

21 Assessing the Fit of Item Response Theory Models

Hariharan Swaminathan


Handbook of Statistics

Cite this paper

Downloaded from [Academia.edu](#) 

[Get the citation in MLA, APA, or Chicago styles](#)

Related papers

[Download a PDF Pack](#) of the best related papers 



[Fitting Item Response Theory Models to Two Personality Inventories: Issues and Insights](#)
KIM YIN CHAN

[Performance of the Generalized SX 2 Item Fit Index for Polytomous IRT Models](#)

Taehoon Kang

[A generalized dimensionality discrepancy measure for dimensionality assessment in multidimension...](#)

Dubravka Svetina

Assessing the Fit of Item Response Theory Models

*Hariharan Swaminathan, Ronald K. Hambleton and
H. Jane Rogers*

1. Introduction: Models and assumptions

Item response theory provides a framework for modeling and analyzing item response data. The advantages of item response theory over classical test theory for analyzing mental measurement data are well documented (see, for example, Lord, 1980; Hambleton and Swaminathan, 1985; Hambleton et al., 1991). However, item response theory is based on strong mathematical and statistical assumptions, and only when these assumptions are met can the promises and potential of item response theory be realized.

Item response theory postulates that an examinee's performance on a test depends on the set of unobservable "latent traits" that the examinee possesses in common with the other examinees. Once these traits are defined, an examinee's observed score on an *item* is regressed on the latent traits. The resulting regression model, termed an *item response model*, specifies the relationship between the item response and the latent traits, with the coefficients of the model corresponding to parameters that characterize the item. It is this item-level modeling that gives item response theory its advantages over classical test theory.

Typically, an examinee's performance on an item is scored on a discrete scale. The most common scoring scheme is dichotomous, i.e., right/wrong, or (1/0). Polytomous scoring has become popular recently as a result of the emphasis on performance assessment and the desire to assess higher level thinking skills. In this case, the examinee's performance on an item or task is scored on an ordinal scale. In general, if U_i is the response of an examinee to item i , then $U_i = k_i$ where k_i is the category assigned to the response; $k_i = 0$ or 1 for dichotomously scored items, and $k_i = 0, 1, 2, \dots$, or $(K_i - 1)$ for polytomously scored items. The number of score categories K_i need not be the same across items. Such mixture scoring has become common in recent years, where some items are scored dichotomously while others are scored polytomously with differing numbers of response categories.

The probability that an examinee responds in category k is specified by the item response model or item response function. The item response function is a monotonically increasing function of the latent traits. In the case of a parametric representation, the item response function is a suitably chosen cumulative density function P that is

a function of the underlying set of latent abilities and the characteristics of the item. The common choices for the function $P(x)$ are the normal ogive function $\Phi(x)$ and the logistic function $L(x)$, where

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt,$$

and

$$L(x) = \frac{e^x}{1 + e^x}.$$

Given the simplicity of the logistic function, it is more commonly employed in item response theory applications. The two functions have close resemblance to each other; in fact, $|L(1.7x) - \Phi(x)| < .01$ for all x . Consequently, the constant 1.7 is frequently included in the logistic formulation of the item response model to maximize the similarity between the two functional forms. If a nonparametric representation is chosen, the item response function needs only to be a monotonically increasing function of the latent trait. For a discussion of nonparametric item response theory, the reader is referred to [Sijtsma and Molenaar \(2002\)](#). We shall only deal with parametric item response theory in this chapter.

In modeling the probability of a response, it is commonly assumed that the *complete* latent space is unidimensional; that is, one underlying latent trait, θ , is sufficient to explain the performance of an examinee on the set of items that comprises the test. Multidimensional item response models have been proposed ([Ackerman, 2005](#); [Hattie, 1985](#); [Reckase et al., 1988](#)), but these models have not reached the operational level or practical feasibility of unidimensional models.

For dichotomously scored items, under the assumption that the latent space is unidimensional, the probability of a correct response to item i is given by the three-parameter logistic model as

$$\begin{aligned} P(U_i = 1 | \theta, a_i, b_i, c_i) &= c_i + (1 - c_i) \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \\ &\equiv c_i + (1 - c_i) L[a_i(\theta - b_i)]. \end{aligned} \quad (1.1)$$

The parameters a_i, b_i, c_i are the parameters that characterize the item: b_i is the *difficulty* level of the item, a_i is the *discrimination* parameter, and c_i is the lower asymptote, known as the pseudo chance-level parameter, which is a reflection of the probability that an examinee with a very low level of *proficiency* will respond correctly to the item by chance. The model given in (1.1) is known as the three-parameter model, and is often used with multiple-choice items.

The two-parameter item response model is obtained by setting $c_i = 0$. Thus the two-parameter model has the form

$$P(U_i = 1 | \theta, a_i, b_i) = \frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \equiv L[a_i(\theta - b_i)]. \quad (1.2)$$

The one-parameter model is obtained by setting $a_i = 1, c_i = 0$, and thus has the form

$$P(U_i = 1 | \theta, b_i) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \equiv L[(\theta - b_i)]. \quad (1.3)$$

While the one-parameter model can be viewed as a special case of the three-parameter or two-parameter models, it was derived independently by Rasch (1960), based on principles of *objective measurement*. Hence the one-parameter model is commonly known as the Rasch model.

In the situation where items are scored on an ordinal scale, the probability that an examinee's score on item i falls in category s ($s = 0, 1, 2, \dots, K_i - 1$) is given by the model (Samejima, 1969):

$$P(U_i \geq s | \theta, a_i, b_{si}) = \frac{e^{a_i(\theta - b_{si})}}{1 + e^{a_i(\theta - b_{si})}}, \quad (1.4)$$

where b_{si} is the boundary for category s . Clearly $P(U_i \geq 0 | \theta, a_i, b_{si}) = 1$ and $P(U_i > K_i - 1 | \theta, a_i, b_{si}) = 0$. It follows that

$$P(U_i = s | \theta, a_i, b_{si}) = P(U_i \geq s | \theta, a_i, b_{si}) - P(U_i \geq s + 1 | \theta, a_i, b_{si}).$$

An alternative model for ordinal response data is the *partial credit model* (Andrich, 1978; Masters, 1982). The partial credit model is a generalization of the Rasch model for ordered responses and can be expressed as

$$P(U_i = s | \theta, b_{ri}) = \frac{e^{\sum_{r=1}^s (\theta - b_{ri})}}{1 + \sum_{r=1}^{K_i-1} e^{\sum_{j=1}^r (\theta - b_{ji})}} \quad (s = 0, 1, \dots, K_i - 1), \quad (1.5)$$

with

$$P(U_i = 0 | \theta, b_{ri}) = \frac{1}{1 + e^{\sum_{r=1}^{K_i-1} (\theta - b_{ri})}}.$$

The partial credit model assumes constant discrimination across items. The *generalized partial credit model* (Muraki, 1992) permits varying discrimination parameters across items and is given as

$$P(U_i = s | \theta, a_i, b_{ri}) = \frac{e^{a_i \sum_{r=1}^s (\theta - b_{ri})}}{1 + \sum_{r=1}^{K_i-1} e^{a_i \sum_{j=1}^r (\theta - b_{ji})}} \quad (s = 1, \dots, K_i - 1), \quad (1.6)$$

and for $s = 0$,

$$P(U_i = 0 | \theta, a_i, b_{ri}) = \frac{1}{1 + e^{a_i \sum_{r=1}^{K_i-1} (\theta - b_{ri})}}.$$

More recently, Tutz (1997) and Verhelst et al. (1997) have described alternate models for the analysis of partial credit data. These models can be viewed as the item response theory analog of the continuation ratio model (Agresti, 1990). The models introduced by Tutz (1997) and Verhelst et al. (1997) have the advantage over the partial and the generalized partial credit models in that the item parameters for each category can be interpreted independently of other category parameters. A detailed analysis of the continuation ratio model and related models is provided by Hemker et al. (2001).

While the procedures for fitting the dichotomous and polytomous item response models to data have been well established, the procedures for checking the assumptions

and assessing the goodness of fit of the chosen item response model have not received the same level of attention. As pointed out above, item response theory is based on strong mathematical and statistical assumptions, and only when these assumptions are met at least to a reasonable degree can item response theory methods be implemented effectively for analyzing educational and psychological test data and for drawing inferences about properties of the tests and the performance of individuals.

Checking model assumptions and assessing the fit of models to data are routine in statistical endeavors. In regression analysis, numerous procedures are available for checking distributional assumptions, examining outliers, and the fit of the chosen model to data. Some of these procedures have been adapted for use in item response theory. The basic problem in item response theory is that the regressor, θ , is unobservable; this fact introduces a level of complexity that renders a procedure that is straightforward in regression analysis inapplicable in the item response theory context.

Assessing model fit in item response theory requires the following two steps:

- (i) Checking the underlying assumptions such as unidimensionality.
- (ii) Assessing the agreement between observations and model predictions.

These two steps are described next. The important area of person-fit assessment procedures, however, is not included in this chapter. Interested readers are referred to the review by [Meijer and Sijtsma \(2001\)](#).

2. Checking the assumption of unidimensionality

When unidimensional models are fitted to test data, the validity of the assumption that the *complete* latent space is unidimensional is obviously critical. There are several approaches that can be used for investigating the assumption of unidimensionality. These include linear and nonlinear factor analysis approaches as well as nonparametric approaches that examine the assumption of local independence.

2.1. Linear factor analysis

A popular approach to investigating this assumption is to apply a linear factor analysis/principal components procedure. In this approach,

- (1) the matrix of inter-item correlations is obtained;
- (2) the percent of variance explained by the largest eigenvalue along with the point where a break occurs in the plot of the eigenvalues, or the *scree plot*, are examined;
- (3) based on the above considerations a determination is made regarding the dimensionality of the item response data.

In addition to its obvious subjectivity, this approach has several drawbacks. When the item responses are discrete, the inter-item correlations will be small. This is a consequence of the fact that discrete item responses have nonlinear relationships with the underlying ability continuum. This fact is particularly evident when the correlations are computed using phi-coefficients. As a result, even when the model is unidimensional,

the percent of variance accounted for by the dominant dimension as indicated by the largest eigenvalue will be relatively small. Tetrachoric correlations provide an improvement over phi-coefficients, but even using tetrachorics, the magnitude of the inter-item correlations will be small. Simulation studies by the authors have shown that in unidimensional item response theory applications, the largest eigenvalue of the matrix of tetrachoric correlations will typically account for only about 25 to 35 percent of the variance. Despite this problem, an investigation of the eigenvalue plot is often a useful first step in the study of test dimensionality.

One possible approach to resolving the problem of the percent of variance accounted for by the dominant eigenvalue is to fit an item response model to the data and then, using the item parameter estimates and a suitably chosen distribution of the trait θ (normal, uniform, or skewed), generate item response data. Once the data are generated, the inter-item correlation matrix is obtained and the percent of variance accounted for by the largest eigenvalue is computed. The purpose of this simulation is to obtain an indicator of the magnitude of the largest eigenvalue to be expected in the unidimensional case. The eigenvalue obtained from analysis of the real item responses can be compared with this “ideal” and a decision made regarding the unidimensionality assumption.

A similar simulation procedure can be used with polytomously scored response data. In this case, the matrix of polyserial correlations is computed with the real and simulated data, the eigenvalues extracted, and a decision made about the assumption of unidimensionality. A more refined approach is to examine the residuals after factor analyzing the simulated and real data using computer software such as LISREL 8 (Joreskog and Sorbom, 2004). The distribution of the residuals for the simulated data will provide a baseline for judging the viability of the unidimensionality assumption for the real data. A problem with this approach is that maximum likelihood procedures will fail if the matrix of tetrachoric correlations is not positive definite.

2.2. Nonlinear factor analysis and item factor analysis

The basic problem with the linear factor analysis approach is its linearity. Item response models are inherently nonlinear; in fact, the item response model is essentially a nonlinear factor model, a fact that has long been recognized. The linear factor approach to analyzing item response data can be thought of only as a first-order approximation that is not appropriate for assessing the dimensionality of item response data except in extreme cases where the item response model is linear. In recognition of this fact, Hambleton and Rovinelli (1986) suggested using the nonlinear factor analysis procedure of McDonald (1967, 1982, 1997). The approach by McDonald is to approximate unidimensional as well as multidimensional normal ogive item response functions by Hermite–Chebychev orthogonal polynomials through harmonic analysis. This approach is implemented in the program NOHARM (Fraser, 1988). As statistical procedures for assessing the dimensionality are not given in NOHARM, Gessaroli and De Champlain (1996) suggested a statistic, $X^2_{G/D}$, based on the residual correlations r_{ij} between items i and j . These residuals are transformed using Fisher’s transformation, i.e., $z_{ij} = \tanh^{-1} r_{ij}$, and the

test statistic is computed as

$$X_{G/D}^2 = (N - 3) \sum_{i=2}^n \sum_{j=1}^{i-1} z_{ij}^2,$$

where N is the sample size. This statistic is assumed to have a chi-square distribution with $[1/2n(n-1) - n(1+d)]$ degrees of freedom where n is the test length and d is the number of dimensions. For testing the assumption of unidimensionality, $d = 1$, and the degrees of freedom is $[1/2n(n-5)]$. While it is difficult to justify the distributional assumption on theoretical grounds, Gessaroli and De Champlain (1996) demonstrated through simulations that the statistic maintained the nominal Type I error rate for the unidimensional case, and had good power in detecting two-dimensional data. More recently, Gessaroli et al. (1997) developed a likelihood-ratio type statistic,

$$ALR = \sum_{i=2}^n \sum_j^{i-1} G_{ij}^2$$

with

$$G_{ij}^2 = \sum_{u_i=0}^1 \sum_{u_j=0}^1 p_{u_i u_j} \log \left(\frac{\hat{p}_{u_i u_j}}{p_{u_i u_j}} \right),$$

where $p_{u_i u_j}$ is the observed proportion of examinees with dichotomous responses u_i and u_j for items i, j ; $\hat{p}_{u_i u_j}$ is the model-predicted joint proportion of examinees with responses u_i and u_j . Gessaroli et al. (1997) have suggested that ALR has the same distribution as $X_{G/D}^2$. Maydeu-Olivares (2001) has shown that these statistics do not have the chi-square distributions claimed. However, Finch and Habing (2005) and Tate (2003) examined the behavior of ALR and concluded that while this statistic may not have the expected theoretical distribution, it is nevertheless effective in recovering the number of underlying dimensions.

Item factor analysis, developed by Bock et al. (1988) and implemented in the software TESTFACT, allows the factor analysis to be carried out using either the matrix of tetrachoric correlations or a full-information maximum likelihood procedure. De Champlain and Gessaroli (1998) found high Type I error rates with TESTFACT. Tate (2003), on the other hand, found that TESTFACT was able to accurately confirm the hypothesis of unidimensionality when the data were indeed unidimensional, except in cases of extreme item difficulty.

2.3. Examination of local independence

Lord and Novick (1968) have shown that the assumption that the latent space is complete is equivalent to the assumption of local independence. They argued that if the latent space is complete, no additional traits are necessary to account for the responses to the items, and hence the responses to items will be statistically independent when conditioned on the complete set of latent traits. This assertion is analogous to the assertion in linear factor analysis that the correlations among the observed variables will be

zero after partialing out or conditioning on the complete set of common factors. Hence, testing the assumption of local independence provides a test of the dimensionality of the test data.

Statistical independence is a strong requirement; a weaker assumption is to require that the covariances among the items be zero when the item responses are conditioned on the complete set of latent traits. In particular, if the test is unidimensional, then conditioning on the latent trait will result in the item covariances being zero. Stout (1987, 1990) used this weaker form of local independence to develop a test for “essential” unidimensionality. Stout argued that a test of length n is essentially unidimensional if the average covariance over all pairs of items is small in magnitude when conditioned on the latent trait. This procedure is implemented in the program DIMTEST (Stout et al., 1991).

The basic procedure in DIMTEST is to: (a) Create an *assessment test* made up of a core subset of m items that is unidimensional using expert judgment or factor analytic methods; (b) create a *partitioning test* using the remaining $n - m$ items to score and group the examinees into subgroups. These subgroups are homogeneous with respect to their test score on the partitioning test and hence are “conditioned” on the latent trait. The score

$$Y_{jk} = \sum_{i=1}^m U_{ijk}, \quad j = 1, \dots, J_k,$$

of each examinee in subgroup k on the assessment test is computed, and the variance of these scores, $\hat{\sigma}_k^2$, is obtained. If local independence holds, then

$$\text{Var}(Y_{jk}) = \sum_{i=1}^m \text{Var}(U_{ijk}) = \sum_{i=1}^m \hat{p}_{ik}(1 - \hat{p}_{ik}) \equiv \hat{\sigma}_{U_k}^2.$$

Thus, if the assumption of unidimensionality is met, these two quantities are the estimates of the same variance. Stout (1987) demonstrated that

$$E(\hat{\sigma}_k^2 - \hat{\sigma}_{U_k}^2 | X) \propto \sum \text{Cov}(U_i, U_j | X), \quad (2.1)$$

where X is a score on the partitioning test. Since in the weak form of local independence, the pairwise covariances are zero, a statistic based on $(\hat{\sigma}_k^2 - \hat{\sigma}_{U_k}^2)$ provides a test of essential unidimensionality. Stout (1990) showed that the test statistic can be further refined by correcting for the bias using another subset of unidimensional items, known as *assessment subtest 2*, chosen to be as similar as possible in difficulty to assessment subtest 1.

Hattie et al. (1996) evaluated the performance of DIMTEST using simulated data. They concluded that the Type I error rate was in the acceptable range and that the procedure was sufficiently powerful to reject the null hypothesis when the test data was not unidimensional. The primary drawback with this procedure is the identification of the assessment subtests. This identification is subjective and will affect the test of essential unidimensionality. A further drawback is that the test under investigation must be sufficiently long to support the selection of the assessment subtests. External subtests may be used but these will not always be available. It should be pointed out, too, that

DIMTEST is designed to test the hypothesis that the test is unidimensional – it is not meant for assessing the dimensionality of the test. Finally, Hattie et al. (1996) showed that the test statistic is adversely affected if the tetrachoric correlation matrix is not positive definite, and recommended that the procedure should not be used in this case.

Yen (1984) examined the issue of local independence from the point of view that items based on a common stimulus may violate the local independence assumption. She suggested examining the correlations between residuals $(u_{ia} - P_{ia})$ and $(u_{ja} - P_{ja})$ across pairs of items $i, j = 1, \dots, n, i \neq j$, across examinees $a = 1, \dots, N$, where u_{ia} is the response of examinee a to item i and P_{ia} is the model-based probability of a correct response. The test statistic is $Q_3 = r_{ij}$, the correlation between the residuals. Yen (1984) indicated that the hypothesis of local independence for pairs of items is equivalent to the hypothesis that the correlation between the residuals is zero. One problem with this approach is that residuals may not be linearly related. A second problem noted by Chen and Thissen (1997) is that the item residuals are not bivariate normally distributed, and hence the distribution of residuals will not have the expected mean and variance. Chen and Thissen (1997) showed the empirical Type I error rates for the Q_3 statistic are higher than the nominal Type I error rates. Glas (2005) and Glas and Suarez Falcon (2003), following the development given by van den Wollenberg (1982) for the Rasch model, provided a statistic for testing local independence based on the difference between observed and expected frequencies

$$d_{ij} = N_{ij} - E(N_{ij}),$$

where N_{ij} is the observed number of examinees responding correctly to items i and j in the group of examinees obtaining a score between $k = 2$ and $k = n - 2$, and

$$E(N_{ij}) = \sum_k N_k P(U_i = 1, U_j = 1 | X = k). \quad (2.1)$$

The joint probability is computed using a recursive algorithm based on that provided by Lord and Wingersky (1984) (see the next section for details of this algorithm). Following the statistic given by van den Wollenberg (1982), Glas and Suarez Falcon suggested the statistic

$$S_{3ij} = d_{ij}^2 \left\{ \frac{1}{E(N_{ij})} + \frac{1}{E(N_{i\bar{j}})} + \frac{1}{E(N_{\bar{i}j})} + \frac{1}{E(N_{\bar{i}\bar{j}})} \right\},$$

where $E(N_{i\bar{j}})$ is defined as in (2.1) with the probability that an examinee with score k responds incorrectly to item i and correctly to item j ; $E(N_{\bar{i}j})$ is defined similarly, while $E(N_{\bar{i}\bar{j}})$ is computed with the examinee responding incorrectly to the items. Simulation studies have indicated that the statistic behaves like a chi-square variate with 1 degree of freedom.

Glas and Suarez Falcon (2003) compared the behavior of Yen's Q_3 statistic with the performance of the S_3 statistic and concluded that the Type I error rates for these statistics were below the nominal rate, and that they did not have much power. Their results agreed with previous findings that although local independence was violated, the item response curves were well recovered, and hence robust to violations of the assumption of local independence.

3. Assessing model data fit: Checking model predictions

Assessing whether the test data is unidimensional is only the first step in determining if the assumptions of item response theory are met. Assessment of the fit of the model to data is multi-faceted and must be carried out at the test level as well as the item level. As in any scientific investigation, verification of a theory is most directly carried out by examining the predictions made by the theory. In item response theory, this is accomplished by comparing what the theory predicts with what is observed.

3.1. Assessment of model fit at the test level

Likelihood approaches

The assessment of the fit of model at the test level can be carried out using the likelihood ratio statistic. The probability of a response pattern $U = [u_1 u_2 \dots u_n]$ for the item response data is

$$P(U) = \int \prod_i P_i(\theta)^{u_i} (1 - P_i(\theta)^{1-u_i}) g(\theta) d\theta \quad (3.1)$$

for the dichotomous case; $P_i(\theta)$ is the probability of a correct response for an examinee on item i given by the item response function specified in Eqs. (1.1) to (1.3), U_i is the response of an examinee to item i , and $g(\theta)$ is the density function of θ . The likelihood function is then given by

$$L = \prod_{\text{response patterns}} P(U)^{f_u}, \quad (3.2)$$

where f_u is the frequency of response pattern U .

In the ideal case, a table that contains all response patterns can be constructed, and based on this table, a likelihood ratio statistic can be calculated. Unfortunately, with n items, there are 2^n response patterns and this number gets large rapidly. Hence it is not feasible to carry out the contingency table approach when the number of items is even moderately large.

When $n \leq 10$, and when the number of examinees is large, it is possible to construct a table of frequencies for all response patterns. Let f_1, f_2, \dots, f_{2^n} denote the frequencies for the 2^n response patterns. The likelihood ratio chi-square statistic for the test of fit is

$$G^2 = 2 \sum_{r=1}^{2^n} f_r \log \frac{f_r}{N \hat{P}(U_r)}, \quad (3.3)$$

where $P(U)$ is given by (3.1) and approximated as

$$\hat{P}(U) = \sum_{k=1}^q P(U|X_k) A(X_k).$$

Here, X is a quadrature point and $A(X)$ is the weight corresponding to the density function $g(X)$. The quantity G^2 defined in (3.3) has an asymptotic chi-square distribution

with degrees of freedom equal to $2^n - kn - 1$, where k is the number of item parameters in the model.

When the number of items is large, the procedure given above is not feasible. In this case, traditional likelihood ratio tests are applicable. The fit of the model to the data can be determined by computing $-2 \log \lambda$, where λ is the maximum value of the likelihood function given in (3.2) for the fitted model. Under the null hypothesis that the model fits the data against the alternative that the probability of a correct response is random, $-2 \log \lambda$ is asymptotically distributed as a chi-square variate with degrees of freedom equal to the number of item parameters estimated.

The likelihood ratio statistic described above can also be used to test the fit of the three-parameter model against the two-parameter model, and the fit of the two-parameter model against the one-parameter model. The likelihood ratio $-2 \log(\lambda_A/\lambda_B)$ is computed, where λ_A and λ_B are the maximum values of the likelihood functions (3.1) obtained by fitting the nested models A and B. This ratio is asymptotically distributed as a chi-square variate with degrees of freedom equal to the number of additional parameters estimated in model A in comparison with model B.

In the case of the Rasch model, a likelihood ratio test based on the distribution of item responses conditional on number-right score, the sufficient statistic, is available (Andersen, 1973). This procedure has an advantage over the likelihood ratio tests based on the joint likelihood of item and ability parameters, since in the joint procedure item and ability parameters must be estimated simultaneously, and hence do not yield consistent estimates of item parameters. The marginal maximum likelihood procedure (Bock and Aitkin, 1981) solves this problem and hence there is very little difference between the conditional likelihood ratio procedure and the "marginal" likelihood ratio procedure.

The extension of the likelihood ratio test described above to polytomous item response models is straightforward. The fit of the generalized partial credit model may be compared with the partial credit model (Eqs. (1.5) and (1.6)), for example. The same comparison can be made with the graded response model (Eq. (1.4)) for testing nested hypotheses.

Likelihood ratio tests for testing fit at the item level are also available. These are described in the next section. The likelihood ratio fit statistic at the item level can be added across items to assess model fit at the test level.

Comparison of observed and expected score distributions

The likelihood approach described above is one approach to assessing model data fit. As mentioned earlier, assessing fit is a multifaceted approach and other approaches should be looked at to arrive at a complete picture of model-data fit. One such approach is to compare the observed score distribution with the predicted score distribution. Once the item response model is fitted to the data, it is possible to obtain the predicted number-correct score distribution.

Lord (1980) showed that for a test of length n , the distribution of number-correct score x , $f_n(x|\theta)$ is the compound binomial distribution, and that the relative frequency of a score x can be obtained from the coefficient of t^x in the expansion of the moment generating function $\prod_{i=1}^n (P_i t + Q_i)$ where Q_i is the probability of an incorrect response to item i . For example, for a test of length $n = 2$, the number-correct score $x = 0, 1$,

or 2. These scores occur with relative frequency

$$\begin{aligned}f_2(x = 0|\theta) &= Q_1 Q_2, \\f_2(x = 1|\theta) &= P_1 Q_2 + Q_1 P_2, \\f_2(x = 2|\theta) &= P_1 P_2.\end{aligned}$$

For long tests the computation of the distribution of number-correct scores using the generating function is extremely tedious, and Lord and Novick (1968) concluded that approximations are inevitable in such situations. However, Lord and Wingersky (1984) developed a simple recursive algorithm for computing the frequency distribution of the number-correct score. The algorithm capitalizes on the fact that if the frequency distribution $f_{n-1}(x|\theta)$ is determined, then when the test is increased by one item, a score $x = k$ is obtained on the n -item test if the examinee obtains a score of $x = k$ on the test with $(n - 1)$ items and responds incorrectly to the n th item, or obtains a score of $(k - 1)$ on the test with $(n - 1)$ items and responds correctly to the n th item, i.e.,

$$f_n[x = k|\theta] = f_{n-1}[x = k|\theta] * Q_n + f_{n-1}[x = k - 1|\theta] * P_n. \quad (3.4)$$

It is relatively straightforward to obtain the conditional distribution of number-correct scores using this algorithm even for very long tests. This algorithm readily extends to the polytomous case.

Once this predicted conditional number-correct score distribution is obtained, the marginal number-correct score distribution is obtained by summing over the examinees at their estimated ability values, $\hat{\theta}_a$:

$$f_n(x) = \sum_{a=1}^N f_n(x|\hat{\theta}_a). \quad (3.5)$$

Alternatively, by assuming a density function of θ , $f_n(x)$ is obtained as

$$f_n(x) = \int f_n(x|\theta)g(\theta) d\theta. \quad (3.6)$$

The observed number-correct score distribution is then compared with the model predicted number-correct score distribution and the results are displayed graphically (Hambleton and Traub, 1973). Alternatively, a graphical display of the cumulative distributions may be obtained (see, for example, Hambleton and Han, 2005). A third option is to plot quantiles against each other. The advantage of such Q-Q plots is that if the two distributions are identical, the Q-Q plot will be a straight line. Departures from the straight line are easier to spot than differences in the frequency distributions or the cumulative frequency distributions.

An example of the procedure described above is provided in Figures 1 to 3, where the observed distribution is compared with the frequency distribution obtained after fitting one-, two-, and three-parameter models to an achievement test. The figures indicate that the three-parameter model produces the closest fitting frequency distribution to the observed score distribution. The one-parameter model provides the poorest fit. In these plots, the marginal score distribution is obtained by summing over the estimated

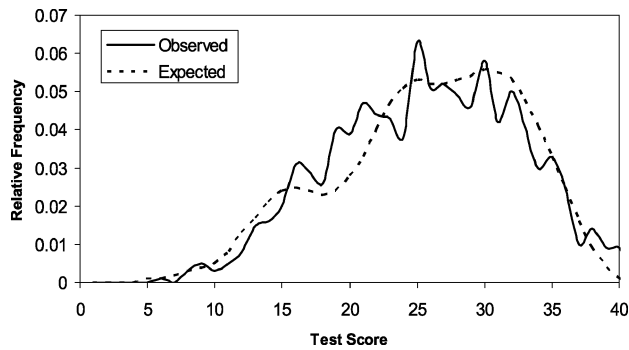


Fig. 1. Observed and expected score distributions based on trait estimates for the one-parameter model.

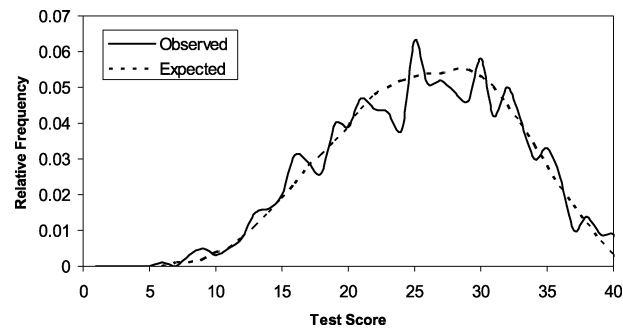


Fig. 2. Observed and expected score distributions based on trait estimates for the two-parameter model.

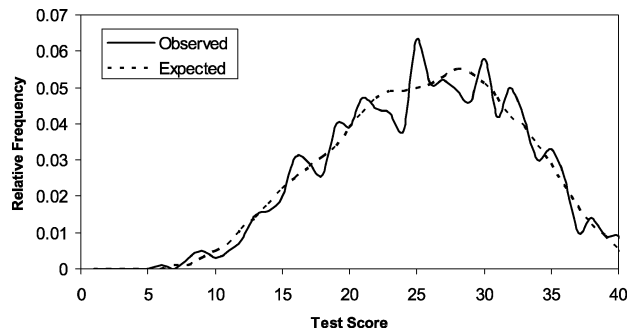


Fig. 3. Observed and expected score distributions based on trait estimates for the three-parameter model.

θ values. The marginal distributions under the assumption that θ is normally distributed are provided in [Figures 4 to 6](#). The differences between the distributions obtained with trait estimates and under the assumption of a normal distribution are small, and the conclusion regarding the fit of the model remains unaltered.

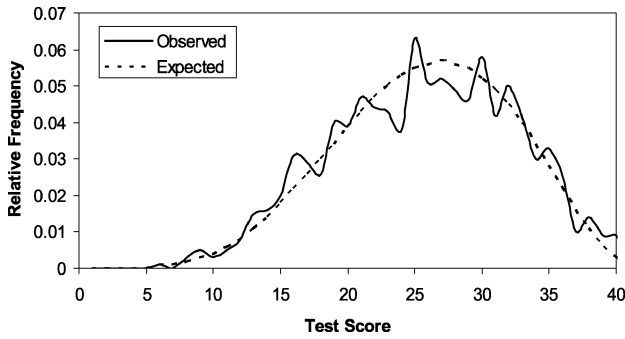


Fig. 4. Observed and expected score distributions assuming a normal trait distribution for the one-parameter model.

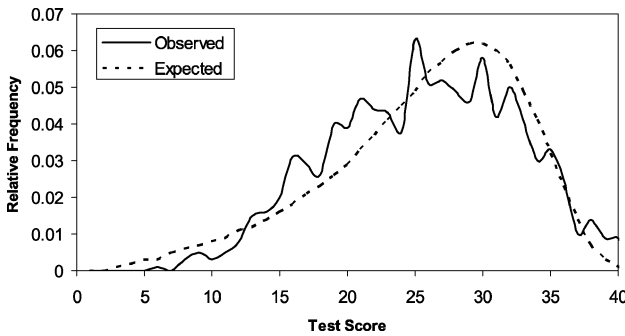


Fig. 5. Observed and expected score distributions assuming a normal trait distribution for the two-parameter model.

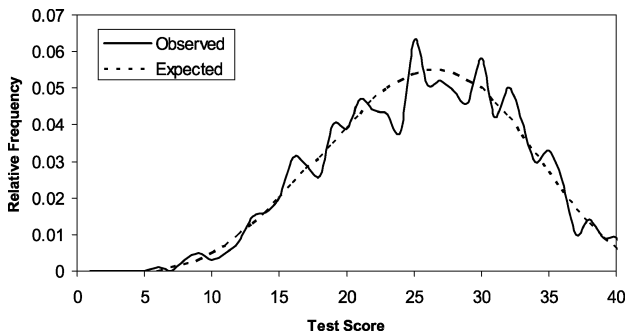


Fig. 6. Observed and expected score distributions assuming a normal trait distribution for the three-parameter model.

The observed and expected distributions may be compared statistically by employing the traditional chi-square test using the statistic

$$X^2 = \sum_{i=1}^m \frac{(f_{oi} - f_{ei})^2}{f_{ei}},$$

where f_{oi} and f_{ei} are observed and expected frequencies. The statistic is distributed as a chi-square with $m - 1$ degrees of freedom, where m is the number of score groups. When f_{ei} is less than one, it is common to combine score categories to achieve a minimum expected cell frequency of at least one. Alternatively, the Kolmogorov–Smirnov (K–S) procedure for comparing two cumulative distributions may be used.

3.2. Assessment of model fit at the item level

Assessment of model fit at the item level takes several forms since the theory may be used to predict a variety of outcomes at the item level. These include comparing the observed probability of a correct response to the model-predicted probability of a correct response at various points on the θ continuum or at observed score levels, examining invariance of item parameters across subgroups of examinees, and examining the invariance of ability estimates for partitions of the test into subsets of items. In one sense, assessing fit at the item level is a more stringent test of model fit than assessing fit at the test level, since it can be expected that not all conditions and requirements will be met.

The primary tools for assessing model fit at the item level are discrepancy measures between observations and expectations (Hambleton and Han, 2005; Hambleton et al., 1991; Rogers and Hattie, 1987). These discrepancy measures lend themselves to both graphical displays and statistical tests of significance. While statistical significance tests have been the mainstay in the social and behavioral sciences, they can often be noninformative in the context of item response theory. Graphical displays, on the other hand, while suffering from a certain degree of subjective interpretation, can often provide meaningful insights into the nature of model–data misfit and have been recommended strongly by Hambleton and Rogers (1990), and Hambleton et al. (1991).

Residual analysis

The most frequently used and intuitively appealing approach to assessing IRT model–data fit is to compare predictions based on the model with what was actually observed. Such analysis of residuals is common in model–fitting enterprises. In the IRT context with dichotomous item responses, the model provides the probability of success for an examinee with a given trait value. This probability is computed using the item parameter estimates obtained from the calibration. In practical terms, this probability is the proportion of examinees at that trait value that answers the item correctly. If a sufficient number of examinees with the same trait value were available, this proportion could be compared with the observed proportion correct at that trait value and a residual computed.

Two problems with this approach occur in reality: First, only estimates of trait values are available, and second, few individuals will have identical trait estimates except in the case of the one-parameter model, where there is a one-to-one correspondence between trait value and total score. The most common empirical solution to this problem is to group examinees according to their trait estimates and treat each group as if all examinees were at the same trait value. The probability of a correct response is computed for examinees in that subgroup using a representative trait value (mean or median, for

example), or by averaging the probabilities of a correct response for examinees in the interval. This result provides the model-based or expected proportion correct for that interval.

In practice, constructing the trait intervals is not trivial. The requirement of sufficient sample size within intervals to obtain reliable estimates of the observed proportion correct must be balanced against that of sufficiently homogeneous subgroups to make the expected proportion correct a representative value for the interval. At the extremes of the trait continuum, there are few examinees, and hence either the range over which the intervals are constructed must be restricted or the intervals broadened at the extremes.

Of the two criticisms raised, the second does not apply to the Rasch model. In the Rasch model, the number correct score is a sufficient statistic for the trait θ and hence there is a one-to-one correspondence between the number-right score and the trait level. It is therefore possible to compare the observed proportion correct at each score value to the expected proportion correct, i.e., the model-predicted proportion correct score, without resorting to the trait estimates. Residual analysis of this nature was first introduced in the context of the Rasch model by Wright and Panchapakesan (1969) and is a tool that is used commonly for assessing item fit in the Rasch model.

Graphical displays are useful in examining the discrepancy between observed and expected proportions correct. An example of a graphical fit analysis is shown in Figures 7 to 9. In these figures, an item from an achievement test was calibrated using one-, two-, and three-parameter models, and the observed and expected proportions correct calculated in the manner described above. In Figure 7, the plot shows that the model fails to fit the data at the lower end of the trait continuum; the observed proportion correct is considerably higher than the expected proportion correct in the lower intervals. Because of the lower asymptote of zero imposed by the one-parameter model, the model cannot account for examinee guessing and hence does not provide adequate fit. In Figure 8, the fit of the model is improved; this is accomplished in the two-parameter model by lowering the discrimination parameter for the item so that the lower asymptote of zero

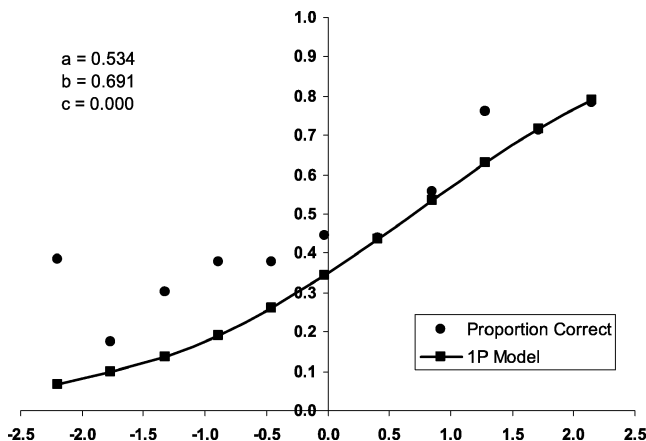


Fig. 7. Fit plot for an item calibrated with the 1P model.

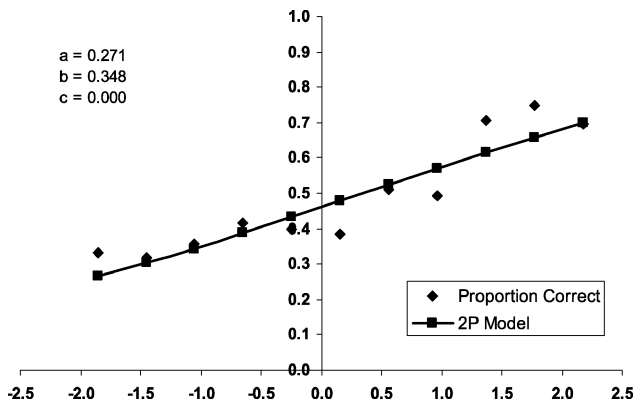


Fig. 8. Fit plot for an item calibrated with the 2P model.

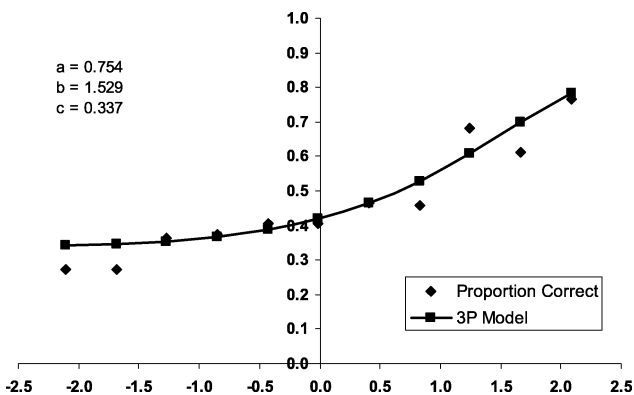


Fig. 9. Fit plot for an item calibrated with the 3P model.

is reached at much lower values of the trait. In Figure 9, where a three-parameter model was fitted, a nonzero lower asymptote is permitted and hence the model is better able to reproduce the observed proportion-correct values. The widely used calibration program BILOG-MG (Zimowski et al., 1996) produces fit plots such as those shown above along with error bands for interpreting departures from the model.

In the case of dichotomous item responses, residuals are obtained by subtracting the expected proportion correct from the observed proportion correct. The residual is generally standardized by dividing by the standard error of the observed proportion correct under the null hypothesis of model-data fit (Hambleton et al., 1991; Wright and Panchapakesan, 1969). Hence, the standardized residual has the form

$$SR_j = \frac{(O_j - E_j)}{\sqrt{\frac{E_j(1-E_j)}{N_j}}}, \quad (3.7)$$

where O_j is the observed *proportion* correct in trait interval j , E_j is the expected *proportion* correct in the interval under the fitted model, and N_j is the number of examinees in the trait interval. In the case of polytomous item responses, residuals can be calculated within each response category. While there is no statistical theory to support the assumption of normality, simulation studies show that the residuals have an approximately normal distribution when the model fits the data (Hambleton et al., 1991).

The distribution of the residuals can be visually inspected to determine the proportion of residuals that are unusually large. The percentage of standardized residuals outside the range $(-2, 2)$ provides an indicator of adequacy of model–data fit. As another interpretational tool, Hambleton et al. (1991) suggested comparing the distribution of standardized residuals to one obtained from analysis of simulated data generated using the item parameter estimates obtained from the calibration of the real data. This distribution provides a baseline for interpreting the extent to which the residual distribution departs from what could be expected when the model fits the data.

Chi-square item fit statistics

Chi-square statistics would seem to be natural tests of the discrepancy between the observed and expected frequencies or proportions computed in a residual analysis. Both Pearson and likelihood ratio statistics have been proposed; these statistics have been standard tools for assessing model fit since the earliest applications of IRT. In its most general form, the Pearson chi-square fit statistic is

$$X^2 = \sum_{j=1}^J \sum_{k=1}^K N_{jk} \frac{(O_{jk} - E_{jk})^2}{E_{jk}}, \quad (3.8)$$

where j indexes the trait interval and k indexes the response category. In the case of dichotomous item responses, this reduces to

$$X^2 = \sum_{j=1}^J N_j \frac{(O_j - E_j)^2}{E_j(1 - E_j)}. \quad (3.9)$$

Variations of this statistic are obtained by varying the numbers of intervals, varying the methods of constructing the intervals (equal frequency or equal width), and varying the methods of obtaining the expected proportion correct (using the mean or median trait value in the interval or averaging the probabilities of response for individuals in the interval). Bock (1972) used J intervals of approximately equal size and used the median trait value in the interval to obtain the expected proportion correct. Yen (1981) specified 10 intervals of equal frequency and used the mean of the probabilities of a correct response as the expected proportion correct. Statistics of this form are generally assumed to have an approximate chi-square distribution with degrees of freedom equal to $J * (K - 1) - m$ degrees of freedom, where m is the number of estimated item parameters in the model. Hence for Yen's statistic (referred to as Q_1), for example, the degrees of freedom are taken as $10 - 3 = 7$ for dichotomous responses under the three-parameter model.

The likelihood ratio statistic has the general form

$$G^2 = 2 \sum_{j=1}^J \sum_{k=1}^K N_j \left(O_{jk} \log \frac{O_{jk}}{E_{jk}} \right),$$

reducing to

$$G^2 = 2 \sum_{j=1}^J N_j \left(O_{jk} \log \frac{O_j}{E_j} + (1 - O_{jk}) \log \frac{1 - O_j}{1 - E_j} \right) \quad (3.10)$$

in the case of dichotomous item responses. This is the item fit statistic provided in BILOG-MG. [McKinley and Mills \(1985\)](#) assumed the degrees of freedom for the distribution to be $J(K - 1) - m$. [Mislevy and Bock \(1990\)](#) argue that since parameter estimation is not based on minimization of the chi-square statistic, no degrees of freedom are lost for the number of item parameters estimated.

A number of problems arise in using chi-square statistics as tests of model-data fit in the IRT context. Principal among them is whether the statistics have the chi-square distribution claimed and if so, whether the degrees of freedom are correctly determined. [Glas and Suarez Falcon \(2003\)](#) note that the standard theory for chi-square statistics does not hold in the IRT context because the observations on which the statistics are based do not have a multinomial or Poisson distribution.

Simulation studies ([Yen, 1981](#); [McKinley and Mills, 1985](#); [Orlando and Thissen, 2000, 2003](#)) have shown that the fit statistics in common use do generally appear to have an approximate chi-square distribution; however, the number of degrees of freedom remains at issue. [Orlando and Thissen \(2000\)](#) argued that because the observed proportions correct are based on model-dependent trait estimates, the degrees of freedom may not be as claimed. [Stone and Zhang \(2003\)](#) agreed with the assessment of [Orlando and Thissen \(2000\)](#) and further noted that when the expected frequencies depend on unknown item and ability parameters, and when these are replaced by their estimates, the distribution of the chi-square statistic is adversely affected. [Yen \(1981\)](#), however, concluded that since trait estimates are based on all items, the loss of degrees of freedom due to the use of trait estimates is negligible in the computation of the fit statistic for a single item.

Another obvious problem with chi-square fit statistics computed as described above is that the trait intervals are arbitrary; different choices of intervals will lead to different values of the fit statistics and potentially, to different conclusions about the fit of individual items. A further issue is that of the minimum interval size needed for a chi-square approximation to be valid (disbelief suspended on all other fronts). The usual recommendation is an expected cell frequency of 5.

Given the forgoing, it is not surprising that alternatives to the chi-square statistics described above have been sought. [Orlando and Thissen \(2000\)](#) addressed the issue of the use of trait estimates to establish the intervals in which residuals are computed. They proposed instead that examinees be grouped according to their number correct score, which is not dependent on the model. The recursive formula described in an earlier section for computing the frequency distribution of number correct score given

trait value (Eq. (3.4)) is used in obtaining expected proportion correct at each score level. The expected proportion correct of examinees with total score k who get item i correct is:

$$E_{ik} = \frac{\int P_i f(k-1|\theta)g(\theta) d\theta}{\int f(k|\theta)g(\theta) d\theta}. \quad (3.11)$$

Once the expected proportions are obtained and the observed proportions calculated from the data, a Pearson chi-square fit statistic (referred to as $S - X^2$ by the authors) or a likelihood ratio statistic (referred to as $S - G^2$) is computed in the usual manner as given by Eqs. (3.9) and (3.10). The two statistics are:

$$S - X_i^2 = \sum_{k=1}^{n-1} N_k \frac{(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})} \quad (3.12)$$

and

$$S - G_i^2 = 2 \sum_{k=1}^{n-1} N_k \left(O_{ik} \log \frac{O_{ik}}{E_{ik}} + (1 - O_{ik}) \log \frac{1 - O_{ik}}{1 - E_{ik}} \right). \quad (3.13)$$

Note that O_{ik} and E_{ik} are observed and expected proportions. The degrees of freedom for the statistic are the number of score levels minus the number of estimated item parameters. If all score levels are used in an n -item test, then the degrees of freedom are $(n-1) - m$, where m is the number of parameters in the model. Score levels may be collapsed from the extremes inwards to ensure a desired minimum expected cell frequency. Orlando and Thissen (2000) argued that a minimum expected cell frequency of 1 is acceptable for the chi-square statistic they proposed.

Orlando and Thissen (2000, 2003) compared their fit statistics with that of Yen and with a likelihood ratio statistic and found that the Type I error rates for the Yen Q_1 and likelihood ratio statistics were unacceptably high in short tests, while the Type I error rates for $S - X^2$ remained close to the nominal level. Type I error rates were somewhat higher for $S - G^2$, and the authors recommend use of $S - X^2$ in preference. In longer tests, the Type I error rates improved for Q_1 and were slightly above the nominal values. Glas and Suarez Falcon (2003) found the false alarm rates for $S - G^2$ to be high in short tests with a 10% of misfitting items.

Stone et al. (1994) and Stone (2000), like Orlando and Thissen (2000), point out that the uncertainty in the θ estimates may be responsible for the departures of the chi-square statistics from their expected behavior. They suggested that since the posterior distribution of θ takes into account the uncertainty of θ estimates, using the posterior distribution of θ to develop a fit measure has clear advantages over procedures that use point estimates of θ .

In the E -step of the $E-M$ algorithm for estimating item parameters using a marginal maximum likelihood procedure (Bock and Aitkin, 1981), the expected number of examinees at trait value X_k is

$$\tilde{N}_{ik} = \sum_{j=1}^N P(X_k | \underline{U}_j, \underline{\xi}) = \sum_{j=1}^N \frac{L(X_k)A(X_k)}{\sum_{k=1}^q L(X_k)A(X_k)}. \quad (3.14)$$

Here, $L(X_k)$ and $A(X_k)$ are the likelihood function and the quadrature weights, respectively, evaluated at quadrature point X_k . Similarly, the expected number of examinees answering the item correctly at trait value X_k is

$$\tilde{r}_{ik} = \sum_{j=1}^N u_{ij} P(X_k | \underline{U}_j, \underline{\xi}) = \sum_{j=1}^N \frac{u_{ij} L(X_k) A(X_k)}{\sum_{k=1}^q L(X_k) A(X_k)}, \quad (3.15)$$

where u_{ij} is the response of examinee j to item i . In their approach, Stone et al. (1994) and Stone (2000) use these *pseudocounts*, \tilde{r}_{ik} and \tilde{N}_{ik} , to assess the fit of the item. If P_k is the value of the item response function at quadrature point X_k , then the “observed” frequency is $O_{ik} = \tilde{r}_{ik}$, and the expected frequency is $E_{ik} = \tilde{N}_{ik} P_{ik}$. Using these values the, the X^2 statistic is computed as

$$X_i^2 = \sum_{k=1}^K \frac{(\tilde{r}_{ik} - \tilde{N}_{ik} P_{ik})^2}{\tilde{N}_{ik} P_{ik} (1 - P_{ik})}. \quad (3.16)$$

The G^2 statistic is constructed similarly.

Unfortunately, as noted by Donoghue and Hombo (2003), with pseudocounts, the contribution of an examinee is to more than one cell. Hence the use of pseudocounts violates the assumption of independence, and consequently X^2 and G^2 do not have chi-square distributions. Stone et al. (1994) demonstrated through simulation studies that the distribution of these quantities can be approximated by a scaled chi-square distribution. However, the degrees of freedom and the scale factor are not known and must be determined either analytically or through a resampling technique. Stone (2000) used the resampling technique to determine the degrees of freedom and the scale factor. In the resampling technique, replicate data sets are simulated using the item parameter estimates of the original data set. For each replication, the item parameters are estimated, the G^2 (and/or X^2) statistic is computed, and the mean and the variance of the empirical distribution of the statistic are calculated. If the scale factor is denoted by λ and the degrees of freedom by ν , then the mean and the variance of the scaled chi-square variate are $\lambda\nu$ and $2\lambda^2\nu$, respectively. From the mean and variance of the empirical distribution, the scale factor and the degrees of freedom are determined.

The major advantage of the Stone procedure is that the pseudocounts are the natural by-products of the E-M algorithm, the procedure employed in BILOG-MG, PARSCALE, and MULTILOG. A further advantage is that the procedure extends effortlessly to polytomous item response models. The major drawback is that the procedure is computationally intensive. A computer program, however, is available (Stone, 2004) for carrying out the fit analysis.

Donoghue and Hombo (1999, 2001, 2003) have derived the asymptotic distribution of the statistic described above under the assumption that item parameters are known. They demonstrated that under this restrictive assumption the statistic behaved as expected.

Stone and Zhang (2003) compared the procedure described above with several procedures including that of Orlando and Thissen (2000), and Bock (1972), and that of Donoghue and Hombo (1999). The Orlando–Thissen procedure and the Stone procedures had nominal Type I error rates while the Donoghue–Hombo procedure had

unacceptably low Type I error rates. On the other extreme, the Bock procedure had abnormally high Type I error rates and identified almost all items as misfitting. These procedures displayed modest power in detecting misfitting items, with the power increasing with sample size in some instances.

While the procedures developed by Stone et al. (1994), Stone (2000) and Orlando and Thissen (2000) provide improvements over procedures that use estimates of θ in constructing fit statistics, a fundamental problem remains with these procedures. As Sinharay (2005a, 2005b) points out, it has been shown by Chernoff and Lehman (1953) that a chi-square test statistic that depends on the estimates of parameters of a model will not have the limiting chi-square distribution. However, more recently Maydeu-Olivares and Joe (2005) have shown that when the parameter vector is estimated using a consistent and asymptotically normal minimum variance estimator, the statistic based on estimated model parameters has a limiting chi-square distribution.

In the computation of the Orlando–Thissen statistic, item parameter estimates are treated as known. While the uncertainty in the θ estimates does not play a part in the Orlando–Thissen procedure, the uncertainty in item parameter estimates does play a role and this affects the limiting distribution of the statistics. A similar criticism can be leveled against Stone (2000) and the Donoghue and Hombo (1999, 2003) procedure. Glas (1998, 1999) and Glas and Suarez Falcon (2003) have criticized these procedures along the same lines for failing to take into account the stochastic nature of the item parameter estimates.

Glas (1998, 1999) developed a Lagrange Multiplier (LM) procedure that takes into account the uncertainty in the item parameter estimates. The LM procedure of Glas has its roots in the classic procedure for finding extreme values of functions under constraints. Using the notation of Glas and Suarez Falcon (2003), let η_1 denote the vector of parameters of the model, and let η_2 denote a vector of parameters added to this model to obtain a more general model; $h(\eta_1)$ and $h(\eta_2)$ are the first-order derivatives of the log-likelihood function. Under the assumption that the parameters η_1 of the model are estimated by maximum likelihood, i.e., $h(\eta_1) = 0$, Glas and Suarez Falcon (2003) showed that the hypothesis $\eta_2 = 0$ can be tested using the statistic

$$LM = h(\eta_2)' \Sigma^{-1} h(\eta_2), \quad (3.17)$$

where Σ is the covariance matrix of $h(\eta_2)$. Details of the computation of $h(\eta_2)$ and Σ for IRT models using the marginal maximum likelihood are given in Glas (1998, 1999) and for the three-parameter model in Glas and Suarez Falcon (2003). Glas (1998, 1999) showed that the LM-statistic has an asymptotic χ^2 -distribution with degrees of freedom equal to the number of parameters in η_2 .

Glas (1998, 1999) and Glas and Suarez Falcon (2003) applied the LM procedure to solve a variety of problems in item response theory including assessing model fit and local independence, and assessing person fit. The major advantage of the LM procedure is that it is computationally straightforward, provides a unified approach for examining the assumptions underlying item response theory, and takes into account the uncertainty in the item parameter and ability estimates in constructing the test statistic. Simulation studies by the authors have demonstrated that the LM procedure compares favorably with the Orlando–Thissen procedure, producing slightly lower Type I error rates and

improvement in power. However, they found the false positive rates alarmingly high in some cases when compared with that of the Orlando–Thissen procedure. They concluded that while the LM procedure was clearly superior to the Yen Q_1 procedure, the Orlando–Thissen procedure had better overall characteristics than the LM procedure.

A further problem with the Glas and Suarez Falcon approach is the specification of the parameters for the general model. In examining the fit of the three-parameter item response model, they defined a general model in which the difficulty parameters differed among subgroups. While such a parameterization is useful in examining differential item functioning or testing the invariance of item parameters in subgroups, this is only one form of misfit. Another issue is the grouping of examinees into score groups. Different groupings produced different results with respect to both hit rates and false positive rates. Despite these problems, the LM method has a sound theoretical basis and offers useful procedures for assessing model fit in the IRT context.

Maydeu-Olivares and Joe (2005) pointed out that assessing goodness of fit in binary latent trait models can be conceptualized in terms of assessing fit in a 2^n contingency table where n is the number of items. Needless to say, as n increases, the number of cells in the contingency table increases exponentially. Since the number of respondents/examinees is usually small in comparison to the number of cells in the table, as n increases the table becomes sparse. The sparseness of the table introduces problems in the estimation of parameters and the assessment of fit. Maydeu-Olivares and Joe (2005) provided a unified framework for assessing fit in a 2^n table. They provided full information procedures that utilize the higher-order joint moments and limited information procedures that use only the first two joint moments for assessing fit in high-dimensional contingency tables. They proposed two statistics, L_r , and M_r where L_r is the statistic based on known model parameters using the first r joint moments (proportions, in this case, and similar to that proposed by Donoghue and Hombo, 2003) while M_r denotes the statistic based on estimated model parameters, again based on the first r joint moments (proportions). Maydeu-Olivares and Joe (2005) show that these statistics, which can be expressed as quadratic forms involving the discrepancies between the observed and expected joint moments weighted by the asymptotic variance–covariance matrix of the sample joint proportions, i.e., Fisher’s information matrix, converge in distribution to a chi-square distribution.

The procedure developed by Maydeu-Olivares and Joe (2005) can be viewed as a major advance in the area of testing goodness of fit in 2^n contingency tables and has implications for assessing fit not only in unidimensional item response models but also in multidimensional item response models. The only limitation in the current presentation is that the procedure is applicable only to binary response data.

3.3. Bayesian procedures

In Bayesian analysis, the posterior distribution of a set of parameters contains all the information about the parameters. If ω denotes the vector of parameters, then the posterior distribution of ω given the observations or data is

$$\pi(\omega|y) = \frac{L(y|\omega)\pi(\omega)}{\pi(y)}, \quad (3.18)$$

where y is the data set (the set of item responses of N examinees to n items). In the IRT context, $\omega = [\theta \ \xi]$, where θ is the vector of trait parameters and ξ is the vector of item parameters. The joint posterior distribution of item and trait parameters contains all the information about these two sets of parameters (Swaminathan and Gifford, 1985, 1986) and takes into account the uncertainty in both item and trait parameters.

One of the objectives in Bayesian analysis is to determine if the model can explain the data adequately (Gelman et al., 1996). Using the approach developed by Rubin (1984), Gelman et al. (1996) provided a detailed framework for conducting *Bayesian Posterior Predictive Model Checking (PPMC)*. Sinharay (2005a, 2005b) and Sinharay and Johnson (2004) applied this procedure for examining model fit in IRT. The model fit approach using *PPMC* as described by Sinharay (2005a, 2005b) requires the determination of the posterior distribution of a replicate data set, y^{rep} , that may be observed if the “mechanism” that generated the observed data y is replicated with the value of ω . Since the values of ω are not known, the posterior distribution of the replicate, y^{rep} can be obtained as

$$\pi(y^{\text{rep}}|y) = \int \pi(y^{\text{rep}}, \omega|y) d\omega = \int \pi(y^{\text{rep}}|\omega)\pi(\omega|y) d\omega, \quad (3.19)$$

i.e., by averaging the likelihood over the values of ω defined by its posterior distribution as given in Eq. (3.18). Unfortunately, the integral in Eq. (3.19) is multidimensional and cannot be evaluated analytically. Following the suggestion of Rubin (1984), K simulations from the posterior distribution of ω , given in Eq. (3.18), are drawn and combined with a draw from the predictive distribution $\pi(y|\omega)$ evaluated at each draw of ω to yield $y^{\text{rep},k}$, $k = 1, 2, \dots, K$. Sinharay (2005a) points out that the Markov Chain Monte Carlo (MCMC) algorithm generates draws from the posterior distribution of ω and hence the first step in *PPMC* is accomplished through the *MCMC* algorithm.

Once the replicate data, y^{rep} , is obtained, the fit of the model to the observed data can be obtained by defining a discrepancy measure $D(y^{\text{rep},k})$ between observed and predicted values. Sinharay (2005b) has suggested such discrepancy measures as (i) the tail area probability, (ii) percent correct for items, (iii) observed score distributions, (iv) item-total correlations, (v) unconditional odds ratio, defined for pairs of items as

$$OR_{ij} = \frac{N_{11}N_{00}}{N_{10}N_{01}},$$

where N_{11} is the frequency of examinees responding correctly to both items, with N_{00} , N_{01} , N_{10} defined similarly, and (vi) the conditional odds ratio, conditioned on the number correct score, summed over score categories. Sinharay (2005b) points out that the last discrepancy measure may also serve to detect departures from unidimensionality and weak local independence. The evaluation of these discrepancy measures is carried out primarily graphically.

Through simulation studies as well by applying these procedures to test data, Sinharay (2005a, 2005b) and Sinharay and Johnson (2004) demonstrated that the Posterior Predictive Model Checking procedures are practical, easy to implement although computationally intensive, and provide valuable information regarding model fit when evaluated using a variety of criteria. These procedures readily extend to polytomous models. The *PPMC* is clearly one of the most promising approaches to assessing model fit.

3.4. Assessing model parameter invariance

Invariance of item and ability parameters across subpopulations of examinees is arguably the cornerstone of item response theory. If the item response model fits the data, then the invariance property will be realized. Thus checking the invariance property of item and ability parameters provides a direct means for verifying or validating the theory. Invariance holds if

$$P(u|\theta, G) = P(u|\theta), \quad (3.20)$$

where P is the probability of a response to an item, u is the observed response to the item, θ is the ability level, and G is group membership or a background variable. Thus, the probability of a response should not depend on any other variable but the ability level of the examinee.

Hambleton and Swaminathan (1985) and Hambleton et al. (1991) recommended that invariance of item parameters be examined graphically. They provided numerous examples of plots of item difficulty estimates across subpopulations of examinees that can be used to assess invariance and hence the fit of the model. If invariance holds, the plot of item parameters, such as difficulty and discrimination, will lie on a straight line. In practice, however, estimates and not parameters are available. Because of sampling fluctuations, these estimates will be scattered about the straight line. Checking for outliers will provide an indication of which items do not satisfy the invariance property.

Statistical procedures for assessing invariance of item parameters across subgroups of examinees fall under the class of procedures for examining differential item functioning (DIF). Available statistical procedures for assessing DIF range from model-based procedures to nonparametric procedures (Camilli and Shepard, 1994; Hambleton et al., 1991; Holland and Wainer, 1993). The most directly applicable procedure is to test the equality of the parameters of the item response model across subgroups (Lord, 1980). Holland and Thayer (1988) demonstrated that under the assumption that the Rasch model fits the data, the Mantel–Haenszel procedure provides a test of the equivalence of the difficulty parameters across subgroups. Thus the nonparametric Mantel–Haenszel procedure provides a direct assessment of the invariance of model parameters in the Rasch model and hence, the fit of the Rasch model to the data.

In general, the invariance of item parameters may be tested directly by including in the item response model parameters that indicate group differences. An example of such an approach was provided by Glas (1998, 2001), and Glas and Suarez Falcon (2003) in the context of assessing the fit of the three-parameter model using the Lagrange Multiplier method. The null hypothesis that difficulty parameters are equal across score groups was tested against the alternative that the difficulty parameter differed across the groups by an additive parameter. Similar tests can be developed for other parameters of interest either using such procedures as the Lagrange Multiplier method or the *MCMC* procedure. These statistical tests combined with graphical displays provide powerful methods of examining invariance of parameters across subgroups.

3.5. Role of simulation studies in assessing fit

The notion of using simulated data for assessing model fit was introduced earlier in the context of assessing dimensionality of a set of items. In fact, simulating data and deter-

mining the empirical distribution of the test statistic play a critical role in the procedure developed by Stone (2000). In the Bayesian *PPMC* procedure, simulations are carried out to obtain the posterior distribution of the replicate data sets. Hambleton and Rogers (1990) and Hambleton et al. (1991) use simulation to generate an expected distribution of standardized residuals assuming model fit for the purposes of interpreting the observed distribution of standardized residuals.

In essence, simulation approaches can be applied with all the procedures described in this chapter. These simulation studies can provide empirical distributions of the test statistics for which theoretical distributions are not available, or when there is sufficient evidence to throw doubt on the validity of the assumed distribution of the test statistic. For example, in comparing the observed score distribution and the expected score distribution, simulations can be carried out to provide a confidence band for the expected score distribution. These simulated distributions are not unlike those obtained using bootstrap procedures. The combination of simulations and graphical displays provides a powerful tool for assessing model fit (see, for example, Hambleton and Han, 2005).

4. Empirical example

Tables 1 to 5 and Figure 11 contain model fit results of the analyses of data from a large scale assessment. A subset of the data containing the responses of 2043 examinees on 32 multiple-choice questions, four short answer binary-scored questions, and six polytomously-scored items (scored 0 to 4) was analyzed for illustrative purposes. One-, two- and three-parameter models were fitted to the dichotomous items and partial credit, generalized partial credit, and graded response models were fitted to the polytomous items. Figure 10 contains the plot of eigenvalues for the correlation matrix of 42 test items. Table 1 provides the classical item indices and IRT item parameter estimates for each model fitted to the dichotomous items, while Table 2 gives item parameter estimates for each model fitted to the polytomous items. Tables 3 and 4 give chi-square fit statistics for the dichotomous and polytomous items, respectively. Fitting multiple IRT models to the same data set and comparing results is an especially useful strategy for becoming familiar with the data, and comparing the relative advantages of one model over another prior to choosing one to address a measurement problem.

Figure 10 highlights clearly that there is one dominant factor underlying the 42 test items. The plot shows that the first factor (corresponding to the dominant eigenvalue of 11.2) accounts for 26.7% of the score variability, almost nine times that accounted for by the second factor; the second, third, fourth, and subsequent factors account for less than 3% each (corresponding to eigenvalues of 1.3, 1.2, and 1.1, respectively. Taken together, these findings support the unidimensionality assumption.

As pointed out earlier, eigenvalues obtained through simulations can provide a useful baseline for assessing the unidimensionality assumption. To this end, random item response data were simulated for the 2043 examinees on the 42 item test, conditioned on the *p*-values of the items in the real data. The eigenvalues of the correlation matrix were then obtained from the simulated data set. The largest three eigenvalues obtained from this analysis were 1.27, 1.26, and 1.24. For the actual data, on the other hand, the

Table 1

Classical difficulty values (p), point biserial correlations (r), and IRT item parameter estimates for 1P, 2P, and 3P models for dichotomous items

Item	Classical indices		1P model ¹	2P model		3P model ²		
	p	r	b	b	a	b	a	c
1	0.91	0.36	-2.24	-1.99	0.87	-2.91	0.76	0.00
2	0.70	0.48	-0.89	-0.86	0.76	-0.52	0.87	0.18
3	0.76	0.40	-1.20	-1.31	0.62	-0.39	0.96	0.38
4	0.84	0.39	-1.68	-1.67	0.71	-1.51	0.70	0.20
5	0.58	0.39	-0.35	-0.41	0.54	0.51	1.35	0.36
6	0.60	0.34	-0.43	-0.60	0.43	0.48	0.90	0.35
7	0.53	0.32	-0.15	-0.21	0.40	0.57	0.64	0.25
8	0.73	0.30	-1.06	-1.51	0.43	0.07	0.77	0.48
9	0.87	0.28	-1.90	-2.38	0.51	-2.27	0.47	0.21
10	0.56	0.55	-0.27	-0.25	0.89	-0.01	1.13	0.13
11	0.67	0.22	-0.78	-1.56	0.28	-0.17	0.37	0.31
12	0.57	0.54	-0.31	-0.28	0.87	-0.06	1.02	0.12
13	0.60	0.54	-0.44	-0.41	0.87	0.06	1.38	0.24
14	0.89	0.43	-2.06	-1.66	1.04	-1.44	1.09	0.24
15	0.88	0.39	-2.03	-1.82	0.86	-1.71	0.81	0.20
16	0.76	0.31	-1.22	-1.66	0.46	-0.01	0.84	0.51
17	0.83	0.49	-1.62	-1.32	1.05	-1.14	1.10	0.16
18	0.73	0.50	-1.03	-0.94	0.84	-0.48	1.10	0.25
19	0.52	0.57	-0.10	-0.10	0.95	0.24	1.79	0.18
20	0.56	0.55	-0.27	-0.25	0.90	-0.08	1.00	0.09
21	0.79	0.37	-1.40	-1.59	0.59	-1.04	0.66	0.27
22	0.71	0.49	-0.93	-0.86	0.80	-0.70	0.84	0.10
23	0.70	0.44	-0.90	-0.95	0.65	-0.66	0.71	0.15
24	0.79	0.42	-1.36	-1.34	0.73	-1.07	0.77	0.17
25	0.84	0.46	-1.69	-1.45	0.93	-1.28	0.93	0.18
26	0.57	0.53	-0.31	-0.29	0.84	0.30	2.26	0.30
27	0.67	0.54	-0.74	-0.65	0.91	-0.38	1.08	0.15
28	0.76	0.44	-1.19	-1.17	0.73	-0.86	0.79	0.17
29	0.70	0.41	-0.89	-0.98	0.61	-0.21	0.87	0.32
30	0.70	0.45	-0.91	-0.93	0.67	0.11	2.28	0.44
31	0.56	0.39	-0.25	-0.30	0.51	0.38	0.87	0.26
32	0.74	0.56	-1.09	-0.90	1.05	-0.47	1.43	0.25
33	0.67	0.48	-0.72	-0.71	0.75	-0.72	0.73	0.00
34	0.67	0.65	-0.70	-0.55	1.31	-0.54	1.29	0.00
35	0.68	0.60	-0.81	-0.65	1.15	-0.64	1.13	0.00
36	0.58	0.61	-0.33	-0.28	1.10	-0.25	1.12	0.00

¹Common a -parameter = 0.714.

² c -parameters for items 33–36 were fixed at zero.

largest three eigenvalues were 11.20, 1.31, and 1.20. Clearly, the evidence to support the assumption of unidimensionality in the actual data is very strong. The second, third, and additional factors in the actual data could not be distinguished from the factors found from the analysis of random error.

Table 2

Partial credit (PC) model, generalized partial credit (GPC) model, and graded response (GR) model item parameters for polytomous items

Item	<i>b</i>	<i>a</i>	<i>d</i> 1	<i>d</i> 2	<i>d</i> 3	<i>d</i> 4
PC model						
37	−0.14	0.71	0.95	0.13	−1.02	−0.07
38	−0.28	0.71	0.19	−0.42	0.50	−0.27
39	−0.09	0.71	0.02	0.98	−1.16	0.15
40	−0.07	0.71	0.96	−0.78	−0.17	−0.01
41	0.02	0.71	0.00	0.36	−0.49	0.14
42	−0.44	0.71	0.80	0.24	−0.12	−0.92
GPC model						
37	−0.15	0.72	0.99	0.10	−1.03	−0.06
38	−0.26	0.82	0.30	−0.38	0.39	−0.31
39	−0.08	0.75	0.08	0.92	−1.14	0.14
40	−0.06	0.73	0.99	−0.79	−0.19	−0.01
41	0.03	0.61	−0.13	0.35	−0.55	0.33
42	−0.44	0.77	0.88	0.21	−0.17	−0.92
GR model						
37	−0.11	1.29	1.20	0.23	−0.53	−0.90
38	−0.28	1.52	0.69	0.14	−0.11	−0.72
39	−0.08	1.43	0.84	0.41	−0.48	−0.77
40	−0.03	1.34	1.07	−0.01	−0.33	−0.73
41	0.04	1.30	0.71	0.23	−0.28	−0.65
42	−0.46	1.28	1.15	0.37	−0.31	−1.20

To further check on dimensionality and to investigate the possibility that the polytomous items might be measuring a different dimension from the dichotomous items, an analysis using the DIMTEST program was performed using the polytomous items as the assessment subtest. A *t*-value of −0.118, with a significance level of .547 was obtained, indicating that the polytomous items were measuring the same construct as the dichotomous items.

A review of the classical item statistics for the binary data in Table 1 highlights two things: Items have a wide range of classical item discrimination indices (.27 to .82) and item difficulties (.52 to .91). Even prior to the IRT analysis, the first finding suggests that a discrimination parameter in the IRT model would likely improve model fit. The second suggests that a “guessing” parameter in the IRT model might also improve the model fit, at least for some of the more difficult items. Though we are cautious in placing too much weight on the item fit chi-square significance test results shown in Table 3, it is clear that the chi-square statistic itself can serve as a summary of model-item data departure, and the number of significant chi-square statistics is substantially smaller for the two-parameter model than the one-parameter model and substantially smaller again for the three-parameter model. These findings suggest that the three-parameter model fits the data better than the two-parameter model, and both models fit the data better than the one-parameter model. The PARSCALE chi-square statistic and the Orlando–Thissen

Table 3

PARSCALE (PSL) and Orlando–Thissen (O–T) chi-square fit statistics for 1P, 2P, and 3P IRT models for dichotomous items

Item	1P model		2P model		3P model	
	PSL	O–T	PSL	O–T	PSL	O–T
1	26.40**	33.45	17.00	28.77	98.62**	145.94**
2	19.16	40.76	15.27	41.56	10.57	41.02
3	17.25	52.13	14.34	45.32	10.64	43.79
4	25.66*	50.94	25.39*	44.58	27.91*	52.55*
5	62.09**	107.63**	48.60**	70.36*	17.95	44.54
6	101.71**	172.75**	34.57*	80.47**	16.25	74.09**
7	100.06**	194.01**	31.31	64.65*	14.79	63.70*
8	88.77**	132.75**	31.15*	61.17	22.45	55.99
9	26.22*	53.89*	19.77	25.50	18.17	24.16
10	40.25**	48.45	24.93	28.40	11.74	31.96
11	176.60**	286.19**	23.00	56.95	16.15	52.07
12	23.69	32.80	8.69	23.32	16.23	23.78
13	42.67**	64.51*	32.53*	57.71*	28.90*	52.78
14	37.94**	52.92**	5.82	22.18	8.00	21.48
15	19.20	36.49	16.27	29.38	23.14*	27.92
16	70.98**	112.99**	24.78	59.43	21.39	49.75
17	55.26**	68.90**	15.87	29.39	12.99	32.08
18	17.74	36.77	13.98	33.14	10.74	29.15
19	85.20**	104.41**	60.02**	86.89**	18.08	42.20
20	29.75*	42.49	15.24	27.88	24.65	28.02
21	19.75	59.04*	10.62	54.74	10.27	56.33
22	23.75	54.03	17.76	53.43	20.61	50.31
23	22.41	52.04	15.66	47.28	15.80	50.68
24	19.16	51.98	10.95	51.36	13.27	53.40
25	33.82**	36.87	9.80	23.50	20.59	23.39
26	88.30**	102.39**	84.51**	96.36**	21.63	48.59
27	40.56**	52.81	8.02	40.80	10.77	40.47
28	10.32	48.54	8.51	48.20	12.46	47.66
29	24.56	61.91*	24.08	47.81	14.62	45.28
30	112.29**	150.62**	110.94**	142.48**	25.12*	45.33
31	61.77**	114.71**	36.72**	62.90*	21.20	55.04
32	59.04**	76.23**	13.16	41.35	9.31	26.14
33	21.59	54.44	22.26	50.51	25.08	53.14
34	143.00**	141.62**	12.08	42.39	17.11	39.08
35	87.50**	101.97**	19.17	41.72	20.31	40.15
36	77.09**	97.27**	22.14	52.85*	13.91	50.73

*Number of intervals for PARSCALE fit statistics set at 20 and collapsed to ensure minimum expected cell frequency of 5.

**Number of intervals for Orlando–Thissen fit statistics determined by number of raw score levels and collapsed to ensure minimum expected cell frequency of 5.

statistic show considerable agreement with respect to the items flagged as misfitting, but sufficient discrepancy to highlight the risks of over-interpreting statistically significant fit statistics. For the polytomous items, adding a discrimination parameter had less of

Table 4

PARSCALE (PSL) and Orlando–Thissen (O–T) chi-square fit statistics for the partial credit (PC), generalized partial credit (GPC), and graded response (GR) models for polytomous items

Item	PC model		GPC model		GR model	
	PSL	O–T	PSL	O–T	PSL	O–T
37	67.81	31.97	66.39	20.54	75.09*	31.19
38	70.15*	45.75	50.83	33.84	63.36*	50.30*
39	110.53**	52.55**	84.94**	42.72*	74.71**	50.28*
40	115.37**	39.53*	95.87**	37.41	81.20**	59.91**
41	106.21**	109.75**	89.64**	92.78**	41.10**	72.75*
42	69.53	26.77*	82.00**	16.49	57.59	24.89

*Number of intervals for PARSCALE fit statistics set at 20 and collapsed to ensure minimum expected cell frequency of 5.

**Number of intervals for Orlando–Thissen fit statistics determined by number of raw score levels and collapsed to ensure minimum expected cell frequency of 5.

Table 5

Distributions of standardized residuals for 1P/PC, 2P/GPC, and 3P/GPC IRT models

SR interval	Percent of residuals in interval		
	1P/PC	2P/GPC	3P/GPC
< -3	2.79	0.63	0.72
(-3, -2)	5.59	3.42	2.79
(-2, -1)	14.23	13.87	12.61
(-1, 0)	27.30	33.51	34.59
(0, 1)	24.99	28.56	32.07
(1, 2)	15.23	14.32	13.51
(2, 3)	6.13	4.50	3.15
> 3	3.78	1.17	0.54

an effect on fit, probably because there was not great variation in the discrimination parameters, as Table 2 shows. Looking at the Orlando–Thissen statistics, the generalized partial credit model appeared to fit the item response data a little better overall than did the graded response model. Table 5 shows the distribution of standardized residuals for the 1P, 2P, and 3P models for the dichotomous items in combination with either the partial credit or generalized partial credit model for the polytomous items. These distributions show the clear improvement in overall fit in going from a 1P/PC model to a 2P/GPC model, and further improvement, though not as dramatic, in going from a 2P/GPC model to a 3P/GPC model.

Figure 11 shows plots of item residuals for binary-scored item 19 for the three-parameter, two-parameter, and the one-parameter model, respectively. The gradual improvement in the model fit can be seen as additional parameters are added to the IRT model. Figure 12 shows residual plots for the partial credit and graded response mod-

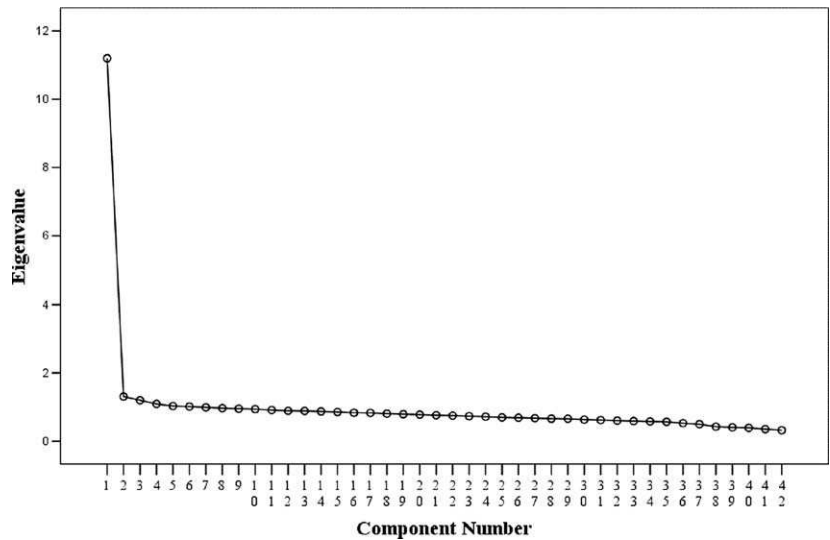


Fig. 10. Plot of the eigenvalues from the item correlation matrix.

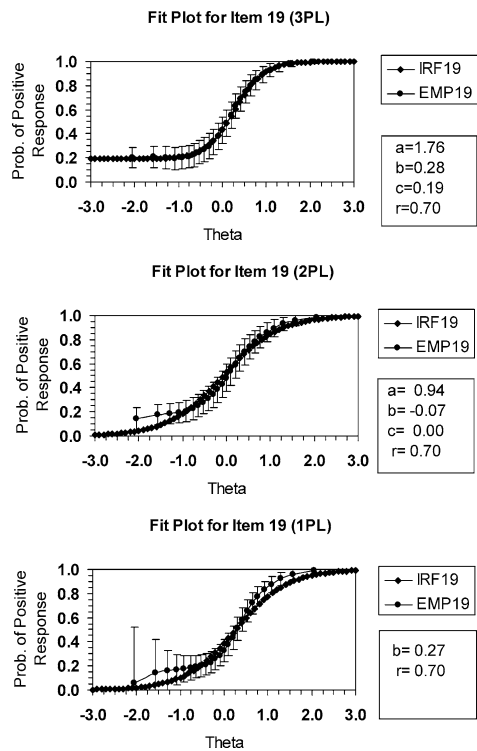


Fig. 11. Residual plots of item 19 (dichotomously-scored item).

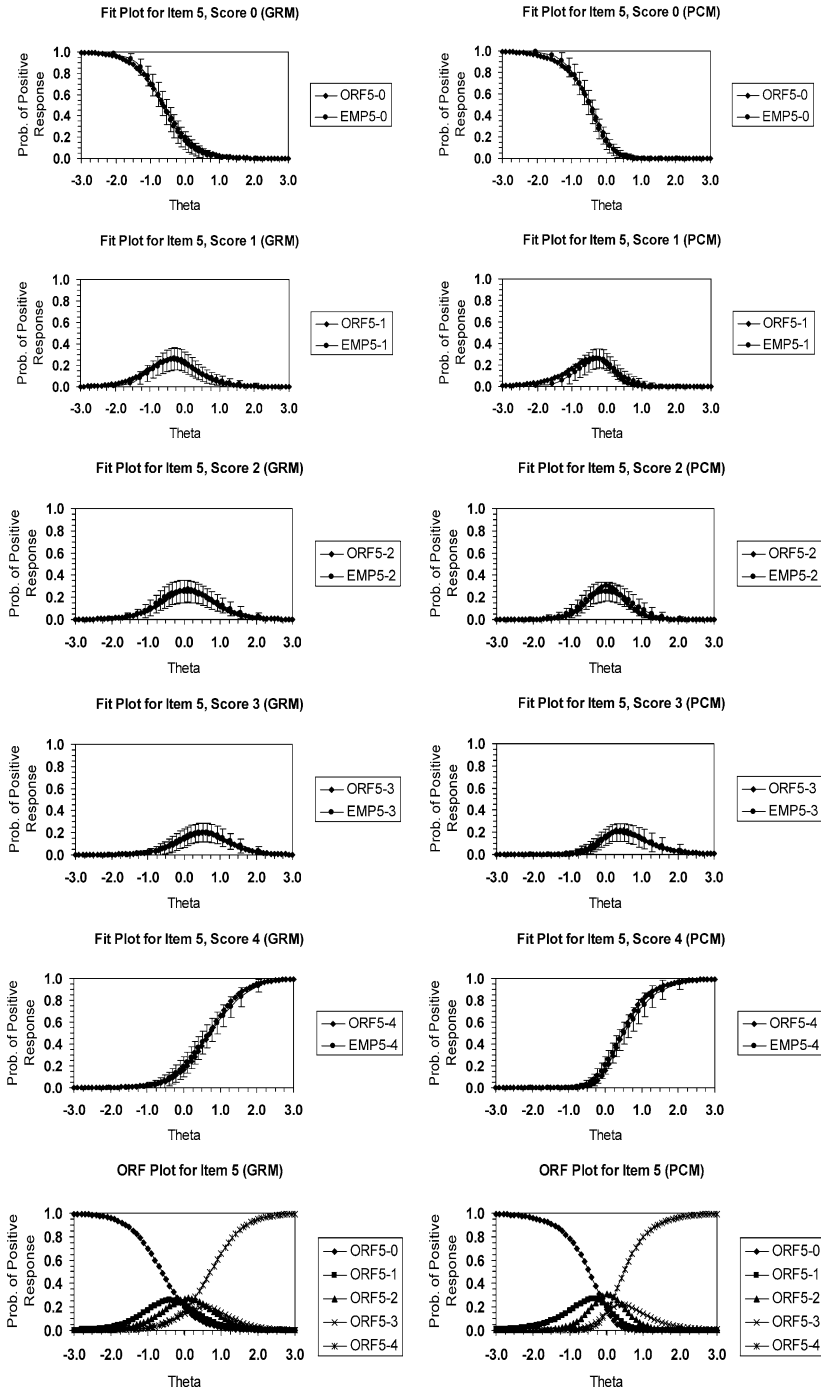


Fig. 12. Residual plots of item 5 (polytomously-scored item).

els for item 41 (item 5 of the polytomously scored items); these plots can be useful in comparing the fit of competing models.

The practical implications of the differences in the model fits would depend on the intended uses of the scores. Figures 1, 2, and 3, and Figures 4, 5, and 6 highlight that the predictions of the test scores are noticeably different from each model, and generally better for models with more item parameters. If the intent was to produce test score norms tables, the results would be very different depending on choice of model, and the most suitable results would come from the more general models. Other empirical investigations could be carried out to investigate model fit and the consequences of model misfit on the intended application such as equating, detecting bias, building tests, and sorting examinees into performance categories.

5. Conclusions

Assessing IRT model fit to item response data is one of the crucial steps before an IRT model can be applied with confidence to estimate proficiency or ability levels of examinees, link tests across administrations, and assess, for example, adequate yearly progress as required by the NCLB legislation. Despite the fundamental importance of this step, assessment of fit receives only the barest of considerations by many test developers and users. Possible reasons for this neglect are the complexity of assessing fit, the lack of understanding of the fit statistics, the absence of comprehensive model fit software, and the fact that the likelihood ratio statistics provided by commercial software often indicate misfit for most of the items with large sample sizes. It is indeed ironic that the highly desirable practice of using large sample sizes to minimize item parameter estimation error results in significant item-fit statistics even when the model fits the data for all practical purposes.

Given the importance of the subject, considerable research has been carried out and advances have been made with respect to fit assessment in item response models, though much of the research is still not being implemented on a wide scale. Some of these model fit procedures are computationally intensive while others are relatively straightforward to implement even without dedicated software, but again, few are being used today.

Hambleton and Han (2005) have suggested a set of activities that practitioners should routinely carry out in the assessment of model fit. These include:

- (i) carrying out routine examination of classical indices such as p -values, item-total correlations, the distribution of scores at both the test and item levels, and factor analysis. These analyses may provide an indication of the appropriate item response model to choose as well as identify potential problems in fitting a model (for example, the need to collapse score categories or the suitability of the unidimensionality assumption);
- (ii) checking the unidimensionality assumption formally and examining possible violations of local independence that may result from groups of items sharing a common stimuli;

- (iii) examining the fit of the model at the test and item levels using graphical as well as statistical procedures;
- (iv) checking invariance of item parameters through plots of item parameter estimates and DIF studies (to detect noninvariance across, say, gender, ethnicity, and language groups).

In a final step, Hambleton and Han (2005) suggest examining the consequences of model misfit on outcomes of interest. Since no model can fit a set of observations perfectly, it is important, if possible through simulation studies at least, to examine the consequence of misfit on such important outcomes as determining proficiency or ability scores, assigning examinees to performance categories, equating test scores, optimizing item selection when implementing computer-adaptive tests, and scale drift, to name a few.

As mentioned earlier, assessing model fit is an important part of the test validation process. It is multifaceted, and as in the verification of any scientific theory, it is an ongoing process where only through the accumulation of empirical evidence can one be confident of the appropriateness of item response theory for the solution of a particular measurement problem. The procedures for assessing model–data fit described in this chapter have the potential for addressing the vexing problem of determining if the measurement procedures used are appropriate for addressing the practical measurement problems faced by practitioners.

Acknowledgements

The authors are indebted to Ms. Yue Zhao for assistance in compiling Section 4 of this chapter.

References

- Ackerman, T. (2005). Multidimensional item response theory. In: Everitt, B., Howell, D. (Eds.), *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, West Sussex, UK.
- Agresti, A. (1990). *Categorical Data Analysis*. Wiley, New York.
- Andersen, E. (1973). A goodness of fit test for the Rasch model. *Psychometrika* **38**, 123–140.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika* **43**, 561–573.
- 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., Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika* **46**, 443–449.
- Bock, R.D., Gibbons, R.D., Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement* **12**, 261–280.
- Camilli, G., Shepard, L.A. (1994). *Methods for Identifying Biased Items*. Sage, Thousand Oaks, CA.
- Chen, W., Thissen, D. (1997). Local independence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics* **22**, 265–289.
- Chernoff, H., Lehman, E.L. (1953). The use of maximum likelihood estimates in chi-square tests for goodness of fit. *Annals of Mathematical Statistics* **25**, 579–586.
- De Champlain, A.F., Gessaroli, M.E. (1998). Assessing the dimensionality of item response matrices with small sample sizes and short test lengths. *Applied Measurement in Education* **11**, 231–253.

- Donoghue, J.R., Hombo, C.M. (1999). Some asymptotic results on the distribution of an IRT measure of item fit. Paper presented at the meeting of the Psychometric Society, Lawrence, KS. June.
- Donoghue, J.R., Hombo, C.M. (2001). The distribution of an item-fit measure for polytomous items. Paper presented at the meeting of the National Council of Measurement in Education, Seattle, WA. April.
- Donoghue, J.R., Hombo, C.M. (2003). A corrected asymptotic distribution of an IRT fit measure that accounts for effects of item parameter estimation. Paper presented at the meeting of the National Council of Measurement in Education, Chicago. April.
- Finch, H., Habing, B. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and allocating items. *Journal of Educational Measurement* **42**, 149–169.
- Fraser, C. (1988). NOHARM II: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. University of New England, Centre for Behavioural Studies, Armidale, NSW.
- Gelman, A., Meng, X., Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–807.
- Gessaroli, M.E., De Champlain, A.F. (1996). Using an approximate chi-square statistic to test the number of dimensions underlying the responses to a set of items. *Journal of Educational Measurement* **33**, 157–179.
- Gessaroli, M.E., De Champlain, A.F., Folske, J.C. (1997). Assessing dimensionality using likelihood-ratio chi-square test based on a non-linear factor analysis of item response data. Paper presented at the meeting of the National Council on Measurement in Education, Chicago. April.
- Glas, C.A.W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica* **8** (1), 647–667.
- Glas, C.A.W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika* **64**, 273–294.
- Glas, C.A.W. (2001). Differential item functioning depending on general covariates. In: Boomsma, A., van Duijn, M.A.J., Snijders, T.A.B. (Eds.), *Essays on Item Response Theory*. Springer, New York, pp. 131–148.
- Glas, C.A.W. (2005). Assessing fit to IRT models. In: Everitt, B., Howell, D.C. (Eds.), *Encyclopedia of Behavioral Statistics*. Wiley, West Sussex, UK.
- Glas, C.A.W., Suarez Falcon, J.C.S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement* **27** (2), 87–106.
- 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: Lenderking, W.R., Revicki, D. (Eds.), *Advances in Health Outcomes Research Methods, Measurement, Statistical Analysis, and Clinical Applications*, Degnon Associates, Washington.
- Hambleton, R.K., Rogers, H.J. (1990). Using item response models in educational assessments. In: Schreiber, W.H., Ingekamp, K. (Eds.), *International Developments in Large-Scale Assessment*. NFER-Nelson, Windsor, UK, pp. 155–184.
- Hambleton, R.K., Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement* **10**, 287–302.
- Hambleton, R.K., Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer Academic Publishers, Boston, MA.
- Hambleton, R.K., Traub, R.E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology* **24**, 273–281.
- Hambleton, R.K., Swaminathan, H., Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Sage, Newbury Park, CA.
- Hattie, J.A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement* **9**, 139–164.
- Hattie, J.A., Krakowski, K., Rogers, H.J., Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement* **20**, 1–14.
- Hemker, B.T., van der Ark, L.A., Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika* **66**, 487–506.
- Holland, P.W., Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In: Wainer, H., Braun, H.L. (Eds.), *Test Validity*. Lawrence Erlbaum Publishers, Hillsdale, NJ, pp. 129–145.

- Holland, P.W., Wainer, H. (Eds.) (1993). *Differential Item Functioning*. Lawrence Erlbaum Publishers, Hillsdale, NJ.
- Joreskog, K.G., Sorbom, D. (2004). LISREL 8 [Computer program]. Scientific Software, Chicago.
- Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Publishers, Hillsdale, NJ.
- Lord, F.M., Novick, M.R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA.
- Lord, F.M., Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement* **8**, 453–461.
- Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika* **47**, 149–174.
- Maydeu-Olivares, A. (2001). Multidimensional item response theory modeling of binary data: Large sample properties of NOHARM estimates. *Journal of Educational and Behavioral Statistics* **26**, 51–71.
- Maydeu-Olivares, A., Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association* **100**, 1009–1020.
- Meijer, R.R., Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement* **25**, 107–135.
- McDonald, R.P. (1967). *Non-Linear Factor Analysis*. Psychometric Monographs, No. 15.
- McDonald, R.P. (1982). Linear versus non-linear models in item response theory. *Applied Psychological Measurement* **6**, 379–396.
- McDonald, R.P. (1997). Normal ogive multidimensional model. In: Van der Linden, W.J., Hambleton, R.K. (Eds.), *Handbook of Modern Item Response Theory*. Springer, New York, pp. 257–269.
- McKinley, R., Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement* **9**, 49–57.
- Mislevy, R.J., Bock, R.D. (1990). BILOG: Item analysis and test scoring with binary logistic models [Computer program]. Scientific Software, Chicago.
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement* **16**, 59–176.
- Orlando, M., Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement* **24** (1), 50–64.
- Orlando, M., Thissen, D. (2003). Further investigation of the performance of $S - X^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement* **27** (4), 289–298.
- Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.
- Reckase, M.D., Ackerman, T.A., Carlson, J.E. (1988). Building unidimensional tests using multidimensional items. *Journal of Educational Measurement* **25**, 193–203.
- Rogers, H., Hattie, J. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement* **11**, 47–57.
- Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172.
- Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores*. Psychometric Monograph, No. 17.
- Sijtsma, K., Molenaar, I.W. (2002). *Introduction to Nonparametric Item Response Theory*. Sage, Thousand Oaks, CA.
- Sinharay, S. (2005a). Practical applications of posterior predictive model checking for assessing fit of unidimensional item response theory models. *Journal of Educational Measurement* **42**, 375–385.
- Sinharay, S. (2005b). Bayesian item fit analysis for unidimensional item response theory models. Paper presented at the meeting of the National Council on Measurement in Education, Montreal. April.
- Sinharay, S., Johnson, M.S. (2004). Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models. Paper presented at the meeting of the National Council on Measurement in Education, San Diego. April.
- Stone, C.A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement* **37** (1), 58–75.
- Stone, C.A. (2004). IRTFIT_RESAMPLE: A computer program for assessing goodness of fit of item response theory models based on posterior expectations. *Applied Psychological Measurement* **28** (2), 143–144.

- 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** (4), 331–352.
- Stone, C.A., Mislevy, R.J., Mazzeo, J. (1994). Classification error and goodness-of-fit in IRT models. Paper presented at the meeting of the American Educational Research Association, New Orleans. April.
- Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika* **52**, 589–617.
- Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika* **55**, 293–325.
- Stout, W.F., Nandakumar, R., Junker, B., Chang, H.H., Steidinger, D. (1991). DIMTEST and TESTSIM [Computer program]. University of Illinois, Department of Statistics, Champaign.
- Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement* **27**, 159–203.
- Tutz, G. (1997). Sequential models for ordered responses. In: van der Linden, W.J., Hambleton, R.K. (Eds.), *Handbook of Modern Item Response Theory*. Springer, New York, pp. 139–152.
- Swaminathan, H., Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika* **50**, 349–364.
- Swaminathan, H., Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika* **51**, 589–601.
- Van den Wollenberg, A.L. (1982). Two new test statistics for the Rasch model. *Psychometrika* **47**, 123–140.
- Verhelst, N.D., Glas, C.A.W., de Vries, H.H. (1997). A steps model to analyze partial credit. In: van der Linden, W.J., Hambleton, R.K. (Eds.), *Handbook of Modern Item Response Theory*. Springer, New York, pp. 123–138.
- Wright, B.D., Panchapakesan, N. (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement* **29**, 23–48.
- Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement* **5**, 245–262.
- Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement* **8** (2), 125–145.
- Zimowski, M.F., Muraki, E., Mislevy, R.J., Bock, R.D. (1996). Bilog-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer program]. Scientific Software International, Chicago.