A Comparison of Several Goodness-of-Fit Statistics

Robert L. McKinley
The University of Toledo

Craig N. Mills
Educational Testing Service

A study was conducted to evaluate four goodness-of-fit procedures using data simulation techniques. The procedures were evaluated using data generated according to three different item response theory models and a factor analytic model. Three different distributions of ability were used, as were three different sample sizes. It was concluded that the likelihood ratio chi-square procedure yielded the fewest erroneous rejections of the hypothesis of fit, whereas Bock's chi-square procedure yielded the fewest erroneous acceptances of fit. It was found that sample sizes somewhere between 500 and 1,000 were best. Shifts in the mean of the ability distribution were found to cause minor fluctuations, but they did not appear to be a major issue.

Item response theory (IRT) is becoming a widely used psychometric tool, with applications ranging from item banking to equating to adaptive testing. IRT models offer many advantages over more traditional test analysis procedures. However, these advantages are gained at the expense of making rather strong assumptions about the nature of the data. It is widely recognized that these assumptions are unlikely to be fully met in practice. Although in some respects IRT models appear to be robust with respect to the violation of these assumptions, it is clear that in many instances the violation of these assumptions has profound implications for the application of IRT methodology.

Because of the strong assumptions required for the use of IRT and the fact that the advantages associated with the use of IRT may not be fully realized if these assumptions are not met, it is important that prospective users of IRT methodology assess the appropriateness of IRT for use in intended applications. One way in which this can be done is by conducting a goodness-of-fit study. Broadly defined, a goodness-of-fit study is the evaluation of the similarity between observed and expected (predicted) outcomes. Within the context of IRT, this typically involves (1) estimating the parameters of an IRT model, (2) using those parameter estimates to predict, by way of the IRT model, examinee response patterns, and (3) comparing the predicted response patterns to actual observed examinee response patterns.

A number of procedures have been proposed in the literature for assessing the goodness of fit of IRT models to data. Unfortunately, there is little information available to assist in the selection or evaluation of such procedures. Data are not generally available regarding the performance of the various procedures under different conditions, nor are criteria available for selecting among the competing alternative goodness-of-fit procedures.

The purpose of this research was to investigate a number of goodness-of-fit procedures to assess

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 1, March 1985, pp. 49–57

0146-6216/85/010049-09$1.70

their adequacy for assessing the degree to which the more popular IRT models fit data. This was accomplished by generating simulated test data with known properties. The parameters of the three most popular IRT models—the one-parameter logistic, the two-parameter logistic, and the three-parameter logistic models—were estimated, and several goodness-of-fit procedures were applied to the results. The accuracy with which the procedures identified known fit and misfit were then evaluated.

Goodness-of-Fit Statistics

The focus of this research was on the goodness of fit of specific IRT models, as opposed to specific aspects of fit, such as local independence. Although there have been some procedures proposed for assessing local independence (e.g., Rosenbaum, 1984), these procedures have largely focused on assessing whether any IRT model is appropriate, rather than on whether a specific model fits the data. Note, for instance, that the Rosenbaum procedure not only fails to utilize IRT model parameters, but the procedure precludes the use of model parameters. Therefore, these procedures were not included in this study. Rather, only statistics that could be used to evaluate the fit of a specific model, or to compare the fit of two models, were included.

After a review of the literature, four goodness-of-fit procedures were selected for this research. All of the procedures selected lend themselves to chi-square analyses, and allow the statistical testing of fit for individual items and for a test as a whole.

Bock's Chi-Square

Bock's chi-square (BCHI) procedure (Bock, 1972) involves computing a chi-square statistic for each item in the following manner. First, the ability scale is divided into J intervals so that roughly equal numbers of examinees are placed into each interval. Each examinee is then assigned to one of the $2 \times J$ cells on the basis of the examinee's ability estimate and whether the examinee answered the item of interest correctly or incorrectly. For each interval the observed and predicted proportion-correct and proportion-incorrect scores are computed and used to form a chi-square statistic. The predicted value for an interval for a given item (E_{ij}) is computed using the median of the ability estimates falling within the interval and the item parameter estimates for that item using the IRT model. The BCHI statistic for item i is given by

$$\text{BCHI}_i = \sum_{j=1}^{J} \frac{N_j (O_{ij} - E_{ij})^2}{E_{ij} (1 - E_{ij})} , \quad (1)$$

where O_{ij} is the observed proportion-correct on item i for interval j, and
N_j is the number of examinees with ability estimates falling within interval j.

To test the significance of an item's lack of fit, $J - m$ degrees of freedom are used, where m is the number of item parameters estimated.

Yen's Chi-Square

Yen's chi-square (YCHI) procedure (Yen, 1981) is the same as the BCHI procedure with two exceptions. First, the YCHI procedure uses 10 intervals, whereas the BCHI procedure does not specify a specific number of intervals. Second, the predicted score E_{ij} is computed as the mean of the predicted probabilities of a correct response for the examinees within an interval. The YCHI statistic is given by Equation 1 with $J = 10$. The degrees of freedom are $10 - m$.

Wright and Mead Chi-Square

The Wright and Mead chi-square (WCHI) procedure was first proposed by Wright and Mead (1977). It is identical to the YCHI procedure with three exceptions. First, the procedure is based on number-correct score groups (i.e., the one-parameter logistic model) rather than on intervals of the ability scale. Second, rather than using 10 intervals, the WCHI procedure requires that six or fewer score groups be used. This is accomplished by collapsing adjacent number-correct score groups until there are six or fewer groups, while maintaining a roughly uniform number of examinees across groups. Third, the chi-square statistic that is computed is modified to correct for the theoretical variance of

the predicted probabilities of a correct response within a score group (due to examinees of different abilities being in the same interval).

To use this procedure with IRT models other than the one-parameter logistic model, Yen (1981) modified it by substituting the grouping method of the YCHI procedure for the number-correct grouping approach. This resulted in within-group ability estimate variances similar in magnitude to those produced by the number-correct score group method using the one-parameter data. The WCHI statistic is given by

$$\text{WCHI}_i = \sum_{j=1}^{J} \frac{N_j (O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij}) - s_{p_j}^2} \quad , \tag{2}$$

where

$$s_{p_j}^2 = \frac{1}{N_j} \sum_{k \epsilon j}^{N_j} (P_i(\theta_k) - E_{ij})^2 \quad . \tag{3}$$

$P_i(\theta_j)$ is the predicted proportion correctly answering item i in score group k, and the other terms are as previously defined. The degrees of freedom are $J - m$.

Likelihood Ratio Chi-Square

The likelihood ratio chi-square (LCHI) procedure follows much the same pattern as the YCHI procedure. The ability scale is divided into 10 intervals in such a way as to result in roughly equal numbers of examinees having ability estimates falling within the intervals. The examinees are sorted into 1 of 20 cells based on their ability estimates and their item responses. A 10×2 contingency table is formed, and a likelihood ratio chi-square statistic (Bishop, Fienberg, & Holland, 1975) is computed.

The LCHI statistic is given by

$$\text{LCHI}_i = 2 \sum_{j=1}^{20} O_{ij} \ln \left(\frac{O_{ij}}{E_{ij}} \right) \quad , \tag{4}$$

where $\ln(x)$ is the logarithm to the base e of x, and the remaining terms are as previously defined. The degrees of freedom are $10 - m$.

Method

The Simulation of Test Data

In all, 36 tests were simulated, each composed of 75 items. Nine tests were simulated to fit each of four models: (1) the one-parameter logistic (1PL) model, (2) the two-parameter logistic (2PL) model, (3) the three-parameter logistic (3PL) model, and, (4) a two-factor linear (LIN) model. The nine tests simulated for each model were composed of three tests at each of three sample sizes—500, 1,000, and 2,000 cases. The three tests with a given sample size varied on the mean ability of the simulated examinees. There was a low ability group (ability centered about one standard deviation below the mean item difficulty), a centered ability group (ability centered at the mean of the item difficulties), and a high ability group (ability centered about one standard deviation above the mean item difficulty).

The item parameters used to simulate the one-, two-, and three-parameter data were selected as follows. All of the item parameters were selected from uniform distributions having the ranges shown in Table 1. The same b values were used for all datasets. The same a values were used for all two- and three-parameter datasets. For all one-parameter datasets a value of 1.0 was used for all a values. For all one- and two-parameter datasets a value of 0.0 was used for all c values.

The θ parameters used for the one-, two-, and three-parameter data were selected as follows. All θs were randomly selected from a standard normal distribution. First, 500 θs were selected and used for the $N = 500$ datasets. For the $N = 1{,}000$ datasets an additional 500 θs were selected and combined with the 500 θs previously selected. Likewise, for the $N = 2{,}000$ datasets an additional 1,000 θs were selected and combined with those already selected. Note that the low ability groups were simulated by subtracting 1.0 from all of the θs, whereas the high ability groups were simulated by adding 1.0 to all of the θs.

The LIN data were generated using the procedure described by Wherry, Naylor, Wherry, and Fallis (1965), which is based on the linear factor analysis model. The procedure forms a multidimensional variable as a weighted sum of independent, normally distributed random variables, and then dichotomizes the variable to give the desired proportion correct. Groups with different mean abilities were simulated by shifting the mean of the target proportion-correct scores of the items. The

Table 1
Ranges of Item Parameter Distributions

Model and Limit	Parameter a	Parameter b	Parameter c
1PL			
Lower	1.00	-3.00	0.00
Upper	1.00	3.00	0.00
2PL			
Lower	0.20	-3.00	0.00
Upper	2.00	3.00	0.00
3PL			
Lower	0.20	-3.00	0.10
Upper	2.00	3.00	0.25

target mean total test proportion-correct scores for the three ability groups were $p = .375$, $p = .500$, and $p = .625$. Items 1–37 had factor loadings of .70 on the first factor and .20 on the second factor, whereas items 38–75 had loadings of .20 on the first factor and .70 on the second factor.

Calibration

The data for all conditions were calibrated for the one-, two-, and three-parameter models using LOGIST (Wingersky, Barton, & Lord, 1982). For the one- and two-parameter models, all c values were held constant at 0.0. For the one-parameter model the a values were held constant at .588 (1.0/1.7). However, this value was modified due to the rescaling, and as a result, different values of a were obtained for each dataset.

Analyses

The four goodness-of-fit procedures were applied to each of the simulation datasets. The results were then inspected to determine whether the procedures performed satisfactorily. That is, it was determined whether the procedures could be used to discriminate cases of fit (such as the three-parameter calibration of the one-parameter data) from cases of misfit (such as the one-parameter calibration of multidimensional data).

Two types of errors were investigated in this study. The first type of error involved the erroneous conclusion that the model did not fit the data when, in fact, the data were generated with that model or a model subsuming the calibration model (e.g., the two-parameter calibration of one-parameter data). The second type of error was the erroneous conclusion of fit when the calibration model did not subsume the generation model (e.g., the one-parameter calibration of three-parameter data).

Results

Table 2 reports summaries of the results obtained for the goodness-of-fit procedures for the one-, two-, and three-parameter data. Table 3 contains a summary of the results obtained for the multidimensional data. The values reported in the tables are the proportion of items for which there was significant misfit of the model to the data. Significance levels of .01 and .05 were used for testing the significance of the chi-squares for individual items. Because the two sets of analyses yielded quite similar patterns of results, only the results for the .01 analyses are reported. Under the hypothesis of fit, the proportion of items for which there was misfit should have been approximately .01.

Table 2 summarizes the results for the one-parameter data. Since these data were generated to fit the one-parameter model, it would be expected

Table 2

Proportion of Items Identified as Misfitting 1PL, 2PL, and 3PL Models for Four Generation Models, Three Ability Distributions, Three Sample Sizes, and Four Fit Statistics

Generating Model, Sample Size, and Fit Statistic	LO			CE			HI		
	1PL	2PL	3PL	1PL	2PL	3PL	1PL	2PL	3PL
One-parameter									
N=500									
BCHI	00	01	03	07	07	04	00	03	00
WCHI	00	00	03	05	04	04	03	01	00
LCHI	00	00	00	01	00	00	00	00	00
YCHI	00	01	00	04	07	04	01	03	00
N=1000									
BCHI	00	01	07	07	05	04	03	03	01
WCHI	00	00	05	04	03	04	01	01	01
LCHI	00	00	04	03	01	03	00	00	00
YCHI	00	01	07	05	04	04	03	03	01
N=2000									
BCHI	03	03	07	08	08	12	03	03	03
WCHI	01	01	08	07	08	09	03	04	04
LCHI	01	00	08	00	00	04	00	00	00
YCHI	04	04	05	08	08	09	03	03	03
Two-parameter									
N=500									
BCHI	36	01	11	37	03	04	42	05	01
WCHI	28	04	08	36	04	04	42	05	04
LCHI	35	00	07	40	00	01	47	00	00
YCHI	34	01	03	37	03	04	40	04	01
N=1000									
BCHI	47	03	05	51	01	03	65	03	03
WCHI	45	01	07	53	03	08	64	01	05
LCHI	47	01	07	51	00	05	69	01	01
YCHI	45	01	05	49	01	03	63	03	03
N=2000									
BCHI	65	03	07	65	05	12	77	04	03
WCHI	61	01	13	72	07	09	77	05	03
LCHI	64	01	07	65	00	08	75	01	01
YCHI	61	01	07	63	05	11	77	03	01
Three-parameter									
N=500									
BCHI	43	12	04	53	12	01	55	11	03
WCHI	53	09	03	57	11	03	47	11	01
LCHI	43	09	00	53	03	00	56	10	00
YCHI	44	15	03	53	12	01	53	12	03
N=1000									
BCHI	65	09	05	67	17	04	73	15	07
WCHI	67	15	05	75	19	01	72	16	01
LCHI	69	12	03	68	07	01	72	19	01
YCHI	65	09	04	68	17	04	70	19	04
N=2000									
BCHI	84	20	05	88	33	05	83	39	04
WCHI	85	29	05	89	28	04	83	29	01
LCHI	85	13	00	85	20	01	80	28	00
YCHI	85	20	04	89	33	05	81	37	03

Note. Decimal points omitted.

that all three calibration models would yield fit. As can be seen from Table 2, however, misfit was shown by all of the chi-square procedures. The most misfit was shown for the centered ability distribution and, to some extent, for the largest sample size. It seems clear from an examination of Table 2 that the values are consistently lower for the LCHI procedure than for the other procedures, though the level of significance of the differences is unclear.

Table 2 also summarizes the results obtained for the chi-square procedures for the two-parameter data. For these data, fit was expected for the two- and three-parameter models, but not for the one-parameter model. As can be seen from Table 2, all four procedures showed clear differences between the one-parameter calibrations and the two- and three-parameter calibrations. There is some lack of fit for the two- and three-parameter models, especially for the three-parameter model, but the proportions of items for which there was misfit are dramatically less than for the one-parameter model, regardless of which procedure is considered.

In the cases when fit was expected, the LCHI procedure once again showed consistently lower values than the other procedures. In the cases when misfit was expected, the LCHI procedure performed as well or better than the other procedures for the 500 sample size case, whereas it performed about as well as the others for the larger sample size cases.

The results obtained for the three-parameter data are also shown in Table 2. For these data, only the three-parameter calibration model was expected to yield fit. It was expected that the fit for the two-parameter model would be worse than for the three-parameter model, but not as bad as for the one-parameter model. This is the pattern obtained for all four procedures.

There was some misfit for the three-parameter model, but at relatively low levels. The least misfit was indicated by the LCHI procedure. The LCHI procedure also tended to show less misfit for the two-parameter model than did the other procedures. There were no clear patterns for the one-parameter calibrations.

Table 3
Proportions of Items Generated by the Multidimensional Model Identified as Misfitting 1PL, 2PL, and 3PL Models for Three Ability Levels, Three Sample Sizes, and Four Fit Statistics

Sample Size and Fit Statistic	LO			CE			HI		
	1PL	2PL	3PL	1PL	2PL	3PL	1PL	2PL	3PL
N=500									
BCHI	92	86	91	25	29	35	72	73	83
WCHI	70	63	76	16	13	25	77	61	59
LCHI	93	85	89	21	19	32	76	71	77
YCHI	92	81	84	27	27	36	72	69	79
N=1000									
BCHI	100	97	98	68	63	76	95	87	92
WCHI	85	82	86	53	45	55	93	77	65
LCHI	99	96	95	73	67	76	93	87	87
YCHI	97	93	92	68	60	68	93	85	87
N=2000									
BCHI	100	99	100	92	95	99	99	95	99
WCHI	92	93	82	75	77	92	97	87	82
LCHI	99	97	97	97	97	100	96	95	96
YCHI	100	96	97	92	93	99	99	95	99

Note. Decimal points omitted.

Table 3 shows a summary of the results obtained for the chi-square procedures for the multidimensional data. For these data, misfit was expected for all three calibration models. This was the obtained pattern, though the level of misfit (proportions of items for which there was misfit) was surprisingly low for the 500 and 1,000 sample size cases for the centered ability distribution. This result was fairly consistent across the four procedures. The only consistent difference among the fit procedures for these data was the tendency of the WCHI procedure to indicate less misfit than the other procedures, especially for the two- and three-parameter calibration models. Nonetheless, in no case would these results be interpreted as indicating that the unidimensional models yield adequate fit.

Discussion and Conclusions

Before discussing the results of this study, it will be helpful if a discussion of goodness of fit as an issue is presented. This will provide a rationale for the selection of the statistics for this study and a context for the discussion of the results.

Goodness of fit is generally recognized as an important aspect of any model-based psychological measurement. Certainly, if goodness of fit of a model to data is not established, the validity of the use of the model comes into question.

Within the context of IRT, goodness of fit is a crucial ingredient in the establishment of the appropriateness of any particular model, and of the use of IRT methodology itself. Despite claims for the robustness of IRT models, it is clear that misfit can seriously detract from the validity of IRT-based measurements. Robustness cannot simply be assumed. Rather, it must be established in each new application.

Lack of goodness of fit can occur for several reasons. For example, the assumptions of the model may not be met in the data. IRT, like any other model-based methodology, involves the use of some rather strong assumptions. Most commonly used models assume that the complete latent space is unidimensional, and that local independence therefore holds when unidimensional IRT models are employed. If the data are multidimensional, which is often the case, local independence may not hold.

IRT models also involve strong assumptions about the shape of the function relating performance to latent ability. If the assumed shape of the item response function (or item characteristic curve) is incorrect, such as is the case in which the one-parameter model is used when guessing is a factor in responses, misfit results.

Even if the assumptions of the model are met, misfit can result from inadequacies in the estimation process. This can be the result of sample sizes that are too small, poor estimation algorithms, or a variety of other problems.

Different procedures for assessing goodness of fit tend to focus on different aspects of misfit. Some procedures, for instance, are designed for detecting violations of local independence and monotonicity assumptions (e.g., Holland, 1981; Rosenbaum, 1984). Others are designed for and limited to particular models or estimation algorithms (e.g., Andersen, 1973; Bock & Aitken, 1981; Wright & Mead, 1977; Wright & Stone, 1979). In this study, only statistics that could be applied with any IRT model or estimation algorithm were employed.

For the one-parameter data of this study, all three models were expected to fit the data. That is, the proportions of misfitting items were expected to all be approximately .01 for the three models. As can be seen in Table 2, to a great extent this was the observed outcome. However, there were a number of instances in which the proportions deviated from the .01 region. Overall, it appeared as though the LCHI procedure yielded values consistently closer to .01 than were the values for the other procedures. This would seem to indicate that when the null hypothesis (fit) is true, the LCHI procedure is least likely to result in erroneous rejections of the null hypothesis. This seemed to be true regardless of sample size, distribution of ability, or calibration model.

For the two-parameter data, lack of fit was expected for the one-parameter model calibrations, but not for the two- and three-parameter calibrations. Again, this was the general observed outcome, though there were a few cases when pro-

portions diverged somewhat from the expected value of .01. As was the case for the one-parameter data, the LCHI procedure appeared to result in the fewest erroneous rejections of the hypothesis that the model fit the data. In terms of correct rejections of the hypothesis of fit, the LCHI and BCHI procedures appeared to yield marginally higher proportions than the other two procedures. However, it would be very difficult, on the basis of these data, to select one fit procedure over the others.

For the three-parameter data, misfit was expected for both the one- and two-parameter models, whereas fit was expected for the three-parameter model. The fit of the two-parameter model was expected to be somewhat better than the fit of the one-parameter model. Once again, the observed outcome closely paralleled the expected outcome. As was the case with the one- and two-parameter data, the LCHI procedure yielded the fewest erroneous rejections of fit when the three-parameter data were calibrated using the three-parameter model. In terms of correct rejections of fit, it would be very difficult to identify one procedure as performing better than the others.

All three models were expected to yield lack of fit to the multidimensional data, and that was, in fact, the observed outcome. Although there was no clear pattern in these data, it did appear as though the BCHI procedure might have yielded, on the average, slightly higher proportions of correct rejections of fit than did the other procedures.

Overall, it appeared as though the selection of a goodness-of-fit procedure would differ depending on the type of error considered more serious. The LCHI procedure appeared to be least likely to result in erroneous rejections of fit, whereas the BCHI procedure appeared to yield marginally fewer erroneous acceptances of fit.

The following cautions in the interpretation of these results should be noted. The study addressed only the issue of fit with nonskewed, normal distributions of ability. The results do not generalize beyond this limitation. Nor does this study address the question of fit for tests of lengths shorter than 75 items, though the results do probably generalize to longer tests. It should also be noted that the results for the multidimensional data are based on a linear model involving two roughly equal factors. These results may not generalize to multidimensional data in which the factors are not equal, or in which linearity does not contribute to misfit. These results must be interpreted in the light of these limitations, in which case the results appear fairly clear-cut.

Based on the results of this study, the following conclusions seem appropriate about the use of these goodness-of-fit procedures. First, sample sizes of 500 to 1,000 seemed to yield the best results. A sample of 2,000 seemed to make the fit procedures too sensitive. Second, shifts in the mean of the ability distribution caused minor fluctuations, but did not seem to be a major issue. This does not, however, address the issue of distribution skewness or non-normality. Third, the likelihood ratio chi-square procedure appeared to yield the fewest erroneous rejections of the hypothesis of fit, whereas Bock's chi-square procedure yielded the fewest erroneous conclusions of fit. The procedure of choice apparently depends on which type of error is considered to be the more serious error.

References

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

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge MA: The MIT Press.

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., & Aitken, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika, 46,* 443–459.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79–92.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49,* 425–435.

Wherry, R. J. Sr., Naylor, J. C., Wherry, R. J. Jr., & Fallis, R. F. (1965). Generating multiple samples of multivariate data with arbitrary population parameters. *Psychometrika, 30,* 303–313.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide*. Princeton NJ: Educational Testing Service.

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

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.

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

Acknowledgment

This research was funded by the Department of Education of the State of Louisiana, U.S.A.

Author's Address

Send requests for reprints or further information to Robert L. McKinley, College of Education, The University of Toledo, 2801 W. Bancroft St., Toledo OH 43606, U.S.A.