A Comparison of Item- and Person-Fit Methods of Assessing Model-Data Fit in IRT

Steven P. Reise
University of Minnesota

Many item-fit statistics have been proposed for assessing whether the responses to test items aggregated across examinees conform to IRT test models. Conversely, person-fit statistics have been proposed for assessing whether an examinee's responses aggregated across items are congruent with a specified IRT model. Statistical procedures to assess item fit have differed from those to assess person fit. This research compared a χ^2 item-fit index with a likelihood-based person-fit index. Eight 0,1 data matrices were simulated under the three-parameter logistic test model. Both the likelihood-based and χ^2 fit statistics were then computed for examinees and items, and Type I and Type II error rates were analyzed. With data simulated to fit the IRT model, the χ^2 test overidentified examinees and items as being misfitting, while the likelihood-based fit index held closer to the specified α levels. The two fit indices gave consistent (mis)fit-to-model results in 94 and 97 percent of cases for items and examinees, respectively, across simulations. Under simulated conditions of data misfit, the χ^2 statistic detected misfit at a higher rate than the likelihood-based statistic, indicating that the χ^2 statistic was slightly more sensitive to response pattern aberrancy. However, other considerations led to a recommendation for employing the likelihood-based index in applied fit analyses to evaluate both examinee and item model-data (mis)fit. *Index terms: chi-square index, item fit, item response theory, model fit, person fit, response aberrancy.*

Model-data fit issues are a major concern when applying item response theory (IRT) models to real test data. Historically, the major research concern has been with item fit (e.g., Andersen, 1973; Yen, 1981). More recently, a sizable literature has developed around the issue of assessing person fit (e.g., Levine & Drasgow, 1982; Tatsuoka & Linn, 1983). Although item fit has taken precedence in the research literature, both types of fit analysis are important. Poor item fit indicates that the item parameters have questionable validity (i.e., they do not accurately represent how examinees respond to test items), while poor person fit indicates that the trait level (θ) estimate has questionable validity (i.e., the θ estimate may be a poor indicator of an examinee's position on the latent continuum).

A main purpose of this research was to demonstrate empirically that some methods used to assess item fit can be readily applied to assess person fit, and vice versa. To accomplish this, the performance of a χ^2 item-fit statistic was compared with that of a likelihood-based person-fit statistic when each was calculated for both examinees and items under monte carlo conditions.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 14, No. 2, June 1990, pp. 127–137
© Copyright 1990 Applied Psychological Measurement Inc.
0146-6216/90/020127-11$1.80

The Purpose and Calculation of Item Fit

IRT requires a formal investigation of the manner in which test items function as trait measures. Once a formal model has been defined and item response functions (IRFs) have been estimated, studies must be conducted to verify that the observed data conform to the estimated IRFs. Without item-fit analysis, the researcher cannot be confident whether the advantages accrued by specifying the formal IRT model can or will be realized.

Several key issues are involved in examining item fit:

1. Model selection has been an important motivation behind item-fit studies (Yen, 1981); it is important to identify the most efficient test model that retains the integrity of the observed data.

2. For most IRT models to be applicable, the item pool must produce item responses that approximately meet the unidimensionality assumption. Item-fit analysis has been proposed as a method of locating extraneous dimensions affecting the responses to test items (McKinley & Mills, 1985).
3. Item-fit studies may assist in identifying faulty item construction (e.g., incorrect item keying).
4. Item fit has the potential to indicate errors that occurred in the calibration phase of test development. For example, if a program consistently underestimates the discrimination parameter for highly discriminating items, an item-fit statistic should identify this problem.

Graphical methods (Kingston & Dorans, 1985) and statistical methods (McKinley & Mills, 1985) have been developed for assessing item fit. Most statistical methods rely on calculating a χ^2 statistic. To implement a χ^2 test, examinees are rank-ordered according to estimated θ and then grouped into some fixed or subjectively determined number of categories. Within each grouping, the proportion of examinees responding to the item in the keyed direction is calculated. This observed proportion is then compared to the predicted proportion based on the IRF. For example, the Bock (1972) χ^2_B is given by

$$\chi^2_{B_i} = \sum_{j=1}^{G} \frac{N_j (O_{ij} - E_{ij})^2}{E_{ij} (1 - E_{ij})} \quad , \qquad (1)$$

where i is the item,

j is the interval created by grouping examinees on the basis of their θ estimates,

G is the number of examinee groupings,

N_j is the number of examinees with θ estimates falling in a given interval j,

O_{ij} is the observed proportion of keyed responses on item i for interval j, and

E_{ij} is the expected proportion of keyed responses on item i for interval j based on the IRF evaluated at the median θ estimate within the interval.

High values of χ^2_B diagnose items that are performing differentially with respect to the IRT model (i.e., items that do not fit the model). The significance of χ^2_B is tested using the χ^2 distribution with degrees of freedom equal to the number of groupings (G) minus the number of estimated parameters.

The various χ^2 measures (e.g., Wright & Mead, 1977; Wright & Panchapakesan, 1969; Yen, 1981) are variations of χ^2_B. For example, the statistic implemented by Yen (1981), named Q_1, uses the mean θ within each category to obtain the predicted proportions, rather then the median. Furthermore, Yen's index specifies 10 categories, whereas no fixed number is specified with χ^2_B. Yen also compared several χ^2 fit indices using monte carlo data. An important result of her research was the determination that the Q_1 statistic was approximately distributed as χ^2 with the number of categories minus the number of parameters as the degrees of freedom. Because χ^2_B is similar to the Q_1 index, these important distributional results should transfer to the former statistic.

McKinley and Mills (1985) studied the performance of several χ^2 fit indices in monte carlo research. They found it difficult to choose between the competing methods. In other words, no particular χ^2 fit index was clearly superior to others. However, in their research χ^2_B yielded the smallest number of erroneous acceptances of fit under conditions of simulated model-data misfit.

The Purpose and Calculation of Person Fit

The purpose of a person-fit statistic is to identify examinees whose response patterns are incongruent with the specified response model. Several specific measurement disturbances have been associated with item response aberrancy. For example, Wright (1977) identified response patterns consistent with the hypotheses of "sleeping" and "guessing." Person-fit statistics can further be used to identify response patterns associated with specific educational deficits (Tatsuoka, 1982, 1984) or to extract deviant response patterns to create a more nearly unidimensional data matrix (Tatsuoka & Tatsuoka, 1982). Person-fit statistics also have potential value as response consistency/validity indices in personality measurement (Reise & Waller, 1990).

Of special interest in this research was a person-fit statistic originally proposed by Levine and Rubin (1979) and standardized by Drasgow, Levine, and Williams (1985). Consider the likelihood function that is to be maximized to estimate θ:

$$L|\theta = \sum_{k=l}^{K} \{U_k \, [\ln P_k(\theta)] + (1 - U_k) \times [\ln Q_k(\theta)]\} \quad , \tag{2}$$

where U_k is the 0,1 item response,
$P_k(\theta)$ is the probability of a correct response given θ,
$Q_k(\theta) = 1 - P_k(\theta)$, and
K is the number of items.

The L value represents the height of the likelihood function given the item responses and the item parameters; its first derivative is 0 at the maximum likelihood θ estimate. L is a simple and direct measure of how well a response pattern conforms to the model. High L indicates good fit, whereas relatively low L indicates poor fit. However, the magnitude of L will depend on θ level, and the range of L will depend on test length. These problems can be circumvented because for each cell in the 0,1 persons × items matrix, a standardization can be performed:

$$Z_3 = \frac{\mathrm{L}(\theta) - \mathrm{E}[\mathrm{L}(\theta)]}{\mathrm{SD}[\mathrm{L}(\theta)]} \quad , \tag{3}$$

where E(θ) is the conditional mean (i.e., expected value) and SD(θ) is the conditional standard deviation of the likelihood whose formulas are stated in Drasgow et al. (1985, p. 71). The Drasgow Z_3 index has heretofore been applied only to study person fit (called "appropriateness measurement" by Drasgow et al.). However, their statistical procedure is equally applicable to studying item-fit issues.

To compute Z_3 for items or examinees, the three terms shown in Equation 3 must be computed for each cell in the examinees × items response matrix. Then the three terms in Equation 3 must be summed either down a column (for an item) or across a row (for an examinee) of a data matrix. The Z_3 statistic should be distributed approximately N(0,1) when the data fit the model, and observed values of Z_3 should be independent of θ when calculated for examinees, and with item difficulty when calculated for items.

Negative values of Z_3 when calculated for examinees are associated with response inconsistency (i.e., response patterns that are unlikely, given the measurement model and the θ estimate). Positive Z_3 values indicate that the examinee's response pattern was more consistent than the probabilistic IRT model expected, and 0 is the expectation. For items, a negative Z_3 value indicates an unlikely item response vector, given the estimated item parameters. Positive values of Z_3 for items indicate response vectors that are more orderly than the probabilistic model and the item parameters predict.

Hence, an advantage of the Drasgow et al. (1985) fit statistic is that two types of misfit can be identified: (1) response vectors that are inconsistent with the model (negative Z_3 values), and (2) response vectors that are more consistent than the model predicts (positive Z_3 values). Evaluating misfit with the Z_3 index thus requires two-tailed tests of fit using the normal distribution. By contrast, the χ^2 test is a one-tailed fit statistic and does not make a distinction between under- and overfitting an IRT response model.

χ^2_B can also be used to assess the degree to which examinees respond according to the specified model. Using χ^2_B to examine person fit is similar to conducting a person response curve (PRC) analysis (Lumsden, 1978; Trabin & Weiss, 1979). A PRC is a function with item difficulty on the abscissa and the probability of item response on the ordinate. This function can be determined for each examinee and then compared to the expected PRC based on the IRT model.

If the model fits, the two curves should be close. One method to formally test the degree to which an observed PRC conforms to its expected shape is to compute χ^2_B by interchanging the roles of j and i in Equation 1. In χ^2_B for examinees, j is the interval created by grouping items on difficulty and i is the examinee.

If the observed PRC is different from the expected PRC, this discrepancy should translate into a large and significant χ^2. Wright (1977) and Trabin and Weiss (1979) proposed χ^2 procedures to study person fit.

Problems With Fit Indices

Two problems are characteristic of χ^2 fit measures. The first problem concerns the intervals: How should the intervals be created, and how many should be constructed? The resulting value of the χ^2 statistic will likely be dependent on the decisions made. Second, the number of observations has a strong influence on the χ^2 statistic. With sufficiently large sample sizes or numbers of items, it can be shown that χ^2 values will tend toward significance (indicating misfit) for items or examinees, respectively. Therefore, with χ^2, any IRT model will be rejected if enough data are collected.

Person-fit statistics, in general, are confounded by trait level (Harnisch & Linn, 1981, p. 140). Although conceptually it makes sense that examinees with very high or very low θ estimates have less response aberrancy potential than middle-range θ examinees, strong linear or nonlinear associations between person-fit statistics and θ estimates (or between item-fit statistics and item parameter estimates) vitiate the index. A second problem is that researchers seldom report the null distributions for their proposed person-fit indices. This leaves the user in a predicament when deciding how to interpret the statistic in practice. A final problem is the general nature of person-fit indices: A significant index does not indicate its cause. Note, however, that χ^2 fit indices provide no more or less diagnostic information.

The problems with the χ^2 statistic seem less correctable than those of person-fit statistics. Thus, if a person-fit index performs well for both items and examinees, the problems with χ^2 can be circumvented by eliminating its use. Before such a proposal can be made, empirical data must be generated comparing the two methods. Hence, χ^2_B was compared with Z_3 in data simulated to fit the three-parameter logistic IRT model.

Method

Simulated Datasets

Eight 50 item × 1,000 examinee 0,1 response matrices were generated under the three-parameter logistic model. The true θ generating parameters were distributed N(0,1). The true item generating parameters and the resulting internal consistency (coefficient α) statistics are displayed in Table 1. Internal consistency and model-data fit are distinct concepts, yet coefficient α is shown in Table 1 to gauge the quality of the simulated tests. Datasets 1 through 4 were created with the intention of simulating tests with guessing parameters expected from five-option multiple-choice items, a reasonable spread of item difficulty, and low to moderate internal consistency. Datasets 5 through 8 were generated to have higher internal consistency.

Table 1
Item Discrimination (*a*), Item Difficulty (*b*), Guessing Parameter (*c*), and Coefficient Alpha for the Eight Simulated Datasets

		b			
Dataset	*a*	Shape	Range	*c*	Alpha
1	.75	Uniform	−2.5, 2.5	.20	.75
2	1.00	Uniform	−2.5, 2.5	.20	.78
3	1.25	Uniform	−2.5, 2.5	.20	.83
4	1.50	Uniform	−2.5, 2.5	.20	.84
5	.75	Uniform	−1.5, 1.5	.10	.82
6	1.00	Uniform	−1.5, 1.5	.10	.87
7	1.25	Uniform	−1.5, 1.5	.10	.91
8	1.50	Uniform	−1.5, 1.5	.10	.92

Procedure

To compute χ^2_B for items, examinees were grouped into 10 θ intervals with 100 examinees per interval. Ten intervals were selected because this is a common number used in the item-fit literature (e.g., Yen, 1981). This resulted in 7 degrees of freedom. To compute χ^2_B for examinees, the item difficulties were grouped into five intervals, resulting in 10 observations (items) per interval. Five intervals were selected in order to have at least 10 observations per interval on which to base the observed proportions. Fewer intervals might have been used, but this would have resulted in a loss of degrees of freedom. With five intervals the statistic had 4 degrees of freedom.

The Z_3 index was calculated by adding the terms in Equation 3—down columns for items, or across rows for examinees. It was assumed that Z_3 would be distributed approximately as N(0,1) and thus the Z distribution is the appropriate null distribution. Previous research (Drasgow et al., 1985) demonstrated that Z_3 will be approximately normally distributed. All statistical tests of fit involving Z_3 were two-tailed in order to identify both types of misfit (i.e., under- and overfitting). The risk involved here in using the Z distribution was assumed to be no greater than the risk involved in assuming that χ^2_B is distributed as χ^2 when the model is true.

Four analyses were conducted:

1. For each simulated data matrix the proportions of examinees and items identified as not fitting (Type I errors) under several α levels were determined. The selected α levels were .001, .01, and .05.
2. The linear correlations between the fit indices and θ for examinees, and between the fit indices and the item difficulty for items, were calculated for each dataset. No strong correlations would be expected between these variables.
3. Correlations between χ^2_B values and the squared Z_3 values were computed for examinees and items within each data matrix. (Note that Z_3 must be squared before being compared to χ^2.) Agreement coefficients were also computed by counting the number of cases in which the Z_3 and χ^2_B fit indices led to the same fit-to-model decisions when both fit indices were evaluated at the same α level.
4. It was determined how well each index performed in identifying misfit when the test model was incorrect (Type II errors). To accomplish this, χ^2_B and Z_3 were computed for examinees and items in Datasets 1 through 3 using the true item parameters of Dataset 4. Furthermore, χ^2_B and Z_3 were computed for examinees and items in Datasets 5 through 7 using the true item parameters from Dataset 8. With such manipulations, the tests for fit were conducted with an incorrect model, because the parameters were incorrect. In this case, the tests of fit were conducted with true item discrimination parameters that were always higher than the item discrimination parameters used to generate the data. The extent of misfit imposed by these manipulations was one of degree; the more the item parameters used to test fit deviated from the data-generating parameters, the greater was the imposed misfit.

Although many types of misfitting data can be conceived and the performance of each index under different types of response aberrancy can be compared (Drasgow, 1982; Gafni, 1987), this was not the focus of the present research. Rather, the interest was in a general type of misfit, namely when the item discriminations used to test fit were larger than the item discriminations that generated the data. It is important to note that testing fit using incorrect item parameter values forces the examinees' response patterns to be erroneous as well. If a response pattern was determined to fit the model when fit was tested under a correct and incorrect IRT model, then the fit indices would not be considered very sensitive to the values of the item discrimination parameters used to test fit.

Results

Aberrancy Detection in Data Simulated to Fit the Model

The proportions of examinees identified by the

Table 2
Proportion of Examinees and Items Identified as Aberrant Across the Eight Datasets at Critical Values (CV) for Specified Type I Error Rates, for Z_3 and χ^2_B Statistics

	Z_3			χ^2_B		
Examinees						
CV	±3.29	±2.57	±1.96	18.46	13.27	9.48
Error rate	.001	.01	.05	.001	.01	.05
Dataset						
1	.003	.015	.039	.005	.017	.081
2	.002	.012	.041	.005	.027	.108
3	.002	.012	.042	.014	.028	.096
4	.004	.015	.053	.010	.041	.109
5	.000	.009	.054	.003	.016	.090
6	.003	.016	.047	.007	.029	.091
7	.000	.007	.049	.002	.018	.096
8	.001	.006	.036	.007	.025	.094
Total	.001	.011	.045	.006	.025	.095
Items						
CV	±3.29	±2.57	±1.96	24.32	18.47	14.07
Error rate	.001	.01	.05	.001	.01	.05
Dataset						
1	.000	.020	.060	.000	.060	.200
2	.000	.000	.020	.000	.060	.160
3	.000	.000	.040	.000	.080	.220
4	.000	.000	.040	.000	.060	.200
5	.000	.000	.080	.060	.100	.280
6	.000	.020	.040	.000	.020	.200
7	.000	.000	.040	.000	.040	.200
8	.000	.000	.060	.000	.060	.220
Total	.000	.005	.047	.007	.060	.210

fit indices as misfitting are displayed in the top portion of Table 2. The total proportions across the eight datasets reveal that for examinees, Z_3 rejected at a rate closer to the specified Z_α than χ^2_B. Z_3 rejected at .001, .011, and .045 for α levels .001, .01, and .05. In contrast, χ^2_B rejected at .006, .025, and .095 for the same three α levels. To a small degree, then, χ^2_B overidentified examinees as misfitting under conditions of data simulated to fit the model.

Table 2 also displays the proportions of items identified as aberrant by each index under conditions of data simulated to fit the model. Similar to the results for examinees, χ^2_B rejected at a higher rate than Z_3. Specifically, χ^2_B rejected at a .06 rate when $\alpha = .01$, and at a .21 rate when $\alpha = .05$. Z_3 appeared to reject at close to the specified α level for items. The χ^2_B rejection rates were in agreement with the findings of McKinley and Mills (1985), who reported that χ^2_B, when computed to assess item fit, rejected at .05, .04, and .07 when fit was evaluated at $\alpha = .01$ under simulations with 1,000 examinees.

Neither index appeared to be confounded by trait level or item difficulty, and the two fit indices were only moderately linearly correlated. In terms of the (mis)fit indications, the two indices behaved similarly for examinees and items in most instances.

Relations Between the Indices

Several correlations of interest are displayed in Table 3. The first four columns in Table 3 show the linear correlations between the fit indices and the true parameters. The first two columns indicate that there was essentially no

Table 3
Correlations Between Z_3, χ^2_B, θ, and b,
and Percentage of Fit Decision Agreements

	Correlations							
					Z_3^2 vs. χ^2_B		Percent Agreement	
Data-	θ		b		Exam-		Exam-	
set	Z_3	χ^2_B	Z_3	χ^2_B	inees	Items	inees	Items
1	.03	.00	.15	-.08	.49	.34	98	92
2	-.02	-.04	.00	-.05	.50	.22	98	94
3	.00	.04	.22	-.03	.48	.17	97	92
4	-.03	.00	.24	-.02	.41	.33	96	94
5	-.02	.02	.13	.03	.42	.11	98	90
6	.04	.00	-.19	.03	.50	.05	97	96
7	-.01	.00	.25	.05	.39	.17	98	96
8	.05	.00	-.06	.07	.36	.23	97	94
Total							97	94

correlation between examinee fit and θ for either index. Also, in linear terms, there were no consistent positive or negative trends in the correlational results to warrant rejecting the hypothesis that item fit was uncorrelated with item difficulty. Table 3 further indicates that the two indices were not highly linearly related when calculated for either examinees or items. In a linear sense, then, it appeared that the two indices were functioning differentially.

As indicated in Table 3, across datasets for examinees, the two indices agreed about (mis)fit in 7,806 cases (97%) and disagreed in 194 cases (3%) when both indices were evaluated at $\alpha = .01$. Clearly, the fit decisions made, in a yes/no sense, were quite comparable between the two fit indices. For items across the eight datasets, the two indices made the same (mis)fit decision 374 times (94%) and were incongruent only 26 times (5%) when fit was evaluated at $\alpha = .01$ for both indices.

Figures 1a and 1b display the relation between θ and Z_3, and between θ and χ^2_B, respectively, for Dataset 1 (comparable results were obtained for the other datasets). No strong relationships are evident in these plots. For both fit indices the variance of fit values seems to decrease at the θ extremes. Figure 1c displays the relation between squared Z_3 values and χ^2_B values for examinees in Dataset 1. Figures 2a and 2b demonstrate that neither Z_3 nor χ^2_B values are related to item difficulty in Dataset 1. Figure 2c illustrates the relation between squared Z_3 values and χ^2_B values in the same dataset.

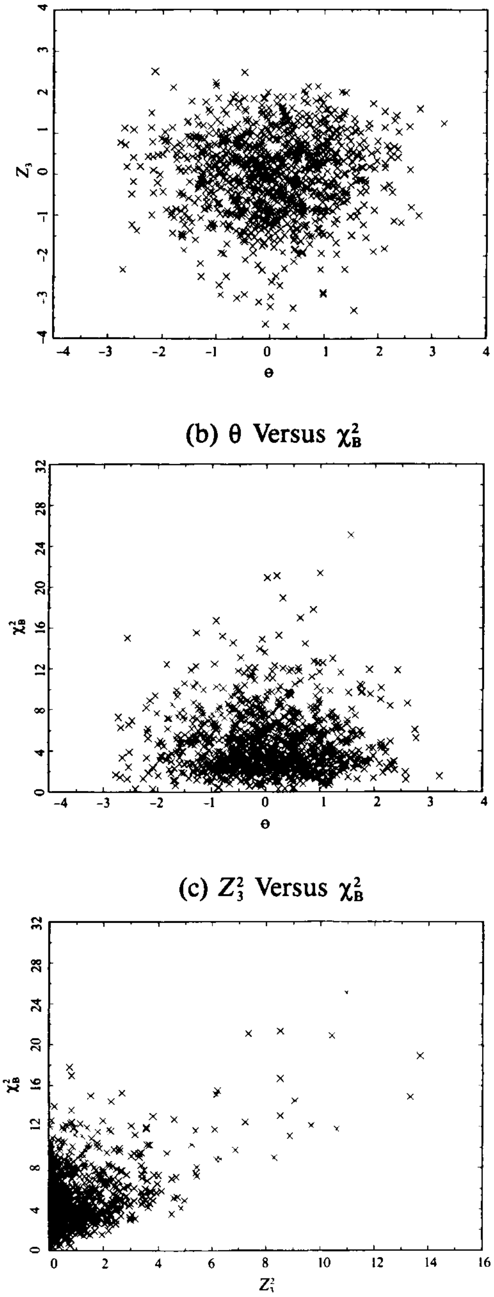
Identifying Simulated Misfit

Table 4 lists the proportions identified as misfitting for examinees and items, respectively, when the Dataset 4 item parameters were used in place of the true generating item parameters in Datasets 1 through 3, and when the Dataset 8 item parameters were used to test fit in Datasets 5 through 7. These results show that the proportion of items or examinees identified as misfitting depended on the degree that the generating item parameters deviated from the Dataset 4 or 8 item parameters. Overall, the χ^2_B fit index identified more items and examinees as misfitting than did the Z_3 fit index.

Using either fit index, item misfit was easier to detect than examinee misfit. This may be due to the type of misfit imposed on the data, and to the fact that there were 20 times (50 vs. 1,000) the number of observations when calculating item fit as opposed to person fit. The power to detect misfit should improve with an increasing number of observations.

The results in Table 4 clarify two important

Figure 1
Scatterplots of Fit Indexes for Examinees in Dataset 1

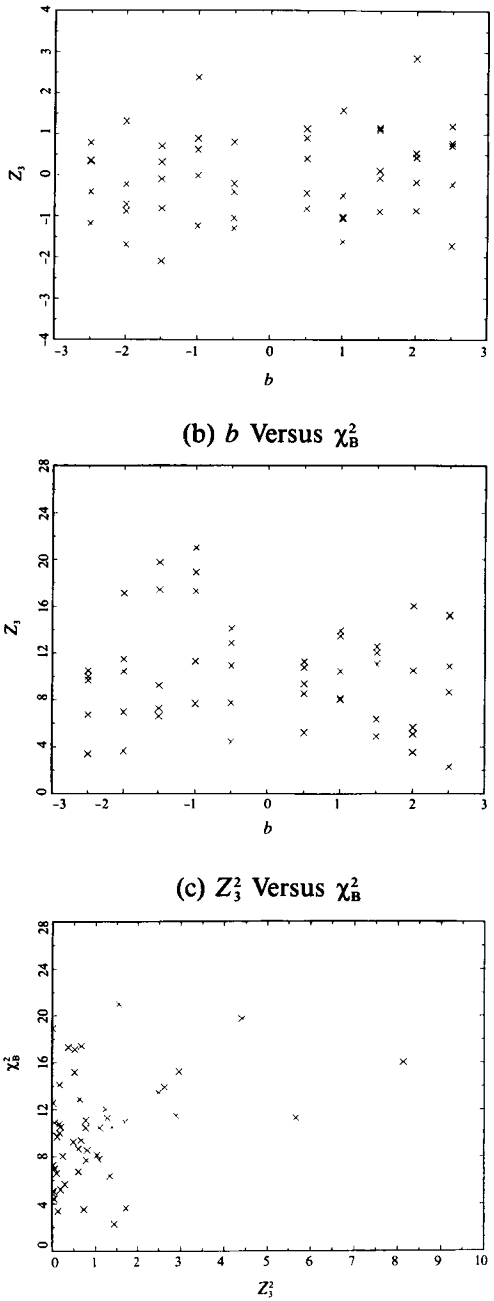


(a) θ Versus Z_3

(b) θ Versus χ^2_B

(c) Z_3^2 Versus χ^2_B

Figure 2
Scatterplots of Fit Indexes for Items in Dataset 1



(a) *b* Versus Z_3

(b) *b* Versus χ^2_B

(c) Z_3^2 Versus χ^2_B

Table 4
Proportion of Examinees and Items Identified as Misfitting When Incorrect Item Parameters Were Used to Test Fit

	Z_3			χ^2_{B}		
Examinees						
CV	±3.29	±2.57	±1.96	18.47	13.27	9.48
Error rate	.001	.01	.05	.001	.01	.05
Dataset						
1	.260	.394	.548	.295	.429	.575
2	.073	.164	.276	.135	.236	.361
3	.019	.045	.104	.042	.100	.188
5	.207	.343	.496	.200	.331	.521
6	.065	.150	.281	.082	.177	.314
7	.012	.036	.098	.025	.065	.161
Items						
CV	±3.29	±2.57	±1.96	24.32	18.47	14.07
Error rate	.001	.01	.05	.001	.01	.05
Dataset						
1	.980	1.000	1.000	.980	1.000	1.000
2	.800	.920	.940	.680	.840	.920
3	.180	.280	.380	.220	.400	.620
5	1.000	1.000	1.000	1.000	1.000	1.000
6	.960	.980	1.000	.940	.980	.980
7	.140	.400	.580	.180	.360	.600

issues. First, neither index is sensitive enough to identify small deviations in item discrimination accuracy. Second, when the item discrimination parameters were incorrect, χ^2_{B} was more sensitive to the misfit that resulted.

Discussion

The data did not clearly support the use of Z_3 over $\chi^2{}_{\mathrm{B}}$ or vice versa. Z_3 held closer to the specified α levels than did χ^2_{B}. On the other hand, χ^2_{B} was more sensitive to misfit under the present conditions. However, it appeared that χ^2_{B} was biased toward rejecting fit. No strong linear relations emerged between the two indices, but the (mis)fit indications they gave, in terms of the percentage of fit decision agreements with respect to examinees and items, were quite comparable. In fact, there were few cases in which one index would indicate misfit and the other would not when the data were simulated to fit the model.

On computational grounds, the Z_3 index is more "efficient" than the χ^2_{B} index. Also, Z_3 allows for two types of misfitting to be differentiated: (1) response vectors that are less consistent than the model predicts, and (2) response vectors that are too consistent with respect to the specified model. However, researchers have not been overly concerned with this latter type of response aberrancy. χ^2_{B} requires several critical computation decisions that can affect the results. Furthermore, with χ^2_{B}, sample sizes much larger than 1,000 will tend to inflate the test statistic and indicate every item to be misfitting. If used for examinees, long tests will tend to inflate the χ^2 statistic. Consequently, it might be appropriate to seek alternative methods to χ^2—both graphical (e.g., Kingston & Dorans) and statistical (e.g., Drasgow et al., 1985)—to assess fit. The data presented here support the use of Z_3 to assess fit for items as well as for examinees.

The results demonstrated that Z_3 was not confounded with b or θ for items and examinees, respectively. The relation between fit values and the other item parameters could not be calculated because discrimination and guessing were held constant within tests. On the other hand, the same positive results were found with χ^2_{B}. However, other researchers have reported relations between χ^2 fit and the item parameters. For example, Koch (1983) noted a tendency for large χ^2 to result for highly discriminating items.

There are several qualifications pertaining to the results presented here. First, true parameters were used to investigate detection rates. The consequences of using estimated parameters were not addressed. Furthermore, only one type of response aberrancy was addressed in this research, namely the special case when all items in the test are less discriminating for the examinee than for the calibration sample.

It was interesting to observe that the fit indices used in this study appeared to have little difficulty identifying items as misfitting when the model was grossly inappropriate (e.g., testing fit in Dataset 1 using the Dataset 4 item parameters), but were problematic when identifying examinees in the same manner. In many cases, examinees were identified as fitting the model (i.e., having an appropriate response pattern) when fit was computed using the incorrect item discrimination parameter. This result should be considered in applied work.

References

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

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

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement, 6,* 297-308.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86.

Gafni, N. (1987). *Detection of systematic and unsystematic aberrance as a function of ability estimate by several person-fit indices.* Unpublished doctoral dissertation, University of Minnesota.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18,* 133-146.

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

Koch, W. R. (1983). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7,* 15-32.

Levine, M. V., & Drasgow, F. (1982). Appropriateness measurement: Review, critique, and validating studies. *British Journal of Mathematical and Statistical Psychology, 35,* 42-56.

Levine, M. V., & Rubin, D. F. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269-290.

Lumsden, J. (1978). Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology, 31,* 19-26.

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

Reise, S. P., & Waller, N. G. (1990). *Applications of response pattern aberrancy analysis to personality assessment.* Unpublished manuscript.

Tatsuoka, K. K. (1982). A latent trait model for interpreting misconceptions in procedural domains. In D. J. Weiss (Ed.), *Proceedings of the 1982 Item Response Theory and Computerized Adaptive Testing Conference* (pp. 322-339). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95-110.

Tatsuoka, K. K., & Linn, R. L. (1983). Indices for detecting unusual patterns: Links between two general approaches and potential applications. *Applied Psychological Measurement, 7,* 81-96.

Tatsuoka, K. K., & Tatsuoka, M. M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7,* 215-231.

Trabin, T. E., & Weiss, D. J. (1979). *The person response curve: Fit of individuals to item characteristic curve models* (Research Report 79-7). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97-116.

Wright, B. D., & Mead, R. J. (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. 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.

Author's Address

Send requests for reprints or further information to Steven P. Reise, Department of Psychology, University of California, Riverside CA 92521, U.S.A.