Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Fall 2015, Vol. 34, No. 3, pp. 39–48

An NCME Instructional Module on Item-Fit Statistics for Item Response Theory Models

Allison J. Ames and Randall D. Penfield, *University of North Carolina at Greensboro*

Drawing valid inferences from item response theory (IRT) models is contingent upon a good fit of the data to the model. Violations of model-data fit have numerous consequences, limiting the usefulness and applicability of the model. This instructional module provides an overview of methods used for evaluating the fit of IRT models. Upon completing this module, the reader will have an understanding of traditional and Bayesian approaches for evaluating model-data fit of IRT models, the relative advantages of each approach, and the software available to implement each method.

Keywords: item response theory, model-data fit, posterior predictive checks

Item response theory (IRT) is a widely adopted measurement framework used in the design, construction, and evaluation of educational assessments. The core principle underlying IRT is the use of a model to specify the probability of observing each scored response category of an item (i.e., correct or incorrect) as a function of the individual's underlying latent ability. The model linking level of latent ability to the probability of response to an item is referred to as an *item response function* (IRF). Examples of IRFs associated with the correct response for three different multiple-choice items are displayed in Figure 1. For each item we see that individuals with low levels of ability have a low probability of correct response to the item. As ability increases, so too does the probability of correct response.

The particular shape and location of an IRF reflects the psychometric properties of the item, such as difficulty, discrimination, and guessing (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980). The different forms of the three IRFs shown in Figure 1 reflect differences in the items' difficulty, discrimination, and guessing. The IRF that is located furthest to right reflects the highest level of difficulty (Item 1), the IRF with the steepest slope reflects the highest degree of discrimination (Item 2), and the IRF having a lower asymptote near .2 reflects the highest degree of guessing (Item 3).

Parametric IRT approaches employ mathematical models to specify the form of the IRFs. Numerous models for specifying IRFs have been proposed (see de Ayala, 2009; van der Linden & Hambleton, 1997), and several of these models are widely adopted in applied testing contexts (Hambleton et al., 1991). In the case of dichotomously scored items (having the scored outcomes of correct and incorrect), the most flexible of the widely adopted models is the three-parameter logistic model (3PL; Birnbaum, 1968). Denoting the response of the nth individual to the ith item by $x_{ni} = 1$ for a correct response and $x_{ni} = 0$ for incorrect response, the 3PL specifies the IRF for correct response using

$$P_{ni1} = c_i + (1 - c_i)\frac{\exp(a_i\,(\theta_n - b_i))}{1 + \exp(a_i\,(\theta_n - b_i))}, \qquad (1)$$

where a_i reflects item discrimination, b_i reflects item difficulty, c_i reflects guessing, and θ represents the individual's level of latent ability. Item 3 of Figure 1 represents a 3PL item having parameters $a_i = 1.0$, $b_i = 1.5$, and $c_i = 0.2$. Other widely used IRF models are constrained versions of the 3PL. Fixing $c_i = 0$ to reflect the absence of guessing yields the two-parameter logistic model, as demonstrated by Item 2 in Figure 1B. A further constraint of fixing $a_i = 1$ yields the Rasch model (Rasch, 1960/1980), represented by Item 1 in Figure 1.

A key assumption of IRT is that each IRF of the scored data accurately reflects the link between an individual's latent ability and item responses, which is commonly described as the IRF being a good fit to the actual data. IRFs that are a good fit to the data can yield appropriate inferences and predictions. IRFs that do not demonstrate good fit to the data run the risk of several undesirable outcomes, including biased ability and item parameter estimates (Wainer & Thissen, 1987; Yen, 1981) that jeopardize the appropriate application of IRT models in such areas as test development, equating, and computer adaptive testing (Kang & Chen, 2008). The consideration of model-data fit is an important step in test development (see Standard 3.9 of the *Standards for Educational and Psychological Testing*; AERA/APA/NCME, 1999), with misfitting items often being discarded from the potential item pool (Sinharay, 2006; Wilson, 2005). Not surprisingly, considerable importance is attached to evaluating model-data fit of IRT models and the procedures for evaluating fit.

Allison J. Ames, University of North Carolina at Greensboro, Educational Research Methodology, Department of Educational Research Methodology ERM Department, School of Education Building, Room #254, Greensboro, NC 27402; ajames@uncg.edu. Randall D. Penfield, University of North Carolina at Greensboro, Educational Research Methodology, 1300 Spring Garden St., Greensboro, NC 27412; rdpenfie@uncg.edu

Practitioners can choose from both graphical (i.e., visual inspection of data) and statistical approaches for evaluating model-data fit, and a wide variety of evidence should be used when evaluating fit of IRT models (van der Linden & Hambleton, 1997, chapter 1). However, a recent survey of testing programs (Sinharay, Haberman, & Jia, 2011) found a fairly limited scope of evidence regarding evaluation of model-data fit. Among those most commonly used approaches, several have considerable limitations, including the tendency to indicate an item as misfitting when it is, in fact, a good fit to the data (DeMars, 2005). In addition, Bayesian methods of assessing model-data fit are becoming increasingly popular, and the growing availability of software for Bayesian methods is increasing the viability of applying these approaches in practical testing contexts. Given these considerations, an overview of available methods will benefit practitioners by providing a more complete toolkit for the evaluation of fit of IRT models.

The purpose of this module is to introduce and illustrate the application of methods used for evaluating IRT model-data fit. We begin with a description of the concept of the residual and graphical evaluations of model-data fit. This is followed by a presentation of traditional methods for evaluating model-data fit of IRT models, and then an introduction to Bayesian methods for evaluating model-data fit. Numerical examples are provided throughout, as well as a discussion of the advantages and limitations of each of the approaches.

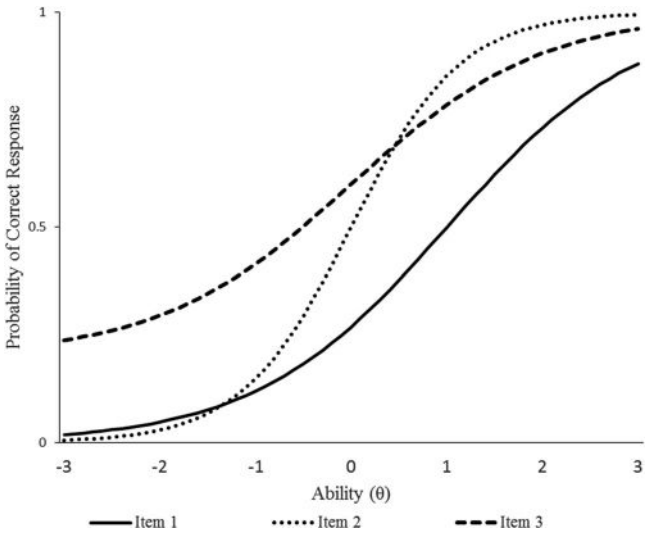

FIGURE 1. IRFs for the correct option of three dichotomous items. For Item 1, $a_1 = 1$, $b_1 = 1$, $c_1 = 0$; for Item 2, $a_2 = 1.75$, $b_2 = 0$, $c_2 = 0$; and for Item 3, $a_3 = 1$, $b_3 = 0$, $c_3 = 0.2$.

Model-Data Fit and Residuals

Methods for evaluating IRT model-data fit are all based on examining how closely observed responses to an item (x_{ni}) match, or fit, those predicted by the item's IRFs (P_{nix}). We begin the discussion of model-data fit using the context of dichotomous items, which narrows the discussion to the IRF for the correct response (P_{ni1}). As the difference between x_{ni} and P_{ni1} increases, the fit of the IRF to the observed data decreases and provides more compelling evidence of a violation of fit. The difference between a particular individual's response to a given item and that predicted by the IRF is referred to as a *residual* (r_{ni}), and is denoted by

$$r_{ni} = x_{ni} - P_{ni1}. \quad (2)$$

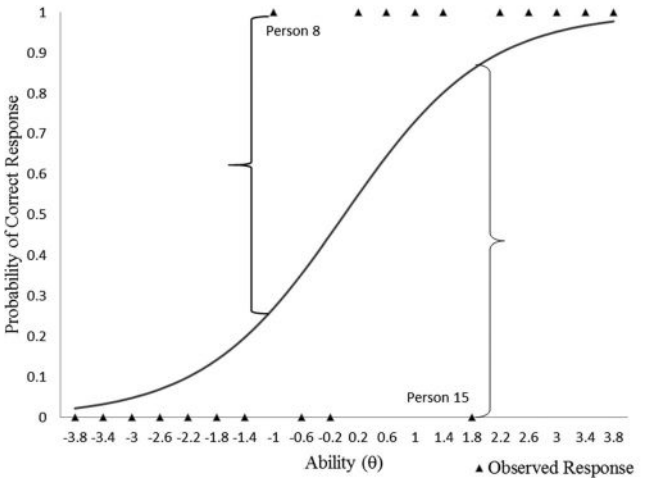

FIGURE 2. Individual-level residuals of 20 responses to a dichotomous item, following the Rasch model with $b_i = 0$.

Figure 2 provides a visual representation of these individual-level residuals for 20 individuals to an item with an IRF for correct response following the Rasch model (where $b_i = 0$). Each individual's observed response ($x_{ni} = 0$ or $x_{ni} = 1$) is shown as a triangle. The residual associated with each individual's response is the difference between the observed response (x_{ni}) and the item's IRF associated with the correct response (P_{ni1}). Figure 2 illustrates that as the distance between the observed response and the IRF increases, the residual becomes larger, and evidence of misfit is more readily seen.

Computed residuals for each of the 20 responses shown in Figure 2 are presented in Table 1. Residuals can be positive or negative. A positive residual indicates the individual performed better than is predicted by the IRF. This is the case with the 8th individual in Table 1 who answered the item correctly ($x_{ni} = 1$), but had a low probability of correct response based upon the IRF ($P_{ni1} = 0.27$) resulting in a residual of $r_{ni} = 1 - 0.27 = 0.73$. Conversely, a negative residual indicates an individual performed worse than predicted by the IRF, as is the case with the 15th individual. This individual answered the item incorrectly ($x_{ni} = 0$) but had a high probability of correct response based upon the IRF ($P_{ni1} = 0.86$) resulting in a residual of $r_{ni} = 0 - 0.86 = -0.86$. Through the magnitude and direction (positive or negative) of the residuals, fit can be evaluated, providing the foundation for several indices that will be described shortly.

An inherent limitation of interpreting fit using individual-level residuals, as described above, is that the individual-level residuals will typically not be zero, even in the presence of relatively good model-data fit. For example, in the case of dichotomous items, the observed responses (x_{ni}) are restricted to values of 0 or 1 and the IRF for correct response assumes values between 0 and 1. It follows that the observed response cannot be equal to the IRF for correct response and thus the individual-level residual for a dichotomous item cannot be zero. Indeed, at levels of θ for which the IRF assumes moderate values near .5, the individual-level residual is unavoidably near .5 when $x_{ni} = 1$, and near -0.5 when $x_{ni} = 0$. To

Table 1. Hypothetical Data Example

Individual	θ	x_{ni}	P_{ni1}	r_{ni}	r_{ni}^2	$P_{ni1} \times (1 - p_{ni1})$	$r_{ni}^2 / P_{ni1} \times (1 - p_{ni1})$
1	−3.8	0	0.02	−0.02	0.00	0.02	0.02
2	−3.4	0	0.03	−0.03	0.00	0.03	0.03
3	−3	0	0.05	−0.05	0.00	0.05	0.05
4	−2.6	0	0.07	−0.07	0.00	0.06	0.07
5	−2.2	0	0.10	−0.10	0.01	0.09	0.11
6	−1.8	0	0.14	−0.14	0.02	0.12	0.17
7	−1.4	0	0.20	−0.20	0.04	0.16	0.25
8	−1	1	0.27	0.73	0.53	0.20	2.72
9	−0.6	0	0.35	−0.35	0.13	0.23	0.55
10	−0.2	0	0.45	−0.45	0.20	0.25	0.82
11	0.2	1	0.55	0.45	0.20	0.25	0.82
12	0.6	1	0.65	0.35	0.13	0.23	0.55
13	1	1	0.73	0.27	0.07	0.20	0.37
14	1.4	1	0.80	0.20	0.04	0.16	0.25
15	1.8	0	0.86	−0.86	0.74	0.12	6.05
16	2.2	1	0.90	0.10	0.01	0.09	0.11
17	2.6	1	0.93	0.07	0.00	0.06	0.07
18	3	1	0.95	0.05	0.00	0.05	0.05
19	3.4	1	0.97	0.03	0.00	0.03	0.03
20	3.8	1	0.98	0.02	0.00	0.02	0.02
Sums					2.14	2.41	13.11

Note. θ = latent ability estimate, x_{ni} = observed response, P_{ni1} = predicted response based off of the IRF, r_{ni} = residual.

circumvent this limitation, one can first group individuals according to specific ranges of ability, referred to as *bins*, and then consider the difference between the observed proportion correct for each bin and the proportion correct predicted by the IRF for each bin. Using this "binned" approach, the residual for bin h is given by

$$r_{hi} = O_{hi1} - P_{hi1}, \tag{3}$$

where O_{hi1} represents the observed proportion of individuals in the hth bin having a correct response to the ith item and P_{hi1} is the probability of correct response for that bin. The bins will be denoted here by $h = 1, 2, \ldots, H$, such that H represents the total number of bins.

An example of using bin-level residuals is shown in Table 2, where binning is based on ability for those individuals in Table 1. For simplicity, 10 bins have been used ($H = 10$), resulting in two individuals per bin. The individuals with the two lowest ability levels were placed in the first bin ($h = 1$), and so on. Examining bin 1, there were no people answering the item correctly and the observed proportion of people answering correctly is 0. The probability of correct response for that bin (P_{1i1}) is subtracted from the observed proportion correct to arrive at the bin-level residual ($r_{1i} = 0 - 0.03 = -0.03$). As with individual-level residuals, these bin-level residuals can also be positive or negative. A positive bin-level residual indicates that as a whole, individuals in the bin tended to perform better than predicted by the IRF and a negative bin residual indicates the individuals in the bin tended to perform worse than predicted. However, unlike the individual-level residuals, the bin-level residuals can approach zero in the presence of good model-data fit, provided sample sizes are sufficiently large.

The difference between values of O_{hi1} and the IRF (P_{hi1}) can be visually inspected to make inferences about model-data fit through use of an empirical IRF, which plots the values of O_{hi1} as a function of θ to provide a visual representation of how the sample proportion correct changes across the θ continuum. Examples of an empirical IRF are shown in Figure 3 where the IRF for the correct response (solid line) is shown in combination with the empirical IRF (dashed line) in four different conditions. When the bin-level residuals are small, and distributed randomly about the IRF, the empirical IRF will follow a pattern consistent with the form of the modeled IRF, and we can conclude the model is a good fit of the data (Hambleton et al., 1991). The top panel of Figure 3 represents just such a situation. In the second panel of Figure 3, we see an item that appears to fit individuals with moderate and high ability, but has large residuals for individuals at the lowest ability range. Inferences drawn for individuals with higher ability seem appropriate with this item, but not for individuals at the lower end of the ability continuum. Examining the shape of the empirical IRF against the IRT model IRF in the second panel illuminates a possible source of the model-data misfit. It appears guessing is present in the item responses, as indicated by the empirical IRF having a lower tail not near zero, but which is not appropriately modeled. The third panel illustrates poor model-data fit at the higher range of ability, and the bottom panel illustrates misfit across the entire range of individuals.

A comprehensive discussion of evaluating fit via visual inspection of residuals can be found in Hambleton et al. (1991). Whereas the visual inspect of residuals can provide useful information in making inferences about the quality of model-data fit, some subjective interpretation is required to determine whether model-data fit is acceptable. To remove some of this subjective element, both individual-level and bin-level residuals are often quantified and used in statistical tests and indices. There are a variety of such measures with the most commonly used and readily available measures presented below.

Traditional Methods for Evaluating Model-Data Fit

Several different statistical approaches have been developed for evaluating fit, all of which involve the concept of the

Table 2. Bins Based on Ability Level, Dichotomous Item

				Bock's χ^2 and Yen's Q_1			G^2			
Bin	N_{hi1}	N_{hi}	O_{hi1}	P_{hi1}	r_{hi}	Element	$\bar{\theta}_{hi}$	P_{hi1}	$N_{hi} - N_{hi1}$	Element
1	0	2	0	0.03	−0.03	0.05	−3.6	0.03	2	0.05
2	0	2	0	0.06	−0.06	0.12	−2.8	0.06	2	0.12
3	0	2	0	0.12	−0.12	0.27	−2	0.12	2	0.25
4	1	2	0.5	0.23	0.27	0.79	−1.2	0.23	1	0.34
5	0	2	0	0.40	−0.40	1.35	−0.4	0.40	2	1.03
6	2	2	1	0.60	0.40	1.35	0.4	0.60	0	1.03
7	2	2	1	0.77	0.23	0.61	1.2	0.77	0	0.53
8	1	2	0.5	0.88	−0.38	2.71	2	0.88	1	0.87
9	2	2	1	0.94	0.06	0.12	2.8	0.94	0	0.12
10	2	2	1	0.97	0.03	0.06	3.6	0.97	0	0.05
					Sum = 7.43				2 * Sum = 8.77	

Note. N_{hi1} = number of individuals responding correctly in bin h to item i, N_{hi} = number of individuals in bin h for item i, O_{hi1} = proportion of correct responses in bin h, = predicted proportion of correct responses in bin h to item i, r_{hi} = bin-level residual, $\bar{\theta}_{hi}$ = average ability estimate in bin h for item i.

residual. While the statistical approaches are numerous in number, they all follow one of two general approaches: a chi-square approach and a likelihood-ratio approach. We first present these two general approaches and then describe specific fit indices in the context of these two general approaches.

The chi-square approach to evaluating model-data fit in dichotomous items is given by the general form of

$$\chi_i^2 = \sum_{h=1}^{H} N_{hi} \frac{(r_{hi})^2}{P_{hi1}(1 - P_{hi1}}, \tag{4}$$

where N_{hi} represents the number of people in bin h responding to item i. The chi-square statistic in (4) represents the sum of squared, standardized residuals, where the standardized residual is $r_{hi}/\sqrt{P_{hi1}(1 - P_{hi1})}$.

Notice that the chi-square approach embeds the bin-level residuals directly in the equation, such that as the residuals increase, so too do the values of the chi-square statistic.

The likelihood-ratio based statistic for dichotomous items is given by the form

$$LR_i = 2\sum_{h=1}^{H}\left[N_{hi1}\ln\left(\frac{N_{hi1}}{N_{hi}P_{hi1}}\right) + N_{hi0}\ln\left(\frac{N_{hi0}}{N_{hi}(1 - P_{hi1})}\right)\right], \tag{5}$$

where N_{hi1} and N_{hi0} represent the number of people in bin h answering item i correctly and incorrectly, respectively. While at first glance the formula for LR_i does not appear to involve residuals, some minor algebraic manipulation reveals that the natural log of residuals is involved. The term $\ln(\frac{N_{hi1}}{N_{hi}P_{hi1}})$expands to $\ln(N_{hi1}) - \ln(N_{hi}P_{hi1})$, which is the natural log of observed bin-level correct responses less the natural log of expected bin-level correct responses.

All of the widely adopted indices and tests of model-data fit adopt either the chi-square or the likelihood-ratio approach, shown above in (4) and (5). The differentiating properties of the various indices and tests of model-data fit are based on two primary dimensions. The first dimension is the manner in which bins are defined. The bins can be defined in several different ways, from having each individual serving as a unique bin, to having bins defined according to a particular number of individuals. The second dimension is the manner in which P_{hi1} is computed. Figure 4 presents a taxonomy of fit statistics according to the two dimensions described above. This taxonomy is intended to help the reader appreciate the subtle differences between the various fit statistics in frequent use in IRT. Each of these fit statistics is described below with particular attention given to the statistic's properties with respect to each of these two dimensions.

Yen's Q_1 and Bock's χ^2

Yen's Q_1 (1981) and Bock's χ^2 (1960) follow the chi-square approach of (4), but differ with respect to how each defines the bin and how they compute the expected probability of correct response for each bin. Yen's Q_1 fixes the number of bins to 10 ($H = 10$), whereas Bock's χ^2 allows for any number of bins. With respect to computing P_{hi1} for each bin, Yen's Q_1 uses the average value of P_{hi1} for the individuals in the bin and Bock's χ^2 uses the median value of P_{hi1} for the individual in the bin.

Both Yen's Q_1 and Bock's χ^2 evaluate the null hypothesis of perfect model-data fit. These statistics are distributed as chi-square with degrees of freedom (df) equal to the number of bins (H) less the number of model parameters estimated by the model. For instance, the 3PL estimates three item parameters resulting in $df = H - 3$. Yen's Q_1 and Bock's χ^2 are not standard output in any commercial IRT software packages, despite being presented in a range of widely cited sources (see Kang & Chen, 2008; Orlando & Thissen, 2000; Reise, 1990; Sinharay, 2005; Stone & Zhang, 2003).

One common criticism of Yen's Q_1 and Bock's χ^2 is their use of ability estimates ($\hat{\theta}$) for creating bins of people. As Yen (1981) discusses, a poorly fitting model could result in biased ability estimates. Thus, binning on biased ability estimates may provide an invalid item-fit statistic. Further, model-dependent $\hat{\theta}$ values prevent the true distribution of the chi-square approach from being known, with uncertainty in the correct number of degrees of freedom for these tests (Orlando & Thissen, 2000). Another limitation is the manner in which bins are created (with arbitrary cut points to place approximately equal numbers of individuals in each bin), resulting in a sample-dependent statistic (Orlando & Thissen, 2000). Very high Type I errors were also found with these statistics, particularly for short test lengths (Stone & Hansen, 2000; Stone, Mislevy, & Mazzeo, 1994; Stone & Zhang,

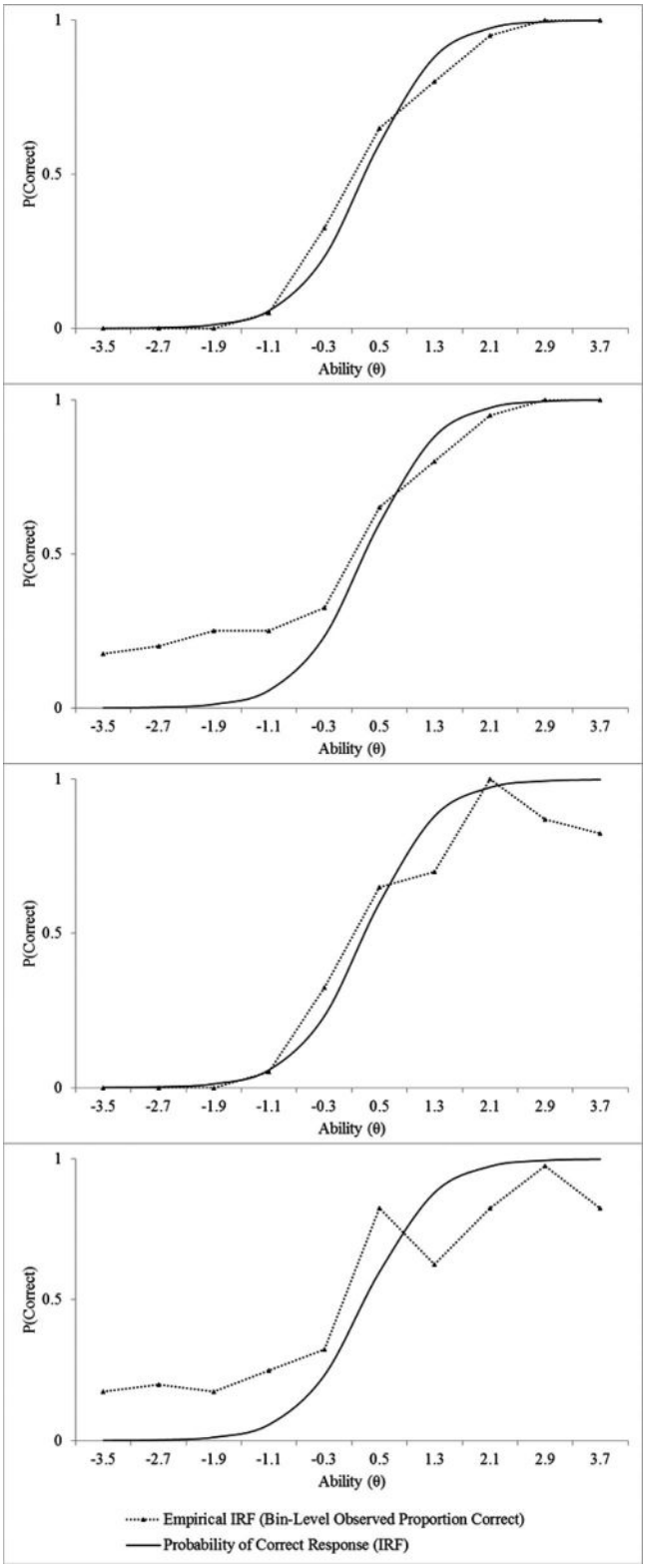

FIGURE 3. Empirical IRFs for an item following the Rasch model.

2003). These limitations highlight the need for alternative approaches for creating bins that are not based on $\hat{\theta}$.

Numeric example. To provide a numeric example of Yen's Q_1 and Bock's χ^2, let us consider the data presented in Table 1, which have been grouped into 10 bins in Table 2 (which follows the specific guidelines of Yen's Q_1) resulting in two individuals per bin ($N_{hi} = 2$). Applying the chi-square approach of (4) requires the following information for each bin: (a) the observed proportion correct, O_{hi1}; (b) the expected proportion correct based on the IRF for correct response, P_{hi1}; and (c) the number of individuals in each bin, N_{hi}. Each of these values is reported in the left side of Table 2. It is germane to note that although Yen's Q_1 and Bock's χ^2 determine the expected proportion correct differently (mean value of P_{hi1} vs. the median value of P_{hi1}), with only two individuals per bin in our example, the mean and median values of P_{hi1} in each bin are identical. For each of the 10 bins, Table 2 presents the associated residual (r_{hi}) and the squared, standardized residual. The resulting value of Yen's Q_1 and Bock's χ^2 is the sum of the elements in the column of squared standardized residuals. In this example, the values of Yen's Q_1 and Bock's χ^2 are identical, both equaling 7.43. Because the Rasch model was used (having only a single item parameter estimated) and employing 10 bins ($H = 10$) we have $df = 10 - 1 = 9$, which results in a critical value of 16.9. The value of Yen's Q_1 and Bock's χ^2 (7.43) is less than this critical value, resulting in insufficient evidence to reject the null hypothesis that the data fits the model. This leads us to conclude that the model is a reasonable fit to the data.

The G^2 Statistic

The G^2 statistic (McKinley & Mills, 1985) adopts the likelihood-ratio approach defined in (5). In applying the G^2 statistic, individuals are binned according to their ability estimate and any number of bins, similar to the approach taken by Bock's χ^2. The value of P_{hi1} for bin h is defined as the probability of correct response at the average value of $\hat{\theta}$ for the individuals in the bin. G^2 is distributed as chi-squared, with degrees of freedom equal to the number of bins, H. Commonly used calibration software programs BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) and PARSCALE (Muraki & Bock, 1997) use G^2 as the standard model-data fit statistic. As with Yen's Q_1 and Bock's χ^2, the use of ability estimates to create bins for computation of G^2 has drawn criticism (DeMars, 2005).

Numeric example. An example of the computation for G^2 is presented in Table 2. For each bin, G^2 finds the average value of $\hat{\theta}$ for the two individuals in each bin, and obtains P_{hi1} for each bin. For the first bin, the average estimated ability is $(-3.8 - 3.4)/2 = -3.6$, P_{1i1} is determined to be 0.03, and the element of G^2 for the first bin is 0.05. This is repeated for each bin, then summed, and multiplied by 2 to arrive at $G^2 = 8.44$. For $H = 10$ degrees of freedom, the chi-square critical value is 18.31, leading to the conclusion that the model-data fit is adequate.

$S - X^2$ and $S - G^2$

As previously discussed, Yen's Q_1, Bock's χ^2, and G^2 have the undesirable characteristic of relying on model-dependent θ estimates. One approach for creating bins that are θ-independent is to define bins according to observed test scores (e.g., summated scores) rather than θ estimates (Orlando & Thissen, 2000). For instance, if an assessment has 15 items, there will be 14 bins representing those earning a summated score of 1, 2, . . . , 12, 13, 14 on the assessment. The bins range from 1 to $I - 1$ because the probability of correct

Statistic	Approach	Definition of Bins	Definition of P_{hi1}
OUTFIT	chi-square	One individual per bin	Probablity of correct response at individual's value of $\hat{\theta}$
INFIT	chi-square	One individual per bin	Probablity of correct response at individual's value of $\hat{\theta}$
Yen's Q_1	chi-square	10 bins based on IRT ability estimates	The mean value of P_{hi1} for the individuals in the bin
Bock's χ^2	chi-square	Any number of bins based on IRT ability estimates	Probablity of correct response at the median value of $\hat{\theta}$ for the individuals in the bin
G^2	likelihood ratio	Any number of bins based on IRT ability estimates	Probablity of correct response at the mean value of $\hat{\theta}$ for the individuals in the bin
S-X^2	chi-square	Summated score	Obtained using a recursive algorithm
S-G^2	likelihood ratio	Summated score	Obtained using a recursive algorithm

FIGURE 4. Taxonomy of traditional model-data fit statistics.

response for an item is always zero at a summated score of 0 (in which a person answers no items correct) and is always 1 at a summated score of I (in which a person answers all items correct). Applying this approach to the chi-square form of Equation (4) yields the $S - X^2$ statistic, and applying this approach to the likelihood-ratio approach of (5) yields the $S - G^2$ statistic. For both $S - X^2$ and $S - G^2$, $df = I - 1$ less the number of model parameters estimated by the model. For instance, for a 15-item test, with the item of interest estimated using a 3-PL, $df = 15 - 1 - 3 = 11$.

Orlando and Thissen (2000) argued that because the expected proportion of correct responses from the IRF is based on model-dependent ability estimates, the statistic's distribution is unclear and conclusions drawn from these statistics may be invalid. To address this issue, $S - X^2$ and $S - G^2$ compute P_{hi1} using a recursive algorithm. The reader is referred to Lord and Wingersky (1984) for a description of this recursive algorithm. The software program IRTPRO (Cai, Thissen, & du Toit, 2011) provides the values of P_{hi1} computed via this algorithm required for the calculation of $S - X^2$ and $S - G^2$ and IRTPRO will compute values of $S - X^2$ for the practitioner.

Numeric example. To illustrate the computation of $S - X^2$ and $S - G^2$, consider an assessment composed of five dichotomous items, each item following a Rasch model. This example does not follow from Tables 1 and 2 because total scores are needed for computation of $S - X^2$ and $S - G^2$ and Tables 1 and 2 reflect data for only one item. Computation of the expected proportion correct for this item (P_{hi1}) results in 0.82, 0.97, 0.99, and 0.99 for observed score bins 1, 2, 3, and 4, respectively, with the observed proportion corrects 0.70, 0.90, 0.95, and 0.98 for observed score bins 1, 2, 3, and 4, respectively. Assume that 10 individuals scored in each of the possible observed score bins ($N_{ih} = 10$ for all h). Thus, $S - X^2$ is computed

$$S - X^2 = 10\frac{(0.7 - 0.82)^2}{0.82\,(1 - 0.82)} + 10\frac{(0.9 - 0.97)^2}{0.97\,(1 - 0.97)} + 10\frac{(0.95 - 0.99)^2}{0.99\,(1 - 0.99)} + 10\frac{(0.98 - 0.99)^2}{0.99\,(1 - 0.99)} = 9.55. \quad (6)$$

This value is then compared to a chi-square with 3 degrees of freedom ($df = I - 1 - 1 = 5 - 1 - 1$). Because the computed value of $S - X^2$ is greater than the critical value (7.81), we reject the null, concluding there is evidence in favor of model-data misfit.

OUTFIT and INFIT

Two related fit indices that follow the chi-square approach of Equation (4) are *OUTFIT* and *INFIT* (Wright & Panchapakesan, 1969). While these approaches are based on the same general form of Yen's Q_1 and Bock's χ^2, they adopt a notably different approach to bin definition. Both *OUTFIT* and *INFIT* assign only one individual per bin, such that each individual serves as a unique bin (thus, $H = N$). Because of this, the residual adopted in (4) for the chi-square approach is the individual-level residual. In addition, because there is a single individual per bin, the observed proportion correct for each bin is simply the individual's scored response to the item (i.e., 0 or 1), and the value of P_{hi1} is the value of the item's IRF for correct response at the individual's estimated ability level. *OUTFIT* and *INFIT* have been applied primarily in the context of the Rasch model (Masters, 1982; Rasch, 1960/1980).

While *OUTFIT* and *INFIT* share the property of having one individual per bin, and thus adopting the individual-level residual of (2), they differ in how much weight they assign to each individual. *OUTFIT* is computed using

$$OUTFIT_i = \frac{1}{N}\chi_i^2, \quad (7)$$

where χ_i^2 is the general chi-square form represented in (4) with one individual per bin ($N_{hi} = 1$ for all h) and N represents the total number of individuals in the sample. Because *OUTFIT* divides χ_i^2 by N, *OUTFIT* is not actually a chi-square statistic, but rather an index of the magnitude of lack of fit that can be interpreted as the typical squared, standardized residual in the sample.

Values of *OUTFIT* close to 1 indicate good model-data fit and values much greater, or much less, than 1 indicate the model-data fit is problematic. One suggestion has been to flag an item as misfitting if *OUTFIT* is less than .5 or greater than 1.5 (de Ayala, 2009, pg. 53). The .5 and 1.5 suggestions are general heuristics roughly based on critical ranges suggested by Wright and Panchapakesan (1969), but researchers have

found applying these heuristics might result in finding misfit too frequently with large samples (de Ayala, 2009, pg. 53). Wu and Adams (2013) also provide informative recommendations regarding how *OUTFIT* and *INFIT* are interpreted.

The software Winsteps (Linacre, 2014) produces *OUTFIT* as well as its transformation to a t statistic. A transformed t statistic less than -2 or greater than 2 would indicate misfit (Bond & Fox, 2007; de Ayala, 2009). However, with very large data sample sizes, these t statistics have very little tolerance for any deviation of item responses from the IRF predictions and should be interpreted cautiously (Bond & Fox, 2007).

OUTFIT assigns each individual the same weight in its computation, which can be a limitation because it can be heavily impacted by the potential of very large individual-level residuals of people for which the probability of correct response is considerably low or considerably high. *INFIT* addresses this limitation by assigning more weight to individuals having an ability level (θ) closer to the item difficulty value (b_i). An individual whose ability is close to the item's difficulty should give better insight into that item's performance than an individual who has ability that is substantially different than item difficulty. The weight assigned to each individual-level residual is equal to the Rasch model information function (see Hambleton et al., 1991 for an accessible description of the information function) at the individual's level of ability, which is given by $P_{hi1}(1 - P_{hi1})$. This leads to *INFIT* being less sensitive to extreme responses than *OUTFIT* (Bond & Fox, 2007). For this reason, stronger consideration typically is given to *INFIT* (de Ayala, 2009). The general heuristic for flagging an item as misfitting using *INFIT* is similar to that used for *OUTFIT*; items are flagged when *INFIT* values are less than .5 or greater than 1.5. Similarly, the transformed t statistic is also provided for *INFIT* in Winsteps, with values less than -2 or greater than 2 indicating poor model-data fit.

Numeric example. To illustrate the computation of *OUTFIT* and *INFIT*, consider again the hypothetical data in Table 1 containing individual-level residuals on an item for 20 people. To compute *OUTFIT*, the residual for individual 1 is computed via $x_{ni} - P_{ni1} = 0 - 0.02 = -0.02$ and then squared $(-0.02)^2 = 0.0004$. This squared residual is standardized by dividing by $P_{ni1} \times (1 - P_{ni1})$ resulting in $0.0004/(0.02 \times (1 - 0.02)) = 0.0196 = 0.0204$. We repeat this process for each bin (where each bin is a particular individual), multiplying by $N_{hi} = 1$ in each bin. *OUTFIT* is computed by summing each squared, standardized residual over all 20 bins (sum = 13.11) and dividing by $N = 20$ to arrive at $OUTFIT = 0.66$.

To compute *INFIT*, each individual's squared residual is summed over the N individuals (sum = 2.14). The product of probability P_{ni1} and $(1 - P_{ni1})$ is also summed over the N individuals (sum = 2.41). Finally, the ratio of the two sums is taken to arrive at $INFIT = .89$. Using the general heuristic approach, we see that both *OUTFIT* and *INFIT* fall inside the appropriate interval (.5 to 1.5), indicating the fit of the model to the data is adequate.

Bayesian Methods for Evaluating Fit

Bayesian methods have become increasingly popular in educational measurement because of their flexibility in evaluating complex models and the increased availability of software for Bayesian estimation. Initial applications of Bayesian methods to IRT focused on parameter estimation (see Kim & Bolt, 2007; Patz and Junker, 1999). Recently, Bayesian methods have extended to the evaluation of model-data fit in IRT (see Hoijtink, 2001; Sinharay, 2005, 2006; Sinharay, Johnson, & Stern, 2006 for examples). In this section, we describe Bayesian methods for evaluating model-data fit.

The Bayesian paradigm is founded on the notion that model parameters (e.g., a_i, b_i, c_i) are treated as unknown, but have a distribution that describes the probability that the parameter equals each possible value. For example, rather than estimating a single point value for b_i of the Rasch model for a particular item, the Bayesian paradigm asserts that b_i can assume any one of many possible values, such that the probability of b_i equaling each value follows a probability distribution. This probability distribution for a parameter is referred to as the *posterior distribution* for the parameter because it is generated after, or posterior to, collecting the data. For example, the right-hand side of Figure 5 presents the posterior distribution for a particular item's value of b_i. The majority of the distribution falls between .5 and 1.7 and we could say with some confidence that b_i for this item is somewhere in that range. In contrast, there is a much smaller chance that b_i for this item comes from the tails of the distribution; we couldn't be very confident in a claim that $b_i = -1.0$, although it is possible.

Posterior Predictive Checks

One of the most flexible approaches for evaluating model-data fit of IRT models is the use of posterior predictive checks (PPC). The concept of PPC is analogous to comparing observed and predicted responses in residual analysis. To conduct the PPC procedure, responses to the item in question are simulated from the posterior distribution (these simulated responses are denoted here by x^{sim}). The data is simulated following the process illustrated in Figure 5 and described below. An ability parameter is randomly sampled from each examinee's ability posterior distribution. This sampled ability value is then used in coordination with an item parameter which is sampled from the item's parameter posterior distribution (or parameters, depending on the model). The combination of the sampled person ability and the sampled item parameter(s) gives rise to a simulated item response. Another set of values is then sampled randomly and used to simulate another data set, which would likely arise from this second set of sampled values. This process is repeated until the desired number of x^{sim} data sets are generated. Next, a comparison is made between the simulated data set and the observed data. If the simulated data sets are similar to the observed data, the conclusion is that the model fits the data well (Lynch, 2007).

To draw conclusions regarding how similar the simulated data sets are to the observed data, tests using Bayesian p values are available. Let $T(x)$ be a statistic applied to the observed data. The statistic $T(x)$ could be any of the fit measures previously defined, such as G^2 or $S - X^2$, or other descriptors of the data. The same statistic is then applied to each of the simulated data sets ($T(x^{\text{sim}})$). This results in one value of the statistic for the observed data, $T(x)$, and multiple values for the simulated data, $T(x^{\text{sim}})$, one for each x^{sim}. The Bayesian p value is the proportion of simulated data sets whose function values $T(x^{\text{sim}})$ are greater than or equal to that

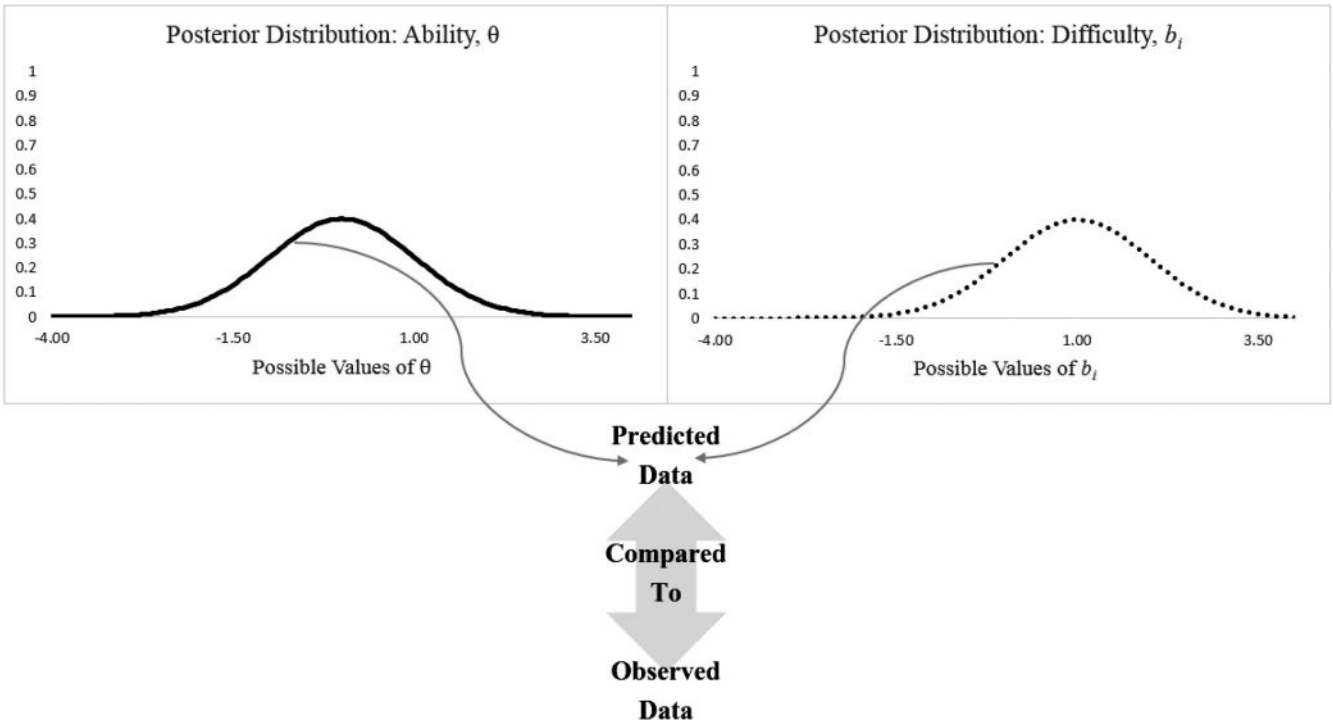

FIGURE 5. An illustration of the posterior predictive checks method.

of the function $T(x)$ applied to the observed data. Values close to 0 or 1 indicate model misfit due to the systematic differences between observed and simulated data. Typically, Bayesian p values less than .05 or greater than .95 are used to flag a misfitting item.

There is no limit on the number of statistics that could be used to obtain Bayesian p values, illustrating the flexible nature of the Bayesian method (Lynch, 2007). Careful consideration should be given to the choice of $T(x)$ used for the PPC approach. For instance, Sinharay and Johnson (2003) found that using the percentage correct (or percent response in a category) as $T(x)$ was unable to detect item misfit. Toribio and Albert (2011) investigated *OUTFIT*, Yen's Q_1, G^2, and $S - G^2$, finding all performed equally well for evaluating fit. Software for implementing Bayesian estimation methods for IRT have been made readily available, but user-friendly software for PPC is still under development, with most users relying on R packages, R script, or the BUGS language.

Numeric example. To illustrate the PPC method, consider a Rasch model fit to the data using Bayesian estimation methods. The item-fit statistic $S - X^2$ was computed for a particular item from the observed data, and the resulting value of $S - X^2$ for this item was 7.35. This is the value of $T(x)$. Next, 100 simulated data sets were generated, and for each of the 100 simulated data sets, $S - X^2$ was computed, yielding 100 values of $T(x^{\text{sim}})$ for the item in question. In this example, 96 of the simulated data sets had $S - X^2$ exceeding $T(x) = 7.35$ of the observed data. The resulting Bayesian p value is given by $96/100 = .96$. This Bayesian p value is close to 1, indicating that the simulated data do not look very much like the observed data, providing evidence that the model is a poor fit of the data for the item in question.

Concluding Remarks

Because there is no unanimously accepted measure of model-data fit, there is a need to evaluate model-data fit using several sources (Sinharay, 2006; van der Linden & Hambleton, 1997). Providing evidence of model-data fit is a necessary but not sufficient step to provide evidence that inferences from educational assessments are being drawn in a valid manner. Demonstrating model-data fit can contribute to a body of validity evidence. Only when the model is shown to be a good fit of the data, can the practitioner be confident in the adequacy of the inferences drawn.

This module presented several approaches to model fit. To guide the understanding of these approaches, a taxonomy has been introduced that makes explicit the differentiating properties of these approaches. A number of examples of each approach have been provided to illustrate its computation. Although this module focused on model-data fit in the context of dichotomous items, all approaches described here can be readily extended to polytomous items. This includes both the visual analysis of residuals, the traditional statistics for evaluating model-data fit, and the Bayesian approaches involving the PPC approach.

The PPC methods are quite flexible, incorporating a wide variety of hypotheses, but computational time can be lengthy due to the required sampling procedures (which involves a Markov chain Monte Carlo algorithm). With the preponderance of complex models, the PPC might be the only readily available option for practitioners (Sinharay & Johnson, 2003). It is our hope that the practitioner now has a

better understanding of the computation, use, and interpretation of item-fit statistics.

Self Test

1. There are two general forms that traditional model-data fit statistics follow. What are these two forms?
2. The individual-level residual can be a difficult measure of model-data fit to interpret, even under good model-data fit. What causes this difficulty in interpretation?
3. What is the difference between Yen's Q_1 and Bock's χ^2 model-data fit statistics?
4. How do *INFIT* and *OUTFIT* define each bin?
5. What is the primary difference between *INFIT* and *OUTFIT*?
6. What is the difference between G^2 and $S - G^2$?
7. Describe the concept of the posterior predictive check for evaluating model-data fit.

Answers to Self Test

1. The two general forms are the chi-square and likelihood-ratio.
2. The individual-level residual can never be zero, and cannot be near zero for individuals having ability near the item's difficulty.
3. Yen's Q_1 and Bock's χ^2 differ with respect to (a) how they define the number of possible bins, and (b) how they compute P_{hi1}.
4. *INFIT* and *OUTFIT* define each individual as a unique bin.
5. *OUTFIT* assigns equal weight to each individual, while *INFIT* assigns weight to individuals in accordance with how close the individual's ability value is to the item's difficulty parameter.
6. While G^2 defines bins according to the estimated θ and P_{hi1} in accordance with the mean value of the estimated θ for the individuals in the bin, $S - G^2$ defines bins according to the summated score and obtains P_{hi1} using a recursive algorithm.
7. PPC is similar to residual analysis. However, with PPC, observed data are compared to data simulated from posterior distributions. A statistic, such as any of the traditional item fit measures, is used to compare the observed and simulated data. When the statistic computed from the simulated data is systematically different than the observed value, we conclude model-data misfit.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological 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. 397–472). Reading, MA: Addison-Wesley.

Bock, R.D. (1960), *Methods and applications of optimal scaling.* Chapel Hill, NC: L.L. Thurstone Psychometric Laboratory.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New York, NY: Routledge.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

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

Hoijtink, H. (2001). Conditional independence and differential item functioning in the two-parameter logistic model. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays in item response theory* (pp. 109–130). New York, NY: Springer-Verlag.

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

Kang, T., & Chen, T. T. (2008). Performance of the generalized $S - X^2$ model-data fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*, 391–406.

Kim, J.-S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*(4), 38–51.

Linacre, J. M. (2014). *Winsteps® Rasch measurement computer program*. Beaverton, OR: Winsteps.com

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, *9*, 49–57.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York, NY: Springer-Verlag.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*, 49–57.

Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating scale data* [Computer software]. Chicago: Scientific Software.

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

Patz, R., & Junker, B. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests* (Copenhagen, Denmark: Danish Institute for Educational Research, 1960). Expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago, IL: The University of Chicago Press, 1980.

Reise, S. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, *14*, 127–137.

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

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*, 429–449.

Sinharay, S., Haberman, S., & Jia, H. (2011). *Fit of item response theory models: A survey of data from several operational tests*. (Research Report ETS RR-11–29). Princeton, NJ: Educational Testing Service.

Sinharay, S., & Johnson, M. (2003). *Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models* (Research Report ETS RR-03-28). Princeton, NJ: Educational Testing Service.

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

Stone, C., & Hansen, M. (2000). The effect of errors in estimating ability on goodness of fit tests for IRT models. *Educational and Psychological Measurement*, *60*, 974–991.

Stone, C., Mislevy, R., & Mazzeo, J. (1994, April). *Classification error and goodness-of-fit in IRT models*. Paper presented at the annual

meeting of the American Educational Research Association, New Orleans, LA.
Stone, C., & 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.
Toribio, S. G., & Albert, J. H. (2011). Discrepancy measures for model-data fit analysis in item response theory. *Journal of Statistical Computation and Simulation*, *81*, 1345–1360.
van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York, NY: Springer.
Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*, 339–368.
Wilson, M. (2005). *Constructing measures*. New York, NY: Taylor and Francis.
Wright, B., & Panchapakesan, N. A. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*, 23–48.
Wu, M. & Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*, 400–413.
Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245–262.
Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG for Windows: Multiple-group IRT analysis and test maintenance for binary items (Version 3.0)* [Computer software]. Chicago, IL: Scientific Software International.

