Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Spring 2014, Vol. 33, No. 1, pp. 23–35

How Often Is the Misfit of Item Response Theory Models Practically Significant?

Sandip Sinharay, *CTB/McGraw-Hill,* and Shelby J. Haberman, *Educational Testing Service*

Standard 3.9 of the Standards for Educational and Psychological Testing *(1999) demands evidence of model fit when item response theory (IRT) models are employed to data from tests. Hambleton and Han (2005) and Sinharay (2005) recommended the assessment of practical significance of misfit of IRT models, but few examples of such assessment can be found in the literature concerning IRT model fit. In this article, practical significance of misfit of IRT models was assessed using data from several tests that employ IRT models to report scores. The IRT model did not fit any data set considered in this article. However, the extent of practical significance of misfit varied over the data sets.*

Keywords: generalized residual, item fit, residual analysis, two-parameter logistic model

According to 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), evidence of model fit should be obtained when an item response theory (IRT) model is used to make inferences from test data. Several tools for evaluation of fit of IRT models have been suggested by such researchers as Bock (1972), Bock and Haberman (2009), Haberman (2009), Orlando and Thissen (2000), Smith, Schumacker, and Joan Bush (1998), Stone and Zhang (2003), Suarez-Falcon and Glas (2003), and Yen (1981). DeMars (2010) and Swaminathan, Hambleton, and Rogers (2006) provided detailed reviews of the literature on evaluation of fit of IRT models. As described in Hambleton and Han (2005) and Sinharay (2008), analysis of fit of IRT models in operational testing consists of examination of item-fit plots and statistics available from commercial IRT software packages, such as PARSCALE (du Toit, 2003). Misfitting items are often removed from the item pool.

"All models are wrong but some are useful" (Box & Draper, 1987, p. 74). Thus, an IRT model that shows misfit can still be used for some purposes. Therefore, several researchers, such as Molenaar (1997), Hambleton and Han (2005), and Sinharay (2005), recommended the assessment of practical significance of misfit of IRT models, which refers to an assessment of the extent to which the decisions made from the test scores are robust against the misfit of the IRT models. If the misfit is not practically significant, then the misfitting IRT model can be used to make relevant inferences from the data. However, other than Sinharay (2005) and Lu and Smith (2007), there have been few attempts of assessment of practical significance of misfit of IRT models for large-scale tests. Practical significance is not considered in model-fit assessment in commercial IRT software packages and hence not during assessment of model fit in operational testing.

In this article, we provide some general guidelines about assessment of practical significance of misfit of IRT models and perform the assessment using data from:

- a test for measuring proficiency in English,
- three subject areas from a state test, and
- two basic skills tests.

To assess the statistical significant of the fit of the IRT models to these data sets, we employed two recently suggested methodologies: residual analysis to assess item fit (Bock & Haberman, 2009; Haberman, Sinharay, & Chon, 2013) and generalized residual analysis (Haberman, 2009; Haberman & Sinharay, 2013).

The next section includes a description of the two methodologies used in this article to assess fit of IRT models and then includes some ideas regarding the assessment of practical significance of misfit. The Examples section includes the results of assessment of fit of the IRT model for the operational tests. The last section includes the conclusions and recommendations.

Methods

Residual Analysis to Assess Item Fit

To assess item fit, Bock and Haberman (2009) and Haberman et al. (2013) employed a specific form of residual analysis that involves a comparison of two approaches to estimation of the item response function. The analysis leads to residuals which, after being standardized, were proved to have an approximate standard normal null distribution for large samples (Haberman et al., 2013). More details on this analysis can be found in the appendix.

If the model does not fit the data and the sample is large, then several residuals will be significantly larger or smaller than can be expected based on the standard normal distribution. Haberman et al. (2013) showed in a detailed simulation study that for large samples these residuals follow the

Sandip Sinharay, CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940; sandip_sinharay@ctb.com. Shelby J. Haberman, Educational Testing Service, 12T, 660 Rosedale Road, Princeton, NJ 08541; shaberman@ets.org.

standard normal distribution much more closely than the standardized residuals of Hambleton, Swaminathan, and Rogers (1991).

One can create plots of item fit using the above residuals, as shown in Haberman et al. (2013) and as described in the appendix.

Figure 1 shows examples of such plots for two items. For each item, the examinee ability θ is plotted along the x-axis, the dotted line denotes the estimated item characteristic curve for the item, and the two solid lines denote a pointwise 95% confidence band. A dot outside this confidence band would indicate a statistically significant residual. These plots are similar to the plots of item fit provided by IRT software packages, such as PARSCALE (du Toit, 2003), except that the plots in Figure 1 are accompanied by a theoretical proof regarding the null asymptotic distribution of the residual. In Figure 1, the left panel corresponds to an item for which no statistically significant misfit is observed (the dotted line almost always lies within the 95% confidence band) and the right panel corresponds to an item for which substantial misfit is observed. The software program described in Haberman (2013) was used to compute these residuals. The program is available on request for noncommercial use.

Generalized Residual Analysis

Generalized residual analysis for assessing the fit of IRT models was suggested by Haberman (2009) and Haberman and Sinharay (2013). In this analysis, a test statistic T, its estimated mean $\hat{E}(T)$, and an estimated standard deviation s_D of the difference $T - \hat{E}(T)$ are computed under the assumption that the IRT model fitted to the data provides a perfect fit. One then computes a generalized residual

$$g = \frac{T - \hat{E}(T)}{s_D}. \qquad (1)$$

More details on these residuals can be found in the appendix. Haberman (2009) and Haberman and Sinharay (2013) proved that if the IRT model fits the data adequately and the sample is large, the distribution of g is well-approximated by the standard normal distribution. Thus, a statistically significant value of the generalized residual g indicates that the IRT model does not adequately predict the statistic T. The method is quite flexible. Several common data summaries such as the item proportion correct, proportion simultaneously correct for a pair of items, and observed score distribution can be expressed as the statistic T. It is possible to create graphical plots using these generalized residuals. For example, one can create a plot showing the values of T and a 95% confidence interval given by $\hat{E}(T) \mp 1.96 s_D$. A value of T lying outside this confidence interval would indicate a generalized residual significant at 5% level. The software program described in Haberman (2013) was used to perform the computations for the generalized residuals.

The Reason Behind the Choice of the Two Aforementioned Methodologies

Because this article concerns assessment of practical significance of model misfit and not the assessment of any particular method to assess model misfit, any IRT model fit technique could have been used before our assessment of practical significance of model misfit. We employed the two aforementioned IRT model fit methods because of the following:

- Both of these techniques have the advantages of being forms of residual analysis.
- These techniques are supported by theoretical proofs, which are mostly lacking in the IRT model fit literature. The relevant statistics follow an asymptotic $\mathcal{N}(0, 1)$ distribution when the IRT model is a good fit to the data.
- These techniques are not computationally intensive.
- The technique of Haberman (2009) offers a framework to assess several aspects of IRT model misfit.
- In simulation studies performed in Haberman et al. (2013) and by us, both these techniques were found to have satisfactory Type I error rate and power.

The IRT model fit statistics available in IRT software packages were not used here. Several of these have limitations such as uncertain sampling distribution (see, e.g., Hambleton & Han, 2005). Also, researchers such as Chon, Lee, and Dunbar (2010), Chon and Sinharay (in press), and Glas and Suarez-Falcon (2003) found the Type I error rates of such statistics too high.

Assessment of Practical Significance of Misfit of IRT Models

Assessment of practical significance of misfit of IRT models involves the determination of the extent to which the decisions made from the test scores are robust against the misfit of the IRT models. In performing this assessment, one needs to have the two following pieces of information:

- Information on how the IRT model is used to compute the reported scores.[1] For example, in several assessments, including those considered in this article, an IRT model is used to equate the raw scores, while in some other assessments an IRT model is used for *pattern scoring*.
- Information on how the reported scores are used and/or what decisions are made from the test scores. For example, scores on the TOEFL iBT® test are used for admissions decisions in colleges and universities and in satisfying visa requirements in Australia and the United Kingdom (http://www.ets.org/toefl/ibt/about), while scores on the National Assessment of Educational Progress are used to determine what America's students know and can do in various subject areas.

Assessment of practical significance of misfit of IRT models should involve application of a variety of techniques—a test statistic or a model fit plot can rarely provide much insight on the practical significance of model misfit. Ideally, this assessment would involve an examination of the agreement between the decisions from the test scores and the *ideal decisions*, which are decisions from the test scores under a perfectly fitting model. Unfortunately, this examination is impossible because of the lack of a model that perfectly fits item response data; for example, Sinharay, Haberman, and Jia (2011) found in a survey of data from several operational tests that the operational IRT model did not fit any of those data. Also, if a perfect model existed, the testing company would probably have used that. One way around is to compute test scores under a model–data combination with better fit and replace the *ideal decisions* by decisions from these scores. A model–data combination with better fit could be obtained by the use of a more general IRT model, exclusion of a few misfitting items from the item pool, combining unpopular score categories of polytomous items with popular ones, exclusion of a few examinees from the data set, and so forth. For example, for the basic skills tests to be discussed later, the scores are

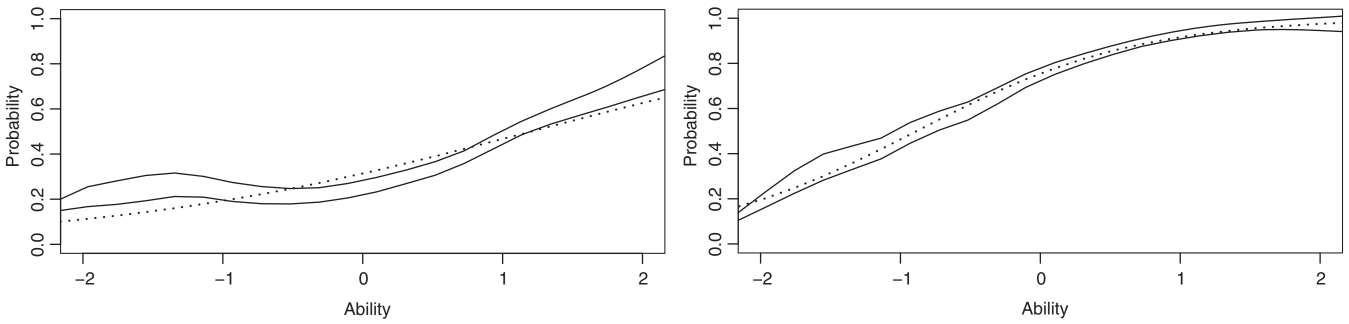

FIGURE 1. Examples of plots of item fit.

used in making pass–fail decisions (where a "pass" decision would lead to the licensing of the examinee concerned) and the operational IRT model was found to provide a poor fit to a few items; the agreement was computed between the operational pass–fail decisions and pass–fail decisions obtained after omitting a few misfitting items. The extent of this agreement provided an idea regarding the practical significance of the observed item misfit. If the decisions from the operational test scores agree well with the decisions from the test scores obtained under better model fit, we would conclude that the model misfit is not practically significant and proceed with the existing model.

If, however, there is a poor agreement between the decisions from the test scores and the decisions from the test scores under better model fit, we would conclude that the model misfit is practically significant. There are several possibilities for the next step in this case. If the misfit is found in the development stage, for example, during a pilot study, the investigator may want to change several items or use a more general psychometric model and recompute the scores. However, if the misfit is found during a regular administration of an operational test, the psychometric model often is decided upon by an agreement between the testing company and the client for whom the test was prepared and cannot be changed. In such cases, the only option is to try to improve fit by, for example, omitting a few misfitting items (while making sure that the remaining items lead to a score that is adequately reliable), combining unpopular score categories of polytomous items with popular ones, or, less preferably, omitting a few misfitting persons, and to use the scores obtained under improved fit.

The flow chart in Figure 2 shows the steps involved in a thorough IRT model-fit assessment. Assessment of both statistical significance and practical significance are included in the steps.

In some applications of IRT models, for example, for tests whose scores have multiple uses, one may have to assess the practical significance of misfit in multiple ways, one for each type of use. In some cases, assessment of practical significance of misfit is quite difficult or even impossible. One example of such cases would be tests whose scores form a part of a decision-making process that cannot be easily quantified. For example, SAT® scores are used by colleges to make admissions decisions, but other information such as essays and recommendations also play a role. Therefore, it would be quite difficult to assess the practical significance of misfit for SAT. Another example would be when it is quite difficult or impractical to obtain decisions under a better model fit. Suppose one finds multidimensionality involving several items in a test for which a unidimensional IRT model is used for IRT true score equating; it is not clear how one can obtain a true score equating of the overall score under a multidimensional IRT model in this case—so the assessment of practical significance of the misfit is quite difficult in this case. Computerized-adaptive tests are another set of example for which assessment of practical significance of misfit is quite difficult because it is impossible to know what an examinee would have scored if a particular item she answered were omitted.[2]

Assessment of the practical significance of misfit may involve several levels of analysis. Consider a test for which an IRT model is used to obtain an equated score that is later used for making pass–fail decisions and the IRT model was found to misfit several items. An investigator could assess the practical significance of misfit for this test by comparing the operational equating conversion with an equating conversion obtained after omitting the misfitting items. However, another investigator could go further and, to assess the practical significance of misfit, compare the operational pass–fail decisions with pass–fail decisions obtained after omitting a few misfitting items. It is possible to go even further and consider the question "How did the model misfit affect the consequences of the pass–fail decisions?" For example, for a teacher certification examination, the consequences of the pass–fail decisions would involve future student performance, the future salary of those who wrongly passed and who wrongly failed. However, data are rarely available to answer those questions. In addition, a judgment is often needed during the assessment of the practical significance of misfit on how large is large. For example, as implied by Hambleton and Han (2005), if the percentage of candidates on a state assessment judged as proficient is 60% with the operational IRT model and 52% with a better-fitting IRT model, the model misfit is practically significant. But suppose the latter percentage is 58% instead of 52%. Then, whether the misfit is practically significant depends on what the investigator judges as a large difference. Thus, in some sense, assessment of the practical significance of misfit is a never ending process and even a Sisyphean task, somewhat similar to validation that is a never-ending process (e.g., Messick, 1980). However, users of IRT models should not be deterred by that fact and should try their best to perform at least some assessment of the practical significance of misfit (just like those performing validation try their best to collect validity evidence).

Examples

We next describe the results of IRT model fit analyses using data from the above-mentioned operational tests. For each

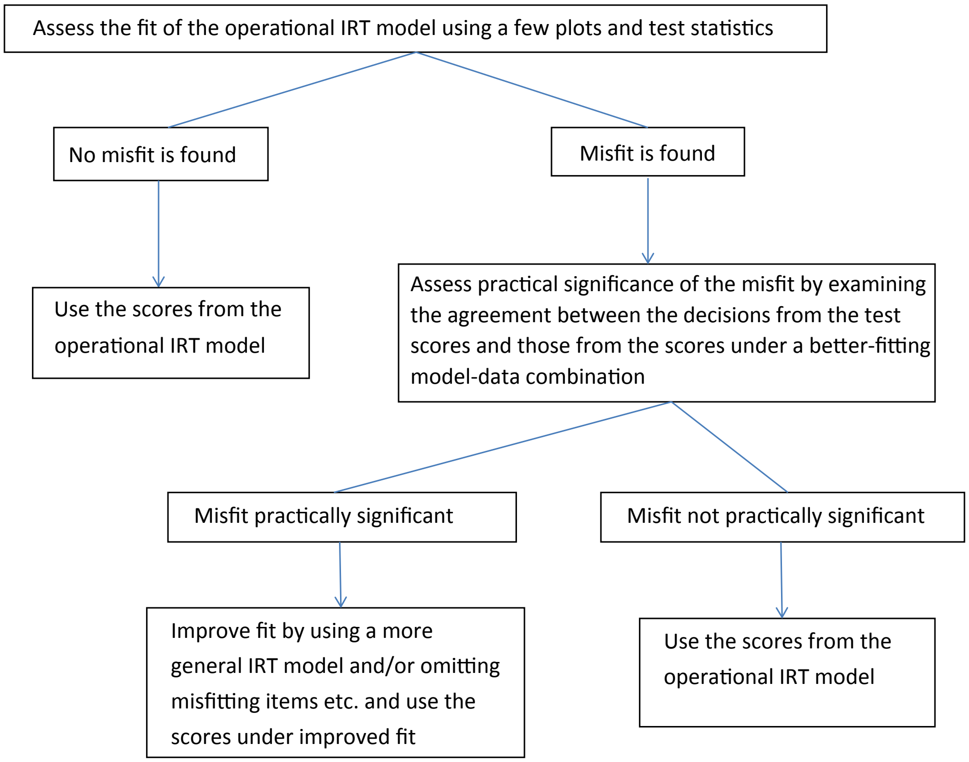
Assess the fit of the operational IRT model using a few plots and test statistics
No misfit is found
Misfit is found
Use the scores from the operational IRT model
Assess practical significance of the misfit by examining the agreement between the decisions from the test scores and those from the scores under a better-fitting model-data combination
Misfit practically significant
Misfit not practically significant
Improve fit by using a more general IRT model and/or omitting misfitting items etc. and use the scores under improved fit
Use the scores from the operational IRT model

FIGURE 2. The process of assessment of practical significance of IRT model misfit.

test, the data are described first. The (statistical significance of) fit of the IRT model is then assessed, and finally the practical significance of the model misfit is assessed.

Example 1: A Test for Measuring Proficiency in English

Data and model-fit assessment. Let us consider data from two forms of a part of a special administration of a test for measuring English proficiency. Let us refer to these forms as the old form and the new form. Administrations of this part usually involve three 3-category polytomous items in addition to 39 dichotomous items. For the part considered here, IRT true score equating using the two-parameter logistic (2PL) model and the generalized partial credit model (GPCM) is employed to equate the raw score on a new form to the raw score on a reference form and then to an operational scale. The scale scores are used to make various decisions, such as admissions decisions in educational institutions. In the special administration considered here, raw scores on the new form could be equated to raw scores on the old form using an external anchor test that had 28 items. In typical administrations of the test, it is very unusual for two forms to have so many common items. For both of these forms, the sample size was about 1,500.

The operational items were combined with the anchor items and the IRT models was fitted to the combined data. We fitted the 2PL model to the binary items and the GPCM to the polytomous items.

Figures 3 and 4 show the item-fit plots for only the items for which substantial misfit was found. Each of these figures include 16 items, which constitute about 23% of the items. In these plots, the anchor items are treated as items 43–70 for the two forms.[3]

Considerable item misfit is evident in Figures 3 and 4.

Assessment of practical significance of misfit. As described earlier, an IRT model is operationally used to equate the raw scores for this test. Therefore, a natural way to assess the practical significance of misfit would be to examine if the removal of the severely misfitting items from the anchor item set leads to a difference in the equating conversion. If we consider only the items in the anchor item set (items 43–70) in Figures 3 and 4, there is substantial misfit for items 44, 49, 56, 61, and 70 in both forms. The results of omitting these five items from the anchor item set is shown in Figure 5. The upper panel of the figure shows the following two equating functions for each raw score:

- the one computed using all anchor items, and
- the one computed after omitting the above-mentioned five misfitting items from the anchor item set, that is, using the remaining 23 anchor items.

The lower panel of Figure 5 shows the differences between the equating functions. The operationally used IRT true score equating method using Stocking-Lord algorithm (Kolen & Brennan, 2004) was employed in creating Figure 5.

To interpret the differences between equating functions, we will use the difference that matters (DTM) criterion suggested by Dorans and Feigenbaum (1994). The DTM is a difference in equating conversions that can be considered practically large. For conversions of raw scores, Dorans and Feigenbaum (1994) recommended a DTM of 0.5. According to Figure 5, the differences between the equating functions for the data set are less than the DTM except for very low scores. Thus, the item misfit observed in Figures 3 and 4 does not seem to affect the equating. One could hence conclude that the IRT model misfit is not practically significant.

However, it is possible to go one step further regarding the assessment of practical significance of misfit in this example. To do so, first, we computed for each new-form examinee two

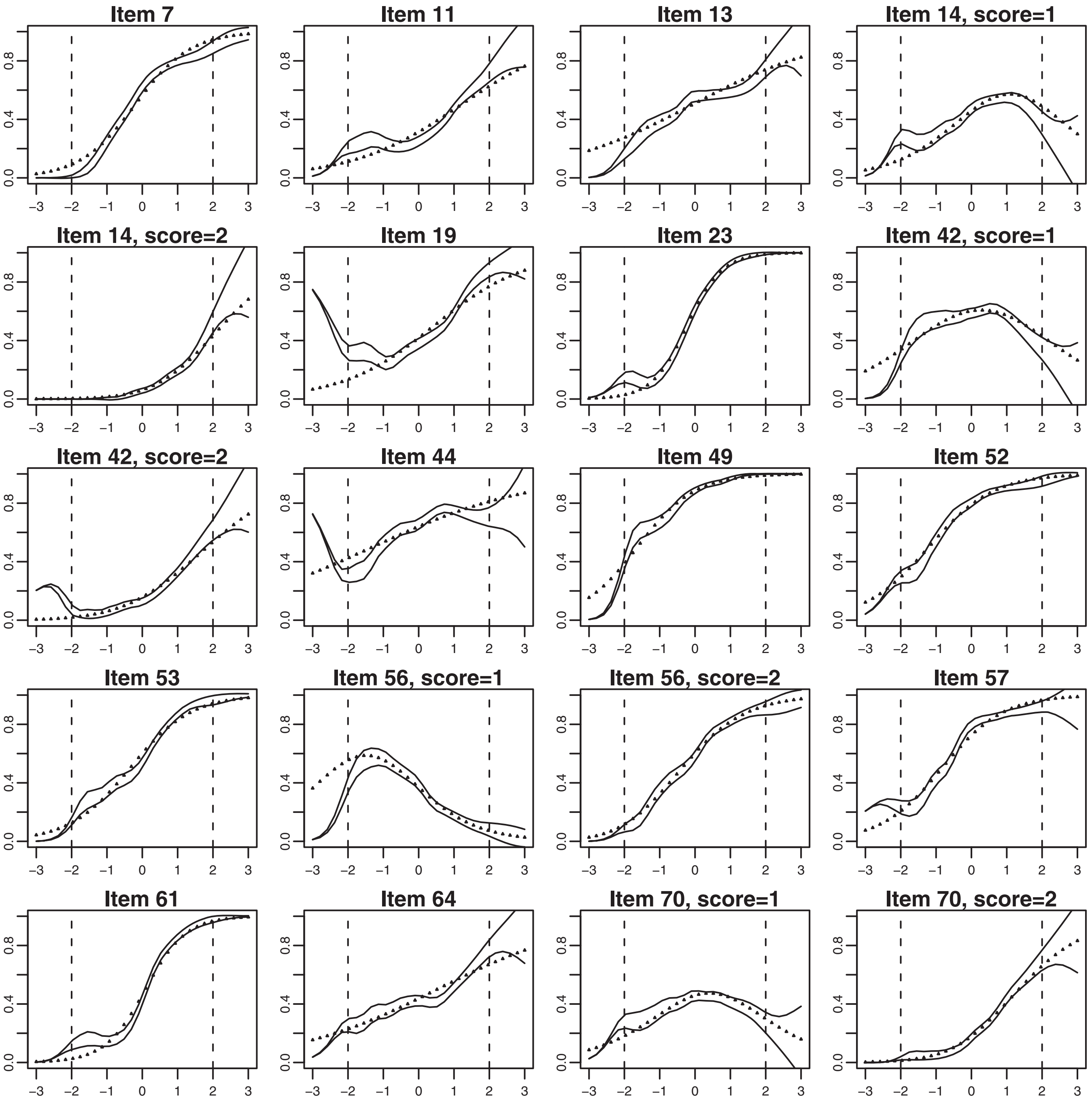

FIGURE 3. Plots of item fit for the test of English proficiency—old form.

scaled scores, one using each of the two raw-to-raw equating conversions that are plotted in the top panel of Figure 5. Computation of both of these scale scores involved the use of the operational raw-to-scale conversion for the old form and the raw-to-raw equating functions from the IRT models. As mentioned earlier, the scores from this test are used by the test score users to make various decisions. The cut scores for such decisions are mostly at the upper half of the score scale.[4] To analyze how the decisions may be affected by misfit, we obtained the operational raw-to-scale score conversion for the old form, and compiled a list of the cut scores at which most of the decisions are made on the basis of the test. There was no way to know what cut score actually applied to each examinee. Also, each examinee usually applies to several institutions. Then, we iterated the following steps 1,000 times[5]:

- We generated a random cut score (from the aforementioned list of cut scores) for each examinee who took the new form.
- We computed for each new-form examinee two pass–fail[6] statuses, one using each of the two scale scores computed above. While one of them is the operational pass–fail status, the other can be thought of as the pass–fail status under a better-fitting model–data combination.
- We computed the extent of disagreement between the two pass–fail statuses over the sample.

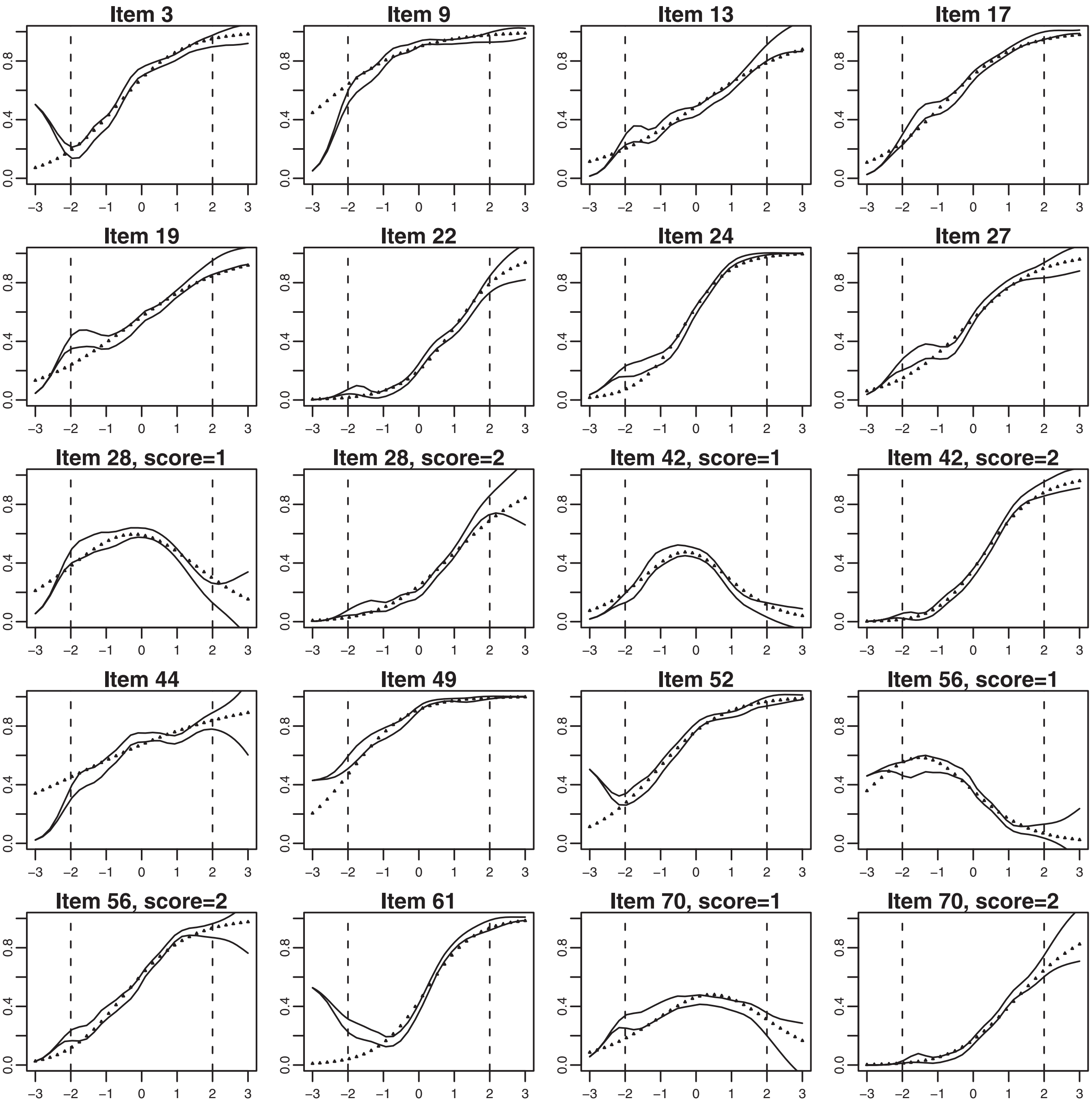

FIGURE 4. Plots of item fit for the test of English proficiency—new form.

The average value of 0.0003% from the above iterations indicates the extent of disagreement, on average, in the pass–fail statuses from the operational score and the score from a better-fitting model–data combination. This number points to, for all practical purposes, a negligible practical significance of the item misfit observed in Figures 3 and 4.

Example 2: Three Subject Areas from a State Test

Data and model-fit assessment. Let us consider the responses of examinees to two forms each of three subject areas of a state test. The scores from tests describe what students should know and be able to do in each grade and subject tested and measure school students' progress toward achieving the academic standards adopted by the state in several subjects. For each subject area, the two forms available will be referred to as the *new form* and the *old form*, respectively, depending on their date of administration. These tests include only multiple choice items. IRT true score equating using the 1PL model and a normal ability distribution is used to equate the raw score of an examinee on a new form to the raw score on an old form and then to an operational scale. The nonequivalent groups with anchor test design with an internal anchor design is used. There are respectively 65, 60, and 75 operational items on a form for the three subjects. Among

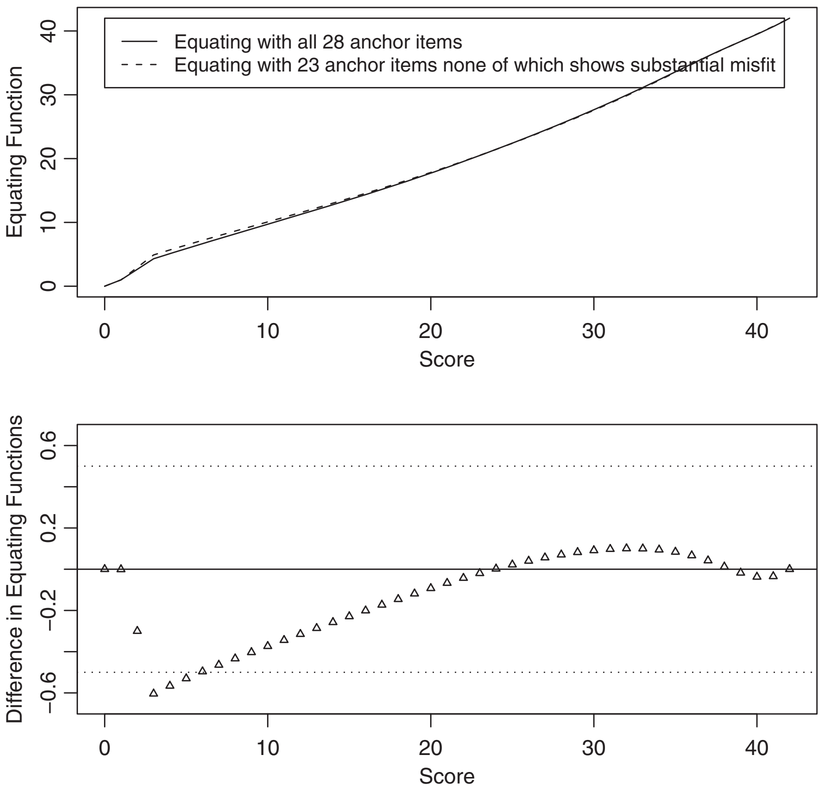

FIGURE 5. Effect of removal of misfitting anchor items on equating for the test of English proficiency.

them, 30, 27, and 29 items, respectively, are internal anchor items. The six data sets included responses of between 31,000 and 75,000 examinees. Thus, these data sets are larger than any other data sets analyzed in this article. The 1PL model was fitted to these data sets using the marginal maximum likelihood method and a normal ability distribution. We also fitted to these data the 3PL model with the restriction that all the guessing parameters are equal. The common guessing parameter is estimated from the data. We will refer to this model as the restricted 3PL model.[7]

Figure 6 represents the generalized residuals for the raw score distribution from the fit of the two above-mentioned IRT models for the two forms of Subject 1. The raw score distribution refers to the distribution of the raw scores and specifies the proportion of examinees who obtained the raw scores of 0, 1, 2, . . . , 65. To compute the generalized residuals for the raw score distribution, the T of Equation 1 was assumed to be equal to the proportion of examinees who obtained the raw scores of 0, 1, 2, . . . , 65. The 1PL model is used for equating of the raw scores of this test and raw scores are sufficient statistics for the ability parameters under the Rasch model. Therefore, it was appropriate to examine the fit of the 1PL model to the raw score distribution. In the figure, the solid line joins the points indicating the observed number of examinees at each score point (T in Equation 1). The dotted line represents the corresponding expected value ($\hat{E}(T)$). The dashed lines represent a corresponding 95% confidence interval given by $\hat{E}(T) \mp 1.96 s_D$. At any score point, a misfit of the model is indicated if the solid line lies far from the dotted line and outside the 95% confidence interval.

The figure shows evidence of severe misfit of the 1PL model to the raw score distribution for both the forms. For example, for the new form, 36 out of the 66 generalized residuals were larger than 4 in absolute value, with the largest of them being about 60 (for the raw score of 6). The observed score distribution appears skewed to the left. At the low end of the score distribution, all but a couple of the observed values lie outside the 95% confidence interval. Thus, Figure 6 suggests that the 1PL model does not adequately fit the observed score distribution of the data.

It is a common belief that examinees, especially low-scoring ones, often randomly guess the answer in multiple choice tests (see, e.g., Birnbaum, 1968, pp. 303–305). If examinees guess answers, then the 1PL model, which does not allow guessing, is likely not to fit the data well, especially at the low end of the score distribution. The restricted 3PL model, which allows a guessing parameter, performs much better than the 1PL model in fitting the raw score distribution. For example, for scores below 25, the expected values are much closer to the observed values for the restricted 3PL model compared to the 1PL model. However, the restricted 3PL model also does not provide a perfect fit to the raw score distribution.

Item-fit plots (Haberman et al., 2013) show evidence of misfit of the 1PL model (plots not shown) for almost all the items. For example, for item 25, there were 26 residuals out of 30 that are larger than 5 in absolute value. Item-fit plots (Haberman et al., 2013) show evidence of misfit of the restricted 3PL model as well for almost all the items. However, the item-fit plots for the restricted 3PL model show better fit than those for the 1PL model for most items. Let us consider Figure 7, which shows the item-fit plots for the two IRT models for the new form of Subject 1 for two items: items 25 and 56.

The top row of the figure shows plots for the 1PL model while the bottom row shows plots for the restricted 3PL model. While neither IRT model is adequate for these items, the dotted line is much closer to the 95% confidence interval (the

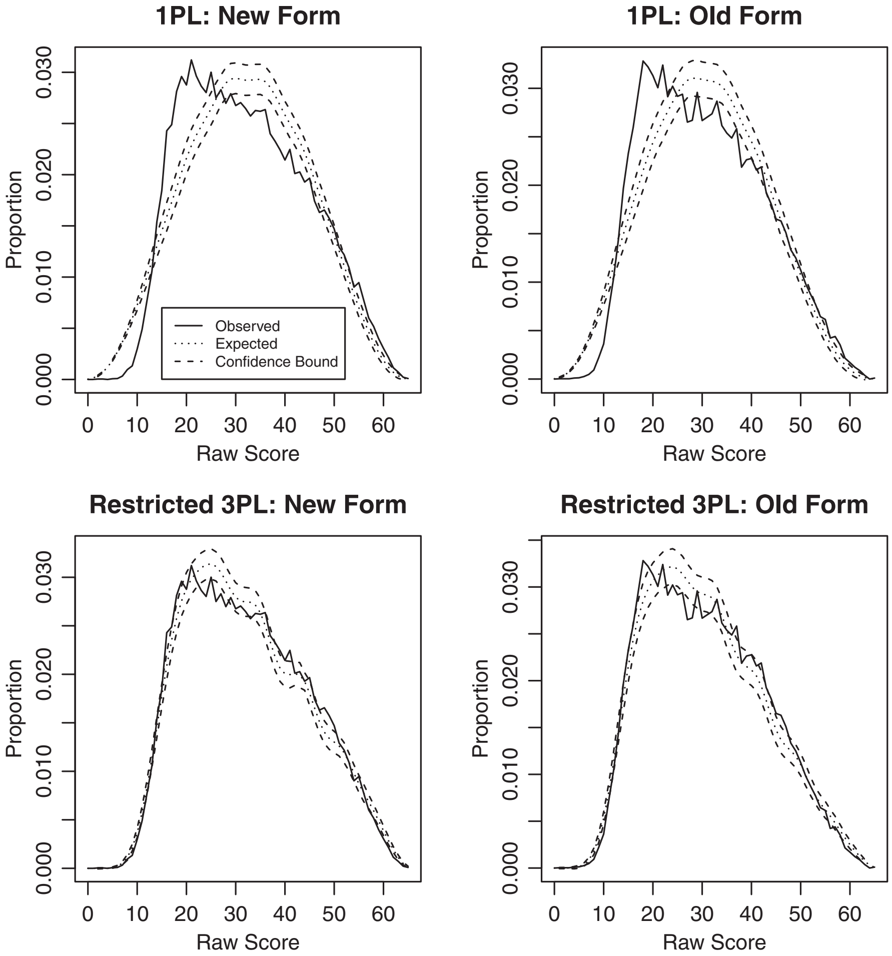

FIGURE 6. Fit of the 1PL model (top) and the restricted 3PL model (bottom) to the raw score distribution for two forms of the state test.

two solid lines) for the restricted 3PL model, indicating better fit for the latter model compared to the former model. The confidence band is very narrow for each plot in Figure 7. Large sample sizes lead to small standard errors for these data sets, which in turn lead to narrow confidence bands and often large residuals.

Assessment of practical significance of misfit. Neither the 1PL model nor the restricted 3PL model adequately fits the data from the state test. In fact, Figures 6 and 7 show that the extent of misfit is severe, especially compared to the other data sets considered in this article. However, given the large sample sizes of these data sets, misfit is expected. Therefore, especially for data sets of such large sizes, the relevant question is not “Does the IRT model fit the data?,” but “Is the misfit of the IRT model practically significant?”

Because the IRT model is used to perform IRT true score equating for this state test, a natural way to assess the practical significance of misfit is to examine if the choice of the restricted 3PL model, which fits the data better than does the operationally used 1PL model, leads to a difference in equating of the test. The top three panels in Figure 8 show the differences between the IRT true score equating functions from the 1PL model and the restricted 3PL model for the three subjects on the raw score scale. The Stocking-Lord method (see, e.g., Kolen & Brennan, 2004) was used to compute these equating functions. The operationally used anchor items were employed in these equatings. The differences between the equating functions from the two IRT models exceed the DTM of 0.5 (recommended by, e.g., Dorans & Feigenbaum, 1994) for some raw score for each of the three subjects. The bottom three panels in Figure 8 show a histogram of the differences between the raw-to-scale conversions from the 1PL model and the restricted 3PL model for the three subjects. For each of the three subjects, the difference between the maximum and minimum scale score is several hundreds.[8] Thus, it seems that the choice of the better-fitting restricted 3PL model over the 1PL model would occasionally have some impact on equating for the state test considered.

In this case, it is possible to assess further the practical significance of misfit. The scores from this state test are primarily used in determining Adequate Yearly Progress, which is used to meet the requirement of the federal Elementary and Secondary Education Act that all students score at the proficient level or above by 2014. To analyze how the classifications of the students into proficient or nonproficient are affected by misfit, we obtained the operational raw-to-scale score conversion for the old form and obtained the score required to be classified as proficient for each subject. We then computed for each new-form examinee two scaled scores, one under each of the two above-mentioned IRT model. Computation of both of these scale scores involved the use of the operational raw-to-scale conversion for the old form and the raw-to-raw equating functions from the IRT models. We then computed

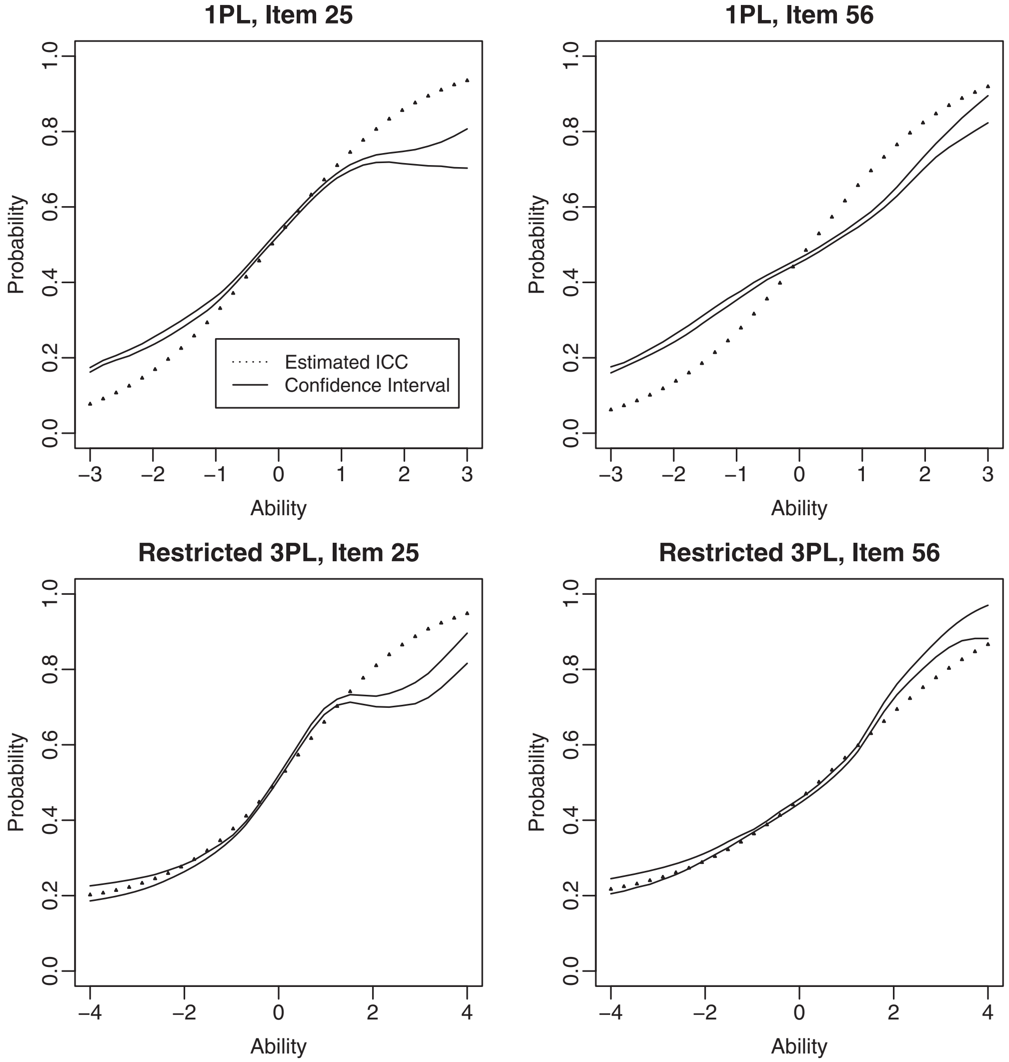

FIGURE 7. Plots of item fit for two items for the state test, Subject 1—new form.

for each new-form examinee two proficiency statuses, one using each of the two IRT models. While one of them is the operational proficiency status, the other can be thought of as the proficiency status if the IRT model fitted the data better. We found that there was no disagreement in the proficiency statuses from the two IRT models for two of these subjects. For the third subject, there was a disagreement for 2.4% of the examinees, all of whom were operationally classified as proficient but would not have been classified as proficient if the restricted 3PL model were used. Thus, it can be concluded for the state test that the misfit of the operationally used 1PL model can occasionally have practical significance.

Example 3: Two Basic Skills Tests

Data and model-fit assessment. Data were obtained from one form each of two tests belonging to a series of basic skills tests. The tests include 40 and 38 multiple-choice items, respectively. These tests are computerized but not adaptive. The sample sizes for these data sets were between 2,000 and 3,000. The test employs a large pool of items that are all calibrated on the same scale using the Stocking-Lord algorithm (e.g., Kolen & Brennan, 2004). The raw scores on a new form are equated to the raw scores on a base form and then to the operational score scale using IRT true score equating (e.g., Kolen & Brennan, 2004) using the 3PL model and the item parameters of the two test forms that were calibrated to be on the same scale. The new form and the base form do not have any items in common. The scaled scores are used for teacher licensing.

In a residual analysis to assess item fit (Haberman et al., 2013), the IRT model is found not to fit several items. For example, significant misfit is observed for nine items for the first test (plots not shown).

Assessment of practical significance of misfit. An IRT model is used to equate the scores on these basic skills tests. Therefore, an assessment of the practical consequences of misfit could involve an analysis of whether the impact of item misfit on the equating is practically significant. IRT true score equating was performed twice for each test to equate the raw scores on the form to those on the base form. The first equating was the operational equating in which the raw score on all the items was equated. In the second equating, the score that was equated was the raw score obtained after removing from the data set the five items that showed the most significant

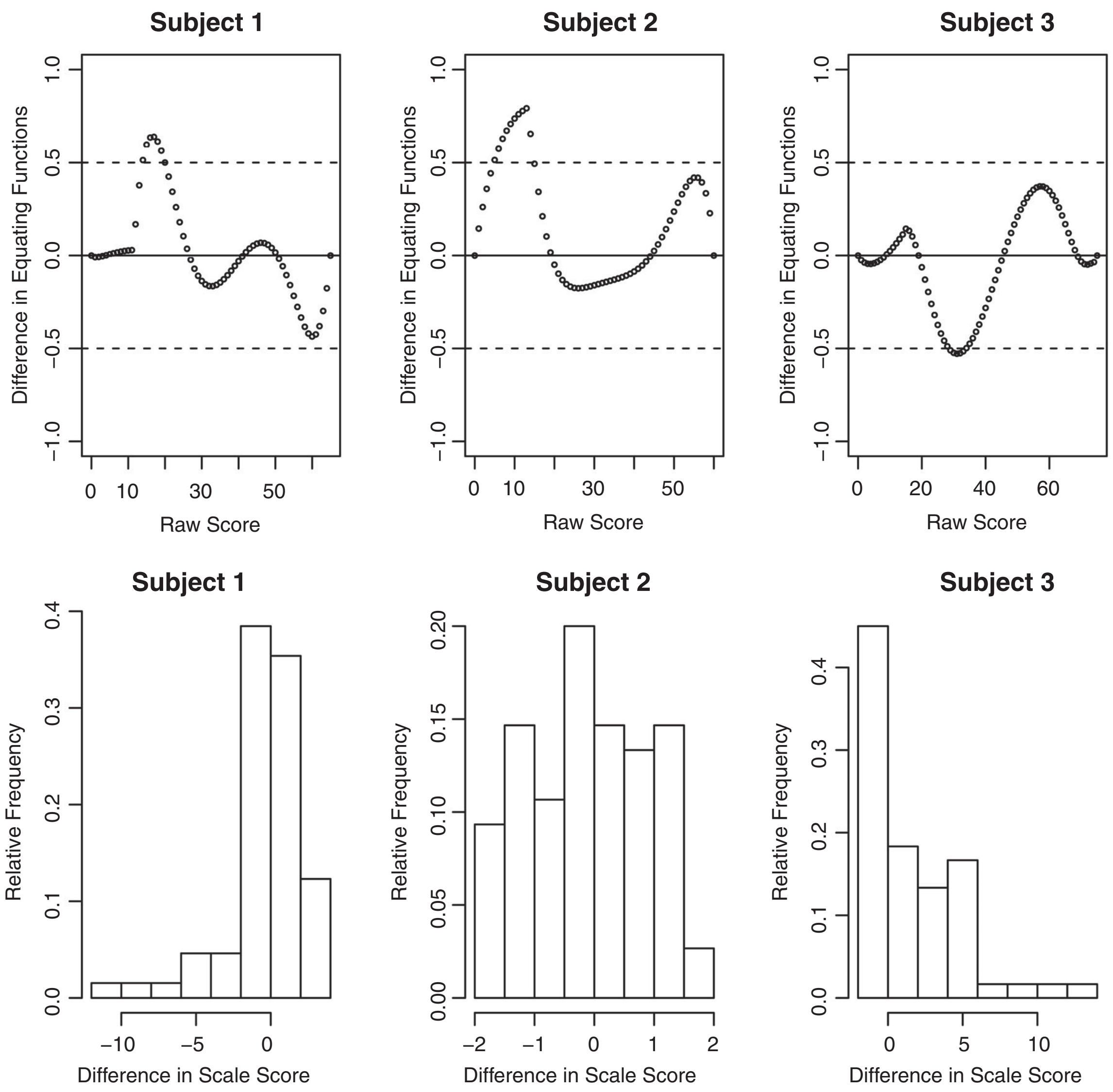

FIGURE 8. The effect of a better-fitting IRT model on equating for the three subjects of the state test.

amount of misfit. We then calculated the following two scaled scores for each examinee:

- E_1: the scaled score from the first equating above.
- E_2: the scaled score from the second equating above.

We computed the correlation coefficient between E_1 and E_2. Figure 9 includes a scatter-plot (using a hollow circle to denote a point) of E_1 versus E_2 for the first test. A diagonal line is also shown for convenience. The plot is very similar for the second test and is not shown.

If the item misfit for the data set is practically significant, then the equating should be severely affected by removal of the misfitting items and the equated scores E_1 and E_2 should not be very highly correlated. However, that is not the case. The correlation and rank correlation between E_1 and E_2 are both 0.99 for both the tests. In Figure 9, the points fall close to the diagonal line. Therefore, the misfit of the items seems to have very little practical significance for these tests.

As with the earlier examples, it is possible to go further regarding the assessment of practical significance of misfit. Let us consider the first test. First, we computed for each examinee two scaled scores for this test, one using the operational items and one using the items remaining after removing the five misfitting items, as described earlier. The scores from this test are used by the test score users for teacher licensing. To assess how the licensing decisions may be affected by misfit, we compiled a list of the cut scores at which the decisions are made on the basis of the test.[9] It was not known what cut score actually applied to each examinee. Also, several examinees apply to multiple institutions so that multiple cut scores apply to each of them. Then, we iterated the following steps 1,000 times:

- We generated a random cut score (from the aforementioned list of cut scores) for each examinee who took the test.
- We computed for each examinee two pass–fail statuses,[10] one using each of the two scale scores computed above. While one of them is the operational pass–fail status, the other can be thought of as the pass–fail status under a better-fitting model–data combination.
- We computed the percent of disagreement between the two pass–fail statuses over all the examinees in the sample.

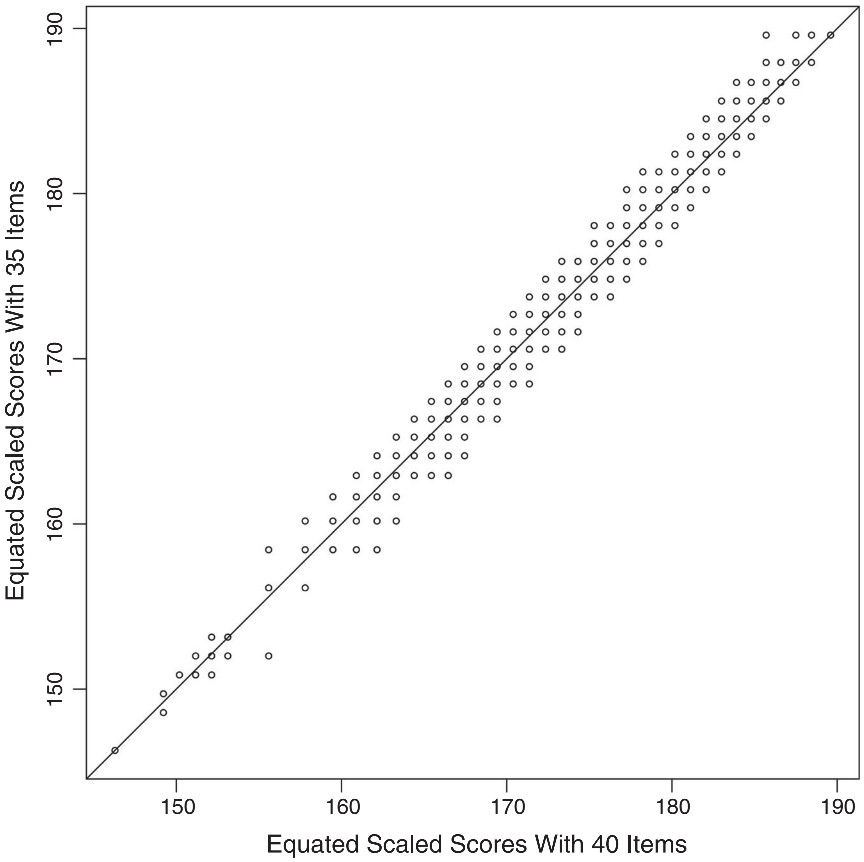

FIGURE 9. Practical significance of misfit for the first basic skills test.

The average value of the percent disagreement is 3.3 from the above iterations. Some disagreement between the two pass–fail statuses is expected because the second scaled score involved the equating of a test that is shorter than the operational test by five items. To examine if the 3.3% disagreement occurred because of randomness rather than misfit, we repeated the above steps after removing five items randomly from the test (as opposed to based on item misfit). The level of disagreement was 2.8% on an average. Thus, the removal of five items based on item misfit rather than randomly leads to an additional disagreement of 0.5% on an average. Similarly, for the second basic skills test, the average level of disagreement was 2.4% when five items were removed based on item misfit and 1.8% when five items were removed randomly. These levels of disagreement probably point to a small practical significance of the item misfit observed for these basic skills tests.

Conclusions

The focus of this article was on the assessment of practical significance of misfit of IRT models using data from several operational tests. The types of analyses performed in this article are rare in the IRT literature and practitioners applying IRT models should find the suggested methods and the data examples useful.

An important finding from this article is that IRT models did not fit any of the data sets considered in this article and severe misfit (in the form of many large residuals) was found for the large data sets. This finding concurs with the statement of George Box that all models are wrong (Box & Draper, 1987, p. 74).

However, the observed misfit was not always practically significant. For example, for the state test example, the operationally used 1PL model was found to be severely inadequate, but a better-fitting model did not lead to any differences in the proficiency statuses of the examinees for two out of the three subjects. Therefore, the 1PL model can be claimed to be useful for these two subjects even though the model was not adequate for the data. An interesting finding in this article is that IRT true score equating was affected very little by the removal of substantially misfitting items from the test or anchor test.[11] This finding, along with the finding of, for example, de Champlain, (1996) that the quality of IRT true score equating is hardly affected by multidimensionality, may indicate that IRT true score equating is robust against realistic violation of the IRT model. It is possible that when the nature of the model misfit is the same in the two equating samples, IRT true score equating is robust.

The results in our study may not be applicable to uses of IRT models that were not examined here. Most importantly, the practical significance of misfit of IRT models in the context of pattern scoring or computerized adaptive testing was not considered here. For these kinds of uses, the IRT model misfit may be practically significant to an extent more than what was observed in this article. There is a need to perform further research about the assessment of practical significance of IRT model misfit for these kinds of uses.

Also, some data sets used here (such as the state test data) did not include items that were found to have poor psychometric properties during pretesting. Therefore, analysis of misfit for pretest data, including the determination of practical significance of misfit, is a possible area for further exploration.

Appendix: Brief Descriptions of the Two Fit Assessment Methodologies Used

Residual Analysis to Assess Item Fit

To assess item fit, Bock & Haberman (2009) and Haberman et al. (2013) employed a specific form of residual analysis that involves a comparison of two approaches to estimation of the item response function. Consider a test with items numbered from 1 to J and examinees numbered from 1 to N. Let response score X_{ij} of examinee i to item j be 0 or 1. For item j, let $\hat{F}_j(\theta)$ denote the estimated item characteristic curve of the item; for example, for the two-parameter logistic (2PL) model,

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

where $\hat{a}_j$ and $\hat{b}_j$ are the respective estimated item discrimination and difficulty parameters for item j.

Then the residual of item j at examinee ability θ is defined as

$$t_j(\theta) = \frac{\bar{F}_j(\theta) - \hat{F}_j(\theta)}{s_j(\theta)}, \qquad \text{(A1)}$$

where $\bar{F}_j(\theta)$ is an estimate of the probability of a correct response by an individual with latent ability θ obtained by the use of a weighted average of the responses X_{ij} for examinees 1 to N and $s_j(\theta)$ is the estimated standard deviation of $\hat{F}_j(\theta) - \bar{F}_j(\theta)$. The weights on the X_{ij}s in the computation of $\bar{F}_j(\theta)$ are proportional to the conditional posterior distribution of the examinee ability given the examinee responses. Haberman et al. (2013) proved that if the model fits the data adequately, then, for each ability level θ, $t_j(\theta)$ follows an approximate standard normal distribution.

One can create plots of item fit using the above residuals. For each item, the examinee ability θ is plotted along the x-axis, the dotted line denotes the values of $\hat{F}_j(\theta)$ from Equation 2, and the two solid lines denote the values of $\bar{F}_j(\theta) - 2s_j(\theta)$ and $\bar{F}_j(\theta) + 2s_j(\theta)$ and form a pointwise confidence band at an approximate level of 95%.

The software program developed by Haberman et al. (2013) was used to compute these residuals. The program is available on request from the authors for noncommercial use.

Generalized Residual Analysis

Generalized residual analysis for assessing the fit of IRT models was suggested by Haberman (2009). Let y denote a possible item response pattern of an examinee. For example, y could be $(0, 1, 1, \ldots, 1)$ for a test with dichotomous items.

Suppose

$p(y) =$ probability of observing the response y,

and

$n(y) =$ the number of examinees in the sample with response pattern y.

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

$$\hat{p}(y) = \frac{n(y)}{N}\cdot$$

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

$$T = \sum_{y\in\omega} d(y)\hat{p}(y)\cdot \qquad \text{(A2)}$$

Then the estimated mean $\hat{E}(T)$ and an estimated standard deviation s_D of the difference $T - \hat{E}(T)$ are computed under the assumption that the true IRT model is the IRT model fitted to the data. One then computes a generalized residual

$$g = \frac{T - \hat{E}(T)}{s_D}. \qquad \text{(A3)}$$

Haberman (2009) proved that if the IRT model fits the data adequately and the sample is large, the distribution of g is well approximated by the standard normal distribution. Thus, a statistically significant value of the generalized residual g indicates that the IRT model does not adequately predict the statistic T. The method is quite flexible. Several common data summaries such as the item proportion correct, proportion simultaneously correct for a pair of items, and observed score distribution can be expressed as the statistic T by defining $d(y)$ appropriately. For example, if

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

then T of Equation 3 becomes the proportion correct for item 1 and the corresponding value of g indicates how well the IRT model predicts the proportion correct for item 1. For another example, if

$$d(y) = \begin{cases} 1 & \text{whenever the components of } y \text{ add up to } r \\ 0 & \text{otherwise,} \end{cases}$$

then T of Equation 3 becomes the proportion of examinees who obtained a raw score of r and the values of g for $r = 0, 1, 2, \ldots.$ indicate how well the IRT model predicts the raw score distribution.

Acknowledgments

The authors wish to express sincere appreciation and gratitude to the editor Derek Briggs, the associate editor Andrew Maul, and the three anonymous reviewers for an extremely quick review process and for several helpful comments. The authors are grateful to Matthias von Davier, Andreas Oranje, and Neil Dorans for their helpful comments and to Behroz Maneckshana, Kevin Larkin, and Yi-Hsuan Lee for their help with the data. The research reported here was partially supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D120006 to Educational Testing Service as part of the Statistical and Research Methodology in Education Initiative. Some of the research reported here was performed when the lead author was an employee of Educational Testing Service. Any opinions expressed in this publication are those of the authors and not necessarily of CTB/McGraw-Hill, Educational Testing Service, Institute of Education Sciences, or U.S. Department of Education.

Notes

[1] We are assuming here that the IRT model is used to compute the reported scores.

[2] Simulations provide one way to estimate this hypothetical score, but simulations have their own problems.

[3] This reordering is only for convenience—the anchor items are actually spread throughout the form.

[4] No further information regarding the cut scores can be provided due to confidentiality restrictions.

[5] Performing this step only once instead of 1,000 times might have made the results influenced by the choice of the cut scores in that step.

[6] Or, “admitted” and “not admitted.”

[7] We faced some problems with convergence when we tried to fit the traditional 3PL model without any restrictions on the item parameters. That is why we chose this restricted 3PL model.

[8] It is not possible to reveal any further information about the score scale due to confidentiality restrictions.

[9] Different states employ different cuts while using scores from this test.

[10] Or “licensed” and “not licensed” statuses.

[11] Though we did not report it earlier, the removal of misfitting items led to hardly any differences in the equating and pass–fail statuses for the state test data.

References

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.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29–51.

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

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York, NY: Wiley.

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.

Chon, K. H., & Sinharay, S. (in press). A note on the Type I error rate of the PARSCALE G^2 statistic for long tests. *Applied Psychological Measurement*. Advance online publication. Available at http://apm.sagepub.com/content/early/2013/11/18/0146621613508307.abstract

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. (2010). *Item response theory*. New York, NY: Oxford University Press.

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: Educational Testing Service.

du Toit, M. (2003). *IRT from SSI*. Lincolnwood, IL: Scientific Software International.

Glas, C. A. W., & Suarez-Falcon, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106.

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: Educational Testing Service.

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm*. ETS Research Report RR-13-32. Princeton, NJ: Educational Testing Service.

Haberman, S. J., & Sinharay, S. (2013). Generalized residuals for general models for contingency tables with application to item response theory. *Journal of American Statistical Association*, *108*, 1435–1444.

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, *78*, 417–440.

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.

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

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

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

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012–1027.

Molenaar, I. W. (1997). Lenient or strict application of IRT with an eye on the practical consequences. In J. Rost & R. Langenheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 38–49). Munster, Germany: Waxmann.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50–64.

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., & Jia, H. (2011). *Fit of item response theory models: A survey of data from several operational tests*. ETS Research Report No. RR-11-29. Princeton, NJ: Educational Testing Service.

Smith, R. M., Schumacker, R. E., & Joan Bush, M. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, *2*, 66–78.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*, 331–352.

Suarez-Falcon, J. C., & Glas, C. A. W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *56*, 127–143.

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.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psycholgoical Measurement*, *5*, 245–262.