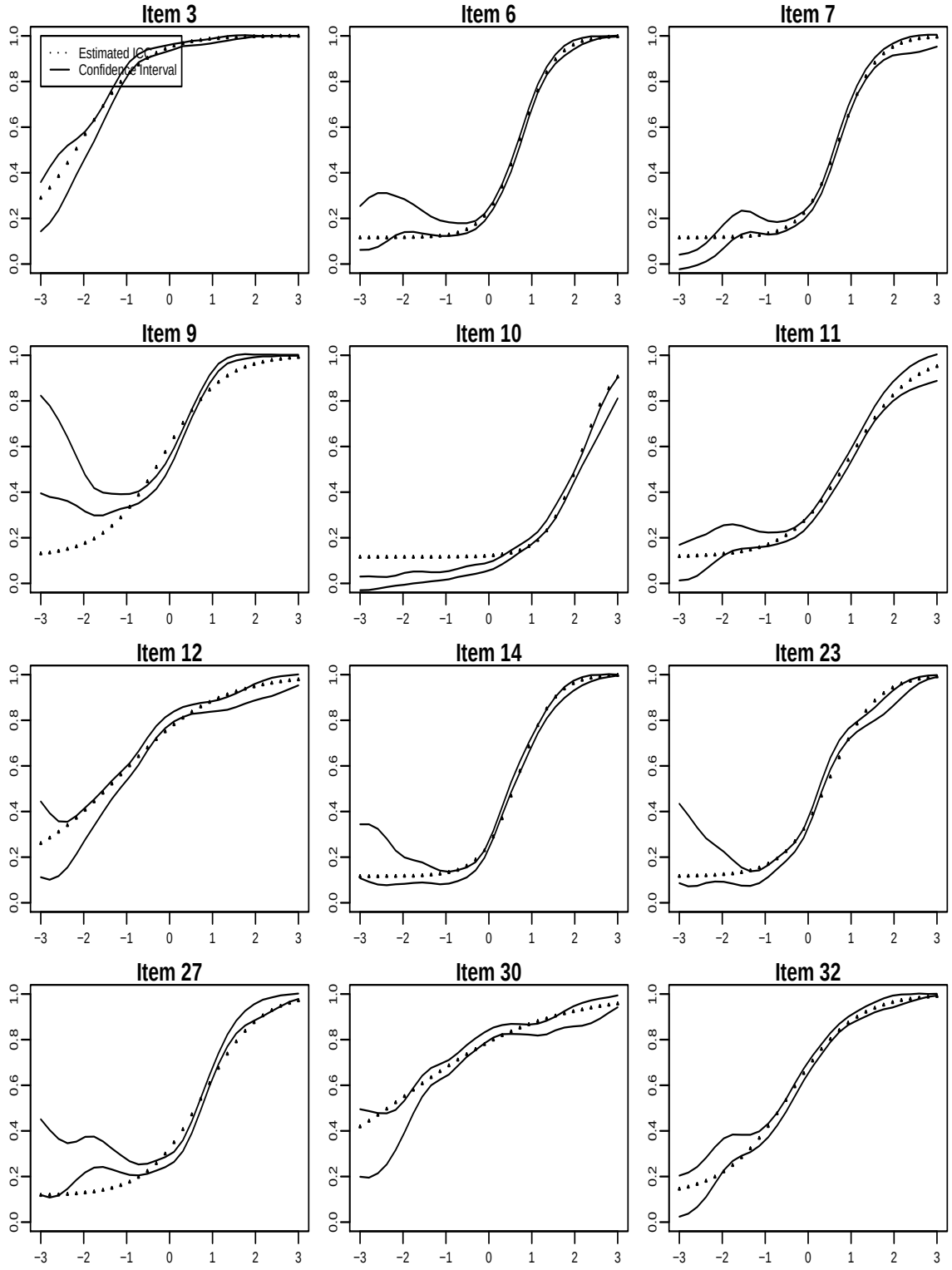

***Figure* 25. Plots of item fit for the test of English proficiency—Second form of Part 1.**

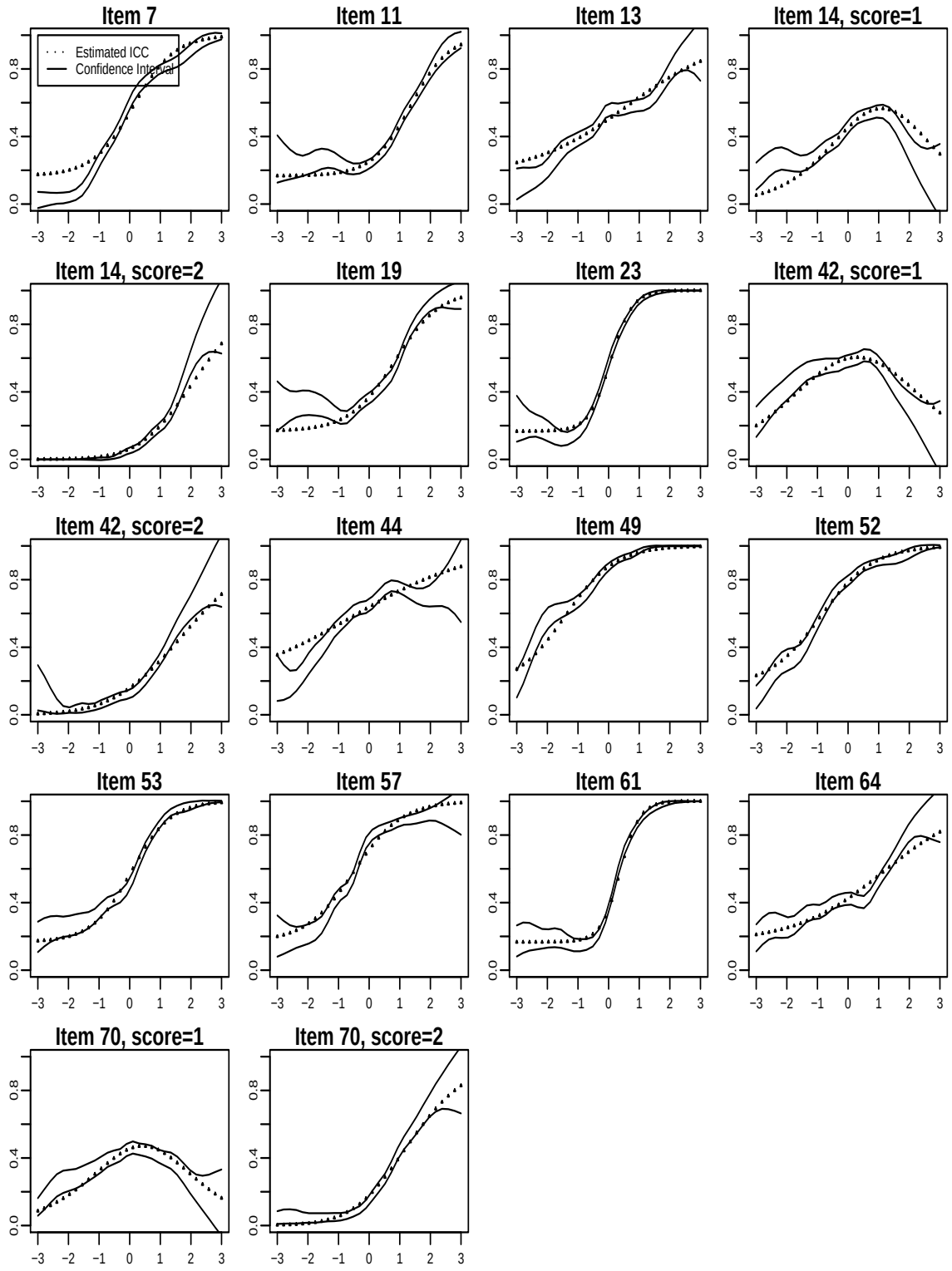

***Figure* 26. Plots of item fit for the test of English proficiency—First form of Part 2.**

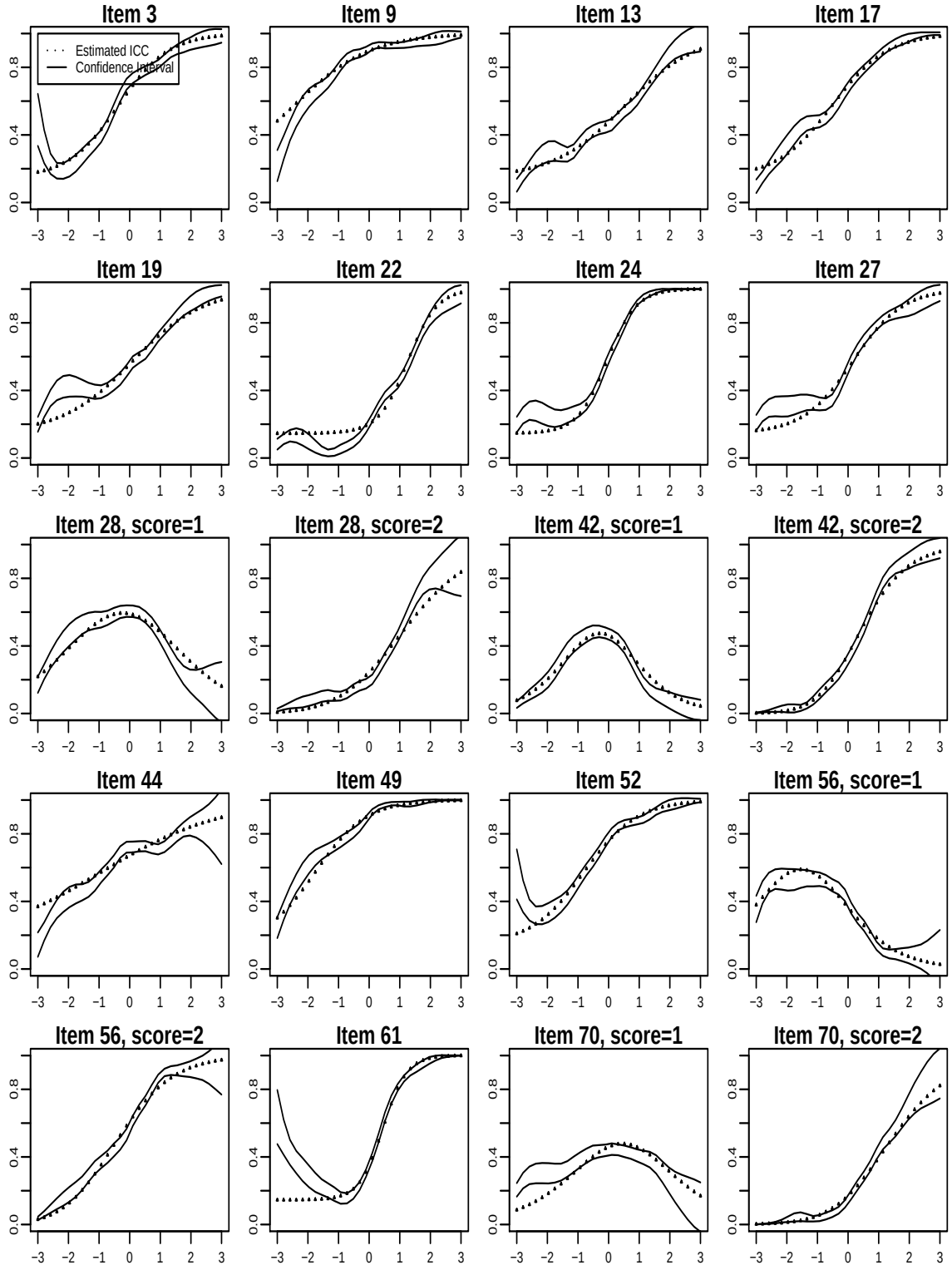

***Figure* 27. Plots of item fit for the test of English proficiency—Second form of Part 2.**

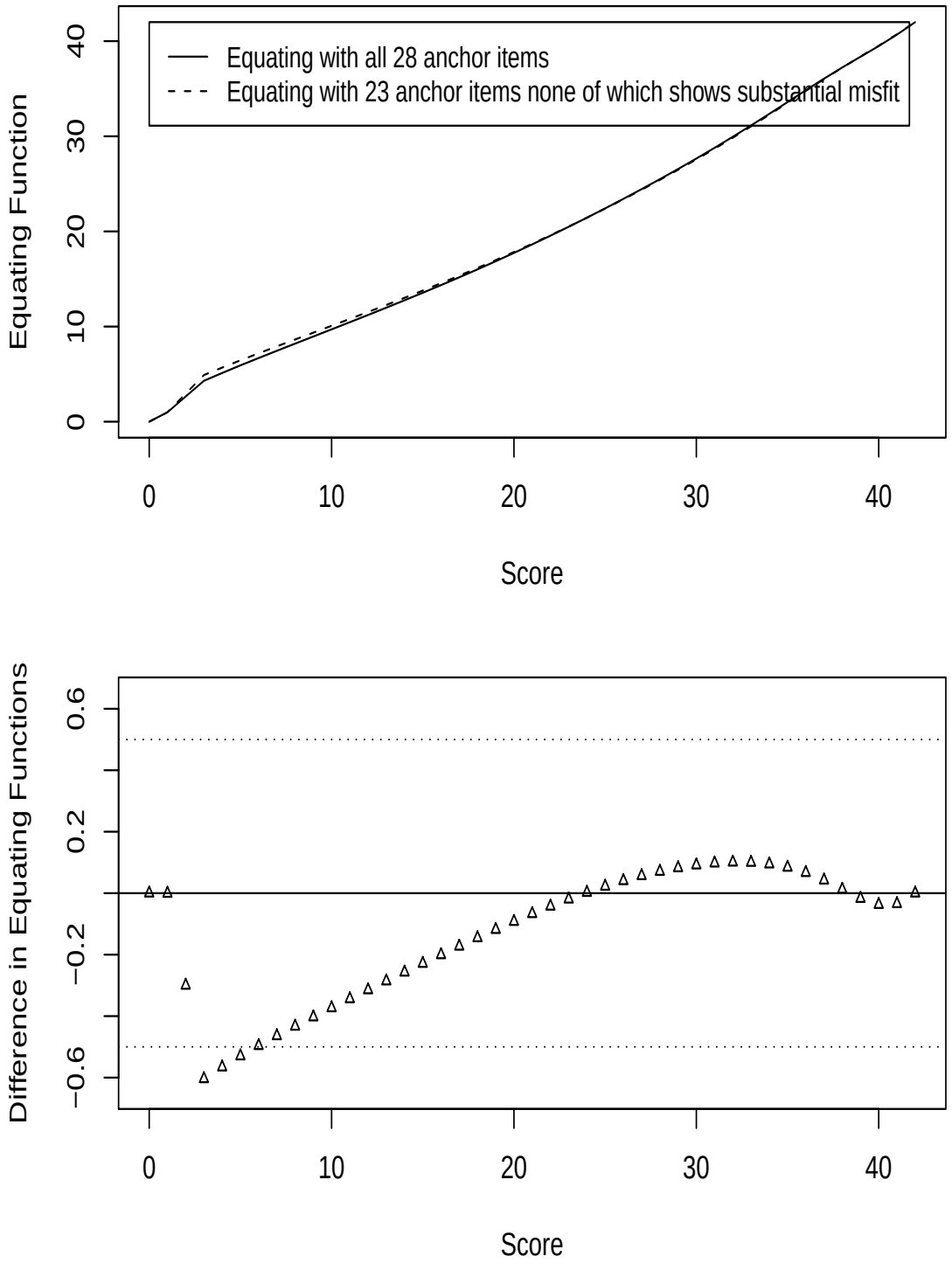

***Figure* 28. Impact of misfit for the test of English proficiency—Part 2.**

- Calculate (a) E_1, the equated score on the full data set of the examinees and (b) E_2, the equated score on the cleaned data set of the examinees. Plot E_1 versus E_2 on a graph (the top panel of Figure 29) and compute the correlation coefficient between E_1 and E_2.

The correlation coefficient (as well as the rank correlation) between E_1 and E_2 is 0.98. For comparison purposes, the bottom panel shows a similar plot made from equating after removing the six misfitting items from the anchor and 10 randomly chosen items from Form 1. The correlation coefficient is 0.98 for this plot as well. Figure 29 shows that the omission of the misfitting items leads to some small differences for some examinees, but the same magnitude of difference is seen overall when the same number of items are omitted randomly. Thus, the item misfit does not seem to have much practical significance.

A third analysis to assess the practical significance of the misfit for Part 2 was performed by noticing the failure of the IRT model to explain the second-order marginal totals in Figure 23. The items of Part 2 are members of five sets, each set having 14 items each. These sets of items are often referred to as testlets (Wainer & Kiely, 1987). The local independence assumption of a unidimensional IRT model is often violated for testlets (see, for example, Wainer & Kiely, 1987). This could be one reason of the severe misfit observed in Figure 23. A more general model that accounts for the dependence within the testlets is a multidimensional IRT model (Haberman, von Davier, & Lee, 2008; Reckase, 1997, 2007). To examine if fitting of such a model leads to any practical change, we fitted a multidimensional generalized partial credit model that is suitable for polytomous items to the data. The model assumes that an r-dimensional random ability vector $\boldsymbol{\theta}_i$ with elements θ_{ik}, $1 \leq k \leq r$, is associated with each examinee i. The element θ_{ik} denotes the ability of the ith examinee on the kth skill. The pairs $(\mathbf{X}_i, \boldsymbol{\theta}_i)$, $1 \leq i \leq n$, are independent and identically distributed, and, for each examinee i, the response variables X_{ij}, $1 \leq j \leq q$, are conditionally independent given $\boldsymbol{\theta}_i$. Suppose the possible scores on item j are 1, 2,...m_j. Let the location parameters of item j are $b_{j1} = 0, b_{j2}, \cdots b_{jm_j}$, $1 \leq j \leq q$. Let $\mathbf{a}_j$ be the r-dimensional item-discrimination vector of item j, $1 \leq j \leq q$. The k-th element of $\mathbf{a}_j$, $1 \leq k \leq r$, denoted as a_{jk}, corresponds to the discrimination of item j with respect to skill k. Given that an examinee has ability vector $\boldsymbol{\theta}$, the probability of a score h on item j is given by

$$P_j(h|\boldsymbol{\theta}) = \frac{\exp\left[\sum_{v=1}^{h}\left(\mathbf{a}_j'\boldsymbol{\theta} - b_{jv}\right)\right]}{\sum_{c=1}^{m_j}\exp\left[\sum_{v=1}^{c}\left(\mathbf{a}_j'\boldsymbol{\theta} - b_{jv}\right)\right]}, \tag{4}$$

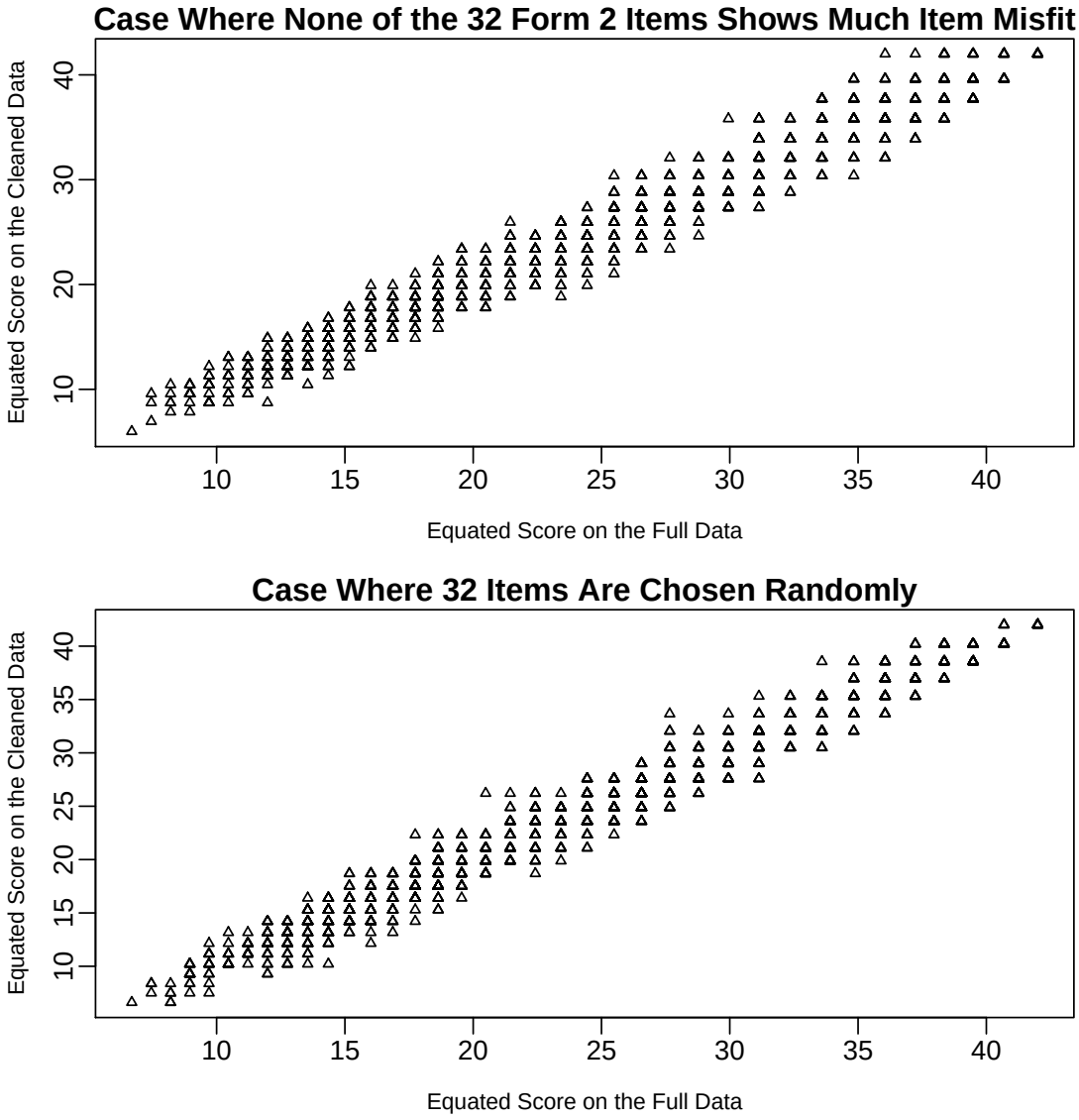

***Figure* 29. Another look at the impact of misfit for the test of English proficiency—Part 2.**

where the vector product

$$\mathbf{a}_j'\boldsymbol{\theta} = \sum_{k=1}^{r} a_{jk}\theta_k.$$

The 2PL model is a special case of the above model. The between-item model (Adams, Wilson, & Wang, 1997) is considered here in which for each j, a_{jk} takes a nonzero value for only one k. Given examinee ability vector $\boldsymbol{\theta}$,

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{j=1}^{q} P_j(x_j|\boldsymbol{\theta}) \tag{5}$$

is the conditional probability that the response vector $\mathbf{X}_i$ is equal to a vector $\mathbf{x}$ with elements x_j, $1 \leq j \leq q$, each of which is equal to either 0 or 1.

In applying the multidimensional IRT model, we assumed that the items in each testlet measure a different skill—that leads to $\boldsymbol{\theta}$ having five components, one for each testlet. After fitting the MIRT model, we computed, for each examinee, the mean of the five estimated values of θ_{ik}s. Because this estimate accounts for the effect of testlets, this mean is a better estimate for the examinee's overall proficiency. We computed the correlation of this estimate with the estimated ability from fitting a unidimensional IRT model. The correlations were 0.99994 and 0.99997, respectively, for the two forms. Thus, the misfit observed in Figure 23 does not seem to be practically significant.

Two Subjects From a Battery of Examinations

We had data on one form each for two subjects from a battery of examinations. For both subjects, true score equating via the 1PL model is used to convert the number-correct raw score of an examinee on a new form to the corresponding raw score on a reference form. The converted raw score is then converted to a scaled score. We fitted the 1PL model and the 2PL model to the two data sets, which had 905 and 3,631 examinees, respectively.

Figure 30 shows histograms for the generalized residuals for the marginal score distribution. The figure does not show much evidence of misfit of the IRT models to the marginal score distribution for Subject 1—few of the generalized residuals lie beyond the 2.5th or 97.5th percentiles. There is slight evidence of misfit for Subject 2, both for the 1PL model and the 2PL model.

Figure 31 shows all pairs of items for which the generalized residual for the statistic p_{11} is statistically significant. The items are sorted according to increasing order of estimated item

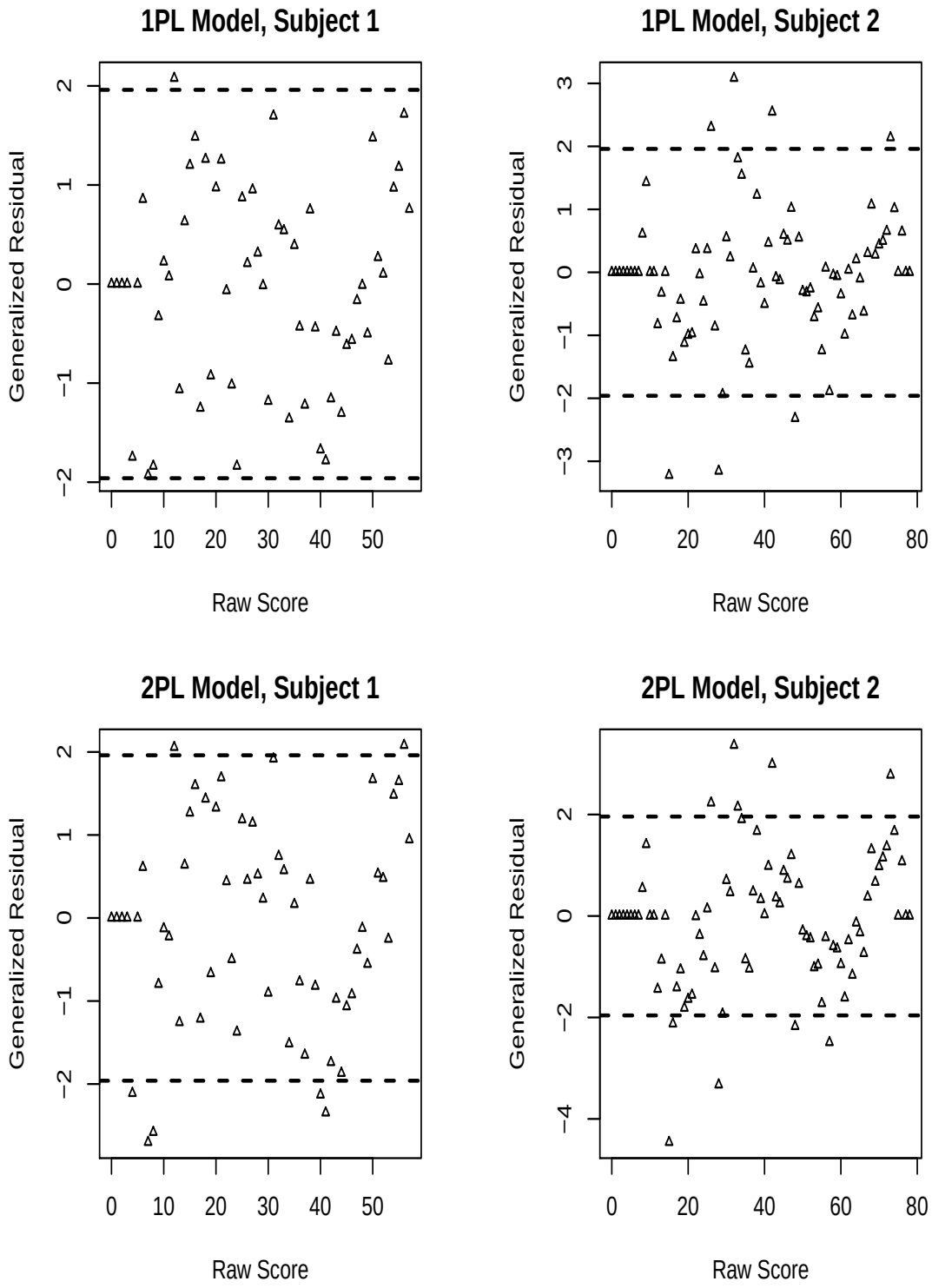

***Figure* 30. Generalized residuals for marginal score distribution for the battery of examinations.**

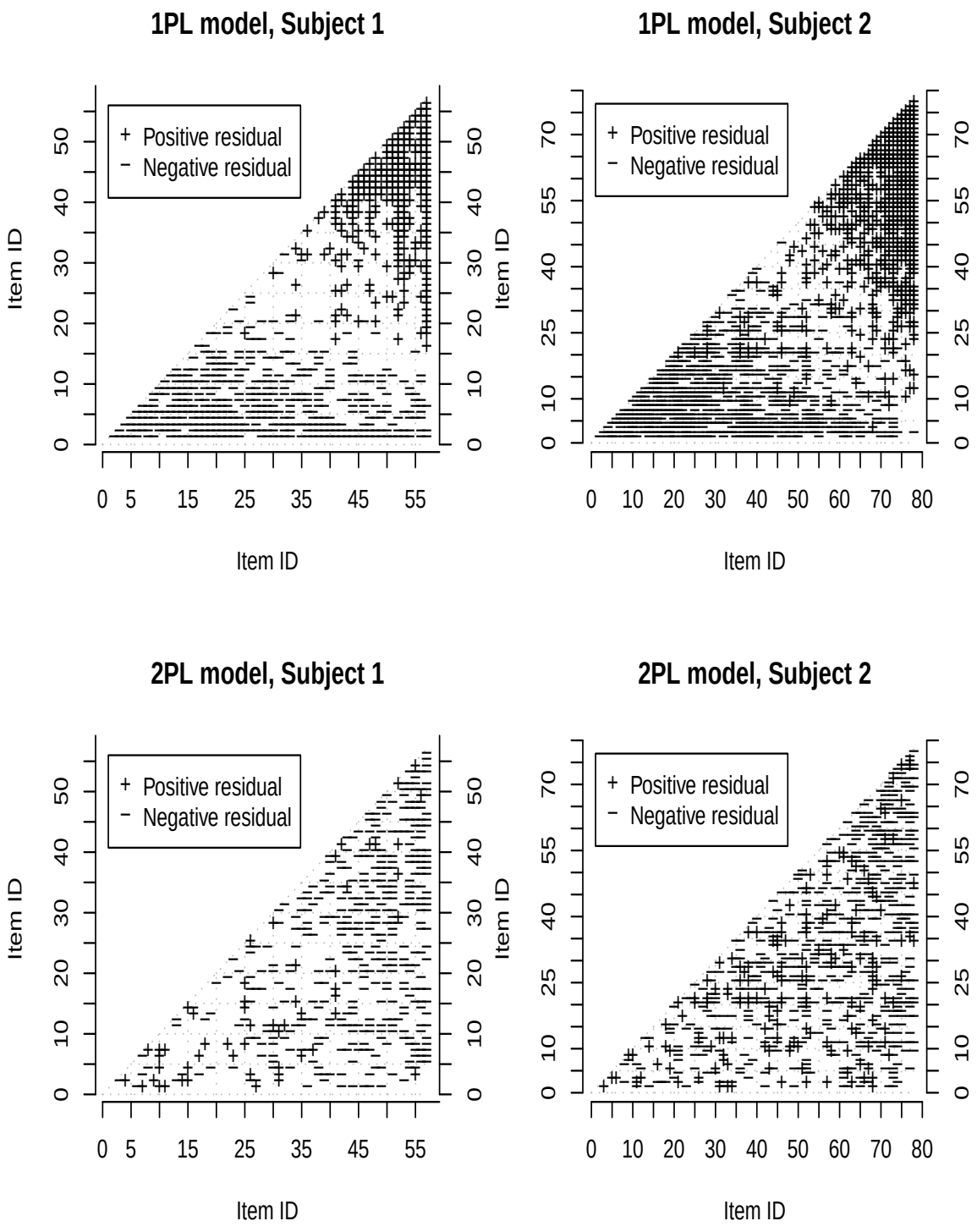

***Figure* 31. The fit of the IRT model to the second-order marginal totals for the battery of examinations.**

discrimination parameters from the 2PL fit. The figure shows substantial evidence of misfit of both IRT models to the second-order marginal totals. The percentage of p-values that are significant at the 5% level are 41 and 42 for the 1PL model and 23 and 22 for the 2PL model. The 1PL model shows more evidence of misfit compared to the 2PL model. In addition, for the 1PL model, several residuals involving items with high discrimination are significantly positive and several residuals involving items with low discrimination are significantly negative. A similar observation was made by Sinharay et al. (2006)—see their Figure 4.

Analysis of item fit shows strong evidence of item misfit for the data set from Subject 2 for both models. For the 1PL model, 58 items were found to have substantial misfit; for the 2PL model, 39 items were found to have substantial misfit—almost all of these items were included in the 58 items for the 1PL model. We also found strong evidence of item misfit for the data set from Subject 1 for the 1PL model—the model shows misfit for 37 items. In contrast, the 2PL model shows misfit for only 9 items for Subject 1. Figures 32 to 35 show the plots of item fit for selected items that had the worst fit. The 2PL model had better fit than the 1PL model even for the items that were found to show misfit for the 2PL model. For example, consider Items 22 and 55 for Subject 1. While both of these items show misfit for both the models, the values of $\bar{I}_j(\theta)$ are further from the confidence bound around $\hat{I}_j(\theta)$ for the 1PL model.

The Practical Significance of Misfit

We noticed substantial amount of misfit of the operationally used 1PL model for these examinations. The 2PL model fits these data better than the 1PL model. Another way to find out if the 2PL model leads to a substantial improvement over the 1PL model is to examine the information-theoretic measure *Minimum Estimated Expected Log Penalty Per Item* (see, e.g., Gilula & Haberman, 2001; Haberman, 2006), henceforth referred to as Penalty. This measure compares models and is based on the logarithmic penalty function for probability prediction (Savage, 1971). For a model fitted to a data set, the Penalty is obtained as

$$\text{Penalty}_{model} = -\frac{\ell}{2nm},$$

where ℓ is the maximum log-likelihood under the model, n is the sample size, and m is the number of items. For example, Penalty_{1PL} is the penalty for the 1PL model and Penalty_{2PL} is the penalty under the 2PL model. Note that as the likelihood increases, the penalty decreases. The values of

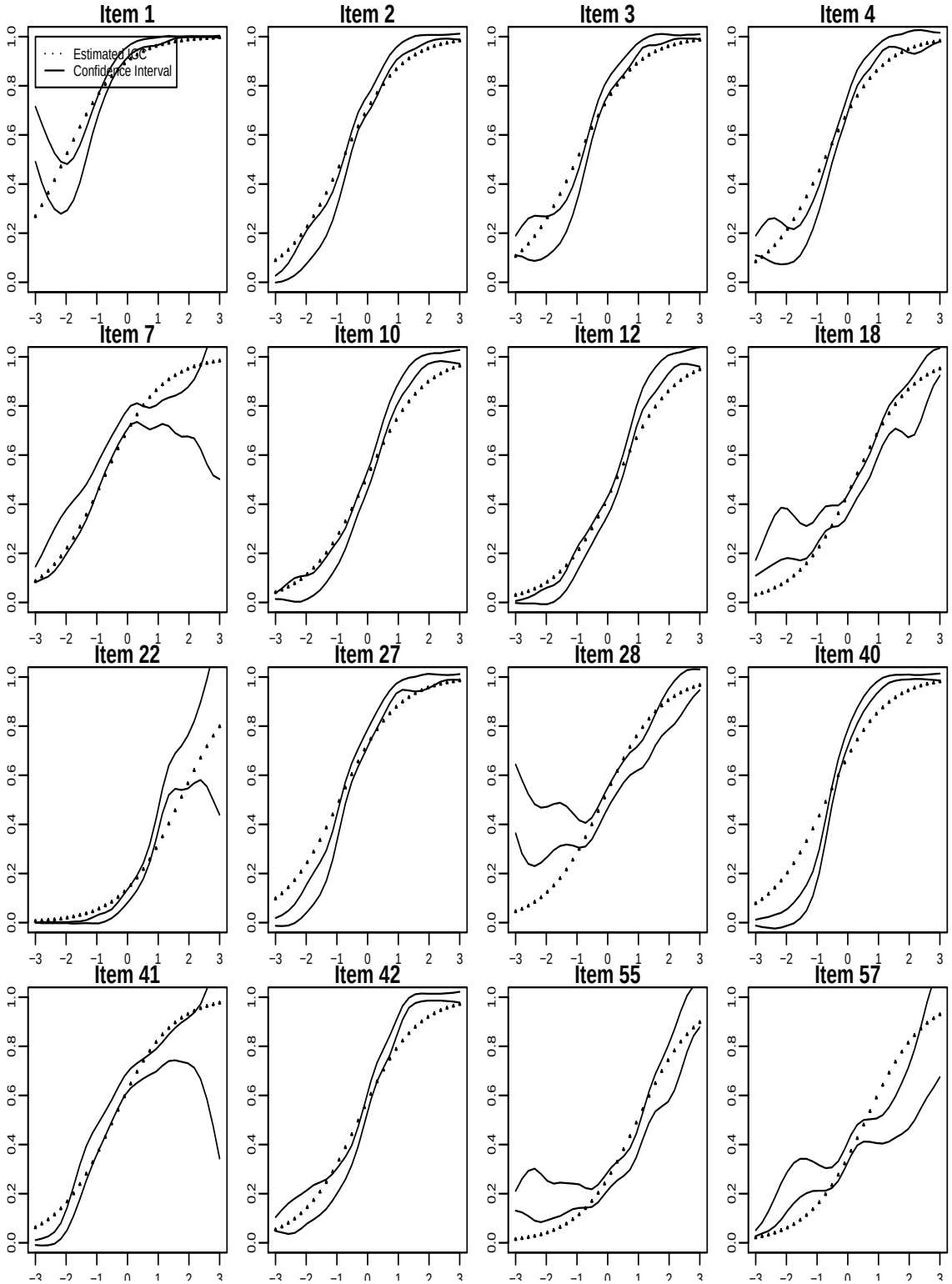

***Figure* 32. Plots of item fit for the battery of examinations—Subject 1 and 1PL model.**

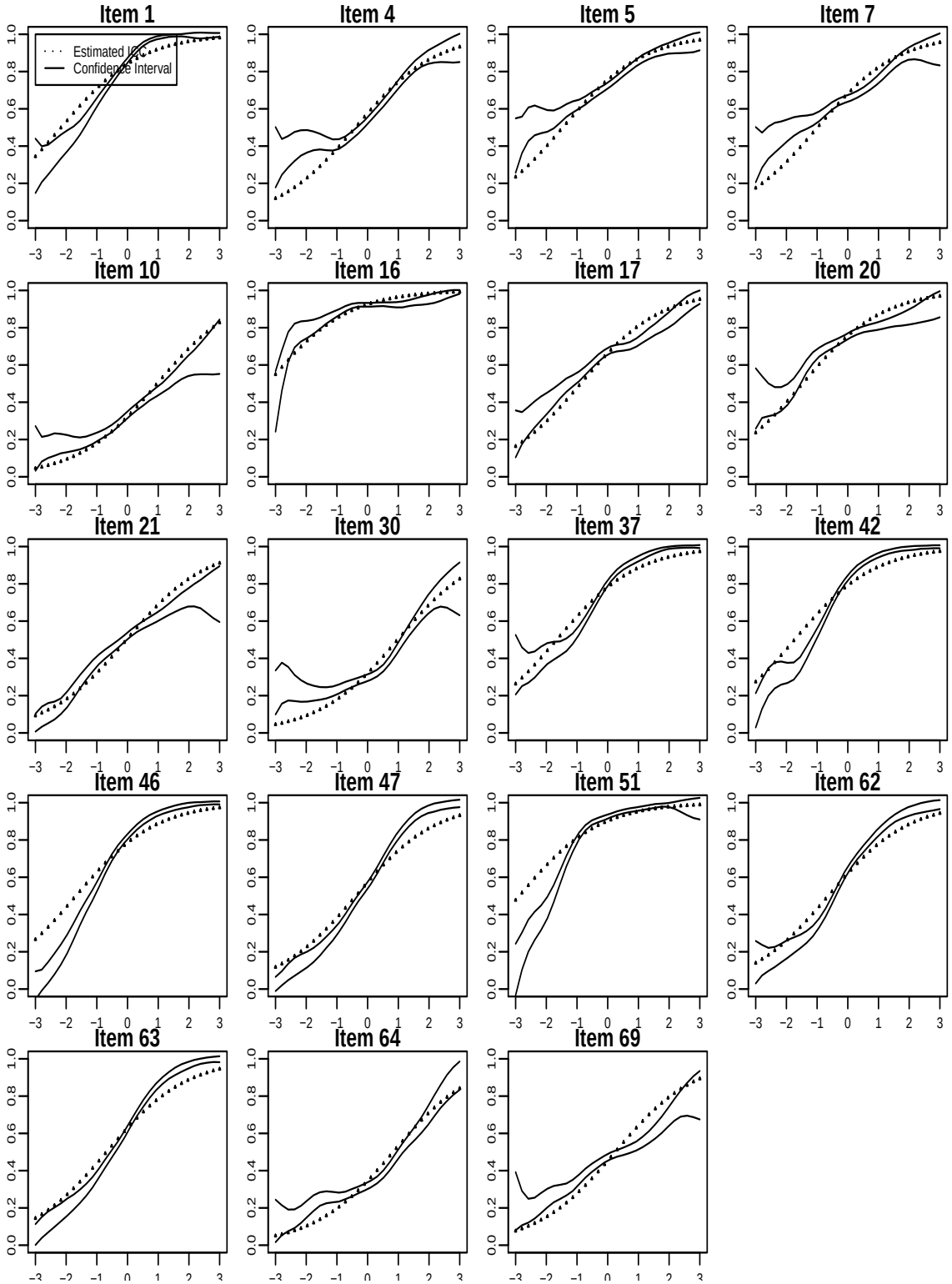

***Figure* 33. Plots of item fit for the battery of examinations—Subject 2 and 1PL model.**

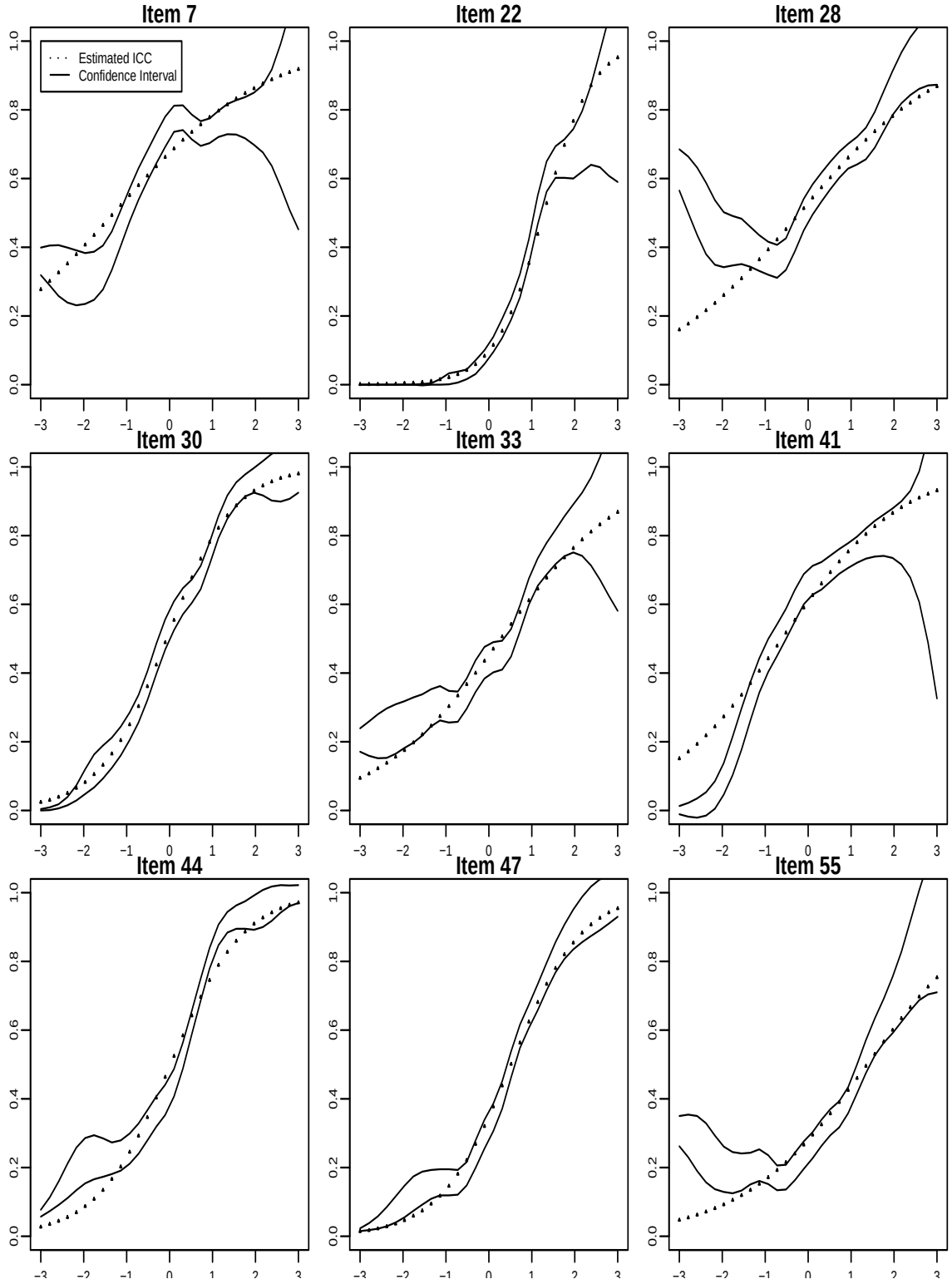

***Figure* 34. Plots of item fits for the battery of examinations—Subject 1 and 2PL model.**

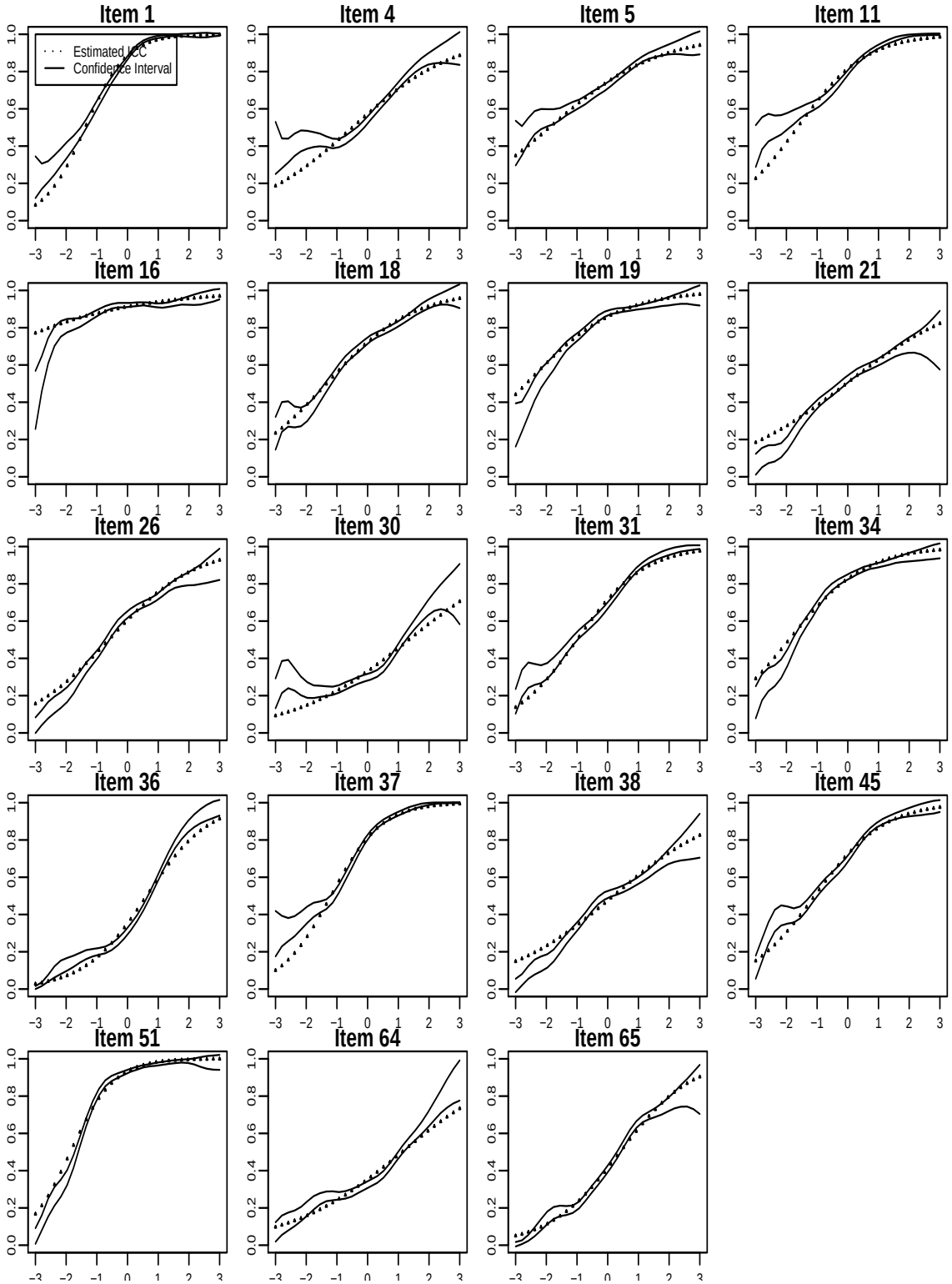

***Figure* 35. Plots of item fit for the battery of examinations—Subject 2 and 2PL model.**

penalty for the 1PL and the 2PL model are given in Table 2.

Table 2

Values of the Penalty Function for the Battery of Examinations

Model	Subject 1	Subject 2
Independence	0.634	0.610
1PL	0.565	0.572
2PL	0.557	0.569

The value of the penalty is also provided for the independence model (Haberman, 2006) that can be used as a baseline model to compare the 1PL model and the 2PL model. In the independence model, the item discrimination parameter a_i is 0 for all items, so that the probability of an answer does not depend on the examinee ability parameter θ. From Table 2, the relative improvement in penalty from the independence model to the 1PL model compared to the relative improvement in penalty from the independence model to the 2PL model is given by

$$\frac{\text{Penalty}_{\text{Independence}} - \text{Penalty}_{\text{1PL}}}{\text{Penalty}_{\text{Independence}} - \text{Penalty}_{\text{2PL}}} = 0.90 \text{ and } 0.93,$$

respectively. This technique for comparing models was used in, for example, Sinharay, Haberman, and Lee (2011). The values of the relative improvement show that the 2PL model is more successful than the 1PL model in explaining the data, but the difference between these two models is small.

A State Test

Next we consider several tests that measure school students' progress toward achieving the academic content standards adopted by a U.S. state in several subjects. These tests describe what students should know and be able to do in each grade and subject tested. We had responses of examinees to two forms each of three subjects of this state test. For each subject, we will refer to the two forms, depending on their date of administration, as the *new form* and the *old form*, respectively. In operational practice, the number-correct raw score of an examinee on a new form is converted to the raw score on the old form using IRT true score equating using the 1PL model with a normal ability distribution, and then to an operational scale. The equating design is the

non-equivalent groups with anchor test (NEAT) design with an internal anchor. The number of operational items on a form for the three subjects are 65, 60, and 75, respectively. Of them, respectively 30, 27, and 29 are internal anchor items.

We fitted the 1PL model and the 2PL model using the marginal maximum likelihood method to these data sets, which had responses of between 31,000 and 75,000 examinees, making them larger than any other data sets analyzed in this paper.

Figures 36 to 38 show the observed and expected marginal score distributions and 95% confidence bounds at each score point for the three subjects. At any score point, an observed proportion that lies outside the 95% confidence bound indicates a significantly large generalized residual. The figure shows strong evidence of misfit of both the 1PL and 2PL models to the marginal score distribution, especially for the first two subjects. Most of the generalized residuals lie beyond the 2.5th or 97.5th percentiles and the observed score distribution mostly lies to the left of the expected score distribution. The observed score distribution of the three subjects appear skewed to the left, slightly bimodal, and skewed to the right, respectively.

Figures 39 to 41 show all pairs of items for which the generalized residual for the statistic p_{11} is statistically significant.

The plots for the 1PL model shows items sorted according to the increasing order of estimated item discrimination parameters from the 2PL fit. In the plots for the 2PL model, the items sorted according to their original order in the test form. The figures show substantial evidence of misfit of both the IRT models to the second-order marginal totals. The percentage of p-values that are significant at the 5% level are between 83 and 90 for the 1PL model and between 71 and 77 for the 2PL model. Thus the 1PL model shows slightly more evidence of misfit to the second-order marginal totals compared to the 2PL model. In addition, for the 1PL model, several residuals involving items with high discrimination are significantly positive and several residuals involving items with low discrimination are significantly negative. A similar observation was made by Sinharay et al. (2006)—see their Figure 4. For the 2PL model, there were fewer positive residuals than negative residuals in both the subjects and there seems to be a speededness effect in Figure 39. The third subject had some reading passages followed by a few items on each passage. Some passage-level dependence is observed in the corresponding Figure 41—for example for the first 3 to 4 items and for Items 23 to 25 of the new form. There are many positive significant residuals among the last few items of Figure 41—this could be a speededness effect, for the last

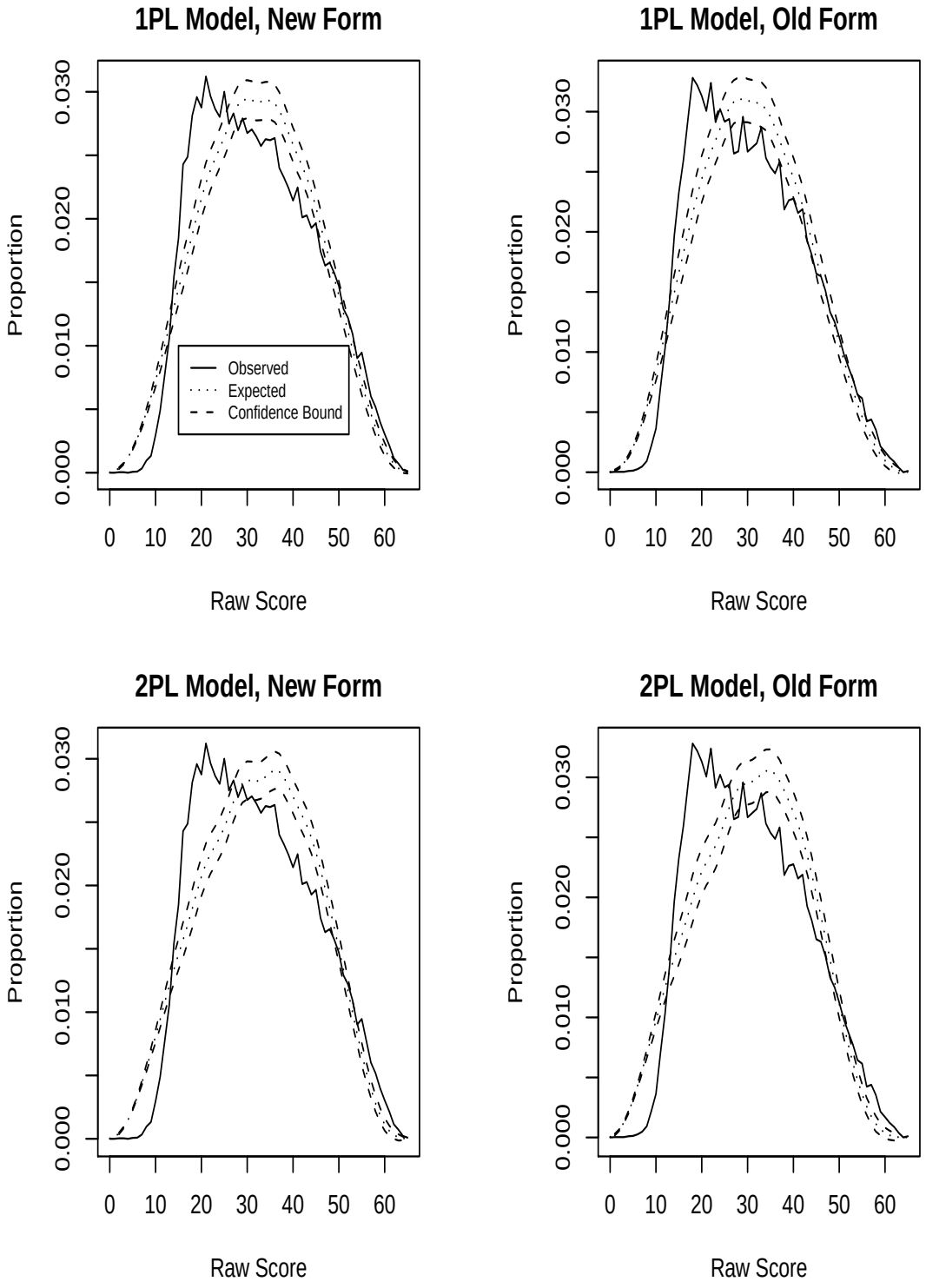

***Figure* 36. Fit of the IRT models to the marginal score distribution for the state test: Subject 1.**

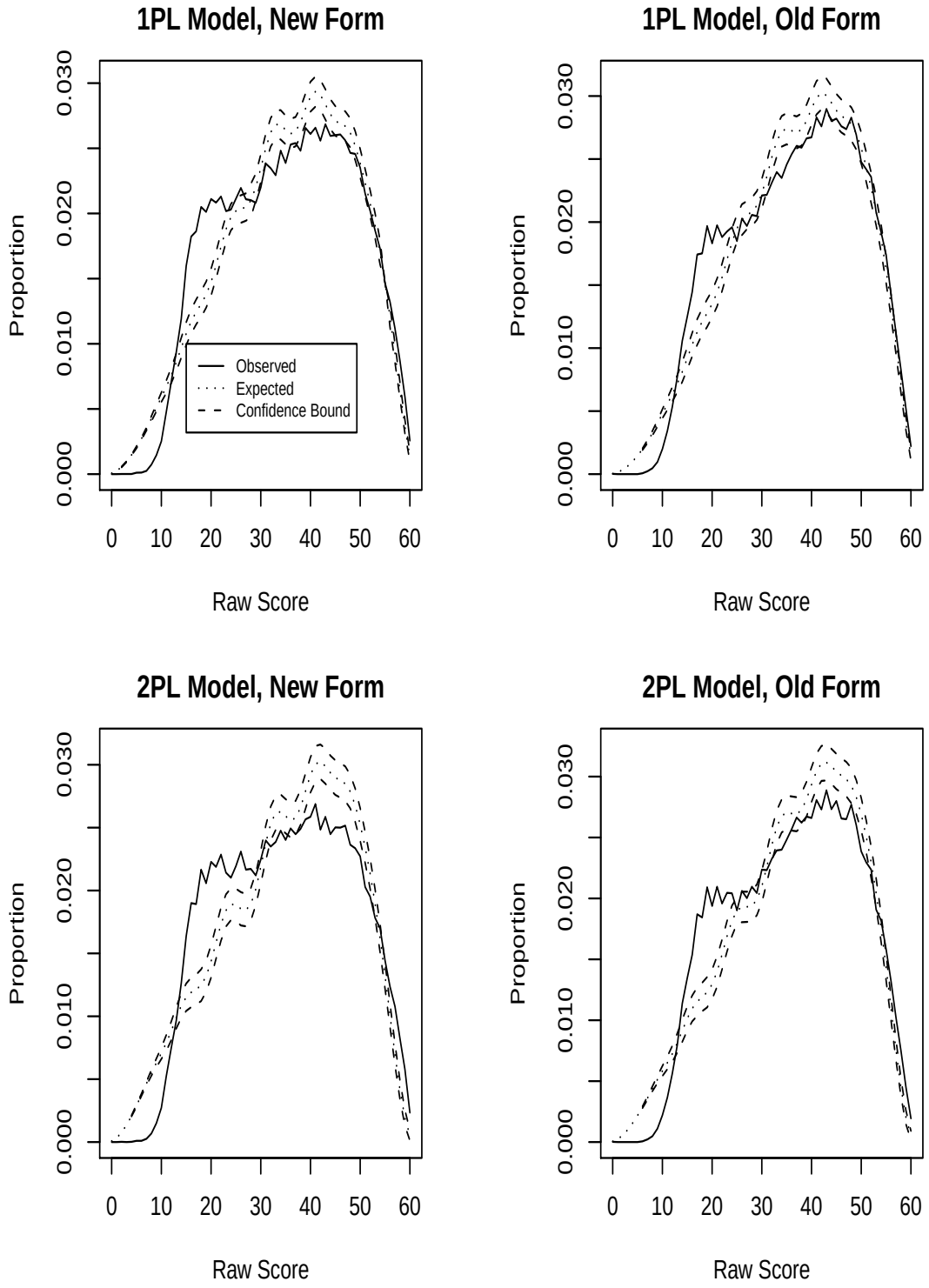

***Figure* 37. Fit of the IRT models to the marginal score distribution for the state test: Subject 2.**

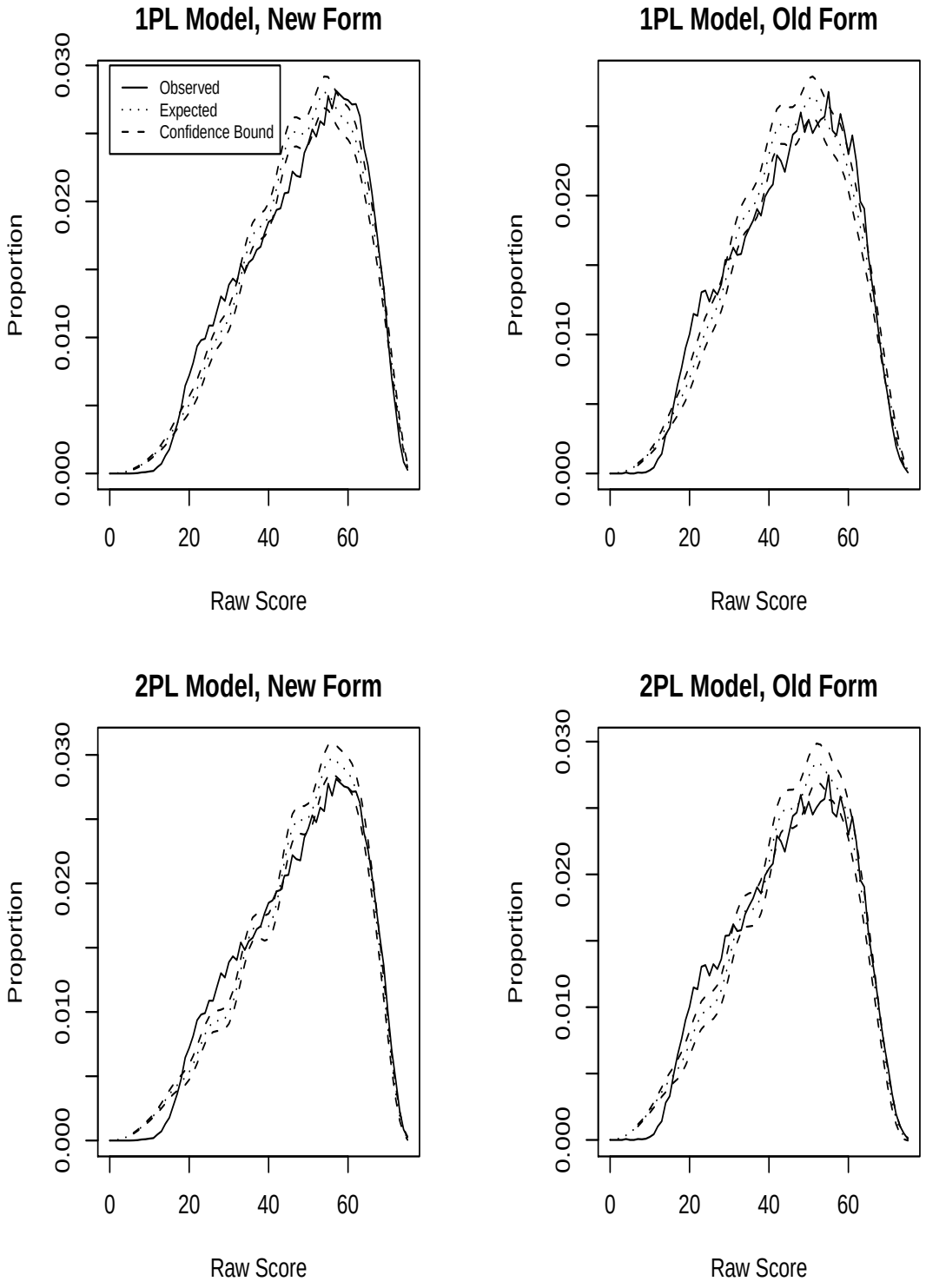

***Figure* 38. Fit of the IRT models to the marginal score distribution for the state test: Subject 3.**

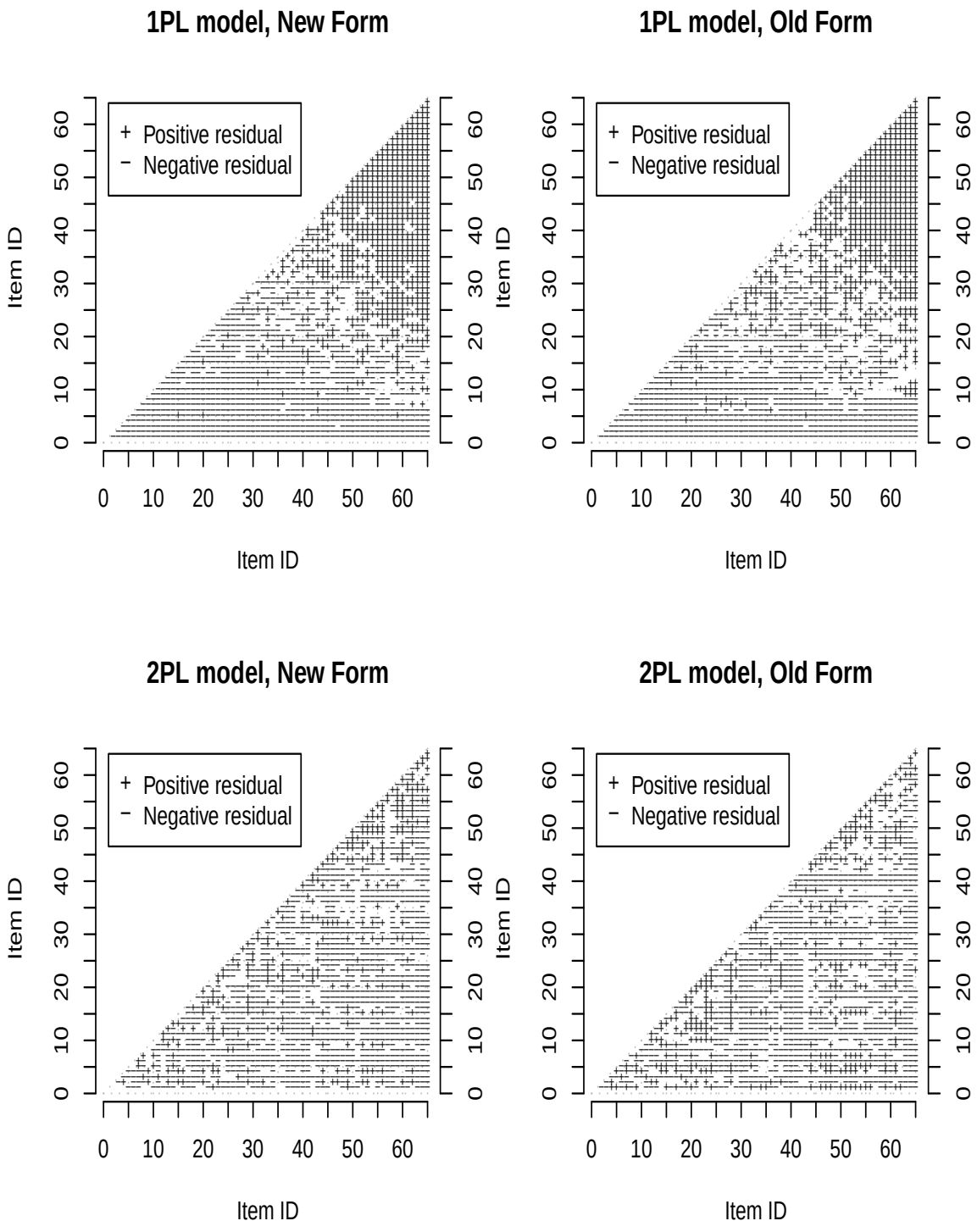

***Figure* 39. The fit of the IRT model to the second-order marginal totals for the state test: Subject 1.**

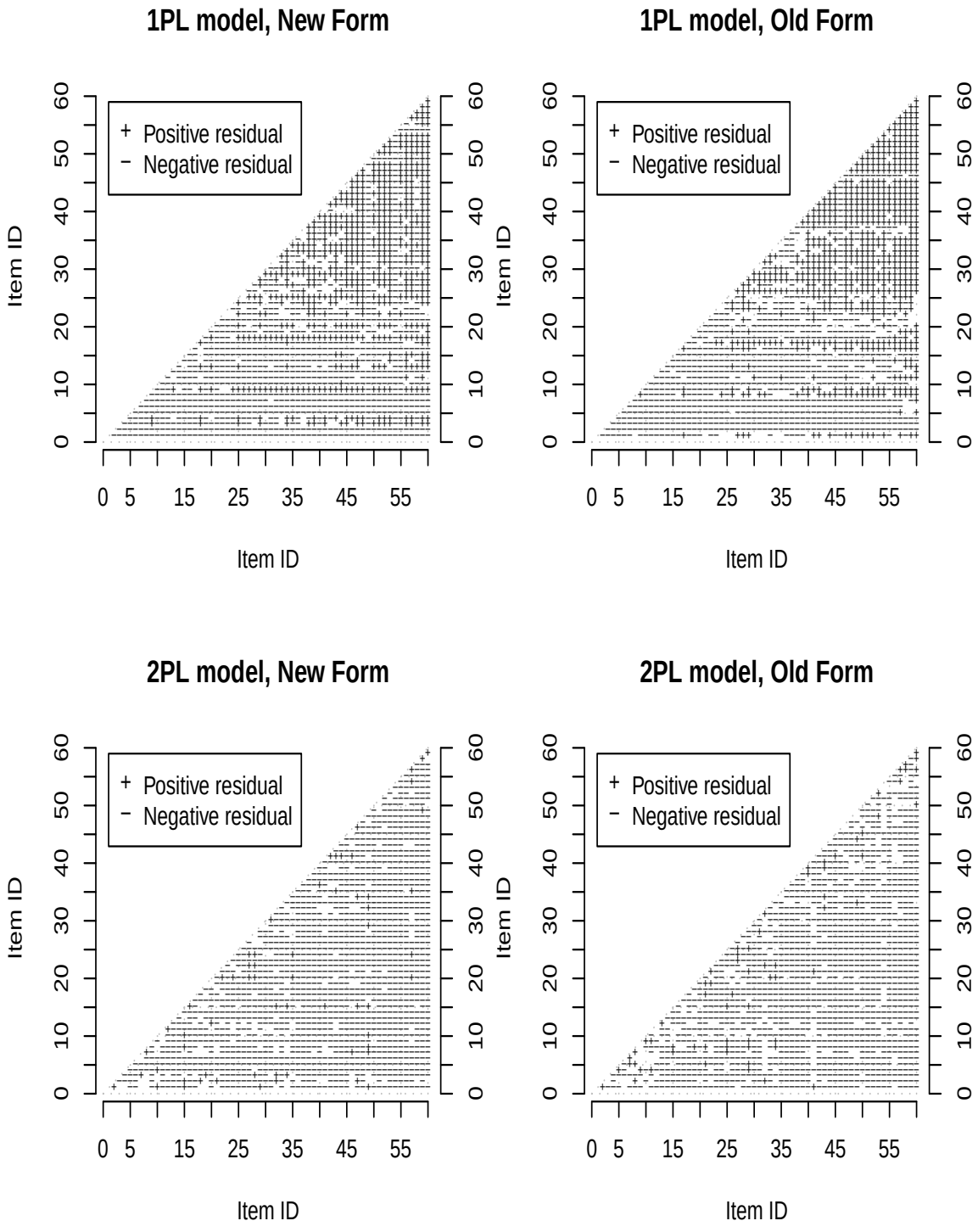

***Figure* 40. The fit of the IRT model to the second-order marginal totals for the state test: Subject 2.**

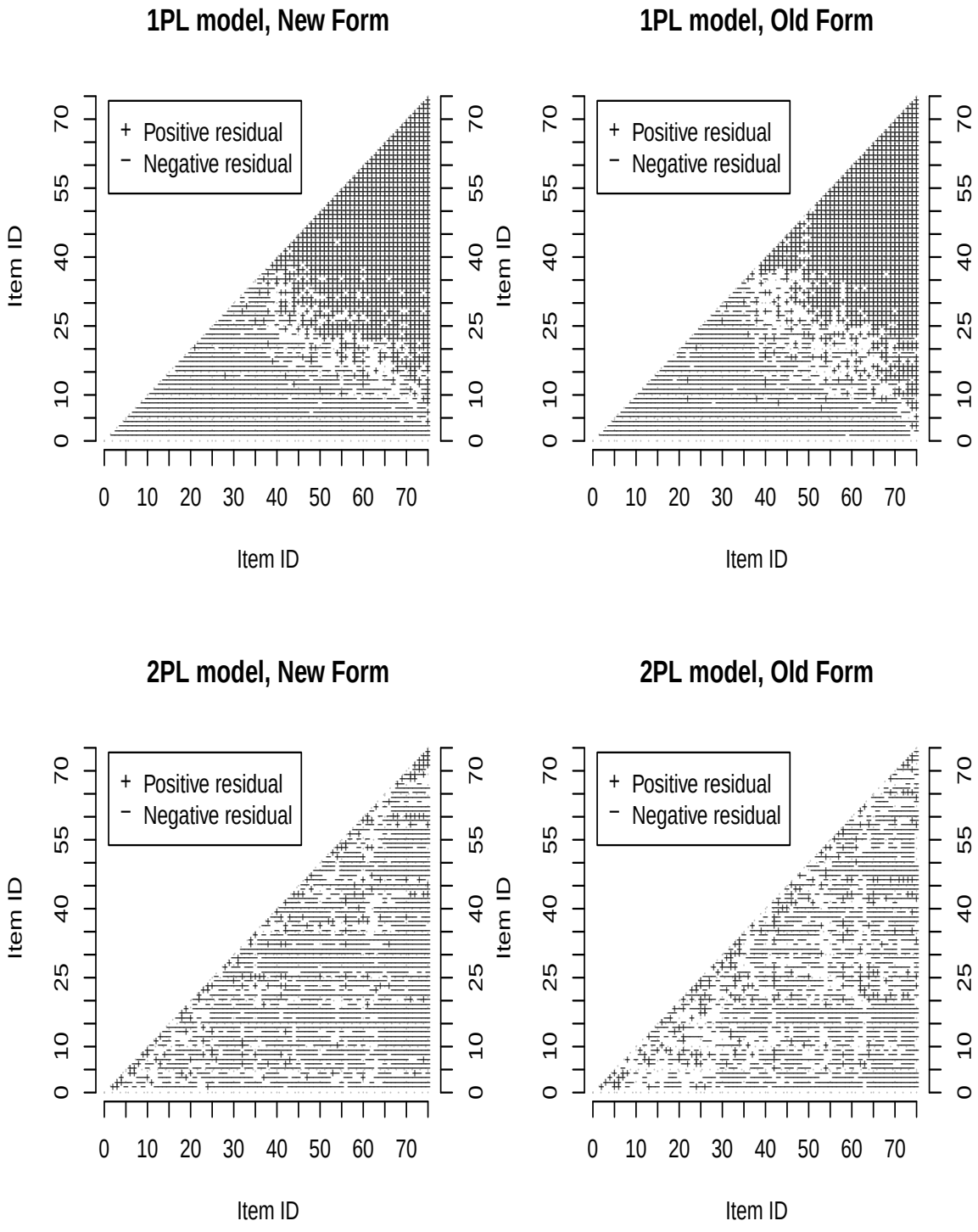

***Figure* 41. The fit of the IRT model to the second-order marginal totals for the state test: Subject 3.**

few items of Subject 3 are discrete items not associated with a common passage.

Almost all items are found substantially misfitting for both the models. Figures 42 to 45 show, for Subject 1, the plots of item fit for items that had the worst fit. Figures 46 to 49 show similar plots for Subject 2. The common items appear at the end in these figures, with their order being the same in both forms of a subject [6]. The figures show that the confidence band around $\hat{I}_j(\theta)$ is very narrow—this is due to the large sample size for these data sets. The 2PL model had better fit than the 1PL model for most items. For example, consider Items 36 and 48 for the new form of Subject 1. While both of these items show misfit for both the models, the values of $\bar{I}_j(\theta)$ are further from the confidence bound around $\hat{I}_j(\theta)$ for the 1PL model.

The Practical Significance of Misfit

The model fit analysis for the state test data sets shows a substantial amount of model misfit. However, given the huge sample size of the data sets, such misfit is not unexpected. It is especially important for this data set to evaluate the practical consequences of misfit.

One way to assess the practical significance of misfit is to examine if the omission of the misfitting items from the anchor test leads to a difference in equating of the raw scores. Figures 42 and 43 show that there is considerable misfit for eight anchor items in both the forms of Subject 1 when the 1PL model is used—these are Items 36, 48, 54, 55, 56, 61, 63, and 65 in both the forms. Figures 46 and 47 show that there is considerable misfit for six anchor items in both the forms of Subject 2 when the 1PL model is used—these are Items 40, 47, 48, 49, 52, and 54 in both the forms. Similarly, for Subject 3, considerable misfit was found for seven anchor items.

The top row of plots in Figure 50 show the impact of omitting these misfitting items (eight for Subject 1, six for Subject 2, and seven for Subject 3) from the anchor set. To obtain the plot, we performed IRT true score equating of the Form 1 score to the Form 2 score for each subject using the 1PL model and the Stocking-Lord method (see, for example, Kolen & Brennan, 2004) twice:

- once using all anchor items and
- once more after omitting the above-mentioned eight (or six) items from the anchor, that is, using the remaining 22 (or 21) anchor items.

These equatings were performed under the assumption that the anchor items were external (if we assume them internal, as done operationally, then a comparison of the equating functions

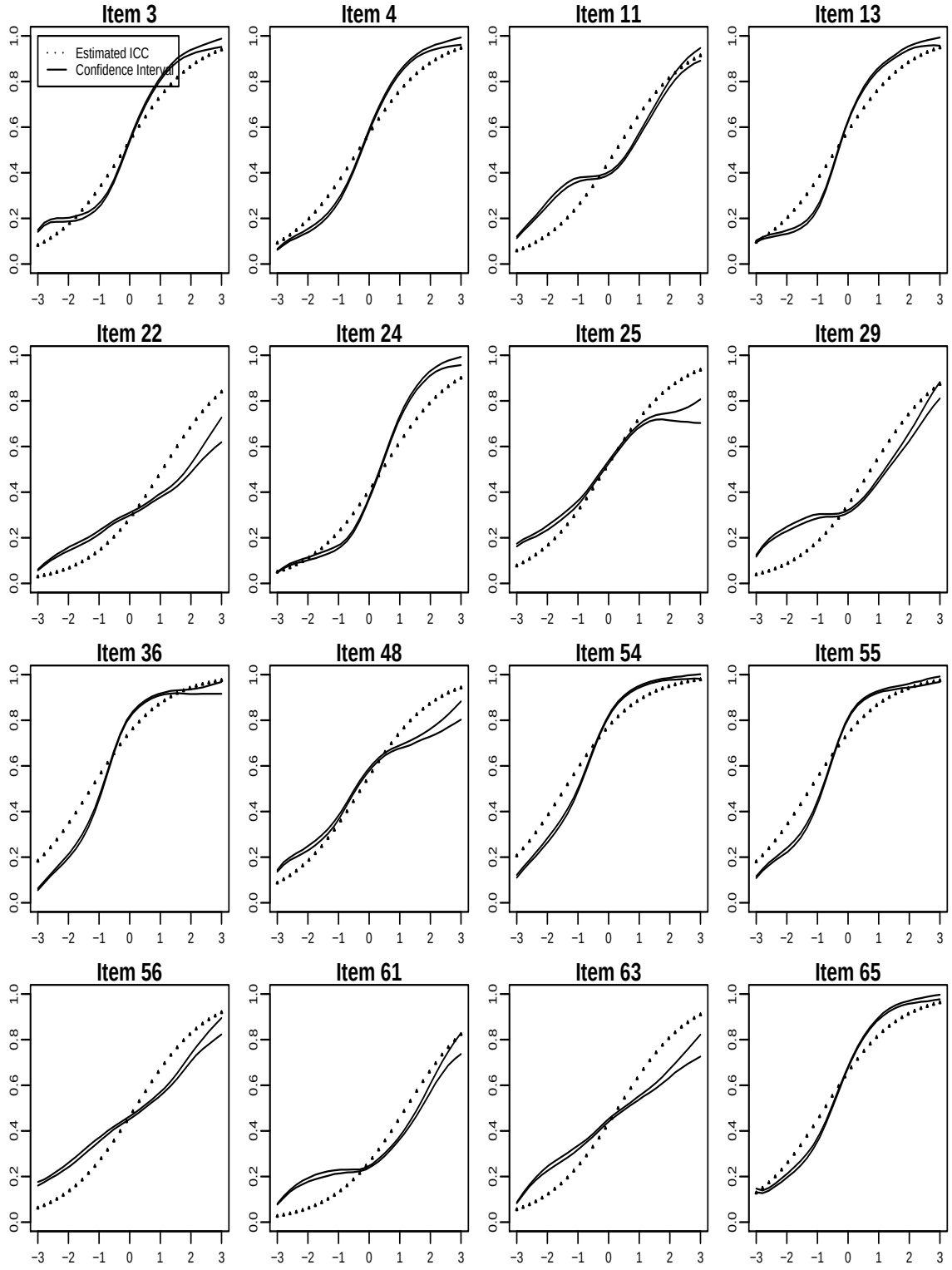

***Figure* 42. Plots of item fit for the state test, Subject 1—New form and 1PL model.**

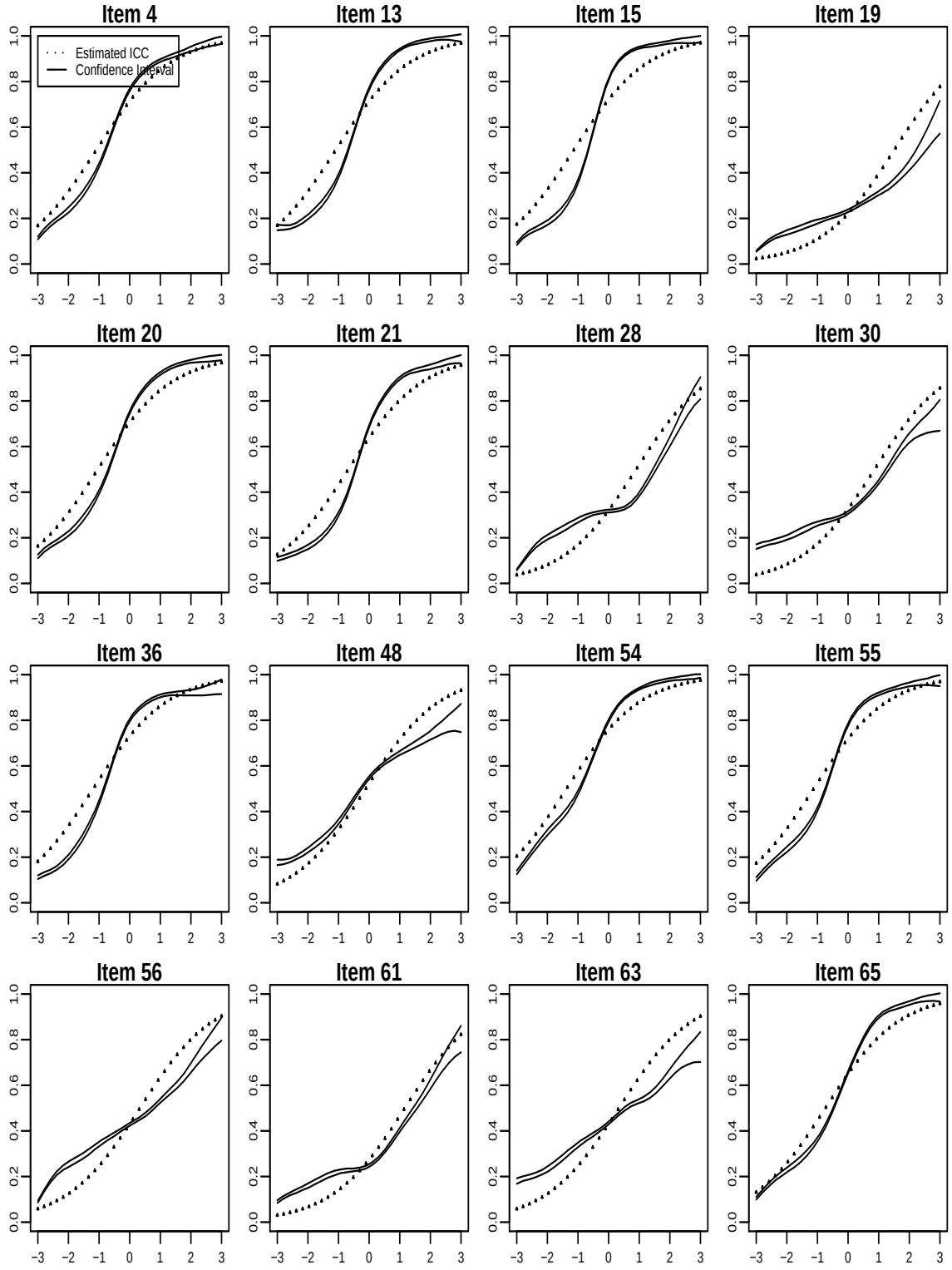

***Figure* 43. Plots of item fit for the state test, Subject 1—Old form and 1PL model.**

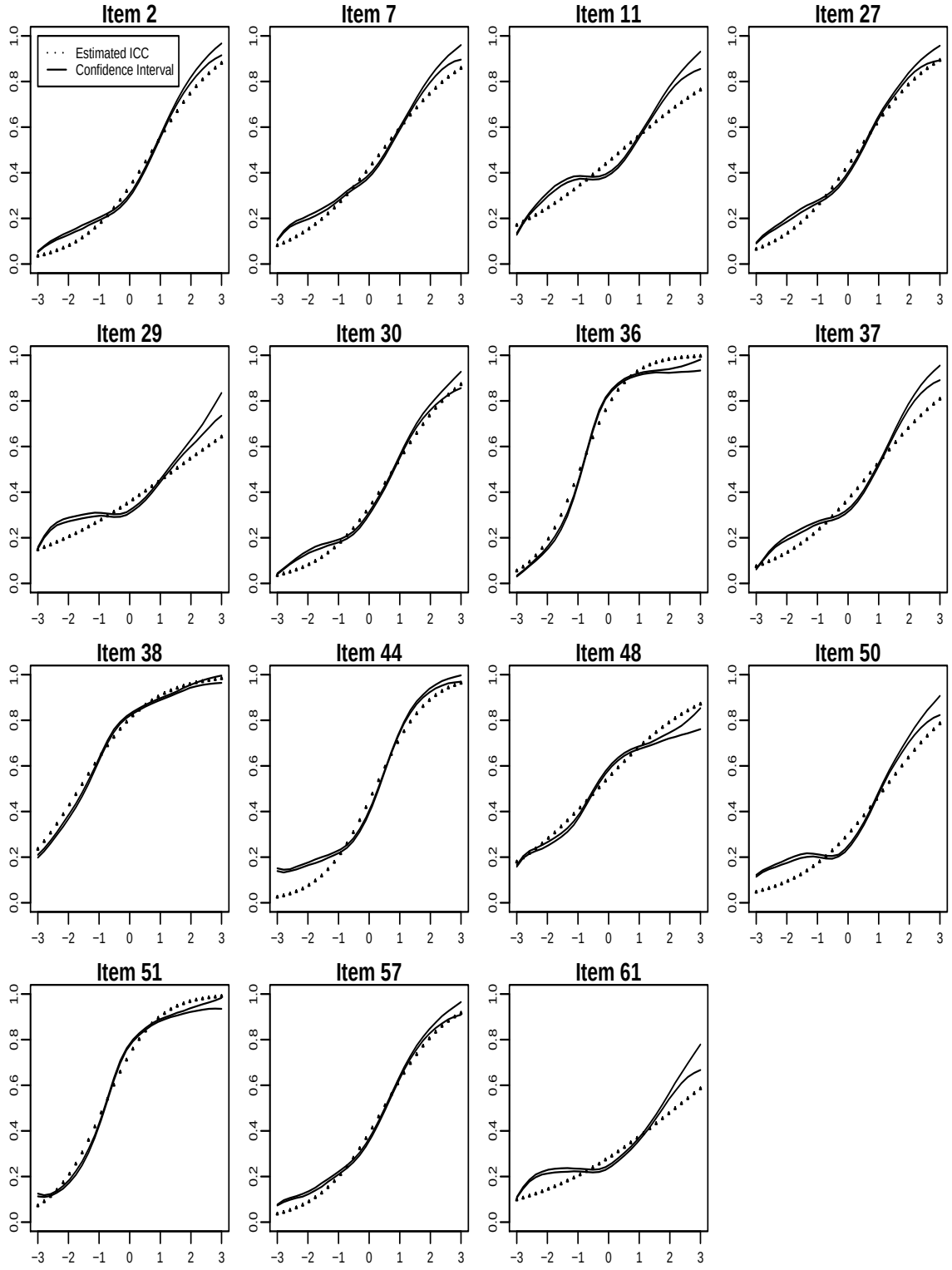

***Figure* 44. Plots of item fit for the state test, Subject 1—New form and 2PL model.**

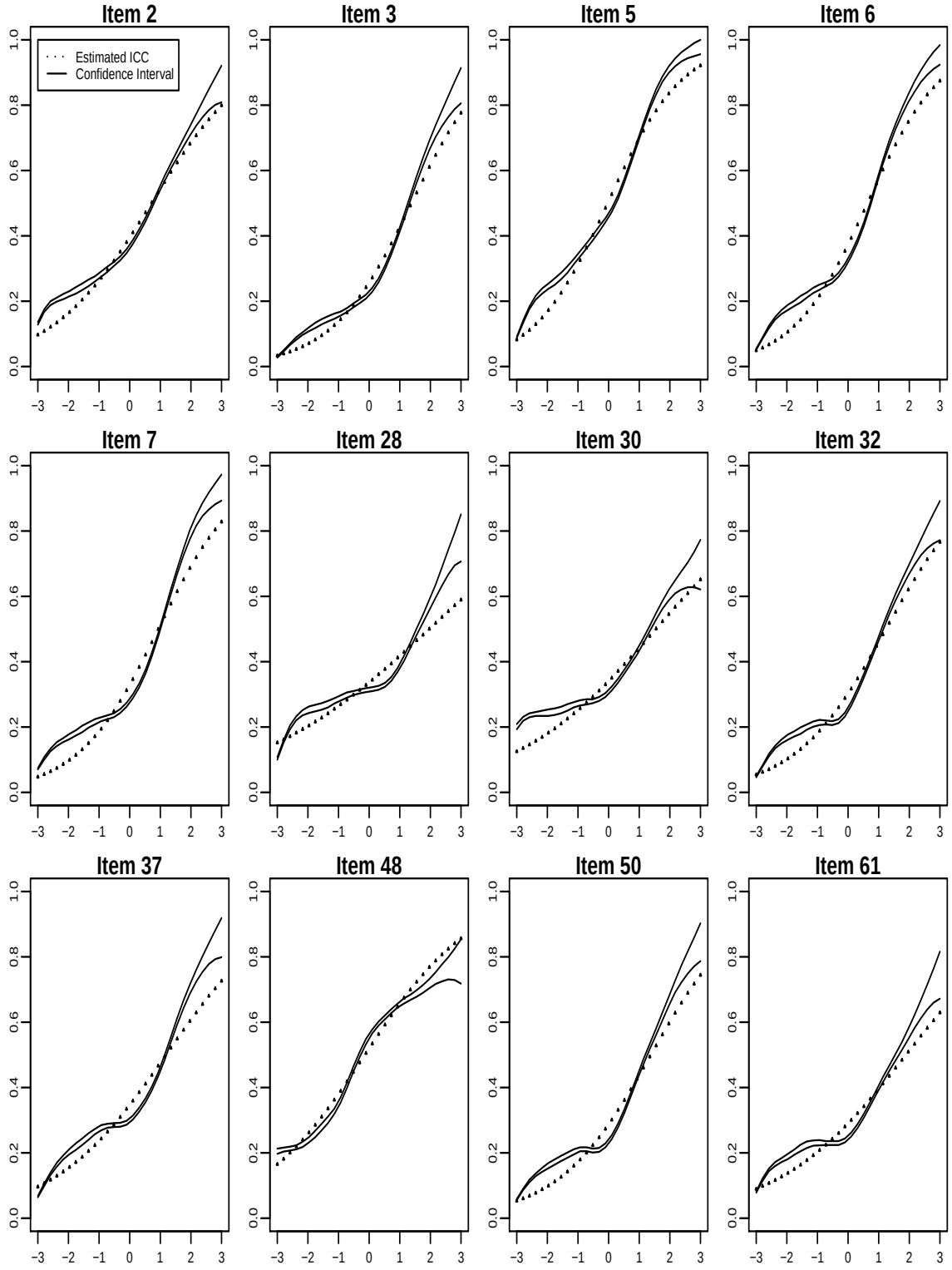

***Figure* 45. Plots of item fit for the state test, Subject 1—Old form and 2PL model.**

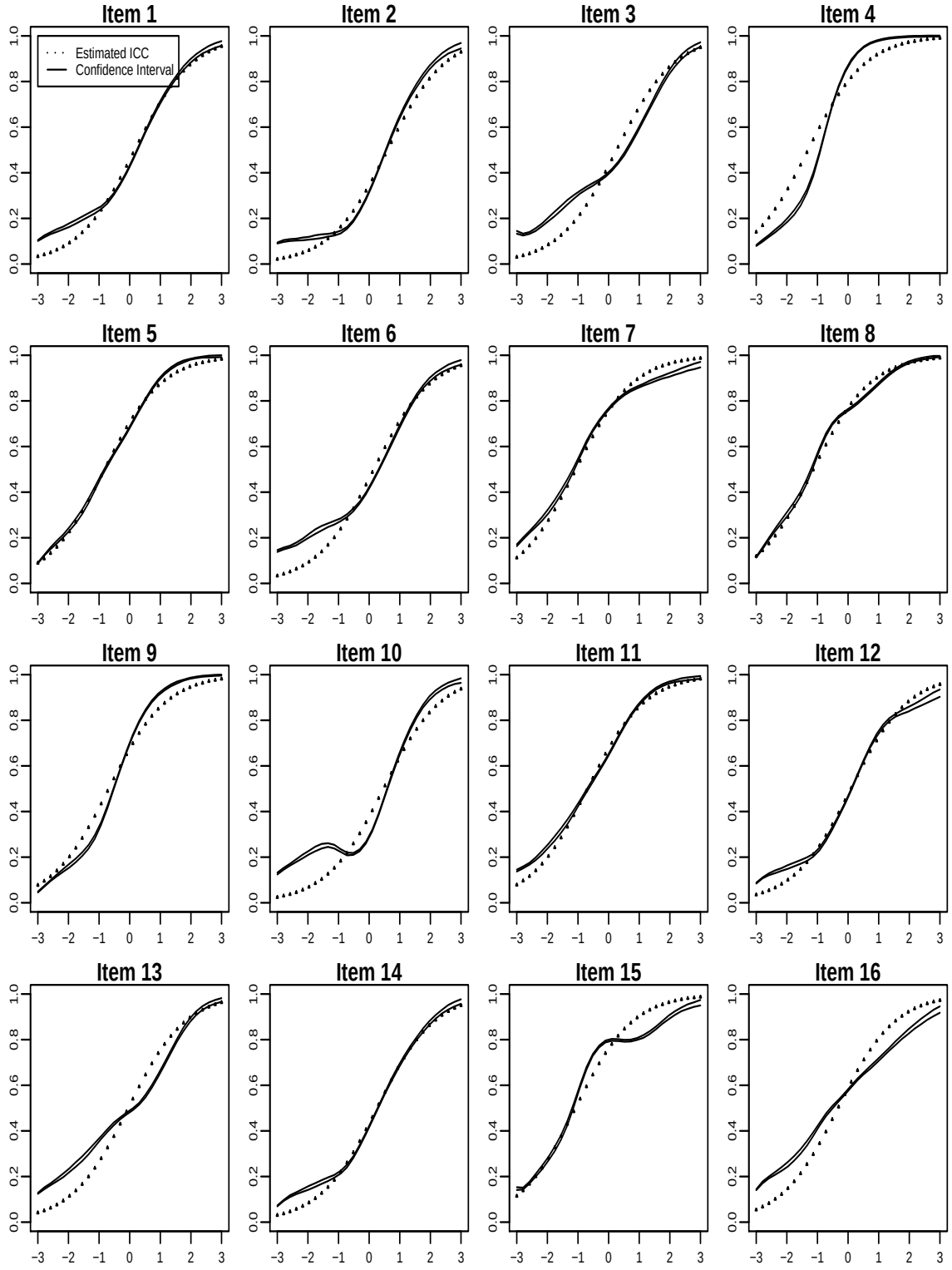

***Figure* 46. Plots of item fit for the state test, Subject 2—New form and 1PL model.**

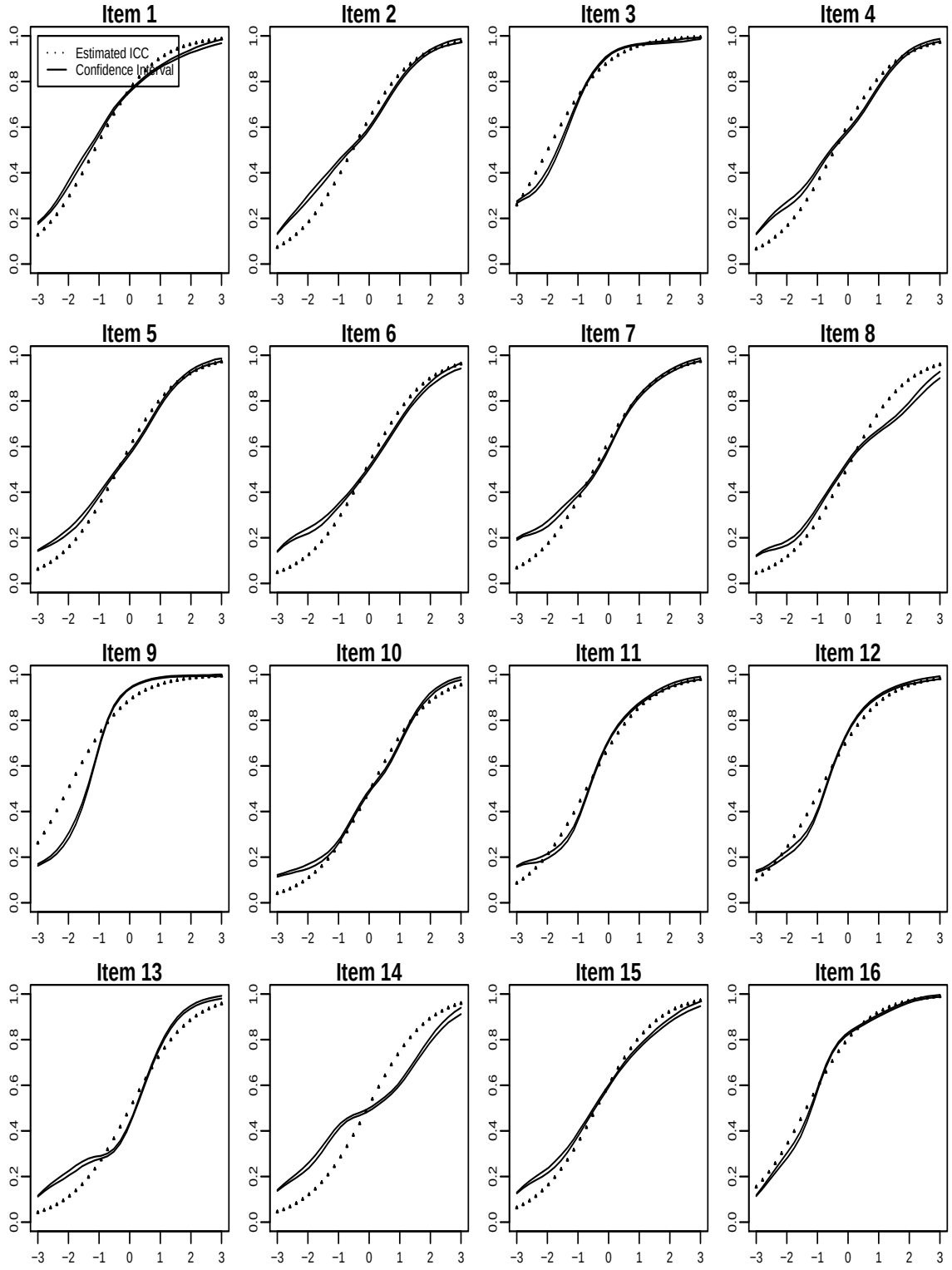

***Figure* 47. Plots of item fit for the state test, Subject 2—Old form and 1PL model.**

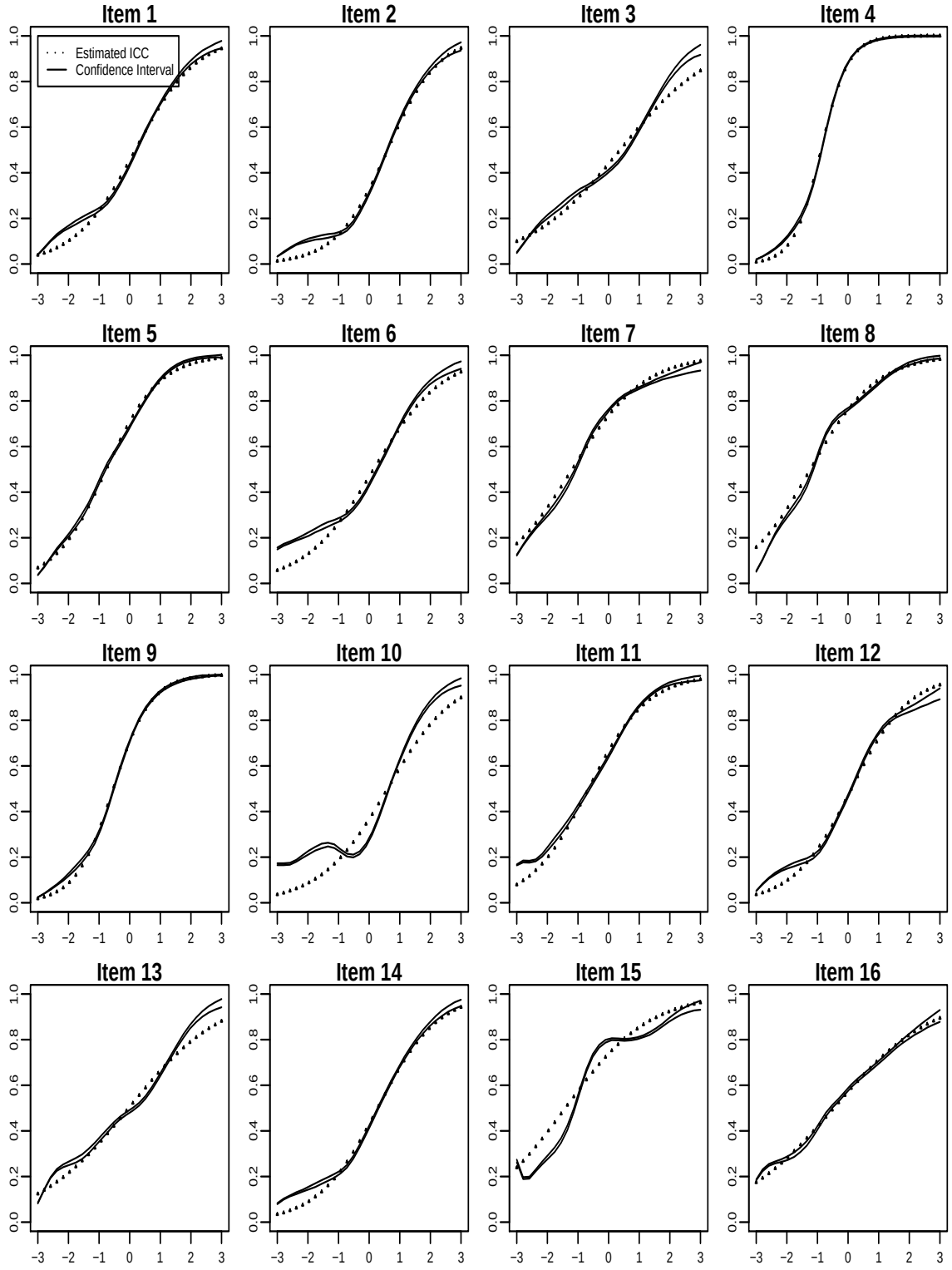

***Figure* 48. Plots of item fit for the state test, Subject 2—New form and 2PL model.**

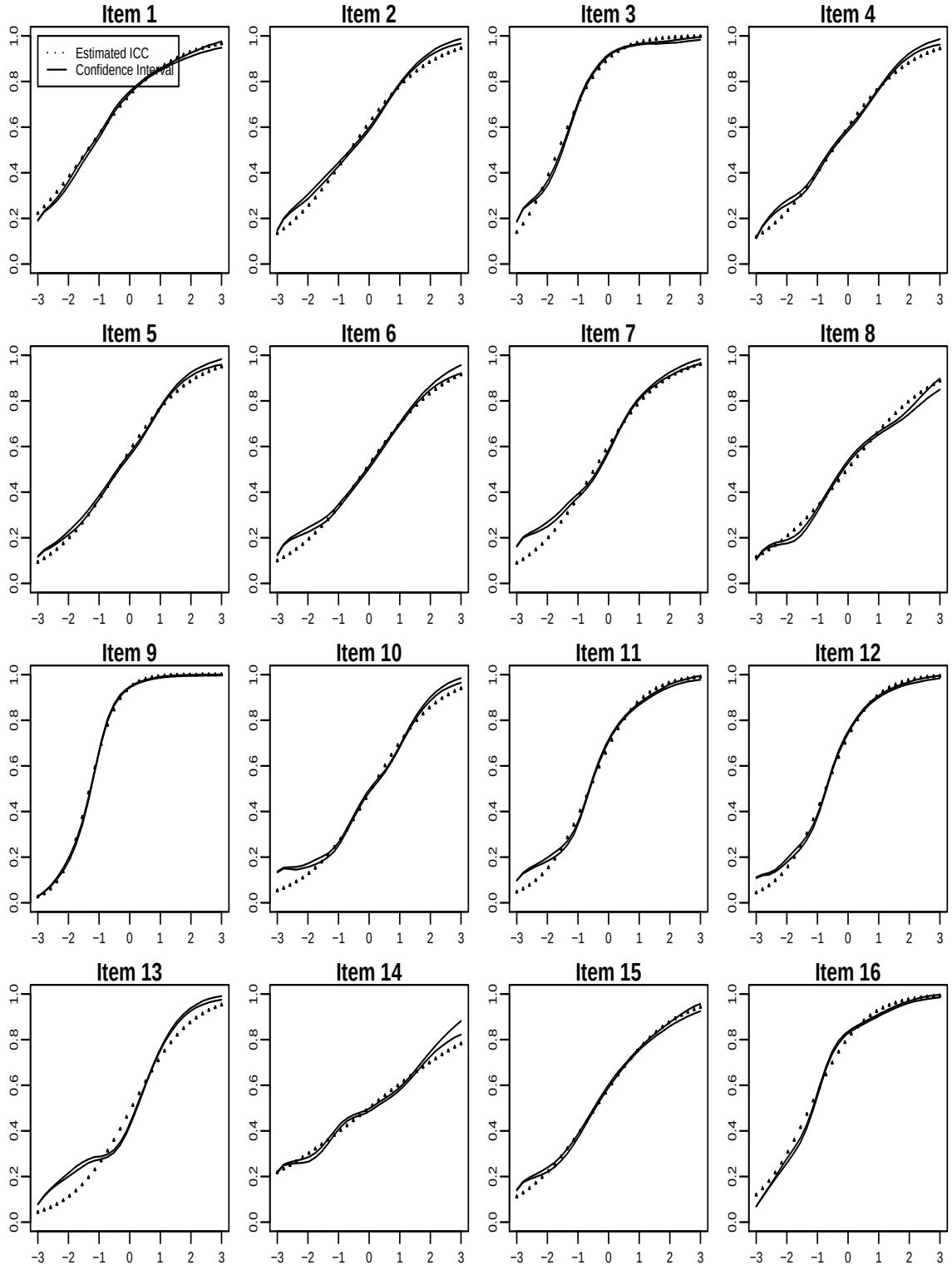

***Figure* 49. Plots of item fit for the state test, Subject 2—Old form and 2PL model.**

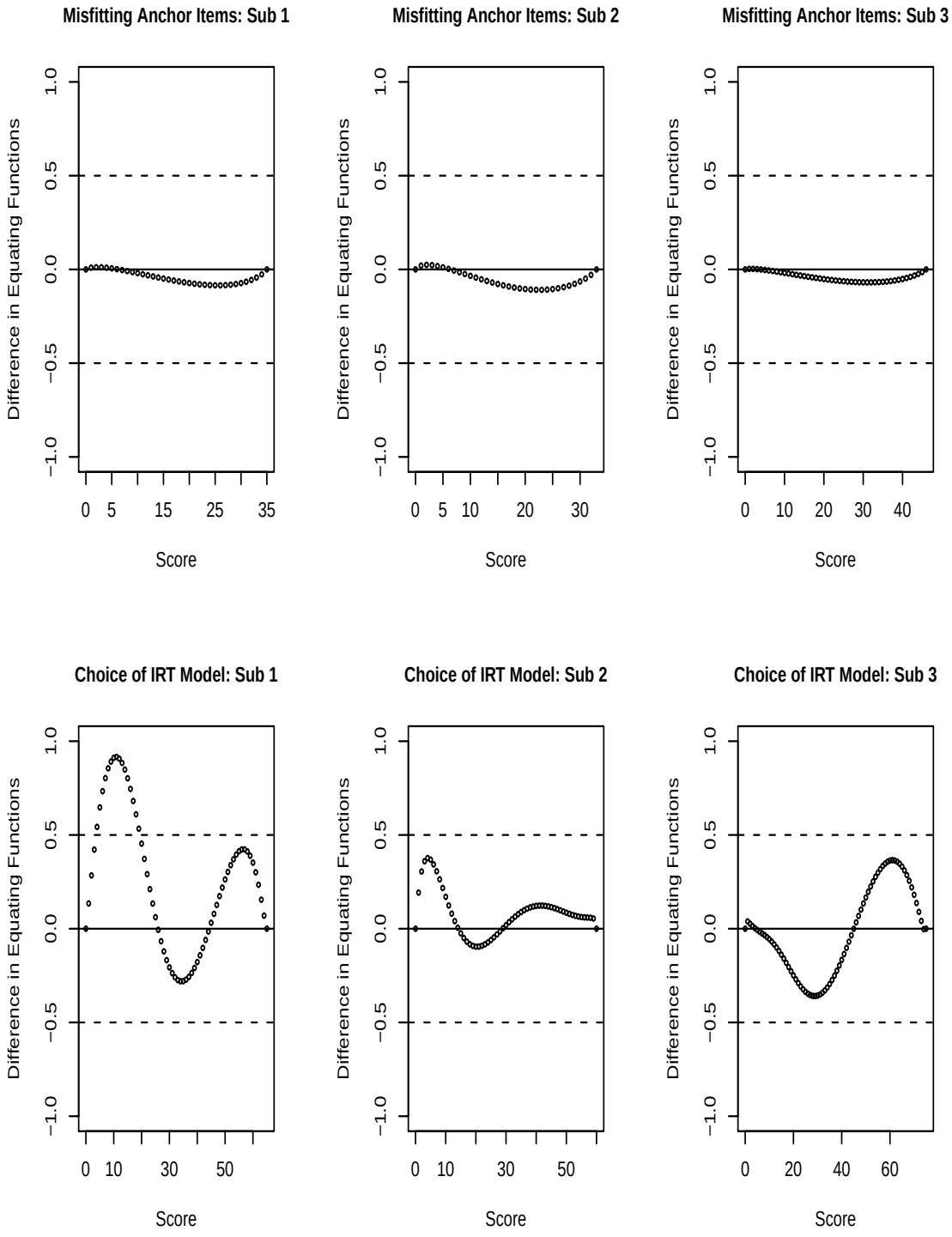

***Figure* 50. Impact of item misfit for the three subjects of the state test**

before and after omitting the misfitting items won't make sense because the test to be equated changes with omission of internal anchor items). The plots in the top row of Figure 50 show the difference between these two equating conversions. The DTM is 0.5 here according to Dorans and Feigenbaum (1994). The differences between the equating conversions are very small and much less than the DTM at all score points for Subject 1. The differences are less than the DTM at all score points for Subject 2, but they are sometimes close to the DTM. Thus, the lack of item fit does not seem to have any practical significance.

The practical significance of misfit due to the choice of the 1PL model compared to the 2PL model can be evaluated by examining the values of the Penalty function for both models for all the data sets. They are given in Table 3.

Table 3

Values of the Penalty Function for the State Test

Model	Subject 1		Subject 2		Subject 3	
	Old form	New form	Old form	New form	Old form	New form
Independence	0.652	0.663	0.646	0.649	0.635	0.618
1PL	0.604	0.611	0.575	0.574	0.575	0.556
2PL	0.597	0.606	0.574	0.571	0.569	0.550

Table 3 shows that the relative improvement in penalty from the independence model to the 1PL model compared to the relative improvement in penalty from the independence model to the 2PL model is 0.87 and 0.91, respectively, for the old form and new form for Subject 1, and 0.99 and 0.96 for the forms for Subject 2, and 0.91 and 0.91 for the forms for Subject 3. Thus, the 2PL model is more successful than the 1PL model in explaining the data, but the difference between these two models is small.

The plots in the bottom row of Figure 50 show the impact of the choice of the IRT model (1PL vs 2PL) on equating. They show for each score point the differences between the IRT true score equating functions obtained using the 1PL model and that obtained using the 2PL model. All the anchor items were used in performing the equatings. There is a horizontal line at 0.5 denoting the DTM criterion (Dorans & Feigenbaum, 1994). The differences between the two equating functions exceed the DTM only for scores around 11 for Subject 1; the difference is close

to 1.0 for several score points. The figure shows that the reported scores of several examinees might change for Subject 1 if the 2PL model is chosen instead of the operationally used 1PL model. The total percentage of examinees who obtained the scores at which the difference between the two equating functions is larger than the DTM is about 20% for Subject 1. The differences between the two equating functions does not exceed the DTM for Subject 2 or Subject 3. Hence, it seems that the choice of the 2PL model over the 1PL model would occasionally lead to a practically significant difference in equating for this test; however, even this practically significant difference is not a huge difference. The 1PL model has often been found to lead to poor equating results (see, for example, Kolen, 1981; Lu & Smith, 2007).

Conclusions

In this paper, we assessed the fit of the IRT models used by several operational testing programs. We used IRT model fit techniques recently suggested by Haberman (2009) and Bock and Haberman (2009).

We found identifiability problems whenever we tried to fit the 3PL model to a data set considered in this study. For the data with which we had problems with the 3PL model, we present results for the restricted 3PL model where we forced the guessing parameters of all the items to be the same.

The generalized residuals were found to be large, most often significantly, for the proportion correct statistic (or 1st order marginal). However, the large residuals are outcomes of the small standard errors of the differences between the observed and expected statistics. The actual difference of the observed and expected proportion correct statistic was almost always very small and hence not practically significant. This finding stresses the need to assess the practical significance of any IRT model misfit.

The generalized residuals for the marginal score distribution were mostly small. The only exceptions were the state tests whose raw score distributions seemed to deviate substantially from a normal distribution.

The generalized residual for the 2nd order marginal total (p_{11}) was often significant. For the basic skills data example, it arose most likely because of speededness. When a 1PL model is fitted to a data set, a large proportion of generalized residuals for p_{11} are statistically significant, for example, as in Figure 31. For several data sets, many of the generalized residuals for the 2nd order

marginal totals were statistically significant and negative, but we do not have a clear explanation. The data sets from the Part 2 of the English proficiency test are examples of that.

The major finding is that the extent of misfit is substantial for almost all the data sets considered in this paper, and especially for the large data sets. This finding echoes the saying of George Box that all models are wrong. The misfit was found to be practically significant for the one basic skills test and in one case for a state test; however, the extent of practical significance was minimal for this state test. The misfit was not practically significant for two basic skills tests, the English proficiency test, the battery of examinations, and in some cases for the state tests. Thus the IRT models, though wrong, were found to be useful for these tests. The practical significance of item misfit, whenever assessed in this paper, was almost negligible—the omission of misfitting items from the test or anchor test never led to a practically significant difference on the IRT true score equating. Together with the finding that the impact of multidimensionality on the quality of IRT true-score equating often appeared to be minimal and of little practical significance (e.g., de Champlain, 1996), our finding shows that the IRT true score equating is quite robust against realistic model violation. It may be possible that the IRT true score equating is robust as long as the nature of the model misfit is the same in the two equating samples. Practical significance could not be assessed for some of the tests considered here.

Note that the extent of practical significance in this study may not generalize to the uses of IRT models that were not examined here. For example, we did not examine the practical significance of misfit in the context of pattern scoring or computer-delivered adaptive testing. The practical significance may be more severe for these kinds of uses. We believe that there is a need to perform further research about the assessment of practical significance of IRT model misfit, especially for these kinds of uses. In addition, for some of the tests such as the state tests, the data we used were already devoid of the items that were found to be unacceptable during pretesting. Therefore, another area of future research is to study the extent of misfit in pretest data and to find ways to determine the practical significance of misfit in pretest data (because, for example, an item that is found poorly fitting in a pretest may not have appeared in an anchor test twice—so some of the analyses we performed to determine practical significance of misfit will not be possible with pretest data).

References

Adams, R. J., Wilson, M. R., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Bock, R. D., & Haberman, S. J. (2009, July). *Confidence bands for examining goodness-of-fit of estimated item response functions.* Paper presented at the annual meeting of the Psychometric Society, Cambridge, UK.

Boughton, K., Larkin, K., & Yamamoto, K. (2004, April). *Modeling differential speededness using a hybrid psychometric approach.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Chon, K. H., Lee, W., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement*, *47*, 318–338.

de Champlain, A. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement*, *33*, 181–201.

DeMars, C. E. (2005). Type I error rates for PARSCALE's fit index. *Educational and Psychological Measurement*, *65*, 42–50.

Dorans, N. J., & Feigenbaum, M. D. (1994). *Equating issues engendered by changes to the SAT and PSAT/NMSQT* (ETS Research Memorandum No. 94-10). Princeton, NJ: ETS.

Gilula, Z., & Haberman, S. J. (2001). Analysis of categorical response profiles by informative summaries. *Sociological Methodology*, *31*, 129–187.

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology*, *18*, 193–211.

Haberman, S. J. (2006). *An elementary test of the normal 2PL model against the normal 3PL alternative* (ETS Research Report No. RR-06-14). Princeton, NJ: ETS.

Haberman, S. J. (2009). *Use of generalized residuals to examine goodness of fit of item response models* (ETS Research Report No. RR-09-15). Princeton, NJ: ETS.

Haberman, S. J., von Davier, M., & Lee, Y. (2008). *Comparison of multidimensional item response models: Multivariate normal ability distributions versus multivariate polytomous distributions* (ETS Research Report No. RR-08-45). Princeton, NJ: ETS.

Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking & D. Revicki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57–78). Washington, DC: Degnon Associates.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement*, *18*, 1–11.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking* (2nd ed.). New York, NY: Springer.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of Royal Statistical Society Series B*, *44*, 226–233.

Lu, Y., & Smith, R. L. (2007, April). *Evaluating the consequences of irt model misfit in equating.* Paper presented at the Annual meeting of the National Council on Measurement in Education, Chicago, IL.

Maydeu-Olivares, A., & Joe, H. (2005). Limited and full information estimation and testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020.

Muraki, E., & Bock, R. D. (1991). *PARSCALE: Parametric scaling of rating data.* Chicago, IL: Scientific Software International.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25–36.

Reckase, M. D. (2007). Multidimensional item response theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Volume 26. Psychometrics* (pp. 607–642). Amsterdam, The Netherlands: North-Holland.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*, 783–801.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*, 375–394.

Sinharay, S. (2008). *A survey of operational practices of several tests that employ item response theory models.* Unpublished manuscript.

Sinharay, S., Haberman, S. J., & Lee, Y. (2011). When does scale anchoring work? a case study. *Journal of Educational Measurement*, *48*, 61–80.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298–321.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2006). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 683–718). Amsterdam, the Netherlands: Elsevier.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York, NY: Springer.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185–201.

Yamamoto, K. (1989). *A hybrid model of IRT and latent class models* (ETS Research Report No. RR-89-41). Princeton, NJ: ETS.

Notes

[1]We imposed this condition because we noticed that for some easy items, the values of both $\hat{I}_j(\theta)$ and $\bar{I}_j(\theta)$ are larger than 0.99 for $0 < \theta < 2$ so that the corresponding residual should not be practically significant even though it is statistically significant.

[2]For example, for Item 2 on reading Form 1, for $\theta = 2.07$, $\hat{I}_j(\theta) = 1$ and $\bar{I}_j(\theta) = 0.998$, so that their difference is quite small; however, the residual is 8.6.

[3]Note that for a pair of binary items, if the generalized residual for the score-category pair $(1, 1)$ is larger than 1.96, we would judge the pair as having a significantly positive association, quite justifiably, according to this rule.

[4]Where we determined whether an item pair has a significantly positive or negative association by the approach described above for the computer-based Science test.

[5]For Part 1, few items in the anchor test showed substantial misfit in both Forms 1 and 2—so evaluation of practical significance was not possible for Part 1.

[6]This reordering is for convenience—the common items are actually spread throughout the form.