Further Investigation of the Performance of $S - X^2$: An Item Fit Index for Use With Dichotomous Item Response Theory Models

Maria Orlando, RAND Health Division, Santa Monica
David Thissen,
University of Carolina, Chapel Hill

This study presents new findings on the utility of $S - X^2$ as an item fit index for dichotomous item response theory models. Results are based on a simulation study in which item responses were generated and calibrated for 100 tests under each of 27 conditions. The item fit indices $S - X^2$ and $Q_1 - X^2$ were calculated for each item. ROC curves were constructed based on the hit and false alarm rates of the two indices. Examination of these curves indicated that in general, the performance of $S - X^2$ improved with test length and sample size. The performance of $S - X^2$ was superior to that of $Q_1 - X^2$ under most but not all conditions. Results from this study imply that $S - X^2$ may be a useful tool in detecting the misfit of one item contained in an otherwise well-fitted test, lending additional support to the utility of the index for use with dichotomous item response theory models. *Index Terms: item response theory*, $S - X^2$, $Q_1 - X$, *model = data fit*, *item fit index.*

Item response theory (IRT) has become a widely used tool for the analysis of items, tests, and people. This methodology offers many advantages over the classical test theory, allowing for a number of advances in testing, including the development of computer adaptive tests, item banking, and test equating. The one-, two-, and three-parameter logistic models (1PL, 2PL, and 3PL) are the three IRT models most commonly used for multiple-choice tests.

The appropriate use of these IRT models requires that a number of assumptions be made about the nature of the data. One of the more basic assumptions made in applications of IRT is that the model accurately represents the data. If this assumption is not met, inferences regarding the nature of the items and tests can be erroneous, and the potential advantages of using IRT are not gained. It is therefore desirable to have methods for checking this assumption in standard applications.

One way to assess the appropriateness of the chosen IRT model is through examination of model-data fit. This examination can be performed at the level of the entire test, as well as at the individual item level. A number of procedures have been developed for the evaluation of goodness of fit for tests and items. Although the assessment of model-data fit for the overall test presents its own set of challenges, the research presented here is concerned with assessment of fit at the level of the individual item.

Generally speaking, the most straightforward means of assessing goodness of fit is by constructing a goodness-of-fit statistic that compares observed data with modeled predictions. In the application of IRT models, assessment of fit at the item level is problematic for models other than the 1PL, or Rasch, family because the underlying ability being measured by the test is defined as a latent variable. For the 1PL model, the total number correct, or summed score, is a sufficient statistic for

Applied Psychological Measurement, Vol. 27 No. 4, July 2003, 289–298
DOI: 10.1177/0146621603253011

the ability being measured by the test, so observed and expected responses to each item within each score group can be directly compared. In contrast, the 2PL and 3PL models' predictions pertain to the patterns of responses rather than the summed scores. Thus, there is no direct way to compare the observed and expected responses.

A number of item-level fit indices for use with dichotomous item response theory models have been developed and studied (see Orlando, 1997, for a more detailed review of currently available fit indices). The common procedure for constructing item fit indices for the 2PL and 3PL models has been to group respondents based on their estimated standing on the latent variable being measured by the test and obtain observed frequencies correct and incorrect for these groups. Use of this procedure means that observed frequencies are not truly observed; they cannot be obtained without first fitting the model. The model-dependent nature of these indices has made it difficult to ascertain the degrees of freedom associated with the resulting χ^2-like statistics and, consequently, the exact nature of their null distributions.

An alternative approach to assessing item fit has been introduced recently (Orlando & Thissen, 2000). In this approach, an item fit statistic is formulated based on the observed and expected frequencies correct and incorrect for each summed score. A method of computing model-predicted joint likelihood distributions for each summed score, described briefly by Lord and Wingersky (1984) and in more detail by Thissen, Pommerich, Billeaud, and Williams (1995), is used to calculate the expected frequencies correct and incorrect for each item for each summed score. The most obvious advantage to this approach is that these expected frequencies can be compared directly to observed frequencies in the data.

This summed score approach was used to form two new indices: $S - X^2$, a Pearson X^2 statistic, and $S - G^2$, a likelihood ratio G^2 statistic. The performance of these two indices was studied and compared to the performance of Pearson X^2 and likelihood ratio G^2 forms of Yen's Q_1index in a simulation study (Orlando & Thissen, 2000). Although the two Q_1-like comparison indices and $S - G^2$ tended to display inflated Type I error rates, the performance of $S - X^2$ was encouraging. The Type I error rates were close to the expected 0.05 and 0.01 levels, and the index displayed moderate power to detect misfit.

The study design employed by Orlando and Thissen (2000) provides only a preliminary indication of the performance of $S - X^2$. Although the study included three different test lengths (10, 40, and 80 items), all conditions used a sample size of 1,000. The type of misfit imposed in that study was also limited. Item misfit was simulated by generating data with the 2PL or 3PL model and calibrating it with the 1PL or 2PL model. Use of this strategy to create model misspecification is common in simulation studies investigating item fit indices, but it provides only one example of misfit; others may occur in real applications. In addition to misspecifying the model for an entire test, it is possible to choose a model that fits most of the items on the test but does not adequately fit one or two items on the test. Ideally, an item fit index would be able to detect this type of misfit.

The purpose of this investigation is to extend the initial evaluation of the summed score approach to constructing item fit indices for use with dichotomous IRT models. Due to the inflated Type I error rates displayed by the likelihood ratio G^2 form of the summed score index, only the Pearson X^2 form is examined in this study. The performance of $S - X^2$ is again compared to that of a Pearson X^2 form of Yen's Q_1 index.

Method

Generating Bad Items

The goal in generating bad items was to obtain item trace lines that could realistically occur in actual data but that would be sufficiently different from the logistic form to be detected as misfitting. To simplify the generation of these items and the subsequent explanation of their form, three nonlogistic items were created based on three variants of the 3PL equation.

The first type of bad item has a nonmonotonic form that is not uncommon in actual data. For this trace line, the probability of a correct response is at guessing level for very low values of proficiency (θ), decreases over moderate values of θ, and increases again as θ increases, creating a trough or dip in the curve. This pattern of responding represents the idea that "a little knowledge can be a dangerous thing." In the context of a multiple-choice test, this would occur when some students of moderate ability are fooled by a particularly good distracter and score at less than chance. Sadler and Lieberman (1994) observed this nonmonotonic pattern of responding in the course of development and refinement of a high school science test aimed to reveal scientific misconceptions. Other examples of correct item response patterns that are best fitted by this parameterization have been identified in an operational military accession test (Thissen & Steinberg, 1984) and in the Armed Services Vocational Aptitude Battery (Thissen, 1986a). A trace line with these characteristics was described by Thissen (1986b), Thissen and Bock (1986), and Wainer and Thissen (1987) and is expressed by

$$T = c\{1/[1 + \exp(1.7a[\theta - (b - d)])]\} + 1/[1 + \exp(-1.7a[\theta - b])], \tag{1}$$

where a, b, and c are as in the 3PL, d is a positive number, and larger values of d create a larger dip in the curve. The top panel of Figure 1 displays the item of this form, referred to as Bad Item 1 in this study. The solid line is the actual trace line for Bad Item 1, and the dashed line is the best-fitting 3PL curve generated from a probit analysis.

The second type of bad item, referred to as Bad Item 2 in this study, is shown as the solid line in the middle panel of Figure 1, with the dashed line representing the best-fitting 3PL curve generated from a probit analysis. This trace line, described by Barton and Lord (1981), differs from the 3PL model in that the probability of answering correctly never reaches 1 but plateaus at some level less than 1. The motivation for this model development was a concern that the 3PL might be severely penalizing high-ability students who simply made a clerical error in answering incorrectly to an easy item. However, this response pattern could perhaps occur for an exceptionally difficult test item. It is expressed by

$$T = c + (1 - c)d/[1 + \exp(-1.7a[\theta - b])], \tag{2}$$

where a, b, and c are as in the 3PL and d is the upper asymptote.

The third bad item, shown in the bottom panel of Figure 1 along with its best-fitting 3PL generated from a probit analysis, exhibits a plateau over middle values of θ but follows a logistic curve before and after the plateau. Bock and colleagues (1973) introduced this parameterization, formed by the addition of two logistic functions, to model growth in recumbent length from birth to age 18. In Bock et al.'s application, the first logistic component described prepubertal growth, and the second component captured the adolescent growth spurt. A trace line reflecting this response pattern is expressed by

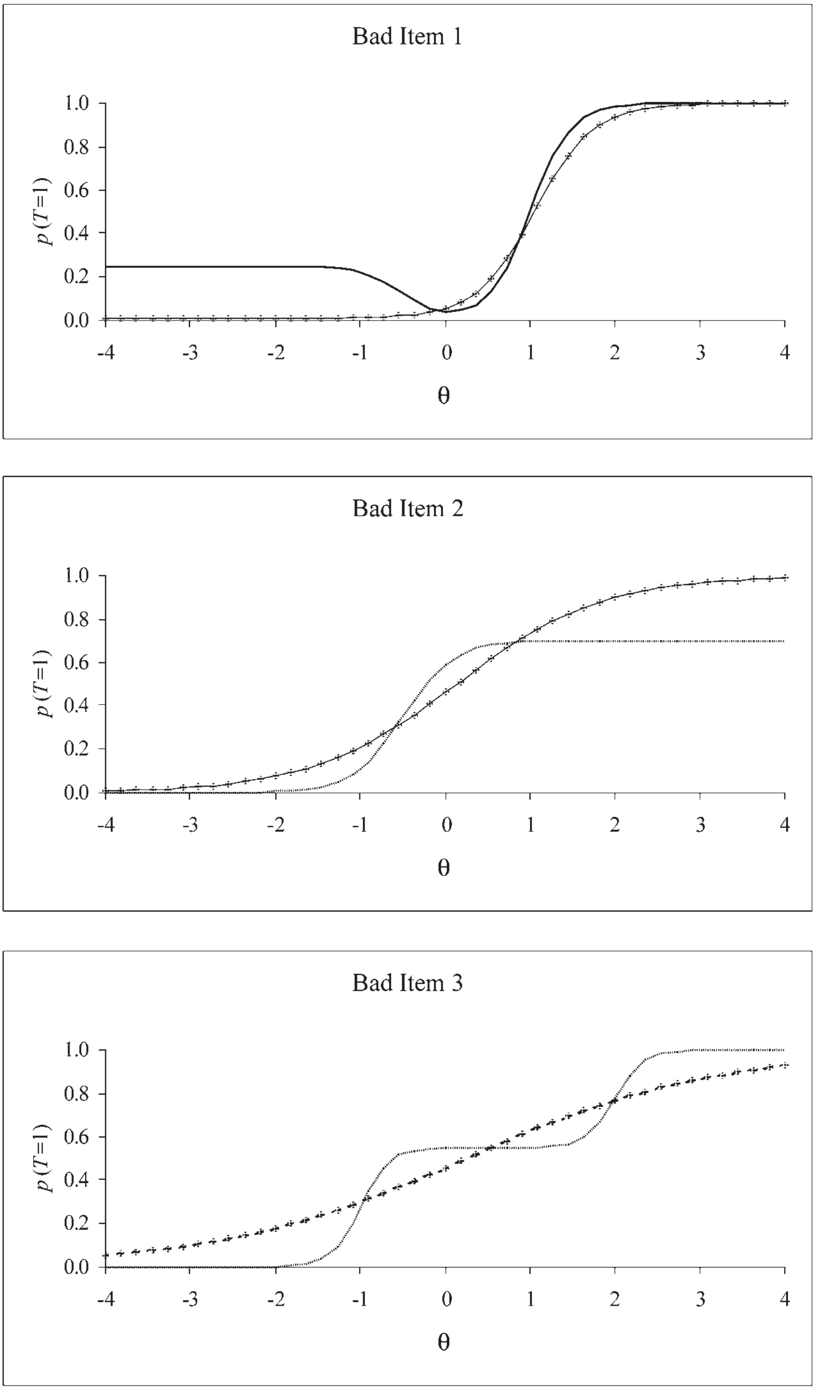
$$T = c + \{x/[1 + \exp(-1.7a[\theta - b])]\} + \{y/[1 + \exp(-1.7a[\theta - (b + d)])]\}, \tag{3}$$

where a, b,and c are as in the 3PL, d is a positive number indicating the length of the plateau, and x and y are the heights of the two functions. For this equation, x, y,and$c$are constrained to sum to unity. Ability test items with curves of this type and with even larger anomalies are described by Ramsay (1991) and Abrahamowicz and Ramsay (1992).

Within the framework of these three equations, an infinite number of possible trace lines can be created given the choice of parameter values. Originally, the parameter values were chosen somewhat casually, based on the appearance of the curve relative to a 3PL. A mini-simulation was performed using these values, and it was found that the curves did not diverge enough from the 3PL to be detected by either index at a rate much greater than chance. Although this finding is not

Figure 1

Actual (Solid) and Best-Fitting Three-Parameter Logistic (Dotted) Trace Lines for the Correct Response to Each of Three Bad Items Used in the Simulation Study



the focus here, it is interesting and rather reassuring to discover that the 3PL does a very good job estimating curves that diverge slightly from the logistic form.

To ensure that the bad items would be sufficiently nonlogistic to create misfitting conditions, a more rigorous approach to parameter selection was adopted to define the three bad item curves. Various combinations of parameters were chosen, and the observed frequencies correct and incorrect for specified values of θ were produced for a normal distribution of θ. These frequencies and θ values were then used in probit analyses, and the relative success of fitting these combinations of parameter values was examined. Final parameter values were chosen based on the results of the probit analyses and the appearance of the curves such that the bad items could conceivably occur in actual data but were sufficiently different from the logistic form to be expected to be detected as misfitting. Final parameters for the three bad items, shown as solid lines in Figure 1, are as follows: for Bad Item 1, $a = 2.5$, $b = 1$, $c = 0.25$, and $d = 1.5$; for Bad Item 2, $a = 2$, $b = 0.5$, $c = 0$, and $d = 0.7$; and for Bad Item 3, $a = 3.5$, $b = -1$, $c = 0$, $d = 3$, $x = 0.55$, and $y = 0.45$.

Simulation Study Design

Item responses were generated and calibrated for 100 tests under each of 27 conditions [(3 bad items) × (3 test lengths) × (3 sample sizes)]. The three nonlogistic (bad) items were created and embedded in otherwise 3PL tests of length 10, 40, and 80 items for samples of size 500, 1,000, and 2,000. Logistic items were generated using an augmented version of GEN (Orlando, 1997). All tests were calibrated in MULTILOG (Thissen, 1991) with the 3PL model. The item fit indices, $S - X^2$and $Q_1 - X^2$,were calculated for each item as described in Orlando and Thissen (2000).

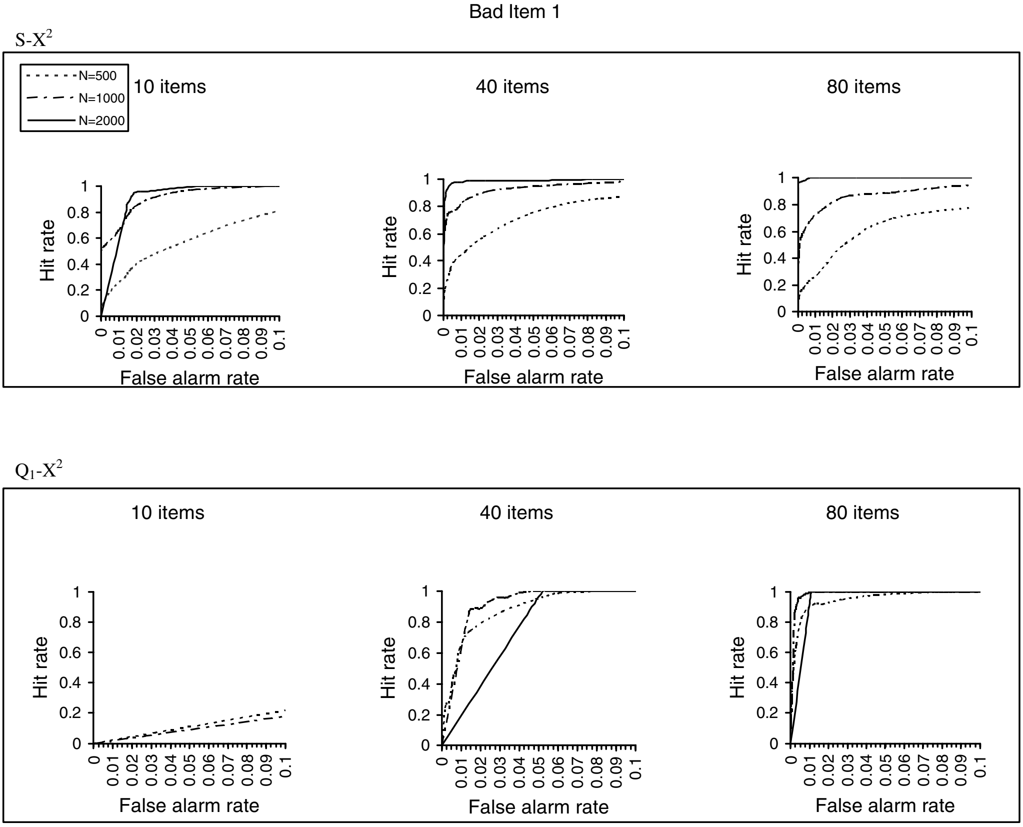
The three bad items were expected to be identified as misfitting by the indices, whereas the other items on the test were expected to be well fitted by the 3PL calibration. Examination of the index values for the three bad items provides an indication of the power of the two indices to detect item misfit when only one item on the test does not fit. Significant index values for the other 9, 39, and 79 items on each test reflect the adverse effect that one misfitting item on a test has on the calibration of the items that are expected to fit. Furthermore, this design allows for the examination of the effect of test length and sample size on the performance of the two indices under these misfit conditions.

Evaluating Performance

Receiver operating characteristic (ROC) curves were constructed (Green & Swets, 1966) for each simulated condition to evaluate the performance of the two fit indices. Each point on the ROC curve represents the ratio of the hit rate to the false alarm rate for the fit index. The shape of the ROC curve can be examined to assess the power/Type I error rate trade-off of the two fit indices in detecting item misfit. For each point on the ROC curve, the value on the horizontal axis is the proportion of "good" items that were erroneously detected as misfitting at a particular cutoff value, and the value on the vertical axis is the proportion of "bad" items that were correctly detected as misfitting at that same value, creating a ratio of power to Type I error. Because false alarm rates higher than 0.1 are not of particular interest in this context, the plots are magnified to show the curve only for false alarm rates between 0 and 0.1 (an entire ROC curve would extend from 0 to 1 on both axes). Perfect performance would result in an ROC curve that nestles into the upper left-hand corner of the ROC plane, and chance performance would result in a curve along the diagonal of the full ROC plane. Better relative performance is indicated by curves that increase vertically more quickly than they increase horizontally or, in other words, curves that reflect higher hit to false alarm ratios.

Figure 2

Receiver Operating Characteristic (ROC) Curves for Evaluation of Performance of $S - X^2$ (Top Panel) and $Q_1 - X^2$ (Bottom Panel) in Appropriately Detecting Misfit of Bad Item 1.



Results

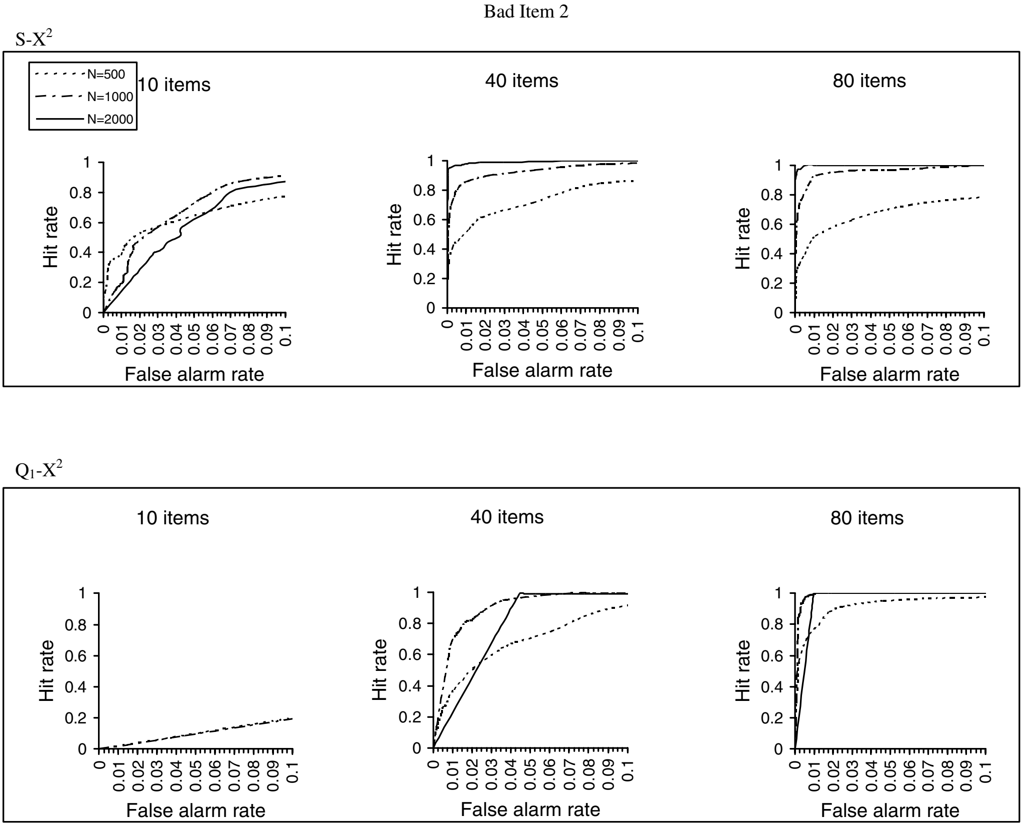
Figures 2 through 4 illustrate the ROC curves evaluating the performance of the two fit indices for each of the three bad items at each sample size and test length. For each bad item type and index, a separate plot was constructed for each test length; separate lines on each of these plots represent the sample size conditions. The top panel of each figure illustrates the performance of $S - X^2$, and the bottom panel illustrates the corresponding performance of $Q_1 - X^2$.

For tests involving Bad Item 1 (see Figure 2), the performance of $S - X^2$ improved with test length and sample size. For all test lengths, the index performed marginally for sample sizes of 500. The $S - X^2$ hit to false alarm ratio was reasonable for sample sizes of 1,000 and favorable for n = 2,000. $Q_1 - X^2$ did not perform well for 10-item tests involving Bad Item 1. For tests with 40 items, the performance of $Q_1 - X^2$ was reasonable for samples of size 500 and 1,000, but interpretation for the sample size 2,000 condition is not clear because hit rates were uniformly 1, and false alarm rates below .05 were not observed. For the 80-item test condition, the $Q_1 - X^2$ hit to false alarm ratio was favorable for all sample sizes.

For tests involving Bad Item 2 (see Figure 3), the performance of $S - X^2$ generally improved with test length and sample size, although the performance in tests with 10 items was marginal and somewhat erratic. For the 40- and 80-item test lengths, the index performed marginally for sample

Figure 3

Receiver Operating Characteristic (ROC) Curves for Evaluation of Performance of $S - X^2$ (Top Panel) and $Q_1 - X^2$ (Bottom Panel) in Appropriately Detecting Misfit of Bad Item 2.



sizes of 500, and the $S - X^2$ hit to false alarm ratio was favorable for the larger sample sizes. The performance of $Q_1 - X^2$ for Bad Item 2 was nearly identical to that of Bad Item 1, with the exception of slightly worse performance for 40-item tests.

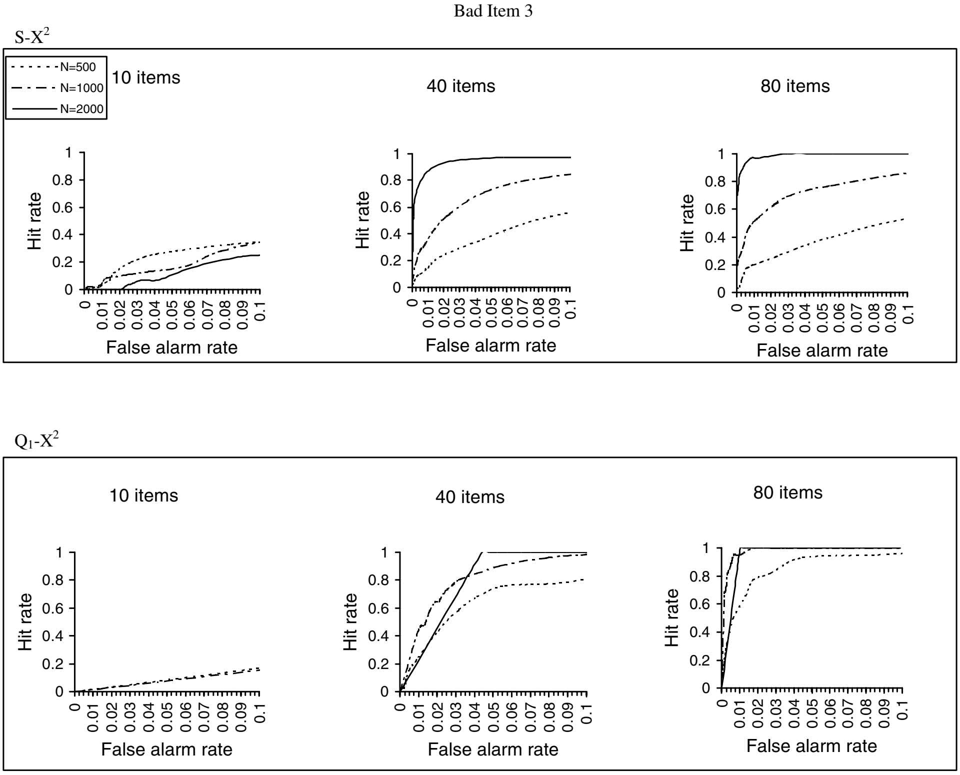
For tests involving Bad Item 3 (see Figure 4), the performance of $S - X^2$ was not acceptable for 10-item tests of any sample size or for sample sizes of 500 at any test length. However, in both the 40- and 80-item conditions, performance of the index was marginal for sample sizes of 1,000 and favorable for sample sizes of 2,000. The performance of $Q_1 - X^2$ for Bad Item 3 was nearly identical to that of Bad Item 2.

Discussion

In an initial evaluation, $S - X^2$ demonstrated promise as a useful tool for assessing item fit in dichotomous IRT models. However, that evaluation was limited in scope; misfit was imposed by using the wrong model to calibrate the entire test (e.g., generating items with the 3PL and calibrating them with the 2PL), and all simulated conditions used a sample size of 1,000. The purpose of this study was to extend the appraisal of $S - X^2$ to include conditions in which only one item in the set is nonlogistic and to assess the performance of the index under different sample sizes. Although the

Figure 4

Receiver Operating Characteristic (ROC) Curves for Evaluation of Performance of $S - X^2$ (Top Panel) and $Q_1 - X^2$ (Bottom Panel) in Appropriately Detecting Misfit of Bad Item 3.



results of this study are not entirely straightforward, $S - X^2$ performed well under many conditions and continued to show advantages over $Q_1 - X^2$, the comparison index.

Based on the results of this study, the $Q_1 - X^2$ statistic does not appear to be useful for 10-item tests. The performance of $S - X^2$, although not entirely consistent under this short test condition, was clearly superior to that of $Q_1 - X^2$. The hit to false alarm ratio of $S - X^2$ was acceptable with large samples for Bad Item 1 and was marginal for Bad Item 2. In addition, supporting analyses indicated that despite the somewhat disappointing hit to false alarm ratio of $S - X^2$ under this short test condition, it still could prove useful in practice. Another perspective on the performance of the indices can be obtained by examining their relative size. In practice, and particularly for short tests, each item has an impact on the fit of all the other items, and test refinement could involve examination of the fit of subsets of items (e.g., removing the worst-fitting item, recalibrating, and reassessing fit). The percentage of times that the value of each index/degree-of-freedom ratio for the bad item was the largest of the set was calculated as an indicator of the indices' relative usefulness in this kind of iterative test refinement. It was found that for 10-item tests, the $S - X^2/df$ ratio was largest for the bad item 56% of the time for samples of 500 and 66% of the time for samples of 1,000 and 2,000. The corresponding percentages for the $Q_1 - X^2/df$ ratio were 8, 9, and 12 (see Table 1).

In the two longer test conditions, the hit to false alarm ratio of $S - X^2$was comparable or superior to that of $Q_1 - X^2$ for sample sizes of 1,000 and 2,000. However, $Q_1 - X^2$ outperformed $S - X^2$in the smallest sample size condition for 80 items. Overall, $S - X^2$appeared to be more affected by

Table 1

Percentage of Times Largest Index/*df* Ratio Value Corresponds to Bad Item for Each Study Condition and Index

		Test Length		
Sample Size	Index	10	40	80
500	$S - X^2$	56	25	19
	$Q_1 - X^2$	8	8	29
1,000	$S - X^2$	66	58	59
	$Q_1 - X^2$	9	16	47
2,000	$S - X^2$	66	90	94
	$Q_1 - X^2$	12	33	61

changes in sample size than did $Q_1 - X^2$. This is most likely due to the construction of the observed and expected cells for each index. For $Q_1 - X^2$, the ability distribution is divided into 10 equally spaced bins across the theta continuum, and observed and expected responses are calculated for each of those bins. In contrast, for $S - X^2$,observed and expected frequencies are calculated for each possible total score. For longer tests and small samples, this creates considerably more sparseness in the frequency tables for $S - X^2$relative to $Q_1 - X^2$ and necessitates a greater degree of collapsing to obtain a minimum expected cell frequency of 0.1. It is possible that the extent of collapsing may create some instability in $S - X^2$.

In addition, for the longest (80-item) tests, the fact that $Q_1 - X^2$ uses response-pattern scores whereas $S - X^2$ is based on summed scores likely accounts for some of the apparent advantage with a sample size of 500 (for all bad items) and 1000 (for Bad Items 1 and 3). With 80 items, there is a great deal of information in each item response pattern, and the likelihood that each examinee is placed in the correct bin for the calculation of $Q_1 - X^2$ is very high. In that situation, the fact that the summed scores used in the computation of the $S - X^2$ index provide less accurate characterization of θ likely places the $S - X^2$ index at a disadvantage when the sample size is relatively small, and so each relatively misplaced examinee has more effect.

This suggests the view that the relative performance of the two goodness-of-fit indices is best viewed on a continuum of test length: For short (10-item) tests, the advantage clearly goes to the $S - X^2$ index because there is little information about individual examinees, and their placement for the computation of $Q_1 - X^2$ in specific bins based on a point estimate of θ is likely flawed. The use of the modeled estimates for the summed scores in the context of the short test is not nearly so great a disadvantage because the computations reflect the lack of precision of the test. At the other end of the continuum, for relatively long tests (and modest sample sizes), the advantage goes to the $Q_1 - X^2$ index, as binning based on point estimates of θ becomes less inaccurate, and so the superiority of response-pattern estimation over summed score estimates is fully realized.

A few limitations should be considered when interpreting the results of this study. First, the parameters used to generate the misfitting items, although chosen carefully, yield only one possible expression of their nonlogistic pattern. Alternative choices for the misfitting item parameters would undoubtedly produce different results. Thus, this study provides an illustration of what may happen under certain conditions and can only be generalized cautiously. A second limitation is that no tests of medium length (e.g., 20 items) were included in this simulation, and inferences from this study cannot be made about the performance of the index for tests of that length. Finally, this simulation examined a specific type of misfit: the presence of one nonlogistic item in a set of logistic items. Results may not pertain to other misfit situations, such as fitting an entire test with the wrong model.

Overall, the results of this simulation lend additional support for the use of $S - X^2$ to detect misfit of items based on dichotomous IRT models, especially for shorter tests. With commonly used sample sizes for the calibration of 3PL items (1,000 or more), even for longer tests, the performance of $S - X^2$ is approximately equal to that of $Q_1 - X^2$, so routine use of the $S - X^2$ index can be recommended under many circumstances.

References

Abrahamowicz, M., & Ramsay, J. O. (1992). Multicategorical spline model for item response theory. *Psychometrika, 57*(1), 5-27.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model* (ETS RR- 81-20). Princeton, NJ: Educational Testing Service.

Bock, R. D., Wainer, H., Petersen, A., Thissen, D., Murray, J., & Roche, A. (1973). A parameterization for individual human growth curves. *Human Biology, 45*(1), 63-80.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Oxford, England: John Wiley.

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

Orlando, M. (1997). *Item fit in the context of item response theory*. Unpublished doctoral dissertation, University of North Carolina, Chapel Hill.

Orlando, M., & Thissen, D. (2000). New item fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50-64.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611-630.

Sadler, P. M., & Lieberman, M. (1994). *An item response theory approach to examining scientific misconceptions*. Unpublished manuscript.

Thissen, D. (1986a). *"Final preliminary" report on item parameter estimation for the 4PL model*. Unpublished manuscript.

Thissen, D. (1986b, April). *Non-monotonic item characteristic curves*. Invited presentation at the annual meeting of the American Educational Association, San Francisco.

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

Thissen, D., & Bock, R. D. (1986). *Parameter estimation for a four-parameter logistic trace line model*. Unpublished manuscript.

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.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*(4), 501-519.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12*(4), 339-368.

Author's Address

Maria Orlando, Ph.D., RAND, 1700 Main Street Box 2138, Santa Monica, CA 90407-2138; phone: (310) 393-0411, ext. 6604; fax: (310) 451-7062; e-mail: Maria_Orlando@rand.org.