Likelihood-Based Item-Fit Indices for Dichotomous Item Response Theory Models

Maria Orlando, RAND Corp.

David Thissen, University of North Carolina at Chapel Hill

New goodness-of-fit indices are introduced for dichotomous item response theory (IRT) models. These indices are based on the likelihoods of number-correct scores derived from the IRT model, and they provide a direct comparison of the modeled and observed frequencies for correct and incorrect responses for each number-correct score. The behavior of Pearson's χ^2 (S-X^2) and the likelihood ratio G^2 (S-G^2) was assessed in a simulation study and compared with two fit indices similar to those currently in use (Q_1-X^2 and Q_1-G^2). The simulations included three conditions in which the simulating and fitting models were identical and three conditions involving model misspecification. S-X^2 performed well, with Type I error rates close to the expected .05 and .01 levels. Performance of this index improved with increased test length. S-G^2 tended to reject the null hypothesis too often, as did Q_1-X^2 and Q_1-G^2. The power of S-X^2 appeared to be similar for all test lengths, but varied depending on the type of model misspecification. *Index terms: chi-square distribution, dichotomous items, goodness-of-fit, item fit, item response theory (item fit), likelihood ratio statistic, Pearson statistic.*

Item response theory (IRT) is a collection of modeling techniques for the analysis of items, tests, and persons. An IRT model for dichotomous item responses generally specifies that the probability of response pattern **x** is

$$P(\mathbf{x}) = \int \prod_i \left\{ T_i(\theta)^{x_i} \left[1 - T_i(\theta)\right]^{(1-x_i)} \right\} \phi(\theta) \partial\theta \ , \tag{1}$$

where

x is the response vector,

$T_i(\theta)$ is the probability of a correct response on item i as a function of the trait θ, and

$\phi(\theta)$ is the population distribution for θ.

The three-parameter logistic model (3PLM),

$$T_i(\theta) = c_i + \frac{1 - c_i}{1 + \exp\left[-a_i(\theta - b_i)\right]} \ , \tag{2}$$

where a_i is the slope parameter, b_i is the location parameter, and c_i is the lower asymptote parameter, is commonly used for multiple-choice items (see Lord, 1980). In some situations, the two-parameter logistic model (2PLM) is used. This model is equivalent to the 3PLM with $c_i = 0$. The one-parameter logistic model (1PLM) additionally restricts $a_i = a$ for all items. IRT applications are based on estimating model parameters (a_i, b_i, and c_i). These estimates are usually obtained by maximizing the likelihood

$$L \cong \prod_{\substack{\text{response} \\ \text{patterns}}} P(\mathbf{x})^{r_x} \ , \tag{3}$$

Applied Psychological Measurement, Vol. 24 No. 1, March 2000, 50–64

where r_x is the frequency of response pattern **x** (see Bock & Aitkin, 1981; Bock & Lieberman, 1970).

When the assumptions of the IRT model hold, it offers many advantages over classical test theory. The model is implicitly assumed to be correct in all applications of IRT, and its usefulness is dependent on the extent to which it accurately reflects the data. Thus, a diagnostic tool should be used to evaluate the degree of model-data fit. This is most commonly achieved by comparing $-2 \times$ log-likelihood for nested models (Thissen, 1991), which is distributed as χ^2 with the appropriate degrees of freedom (DF).

Assessing the Goodness of Fit for Items

Individual items can also be assessed for fit. However, in IRT, this is not as straightforward as assessing overall model fit. The problem is that θ is a latent variable, so model predictions cannot usually be directly compared with observed data. However, Rasch (1960/80) models (1PLMs) are an exception. These models estimate only b_i, assuming that a_i is identical for all items and that c_i is 0. Under these assumptions, the number-correct (NC) score is a sufficient statistic for θ, and the model's predictions can be directly compared with observed data for each score group. Goodness-of-fit statistics for this family of models have been proposed by Andersen (1973), Glas (1988), Rost & von Davier (1994), Wright & Mead (1977), and Wright & Panchapakesan (1969). The behavior of some of these statistics has been studied by Rogers & Hattie (1987), among others.

Goodness-of-fit statistics for the 2PLM and 3PLM have also been constructed. To compute such statistics for these models, the observed data must be arranged so that they can be compared with the predictions of the model. The general procedure for the construction of such measures is to (1) estimate θ and item parameters from a dataset, (2) sort examinees by their θ estimates, (3) form subgroups of the sorted examinees, (4) calculate the proportion of examinees in each subgroup who answered correctly/incorrectly for each item, and (5) compare these "observed" proportions with those predicted by the model using a χ^2-like statistic and/or a graphical representation (Ankenmann, 1994).

One such measure, Yen's (1981) Q_1, has the form

$$Q_{1i} = \sum_{k=1}^{10} \frac{N_k(O_{ik} - E_{ik})^2}{E_{ik}} + \sum_{k=1}^{10} \frac{N_k[(1 - O_{ik}) - (1 - E_{ik})]^2}{1 - E_{ik}} = \sum_{k=1}^{10} \frac{N_k(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})} \,. \tag{4}$$

For item i, the observed proportions (O_{ik}) are obtained by first partitioning the θ scale into 10 intervals such that the number of examinees in each interval is approximately equal. Then, the observed responses for each interval are tallied in a 10×2 contingency table. The expected proportions (E_{ik}) are computed from the model as the mean predicted probability of a correct response in each interval. Yen (1981) showed that under some circumstances, Q_1 is distributed as approximately χ^2 with $10 - m$ DFs, where m is the number of item parameters. Another item-fit measure, Bock's (1972) χ^2, differs from Yen's Q_1 in that the number of intervals varies and the expected proportions are computed using the median (rather than the mean) of the θ estimates within each interval.

McKinley & Mills (1985) constructed a likelihood ratio G^2 statistic based on computations similar to those used for Q_1. Examinees were rank-ordered and partitioned into 10 equal-size intervals according to their θ estimates. The correct and incorrect responses for each interval were tallied, and G^2 was computed:

$$G_i^2 = 2\sum_{k=1}^{10} N_k \left[O_{ik} \ln\left(\frac{O_{ik}}{E_{ik}}\right) + (1 - O_{ik}) \ln\left(\frac{1 - O_{ik}}{1 - E_{ik}}\right)\right], \quad (5)$$

with $10 - m$ DFs.

Several graphical representations of item fit have also been proposed. These are used either in conjunction with a fit statistic or as an exploratory diagnostic for item fit. Hambleton & Swaminathan (1985) suggested a graphical comparison of observed average item performance with the performance predicted by the model. They proposed that this comparison be plotted in two ways—either the observed performance would be plotted with the item response function (IRF), or the residuals would be plotted. Hambleton and Swaminathan demonstrated the usefulness of these methods by comparing the fit of two different models to a particular item. Wainer & Mislevy (1990) displayed similar plots, and Kingston & Dorans (1985) proposed the analysis of item-θ regressions as a diagnostic tool for item fit.

Limitations of Current Item-Fit Measures, and an Alternative

Although current item-fit statistics are useful for detecting various types of model misfit, acceptable Type I error rates for these statistics have not been clearly established. The problem is that these item-fit statistics are not constructed like traditional χ^2 goodness-of-fit statistics. With the exception of some statistics for the Rasch model, current fit measures rely to some degree on θ estimates to obtain the observed proportions, which should be available before the model is fitted. Model-dependent observed proportions make it difficult to ascertain the distribution of the fit statistic. This uncertainty has been attributed to unclear DFs for χ^2 approximations. Yen (1981) noted that the loss of DFs due to θ estimation is "negligible" for any one item.

A second problem with current indices is in the method by which examinees are grouped into intervals. Grouping examinees into equal-size groups is highly sample dependent. The cutoff points, as well as the number of intervals, affect the resulting fit statistic.

A better index could be obtained if examinees were grouped according to some aspect of the observed data instead of estimated θ. The observed proportions correct and incorrect, cross-classified by the NC score on the test, could serve as a table of observed responses, as in the Rasch model (although some collapsing of cells is often necessary to avoid low or zero frequencies). The most accessible expected proportions from the 2PLM and 3PLM are associated with the response patterns, not the NC scores, so direct comparison of these expected proportions to the observed data is not feasible. However, a table of expected responses can be derived from the 2PLM and 3PLM predictions for each item and θ interval. This expected table can then be directly compared to the observed data.

Due to dependencies among tables of observed counts for each item on the same test, the number of DFs associated with a statistic computed from the cross-classification of score groups and item responses remains unclear. However, this method has an advantage over current statistics in that the observed frequencies are solely a function of the data, and examinees do not need to be grouped in an arbitrary and model-dependent manner. This method can also be readily extended to polytomous models. The challenge is to obtain the expected frequencies of correct and incorrect responses for each NC score for each item. Recently, a new method has been developed to compute these frequencies. This method is described below.

The present study had three objectives: (1) to introduce two new item-fit statistics based on this new computational method, (2) to examine the performance of the new indices with fitting and

misfitting data, and (3) to compare the behavior of the new indices with that of Pearson χ^2 and likelihood ratio G^2 indices similar to those currently in use.

Method

Calculating Expected Proportions

Lord & Wingersky (1984) briefly described a method of predicting joint likelihood distributions for each NC score; this method was further developed by Thissen, Pommerich, Billeaud, & Williams (1995). The method uses a recursive algorithm that builds the joint likelihood for each score group, one item at a time.

When constructing an item-fit statistic, the goal is to determine a likelihood for each possible NC score without each item (where the number of likelihoods is the product of the number of items times the number of possible scores). Then the item is added back to obtain the proportion of examinees with NC score k who answered item i correctly. First, the likelihood distributions for each NC score are obtained, then these results are used to determine a likelihood distribution for each NC score without each item.

For dichotomous models, the NC score likelihood distributions are obtained as follows. The likelihoods for NC scores 0 and 1 are set equal to the IRFs for correct and incorrect responses to the first item:

$$S_0^* = 1 - T_1 \tag{6}$$

and

$$S_1^* = T_1 \ , \tag{7}$$

where S_k^* is the interim value for the likelihood for NC score k and T_1 is the IRF for the correct response to Item 1. (Note that this notation suppresses the fact that $S_k(\theta)$ and $T_i(\theta)$ are functions of θ; S_k and T_i refer to functions of θ throughout, not scalar values.) Then, add each item i to the test, computing

$$S_0 = (1 - T_i)S_0^* \tag{8}$$

and

$$S_k = T_i S_{k-1}^* + (1 - T_i)S_k^* \ , \tag{9}$$

for $k = 1, 2, \ \ldots \ , i - 1$, and

$$S_i = T_i S_{i-1}^* \ . \tag{10}$$

After each item is added, the new S_k replaces S_k^* for all scores computed for the previous item.

At the last iteration of the algorithm, the joint likelihoods have been accumulated for each NC score of all the items except the last. The joint likelihood for NC score k for all of the items is then

$$S_k = T_{last} S_{k-1}^* + (1 - T_{last})S_k^* \ , \tag{11}$$

where

S_k is the NC score posterior distribution for score group k,
T_{last} is the IRF for the last item,
S_{k-1}^* is the NC score posterior distribution for score group $k - 1$ without the last item, and
S_k^* is the NC score posterior distribution for score group k without the last item.

A variation of this recursive algorithm provides the model-predicted proportions correct and incorrect for each item for each NC score. After computing Equation 11, the recursive algorithm is repeated for all i. A different item is omitted for each iteration, yielding the joint likelihoods for each score group without item i (S_k^{*i}). Once the NC score likelihoods with and without each item have been obtained, they can be combined with each (omitted) item to arrive at the desired proportion of examinees with score k who responded correctly to i:

$$E_{ik} = \frac{\int T_i S_{k-1}^{*i} \phi(\theta) \partial\theta}{\int S_k \phi(\theta) \partial\theta} \ . \tag{12}$$

The integrals in Equation 12 are approximated using rectangular quadrature over equally spaced increments of θ from -4.5 to 4.5 (see Stroud, 1974, for alternative methods for the numerical evaluation of such integrals).

The Proposed Indices

Previous studies examining the behavior of item-fit statistics have not resulted in clear conclusions about the relative usefulness of X^2 versus G^2. Therefore, both forms of the new fit index are considered here. The proposed X^2 index has the form

$$S\text{-}X_i^2 = \sum_{k=1}^{n-1} N_k \frac{(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})} \ , \tag{13}$$

where the observed proportions (O_{ik}) for item i and NC score group k are computed from the data, and the expected proportions (E_{ik}) are computed using Equations 11 and 12. The summation is over the number of possible NC scores minus 2. (The proportion of examinees who respond correctly and have NC score = 0 is always 0, and it is always 1 for those with NC score equal to the number of items.) The proposed G^2 index has the form

$$S\text{-}G_i^2 = 2 \sum_{k=1}^{n-1} N_k \left[O_{ik} \ln\left(\frac{O_{ik}}{E_{ik}}\right) + (1 - O_{ik}) \ln\left(\frac{1 - O_{ik}}{1 - E_{ik}}\right) \right] \ , \tag{14}$$

where the observed and expected proportions are defined as in Equation 13.

Simulation

Item response data were generated and calibrated for 100 tests under each of six conditions and three test lengths. In three conditions, the calibrating model (CM) had fewer parameters than the generating model (GM). These conditions provided a preliminary examination of power. In three other conditions, the GM and CM were identical. These conditions permitted the Type I error rates of the new indices to be determined. Data were generated for tests of 10, 40, and 80 items under all six conditions. This showed the effect of test length on the performance of the proposed indices. For each of the 18 conditions, there were 100 replications. For each replication, the new indices and those being used for comparison were calculated. The design of the simulation is shown in Figure 1.

Generating the Item Responses

The computer program GEN (Orlando, 1997) was used to generate item responses. This program randomly selects item parameters as follows. First, a_i is randomly selected from a log-normal $(0, .5)$ distribution. Second, b_i is randomly selected from a normal $(0, 1.0)$ distribution. Finally, c_i

Figure 1
Simulation Study Design

Calibrating Model	*Generating Model*		
	1PLM	*2PLM*	*3PLM*
1PLM	Type I	power	power
2PLM	---	Type I	power
3PLM	---	---	Type I

is randomly selected from a logit-normal $(-1.1, .5)$ distribution. These distributions are similar to the distributions of item parameters commonly observed in educational test construction. For all conditions in this simulation, item responses were generated for a sample size of 1,000.

Calibrating the Data

Item responses were calibrated using the computer program MULTILOG (Thissen, 1991). Chen (1995) made three suggestions that were followed here to increase the precision of the calibration. First, the number of EM cycles was increased from the default 25 to 100. Second, the number of quadrature points was increased. For the 10-item tests, 10 quadrature points were used for the 1PLM and 2PLM, and 19 points were used for the 3PLM. For the 40-item tests, 19 quadrature points were used for all three models. For the 80-item tests, 37 quadrature points were used for all models. Third, prior distributions were used for all item parameters to ensure convergence to a solution. For all models, the prior used for b_i was $N(0, 1.5)$. For the 2PLM a_i, the $N(1.1, .6)$ distribution was used. For the 3PLM, the prior for a_i was $N(1.9, 1.0)$. (The average value for the prior distribution of the a_is differed for the 3PLM because the parameterization in MULTILOG uses a multiplicative factor of 1.7 that is not used for either the 1PLM or the 2PLM.) The c_i prior distribution for the 3PLM was the logit-normal $(-1.1, .5)$. The priors used for this calibration were similar to the default priors used in BILOG (Mislevy & Bock, 1986, p. 4–26).

Calculation of the Four Fit Indices

For the new indices (S-X^2 and S-G^2), observed proportions correct and incorrect for each NC score for each item were obtained from observed data. Expected proportions were obtained through the calculation of likelihoods for each NC score as well as the likelihoods for each NC score omitting each item. These likelihoods were then combined with the omitted item (Equation 12) to produce the expected proportions for each item for each NC score.

The indices selected for comparison were a Pearson χ^2 index (Q_1-X^2) [similar to Yen's (1981) Q_1 shown in Equation 4] and a likelihood ratio G^2 index (Q_1-G^2) [similar to McKinley & Mills' G^2 (1985) shown in Equation 5]. These two indices were selected because (1) they appear to be the most common indices used in applied settings (BILOG provides an item-fit measure very similar to Q_1-G^2), and (2) there is ample information regarding the construction and performance of the two indices in the literature.

The observed frequencies for the two comparison indices were calculated by first ordering examinees based on their modal a posteriori (MAP) θ estimates. Examinees were then partitioned into 10 intervals such that approximately 100 fell in each interval. Correct and incorrect responses to each item were then tallied for each interval. The expected frequencies were computed using the modeled expected proportions for the median MAP for each interval.

All four fit indices were calculated from the tables of observed and expected frequencies using the X^2 and G^2 formulas. Calculations were made using the computer program GOODFIT (Orlando, 1997).

Minimum Cell Frequencies

Often, the calculated expected frequencies for correct and incorrect responses are small for a given score interval. Very small expected frequencies in the rows of the tables used to construct these fit statistics could decrease the accuracy of the χ^2 approximation for their distribution. To avoid this problem, an algorithm was developed to evaluate the expected counts. This algorithm began at each end (the highest NC score or score group and the lowest NC score or score group) and collapsed cells toward the middle of the list of scores until all cells had sufficiently large expected counts. The table of observed frequencies was collapsed the same way to be comparable to the collapsed expected table.

Cochran (1952) suggested a minimum cell frequency of 5. This has been advocated in many textbooks, but it has been challenged in recent reports. Roscoe & Byars (1971) suggested that restrictions be placed on the average expected frequency rather than requiring a minimum expected frequency. Larntz (1978) reported that X^2 appeared to achieve the desired frequency of rejection under null conditions in which all expected cell frequencies were greater than 1.0. The G^2 statistic, however, is much more sensitive to small cell expectations. Larntz reported that with many cell expected frequencies between 1.5 and 4.0, the G^2 statistic rejected the null hypothesis too often. Read & Cressie (1988) found that "... the first two moments of G^2 will be larger than those for X^2 when many expected frequencies are in the range 1 to 5" (p. 141).

Based on these discussions, some preliminary analyses were performed in which results obtained with a minimum cell frequency of 4 were compared with results obtained with a minimum frequency of 1. X^2 appeared to perform reasonably well with a minimum frequency of 1, but G^2 did not. Although the performance of G^2 began to improve with a minimum frequency of 4, it was still less than optimal. Further, the more extensive collapsing required to obtain the minimum frequency of 4 meant that several items were omitted from each analysis; too few rows (score groups) remained after collapsing to calculate a positive value for the DF. Based on this information, the main analyses were implemented with a minimum cell frequency of 1, with the expectation that X^2 would perform well and that G^2 would be less than optimal.

Results

Type I Error Rates

Significance levels of .05 and .01 were used to examine the Type I error rates of the indices. Because the general pattern of results was similar for these levels, only results for the .05 level are reported.

Table 1 summarizes the performance of the four indices for three test lengths under conditions in which the CM and GM were identical. For all indices, the DFs were calculated as the number of rows remaining after collapsing minus the number of parameters used to calibrate the item. Optimal performance under these conditions would yield proportions close to .05.

For the new NC score indices (S-X^2 and S-G^2) at all test lengths, Table 1 shows that the rejection rates for S-X^2 were close to .05, though S-G^2 rejected model fit more often than expected. As discussed above, this could be due to the high sensitivity of G^2 to small cell frequencies. There

Table 1
Proportion of Indices Greater Than $p = .05$, With GM and CM Identical (GM = CM) and GM More Complex Than CM (GM > CM)

Index and	GM = CM			GM > CM		
Test Length	1PLM	2PLM	3PLM	3PLM/1PLM	2PLM/1PLM	3PLM/2PLM
$S\text{-}X^2$						
10	.04	.06	.07	.52	.57	.11
40	.06	.05	.06	.58	.52	.13
80	.05	.04	.06	.53	.46	.12
$S\text{-}G^2$						
10	.05	.09	.09	.54	.59	.13
40	.08	.10	.10	.63	.60	.18
80	.12	.12	.13	.63	.60	.20
$Q_1\text{-}X^2$						
10	.95	.95	.96	.98	.97	.97
40	.14	.20	.30	.72	.68	.42
80	.06	.08	.15	.70	.66	.26
$Q_1\text{-}G^2$						
10	.97	.96	.96	.99	.97	.97
40	.21	.28	.37	.74	.69	.51
80	.10	.10	.18	.72	.67	.29

did not appear to be a clear pattern of differences in the performance of these indices due to the number of parameters in the IRT model.

The performance of $S\text{-}X^2$ did not seem to be affected by test length, but $S\text{-}G^2$ performed more poorly for longer tests. This could be due to the larger number of score groups in longer tests, which offered more opportunities for small expected frequencies to appear. For tests with either 10 or 40 items, Table 1 shows that the partitioned score indices were not useful; both $Q_1\text{-}X^2$ and $Q_1\text{-}G^2$ rejected fit far too often. The performance of these indices improved as test length increased, approaching reasonable levels for the 80-item test. $Q_1\text{-}G^2$ also tended to reject fit more often than $Q_1\text{-}X^2$. For the 40- and 80-item tests, the partitioned score indices rejected fit more frequently as the number of parameters increased.

Performance of the NC Score Indices

To determine the degree to which the NC score indices followed a χ^2 distribution with the specified DFs, the first two moments of the indices were examined. If the index is distributed as approximately χ^2, the mean would equal the DFs and the variance would be twice the DFs.

For each test length and Type I condition, $S\text{-}X^2$ and $S\text{-}G^2$ were grouped by their DFs, and the means and variances of each group of statistics were calculated. Summary statistics were computed to indicate whether these moments were significantly different from their expected values. The extent of this significance is indicative of the indices' deviation from a χ^2 distribution.

Results of these analyses [Orlando (1997), Appendix A] generally supported conclusions based on the overall Type I performance of the indices. The mean and variance of the $S\text{-}G^2$ index were consistently larger than expected, often significantly so, though the first two moments of $S\text{-}X^2$ tended to be relatively close to their expected values.

Examination of the first two moments of the $S\text{-}X^2$ index did uncover some information not evident from the overall analysis, which showed no clear difference among test lengths or model

conditions. The mean of $S\text{-}X^2$ suggested that the approximation of the distribution of this index by the χ^2 distribution improved as test length increased, and was less accurate for the 3PLM than for the 2PLM and 1PLM.

Power

Conditions in which the GM used more item parameters than the CM were used to examine the power of the four fit indices. Table 1 also summarizes these results. (The power values for $Q_1\text{-}X^2$, $Q_1\text{-}G^2$, and $S\text{-}G^2$ are not very useful, however, because of the inflated Type I error rates of these indices.) Table 1 reveals that $S\text{-}X^2$ did detect some degree of misfit. It was expected that the 3PLM-1PLM combination would yield greater detection of misfit than the 2PLM-1PLM combination, but that was not always the case. Overall, the 3PLM-2PLM combination detected misfit least often. There appeared to be no difference in the power of this index as a function of test length.

Power of $S\text{-}X^2$ as a function of item parameters. Even if the CM has fewer parameters than the GM, items with moderate values for generating parameters should still have adequate fit. For example, in the 2PLM-1PLM case, a common a_i was fit to all of the items on the test when the data were calibrated. Items with similar true and calibrated a_is should fit well, though items with very different a_is should display greater misfit. If the index correctly rejects misfitting items, the detection of misfit should be predictable, given the values of the true generating parameters. Table 2 summarizes the power of $S\text{-}X^2$ according to pertinent values of the generating parameters.

For the 3PLM-1PLM combination, the percentage of misfitting items was calculated separately for the quartiles of the generating a_i parameter and, within those quartiles, halves of the generating b_i. (Examination of the power for different values of c_i yielded no discernible pattern). For all test lengths, $S\text{-}X^2$ was most likely to detect misfit when the generating a_i was in the fourth quartile (Q4). That is, the highest generating a_is resulted in the greatest detection of misfit. The lowest detection occurred when items with generating a_is were in the second quartile (Q2). The detection of misfit shown for the first and third quartiles (Q1 and Q3) was relatively similar. Q1 resulted in a greater detection of misfit in the 40- and 80-item tests, and Q3 showed greater misfit in the 10-item test.

This similarity of misfit detection when the generating a_i was in Q1 and Q3 is interesting. Because the CM fits a common slope, the extreme low value of a_i in Q1 should result in greater misfit detection than a_i in Q3. The presence of c_i would contribute to the asymmetry of this result. In general, when the 3PLM was the GM and the 1PLM was the CM, the calibrated curve underestimated the proportion of correct responses at the low end of the θ scale where there were few observations. This made it more difficult to detect misfit. Thus, the power to detect misfit when the generating a_i was in Q1 was more similar to conditions in which the generating a_i was in Q3 rather than those in which a_i was in Q4.

Table 2 also shows the percentage of rejections for halves of the generating b_i at each of the quartiles of the generating a_i. The generating a_is and b_is interacted to affect the power of the indices. In almost all cases, for Q1 and Q2 of a_i the indices rejected fit most often when the generating b_i was high. For Q3 and Q4 of a_i, the indices rejected fit more often when b_i was low. For some combinations of generating a_i and b_i values, the misfit of the CM was concentrated in the extremes of the θ scale, where there were few observations and misfit was relatively difficult to detect. For other combinations, the CM tended to adjust in such a way that it underestimated the proportion of correct responses for many observations in the middle of the θ scale. This resulted in obvious lack of fit.

In the 3PLM-1PLM condition, the overall power exhibited by $S\text{-}X^2$ ranged from .52 to .58. This is low, but misleading. The results summarized in Table 2 illustrate that $S\text{-}X^2$ can exhibit high power when the fit is truly poor. The power for the index when the generating a_i was in Q4, where

Table 2
Proportion of NC Score Pearson Indices Greater Than $p = .05$ by Quartiles of Generating a_i and Halves of Generating b_i

Models and Distribution	Test Length		
	10	40	80
GM = 3PLM, CM = 1PLM			
Overall	.52	.58	.53
Quartiles of a_i			
Q1	.48	.56	.55
Q2	.31	.30	.27
Q3	.53	.54	.45
Q4	.76	.89	.86
Halves of b_i within a_i quartiles			
Q1			
Lower	.29	.35	.34
Upper	.67	.78	.76
Q2			
Lower	.25	.12	.09
Upper	.38	.47	.44
Q3			
Lower	.52	.59	.45
Upper	.55	.48	.45
Q4			
Lower	.91	.95	.91
Upper	.64	.83	.81
GM = 2PLM, CM = 1PLM			
Overall	.57	.52	.46
Quartiles of a_i			
Q1	.84	.88	.89
Q2	.27	.24	.20
Q3	.33	.16	.05
Q4	.82	.79	.69
GM = 3PLM, CM = 2PLM			
Overall	.11	.13	.12
Halves of b_i			
Lower	.07	.09	.08
Upper	.15	.17	.15

misfit detection was most pronounced, ranged from .64 to .95. Results of this analysis lend further support to the idea that S-X^2 demonstrated sufficient power to detect misfit when the 3PLM was the GM and the 1PLM was the CM.

Table 2 also summarizes results of the 2PLM-1PLM generation-calibration combination, for quartiles of the generating a_i parameter. The greatest misfit detection rate was for generating a_is in Q1, and Q4 of generating a_is yielded the next largest rate. For the 40- and 80-item tests, generating a_is in Q3 yielded the least misfit; for the 10-item test, generating a_is in Q2 resulted in the least misfit.

These results show that S-X^2 detected misfit when it was expected. When the generating a_i was in either Q1 or Q4, it was relatively extreme. Because the common calibrating a_i is usually a moderate value, the greatest misfit detection rate should be at the extremes. The power values for the .05 level ranged from .69 to .89 when the generating a_i was in Q1 or Q4. However, when the

generating a_i was in Q2 or Q3, the power ranged from .05 to .33. Thus, $S\text{-}X^2$ possessed adequate power to detect misfit when the 2PLM was the GM and the 1PLM was the CM.

For the 3PLM-2PLM combination, Table 1 shows that $S\text{-}X^2$ had greater misfit detection rates when the generating b_i was in the upper half of its distribution than when it was in the lower half. In the latter case, the calibrated 2PLM adjusted by setting b_i slightly lower.

Discussion

Performance of Q_1-X^2 and Q_1-G^2

When comparing the performance of $Q_1\text{-}X^2$ and $Q_1\text{-}G^2$ in this study to the performance of similar indices discussed in the literature (e.g., McKinley & Mills, 1985; Yen, 1981), it should be noted that there are computational differences that might make direct comparison difficult, although $Q_1\text{-}X^2$ and $Q_1\text{-}G^2$ were computed in a way similar to fit indices currently in use. For example, other studies of these indices did not appear to employ a minimum cell frequency requirement. Thus, the cell collapsing implemented in this study was likely different from what has been done in the past. Also, partitioning observations into 10 groups might differ from prior studies. Both of these computational differences clearly affect the value of the indices.

There are also some differences between the design of this study and that of studies that previously examined the performance of Q_1-type indices. First, McKinley & Mills (1985) and Yen (1981) used item parameter and θ estimate distributions different from those used here. Second, Yen's (1981) simulation was based on 36-item tests, and McKinley & Mills (1985) used a 75-item test. The present study simulated test lengths of 10, 40 and 80 items. Both Yen (1981) and McKinley & Mills (1985) had only one replication for each condition, whereas results based on 100 replications per condition were reported here.

Despite these differences, it seems reasonable to regard the behavior of $Q_1\text{-}X^2$ and $Q_1\text{-}G^2$ (as calculated here) as descriptive of the performance of other fit indices similar to Q_1. Although both Yen (1981) and McKinley & Mills (1985) reported favorable results for Q_1, this index might have inflated Type I error rates in practice. In applications of Q_1, the index is often transformed into a z score, and then the frequency distribution of the z-transformed indices is examined. It is not unusual in these applications to observe z scores as large as 30 or more, and it appears that the cutoff value for detection of misfit for these transformed indices is 3.0 rather than 1.96 (A. Fitzpatrick, personal communication, August 8, 1996). Ansley & Bae (1989) (as cited in Ankenmann, 1994, p.22) suggested that these indices are not always approximated well by a χ^2 distribution with the specified DFs. They reported that Yen's statistic appeared to be distributed as a noncentral χ^2 with the noncentrality parameter varying as a function of test length and sample size. Given this information, the performance of the $Q_1\text{-}X^2$ and $Q_1\text{-}G^2$ indices in this study is not surprising.

Because the performance of Q_1 has not been extensively studied for tests with fewer than 36 items, there were no a priori predictions regarding the behavior of $Q_1\text{-}X^2$ and $Q_1\text{-}G^2$ for the 10-item condition. Based on the results of this study, it seems reasonable to conclude that these indices are not appropriate for use with 10-item tests.

Following Yen's (1981) suggestion, $Q_1\text{-}X^2$ and $Q_1\text{-}G^2$ in the 40-item condition were expected to be approximately distributed as χ^2 with $10 - m$ DFs. Yen used Pearson χ^2 significance tests to evaluate Q_1 and was unable to reject the null hypothesis that Q_1 is distributed as χ^2 with $10 - m$ DF. Because Yen did not calculate the number of items for which the index was significant at the .05 and .01 levels, it is difficult to directly compare Yen's results with the results of $Q_1\text{-}X^2$ and $Q_1\text{-}G^2$ in the 40-item condition reported here. Despite the possible computational differences and the differences in study design, the performance of $Q_1\text{-}X^2$ and $Q_1\text{-}G^2$ in the 80-item condition was not dramatically different from results reported by McKinley & Mills (1985) for 75-item tests.

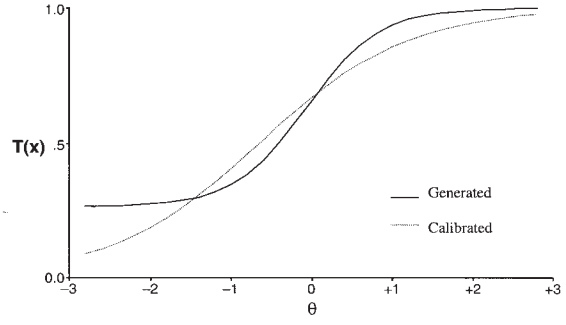
Performance of S-G^2

Because G^2 possesses some desirable properties that X^2 lacks, the performance of S-G^2 was hoped to be acceptable. However, that was not the case. The inflated Type 1 error rates of S-G^2 might be due to the small minimum cell frequency requirement imposed. With a sufficiently large sample size, the minimum requirement could be raised to 5 (see Cochran, 1952). It is possible that performance would improve under these conditions, although it is not clear how large a sample size would be required; it would certainly be much greater than 1,000.

Performance of S-X^2

Although the Type I error rates of S-X^2 were fairly consistent across study conditions, the power of the index varied by condition and test length. The index exhibited a high rate of misfit detection when the 3PLM was the GM and the 1PLM was the CM. Figure 2 shows an example of the generating and calibrated IRFs of an item from an 80-item test under this condition. As shown in Figure 2, the calibrated 1PLM IRF underestimated the proportion of correct responses for very low θs, overestimated for $-1.5 < \theta < 0$, and underestimated for high θs. In this example, a large proportion of the misfit was concentrated at relatively extreme values of θ, where there were few observations. Despite this pattern, the item was detected as displaying misfit. S-X^2 was not very large in this case, but was significant at the .05 level.

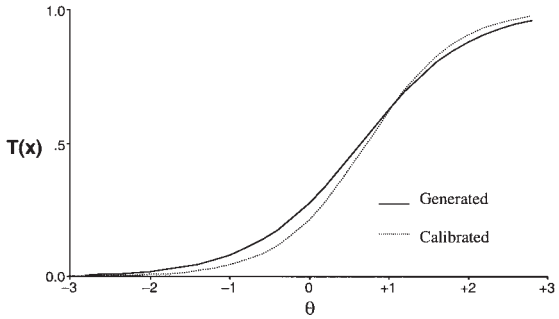
Figure 2
Generated (3PLM) and Calibrated (1PLM) IRFs for an Item From an 80-Item Test [S-$X^2(51) = 70.16$, $p = .04$; Generating Parameters: $a_i = 1.31, b_i = -.06, c_i = .26$; Calibrated Parameters: $a_i = 1.09, b_i = -.65$]



For the 2PLM-1PLM generation-calibration combination, S-X^2 exhibited adequate power to detect misfit. Figure 3 shows an example of the generating and calibrated IRFs of an item from a 10-item test with this combination. In this example, the two IRFs were nearly coincident, but the majority of the misfit was concentrated near the center of the θ continuum. Because the calibrated 1PLM IRF slightly underestimated the proportion of correct responses for middle values of θ, where there were many observations, S-X^2 was just able to detect this item as misfitting at the .05 level. Depending on the location of the misfit, S-X^2 might be able to detect misfit in cases for which the degree of misfit is not pronounced.

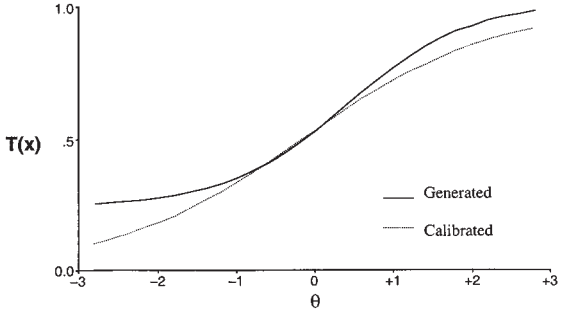
When the 3PLM was the GM and the 2PLM was the CM, S-X^2 did not demonstrate much power. This type of misfit is difficult to detect because the CM is often able to adjust its parameters so that the majority of the misfit is concentrated at the upper and lower extremes of the θ scale. Figure 4

Figure 3
Generated (2PLM) and Calibrated (1PLM) IRFs for an Item From a 10-Item Test [$S\text{-}X^2(8) = 15.68$, $p = .047$; Generating Parameters: $a_i = 1.48$, $b_i = .65$; Calibrated Parameters: $a_i = 1.79$, $b_i = .73$]



illustrates this point. For one item from a 40-item test, the calibrated b_i was much lower than the generating b_i and the calibrated a_i was higher. These adjustments resulted in IRFs that were nearly coincident between $-1 < \theta < 1$. The calibrated IRF failed to fit the generating curve only at θs for which there were very few observations. As a result, this item was not detected as misfitting.

Figure 4
Generated (3PLM) and Calibrated (2PLM) IRFs for an Item From a 40-Item Test [$S\text{-}X^2(26) = 35.40$, $p = .103$; Generating Parameters: $a_i = .76$, $b_i = .40$, $c_i = .23$; Calibrated Parameters: $a_i = .82$, $b_i = -.15$]



$S\text{-}X^2$ had very little power to detect misfit when the 3PLM was the GM and the 2PLM was the CM. Yen (1981) found that Q_1 was often unable to detect that the 2PLM was inappropriate for 3PLM-generated data. McKinley & Mills (1985) also reported low power to detect this type of misfit. 2PLM calibration appears to fit 3PLM-generated data fairly well, so there is not much misfit to detect.

Conclusions

Results of this study indicate that the Pearson χ^2 form of the new index $S\text{-}X^2$ is a promising candidate for detecting item misfit. However, strong conclusions about the performance of $S\text{-}X^2$ are limited by the design of this study. All results were based on a sample size of 1,000 and

do not necessarily extend to larger or smaller samples. The index could become more sensitive as sample size increases, but smaller sample sizes might present problems due to small expected cell frequencies. Also, the index could behave differently for tests with more than 80 items. The performance of $S\text{-}X^2$ under conditions of misfit not discussed here should also be examined, e.g., multidimensionality and nonmonotonicity.

A graphical representation of the observed and expected proportions correct and incorrect would make $S\text{-}X^2$ more useful. With simulated data, it is possible to examine the pattern of misfit by superimposing the calibrated IRF on the generating IRF; examining this display aids in understanding the type of misfit detected for the item. With real data, of course, the true parameters are unknown, so there is no generating IRF to compare to the calibrated IRF. In this case, the relationship between the observed and expected frequencies can provide some insight into the type of misfit being detected. Although it is possible to examine these frequencies directly, informative graphics could provide a useful visual representation of the relationship between them.

The algorithm described by Thissen et al. (1995) for the calculation of NC score likelihoods is not limited to dichotomous items. Because this algorithm has been extended to polytomous items, $S\text{-}X^2$ could also be extended for use with polytomous items. It is unknown whether the performance of this index for polytomous items would be acceptable.

References

Andersen, E. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123–140.

Ankenmann, R. (1994). *Goodness of fit and ability estimation in the graded response model.* Unpublished manuscript.

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., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35,* 179–197.

Chen, W. (1995). Estimation of item parameters for the three-parameter logistic model using the marginal likelihood of summed scores (Doctoral dissertation, University of North Carolina, 1995). *Dissertation Abstracts International, 56/10-B,* 5825.

Cochran, W. G. (1952). The chi-square test of goodness of fit. *Annals of Mathematical Statistics, 23,* 315–345.

Glas, C. A. W. (1988). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525–546.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

Kingston, N., & Dorans, N. (1985). The analysis of item-ability regressions: an exploratory IRT model fit tool. *Applied Psychological Measurement, 9,* 281–288.

Larntz, K. (1978). Small-sample comparisons of exact levels for chi-squared goodness-of-fit statistics. *Journal of the American Statistical Association, 73,* 253–263.

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453–461.

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.(1986). *Bilog: Item analysis and test scoring with binary logistic models.* Mooresville IN: Scientific Software.

Orlando, M. (1997). Item fit in the context of item response theory. (Doctoral dissertation, University of North Carolina, 1997). *Dissertation Abstracts International, 58/04-B,* 2175.

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

Read, T. R. C., & Cressie, N. A. C.(1988). *Goodness-of-fit statistics for discrete multivariate data.* New York: Springer.

Rogers, H., & Hattie, J. (1987). A monte carlo in-

vestigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11,* 47–57.

Roscoe, J. T., & Byars, J. A. (1971). An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association, 66,* 755–759.

Rost, J., & von Davier, M. (1994). A conditional item-fit index for Rasch models. *Applied Psychological Measurement, 18,* 171–182.

Stroud, A. H. (1974). *Numerical quadrature and solution of ordinary differential equations.* New York: Springer.

Thissen, D. (1991). *MULTILOG user's guide: Multiple categorical item analysis and test scoring using item response theory.* Chicago: Scientific Software.

Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19,* 39–49.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen, *Computerized adaptive testing: A primer* (pp. 65–101). Hillsdale NJ: Erlbaum.

Wright, B., & Mead, R. (1977). *BICAL: Calibrating items and scales with the Rasch model* (Research Memorandum No. 23). Chicago: University of Chicago, Department of Education, Statistical Laboratory.

Wright, B., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23–48.

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

Acknowledgments

This research was funded by the Law School Admissions Council. The opinions expressed in this report do not necessarily represent those of LSAC.

Author's Address

Send requests for reprints or further information to Maria Orlando, RAND Corp., 1700 Main Street, Santa Monica CA 90407, U.S.A. Email: Maria_Orlando@rand.org.