DOCUMENT RESUME

ED 242 762 TM 840 175

AUTHOR Mills, Craig N.; McKinley, Robert L.
TITLE An Investigation of the Adequacy of Several Goodness of Fit Statistics.
PUB DATE Apr 84
NOTE 28p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 24-26, 1984).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Variance; Data Analysis; Evaluation; *Goodness of Fit; *Latent Trait Theory; *Mathematical Models; *Statistical Analysis
IDENTIFIERS Chi Square; Likelihood Ratio Criterion

ABSTRACT

A study was conducted to evaluate six goodness-of-fit procedures using data simulation techniques. The procedures evaluated included the weighted average absolute deviation (WAAD), the mean absolute deviation (MAD), Bock's chi-square (BCHI), Yen's chi-square (YCHI), Wright and Mead's chi-square (WCHI), and the likelihood ratio chi-square (LCHI) statistics. These procedures were evaluated using data generated according to three different item response theory models and a factor analytic model. Three different distributions of ability and sample sizes were used. The results of this study yielded the following conclusions: (1) sample sizes of 500 to 1000 seemed to yield the best results; (2) the largest sample size (N=2000) seemed to make the fit procedures too sensitive; (3) shifts in the mean of the ability distribution caused minor fluctuations, but did not appear to be a major concern; (4) the chi-square statistics performed better than did the two non-chi-square statistics; and (5) the likelihood ratio chi-square procedure appeared to yield the best results. (Author/DWH)

An Investigation of the Adequacy of Several Goodness of Fit Statistics[1]

Craig N. Mills[2]
Educational Testing Service

and

Robert L. McKinley
The American College Testing Program

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 only at the expense of making strong assumptions about the nature of the data. It is widely recognized that these assumptions are unlikely to be fully met in practice.

Because of the strong assumptions required for the use of IRT and the fact that the advantages associated with the use of IRT will not be realized if these assumptions are not met, it is important that prospective users of IRT methodology conduct an investigation to assess the appropriateness of IRT for use in the intended application. 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 estimating the parameters of an IRT model, using those parameter estimates to predict, via the IRT model, examinee response patterns, and comparing the 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 is to investigate a number of goodness of fit procedures to assess their adequacy for assessing the degree to which the more popular IRT models fit the 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 (1PL), two-parameter logistic (2PL), and the three-parameter logistic (3PL)] were estimated, and several

[1] Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, April, 1984.

[2] This study was conducted while the first author was affiliated with the Louisiana Department of Education.

TM 840 175

goodness of fit procedures were applied to the results. The accuracy with which the procedures identified known fit and misfit were then compared.

Before the results of this study are presented, a discussion of the goodness of fit statistics selected for this study will be presented. This will be followed by a more detailed discussion of the methodology used in this research. Finally, the results will be presented and discussed.

Goodness of Fit Statistics

After a review of the literature, six goodness of fit procedures were selected for this research. The procedures selected include only those which could be used with any IRT model (or could be modified for use with any model). In addition, only procedures which actually assess fit, as opposed to dimensionality or local independence, were selected. Finally, two types of fit statistics were selected for use in the study: those which lend themselves to chi-square analyses, and those that consider only the magnitude of the difference between observed and predicted performance. A description of each procedure selected follows.

Weighted Average Absolute Deviation (WAAD)

The WAAD procedure (Mills, 1982) requires that the ability scale be divided into intervals into which examinees are sorted on the basis of their ability estimates. The WAAD statistic is then computed for a given item as the weighted mean of the absolute deviations between the observed and predicted proportion-correct scores within the intervals. Interval values are weighted by the number of examinees falling within the intervals. The WAAD statistic is given by

$$WAAD_i = \frac{\sum_{j=1}^{J} N_j \mid O_{ij} - E_{ij} \mid}{\sum_{j=1}^{J} N_j} , \qquad (1)$$

where $WAAD_i$ is the WAAD statistic for item i, J is the number of intervals, N_j is the number of examinees in interval j, O_{ij} is the observed proportion-correct score on item i for examinees in interval j, and E_{ij} is the predicted proportion-correct score on item i for examinees in interval j. E_{ij} is computed using the appropriate IRT model, the model item parameter estimates, and the midpoint of interval j.

Mean Absolute Deviation (MAD)

The MAD statistic for an item is the mean over examinees of the absolute differences between the observed and predicted responses to the item. The MAD statistic is given by

$$MAD_i = \frac{\sum_{j=1}^{N} \left| E_{ij} - O_{ij} \right|}{N}, \qquad (2)$$

where MAD_i is the MAD statistic for item i, E_{ij} is the expected response to item i by examinee j (the probability of the observed response computed from the IRT model), O_{ij} is the observed response to item i by examinee j, and N is the number of examinees.

Bock's Chi-Square (BCHI)

The 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. Each examinee is then assigned to one of 2 x 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 compute a chi-square statistic. The predicted values for an interval are computed using the median of the ability estimates falling within the interval. The BCHI statistic for an item is given by

$$BCHI_i = \sum_{j=1}^{J} \frac{N_j (O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})}, \qquad (3)$$

where $BCHI_i$ is the chi-square statistic for item i, O_{ij} is the observed proportion-correct on item i for interval j, and the remaining terms are as previously defined. To test the significance of an item's fit, J-m degrees of freedom are used, where m is the number of item parameters estimated.

Yen's Chi-Square (YCHI)

The YCHI procedure (Yen, 1981) is the same as the BCHI procedure with two exceptions. First, the YCHI procedure uses ten intervals, whereas the BCHI procedure doesn't 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 the interval. The YCHI statistic is given by (3) with J = 10. The degrees of freedom are 10-m, where m is as previously defined.

Wright and Mead Chi-Square (WCHI)

The WCHI procedure (Wright and Mead, 1977) is identical to the YCHI procedure with three exceptions. First, the procedure is based on number-correct score groups (i.e., the 1PL model) rather than on intervals of the ability scale. Second, rather than using ten 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 which 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 1PL model, it was modified by substituting the grouping method of the YCHI procedure for the number-correct grouping approach. The WCHI statistic is given by

$$WCHI_i = \sum_{j=1}^{J} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij}) - \sigma_{p_j}^{2}} \,, \qquad (4)$$

where $WCHI_i$ is the WCHI statistic for item i, $\sigma_{p_j}^{2}$ is given by

$$\sigma_{p_j}^{2} = \frac{1}{N_j} \sum_{k=j}^{N_j} [\hat{P}_i(\hat{O}_k) - E_{ij}]^2 \,, \qquad (5)$$

$\hat{P}_i(\hat{O}_k)$ is the predicted proportion passing item i in score group k, and the other terms are as previously defined. The degrees of freedom are given by J-m.

Likelihood Ratio Chi-Square (LCHI)

The LCHI procedure follows much the same pattern as the YCHI procedure. The ability scale is divided into ten intervals, and examinees are sorted into one of twenty cells based on their ability estimates and whether or not they correctly responded to the item of interest. A ten by two contingency table is formed, and a likelihood ratio chi-square statistic (Bishop, Fienberg, and Holland, 1975) is computed. The LCHI statistic is given by

$$LCHI_i = 2 \sum_{j=1}^{20} O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right), \qquad (6)$$

where $LCHI_i$ is the LCHI statistic for item i, ln(x) is the logarithm to the base e of x, and the remaining terms are as previously defined. The degrees of freedom are given by 10-m.

Methodology

The Simulation of Test Data

In all, 36 tests were simulate', 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, 1000, and 2000 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 item difficulty), and a high ability group (ability centered about one standard deviation above the mean item difficulty). Table 1 summarizes the data which were generated and provides some descriptive statistics for the various simulated tests.

The item parameters used to simulate the 1PL, 2PL, and 3PL data were selected as follows. All of the item parameters were selected from uniform distributions. These distributions had the following ranges:

1PL Range:	b=-3.0 to 3.0	a=1.0 to 1.0	c=0.0 to 0.0
2PL Range:	b=-3.0 to 3.0	a=0.2 to 2.0	c=0.0 to 0.0
3PL Range:	b=-3.0 to 3.0	a=0.2 to 2.0	c=0.1 to 0.25

The same b-values were used for all datasets. The same a-values were used for all 2PL and 3PL datasets. For all 1PL datasets a value of 1.0 was used for all a-values. For all 1PL and 2PL datasets a value of 0.0 was used for all c-values. Table 2 presents some descriptive statistics for the item parameters used for generating the 1PL, 2PL, and 3PL data.

The ability parameters used for the 1PL, 2PL, and 3PL data were selected as follows. All abilities were randomly selected from a standard normal distribution. First, 500 abilities were selected and used for the 500 sample size datasets. For the 1000 sample size datasets an additional 500 abilities were selected and combined with the 500 abilities previously selected. Likewise, for the 2000 sample size cases an additional 1000 abilities were selected and combined with those already selected. Table 3 shows some descriptive statistics for the ability distributions for the nine tests for each model. Note that the low ability groups were simulated by subtracting 1.0 from all of the abilities, while the high ability groups were simulated by adding 1.0 to all abilities.

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. Table 4 shows the factor loadings and the proportion-correct scores used to generate these data. Note that there are three sets of proportion-correct scores, each with a different mean. Groups with different mean abilities were simulated by shifting the mean of the target proportion-correct scores. The target mean total test proportion-correct scores for the three ability groups are p=0.375, p=0.500, and p=0.625 for the low, centered, and high ability groups, respectively. Items 1-37 had factor loadings of 0.70 on the first factor and 0.20 on the second factor, while items 38-75 had loadings of .20 on the first factor and 0.70 on the second factor.

Calibration

All of the data for all conditions were calibrated for the 1PL, 2PL, and 3PL models using LOGIST (Wingersky, Barton, and Lord, 1982). For the 1PL and 2PL models, all c values were held constant at zero. For the 1PL model the a values were held constant at a value selected by the LOGIST program.

Analyses

The first analysis performed was the application of the six goodness of fit procedures 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 3PL calibration of 1PL data) from cases of misfit (such as the 1PL calibration of multidimensional data).

Results

The statistics used in this study fall into two main types: those which are based on approximately chi-square distributed statistics, and those which are not. The results for the chi-square statistics will be presented first, followed by the results for the remaining two procedures.

Chi-Square Based Procedures

Tables 5, 6, 7, and 8 report summaries of the results obtained for the chi-square based procedures for the one-, two-, and three-parameter data and the multidimensional data, respectively. The values reported in the tables are the proportion of items for which there was significant misfit of the model to the data. A significance level of 0.01 was used for testing the chi-squares for the individual items. Thus, under the hypothesis of fit, the proportion of items for which there was misfit should have been around 0.01.

Table 5 summarizes the results for the one-parameter data. Since these data were generated to fit the 1PL model, it would be expected that all three calibration models would yield fit. As can be seen from Table 5, however, some 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 5 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 6 summarizes the results obtained for the chi-square procedures for the 2PL data. For these data fit was expected for the 2PL and 3PL models, but not for the 1PL model. As can be seen from Table 6, all four procedures showed clear differences between the 1PL calibrations and the 2PL and 3PL calibrations. There is some lack of fit for the 2PL and 3PL models, especially the 3PL model, but the proportions of items for which there was misfit are dramatically less than for the 1PL model, regardless of which procedure is considered.

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

Table 7 summarizes the results obtained for the 3PL data. For these data, only the 3PL calibration model was expected to yield fit. It was expected that the fit for the 2PL model would be worse than for the 3PL model, but not as bad as for the 1PL model. This is the pattern obtained for all four procedures.

There was some misfit indicated for the 3PL 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 2PL model than did the other procedures. There were no clear patterns for the 1PL calibrations.

Table 8 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, although the level of misfit (proportions of items for which there was misfit) was surprisingly low for the 500 and 1000 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 2PL and 3PL calibration models. Nonetheless, in no case would these results be interpreted as indicating that the unidimensional models yield adequate fit.

MAD

Table 9 summarizes the results obtained for the MAD procedure. The values shown are the mean MAD statistics obtained for the various datasets. In order to further investigate these data, four analyses of variance (ANOVAs) were run. The first ANOVA was run on the 1PL data, the second ANOVA was run on the 2PL data, the third was run on the 3PL data, and the fourth was run on the multidimensional data. All four of these ANOVAs followed the same design. Ability distribution and sample size were treated as independent variables, with sample size nested within ability distribution. Calibration model was treated as a repeated measure.

Table 10 summarizes the ANOVA results obtained for the 1PL data. As can be seen from Table 10, the main effect associated with the ability distribution was significant, as was the calibration model effect. The values reported in Table 9 show that the mean MAD statistics increased across ability distributions (low to high), and were slightly higher for the 1PL calibration model, though the differences among the calibration models tend to be masked by rounding.

Table 11 summarizes the ANOVA results for the 2PL data. For these data, only the calibration model effect was significant. The values tended to be largest for the 1PL model and smallest for the 2PL model. This is consistent with the fact that these data were generated to fit the 2PL model.

Table 12 summarizes the ANOVA results for the 3PL data. In this case, both the ability distribution and calibration effects were significant. The values in Table 9 decreased across ability distributions (low to high), and were largest for the 1PL model and smallest for the 3PL model. This is as would be expected for 3PL data.

Table 13 summarizes the ANOVA results for the multidimensional data. For these data, the ability distribution and calibration model effects were significant, as was the calibration model by ability distribution interaction effect. Examination of the values shown in Table 9 reveals that the values tended to be largest for the 1PL model and smallest for the 3PL model, regardless of the distribution of ability. For the low ability distribution, the 2PL values were between the 1PL and 3PL values. For the centered ability distribution, the 2PL values were about the same as the 3PL values, while for the high distribution of ability the 2PL values were about the same as the 1PL values.

Overall, the MAD procedure performed as expected. The calibration model matching the data generation model tended to have the lowest values. However, the procedure showed some sensitivity to sample size and ability distribution. More importantly, despite their statistical significance, the differences among the values shown in Table 9 are so small as to severely restrict their practical usefulness. In addition, using the mean MAD values as a criterion, all these models appeared to fit the multidimensional data better than the unidimensional 3PL data. This seems inconsistent with the purposes of using a goodness of fit procedure and brings into question the value of the MAD procedure.

WAAD

Table 14 shows the mean WAAD statistics obtained for all three calibration models for all sample sizes and ability distributions. In order to further analyze these data, the same four ANOVAs run on the MAD data were run on the WAAD data.

Table 15 summarizes the results of the ANOVA run for the 1PL data. For these data the main effects associated with ability distribution, sample size, and calibration model were significant, as was the calibration model by ability distribution interaction effects. An examination of Table 14 indicates that mean WAAD statistics decreased as sample size increased. Overall, the 2PL values were smallest and the 1PL values were largest, though there were several exceptions to this pattern. For the low distribution of ability, the 1PL and 3PL values were similar, while for the other two ability distributions the 2PL and 3PL values tended to be similar.

Table 16 shows a summary of the ANOVA results for the 2PL data. As can be seen from the table, the ability distribution, sample size, and calibration model main effects were significant. The calibration model by ability distribution interaction effect and the calibration model by sample size interaction effect were also significant. From Table 14 it can be seen that for these data the mean WAAD values for the 1PL model were consistently larger than for the other models. For the low ability distribution the 2PL values were the smallest. For the centered distribution of ability the 2PL values were about the same as the 3PL values, except in the 2000 sample size case, in which the 2PL value was smaller than the 3PL value. For the high ability group the 2PL value was smaller than the 3PL value for the 500 sample size case. For the 1000 and 2000 sample size cases the 3PL model values were smaller.

Table 17 shows a summary of the ANOVA results for the 3PL data. Again, the main effects were all significant, and the calibration model by ability distribution interaction effect and the calibration model by sample size interaction effect were significant. The 1PL mean WAAD statistics were consistently higher than for the other models. For the low ability distribution the 3PL values were smallest for the 1000 and 2000 sample size cases, but the 2PL values were smallest for the 500 sample size case. For the centered and high ability distributions the 3PL values were the smallest.

The results of the ANOVA on the multidimensional data are summarized in Table 18. All of the main effects were significant, as was the calibration model by ability distribution interaction effect. The values were largest for the low distribution of ability and smallest for the centered ability distribution. For the low distribution of ability the 2PL values were largest and the 3PL values were smallest. For the centered distribution the 3PL values were largest and the 2PL values were smallest. For the high ability distribution the 1PL values were largest, and for the 500 and 1000 sample sizes the 3PL values were smallest. For the 2000 sample size case the 2PL values were smallest.

Overall, the WAAD procedure performed less well than did the MAD procedure. The data generation model, when used as the calibration model, did not always yield the best fit. The procedure appears to be overly sensitive to ability distribution effects. In general, the results did not adequately follow the patterns expected.

Discussion

A desirable goodness of fit procedure is one that indicates fit when there is fit and misfit when there is misfit, and does not indicate fit when there is misfit or misfit when there is fit. That is, the procedure should be sensitive to misfit, anr nothing but misfit. The procedures in this study were evaluated with that in mind.

Overall, most of the procedures performed about as expected. They correctly indicated misfit when the calibration model did not match the generation model (and did not subsume the generation model, as would be the case with the 3PL calibration model and the 1PL generation model). On the other hand, most of the procedures seemed overly sensitive to problems caused by different distributions of ability, and one or two of the procedures seemed overly sensitive to sample size.

All things considered, the chi-square based procedures seem of more practical usefulness than the MAD and WAAD procedures. They allow, for instance, significance testing on the individual item basis, which the MAD and WAAD procedures do not allow. With the chi-square procedures, moreover, the proportion of items for which there were significant chi-squares can theoretically be computed and compared to an alpha level to assess fit. The MAD and WAAD statistics can only be used for comparative purposes.

More pertinent to the current study, the differences between fit and misfit were much more clearly discernable with the chi-square procedures than with the MAD and WAAD procedures. With the MAD and WAAD procedures, a visual examination often would have left the impression that there were no real differences between the calibration models that should have yielded fit and those that shouldn't have yielded fit. The chi-square procedures made those distinctions much more pronounced. All things considered, the chi-square procedures seemed superior to the MAD and WAAD procedures.

Among the chi-square procedures, the LCHI procedure seemed to be the most satisfactory procedure. It performed as well as the other statistics at identifying misfit, and it seemed much less sensitive to sample size and ability distribution effects. Overall, the LCHI procedure appeared to be the procedure of choice.

Summary and Conclusions

A study was conducted to evaluate six goodness of fit procedures using data simulation techniques. The procedures were evaluated using data generated according to four different models. Three different distributions of ability were used, as were three different sample sizes.

The following cautions in the interpretation of these results should be noted. This 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, although the results do probably generalize to longer tests. These results must be interpreted in the light of these limitations, in which case the results appear clear-cut.

Based on the results of this study, the following conclusions seem appropriate in regard to the use of these goodness of fit procedures. First, sample sizes of 500-1000 seemed to yield the best results. The largest sample size seems to make the fit procedures too sensitive. Second, shifts in the mean of the ability distribution cause minor fluctuations, but doesn't seem to be a major issue. This does not, however, address the issue of distribution skewedness or nonnormality. Third, the chi-square statistics yield better results than the MAD and WAAD statistics. The likelihood ratio chi-square procedure appearred to yield the best results.

References

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

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

McKinley, R.L. (1983, April). A multidimensional extension of the two-parameter logistic latent trait model. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Mills, C.N. (1982). An investigation of the adequacy of two item response models for calibrating an item bank. Unpublished doctoral dissertation, Univerrity of Massachusetts.

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

Wingersky, M.S., Barton, M.A., and Lord, F.M. (1982). LOGIST User's Guide. Princeton, NJ: Educational Testing Service.

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

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

Table 1
Summary of Generated Tests

Test[1]	Generating Model	Mean[2]	S.D.	KR-20
CE500	1PL	31.18	12.96	0.950
CE1000		30.39	12.65	0.948
CE2000		30.14	12.53	0.947
LO500		19.69	10.79	0.941
LO1000		19.10	10.47	0.939
LO2000		18.81	10.26	0.936
HI500		44.01	13.59	0.953
HI1000		43.23	13.38	0.951
HI2000		42.94	13.35	0.951
CE500	2PL	31.06	12.51	0.942
CE1000		30.32	12.13	0.939
CE2000		30.10	11.97	0.938
LO500		20.50	10.13	0.925
LO1000		19.94	9.80	0.921
LO2000		19.69	9.53	0.917
HI500		43.41	13.10	0.946
HI1000		42.66	12.92	0.945
HI2000		42.42	12.86	0.944
CE500	3PL	38.62	10.67	0.892
CE1000		37.86	10.40	0.886
CE2000		37.57	10.23	0.882
LO500		29.83	9.05	0.843
LO1000		29.29	8.69	0.830
LO2000		28.97	8.47	0.822
HI500		48.74	11.05	0.912
HI1000		47.84	10.95	0.909
HI2000		48.09	10.98	0.910
CE500	LIN	36.95	14.08	0.936
CE1000		36.98	14.09	0.936
CE2000		26.99	14.08	0.936
LO500		26.84	17.67	0.968
LO1000		26.88	17.63	0.968
LO2000		26.82	17.66	0.968
HI500		48.14	13.73	0.941
HI1000		48.25	13.60	0.940
HI2000		48.27	13.57	0.940

[1] Tests are defined as follows: the first two characters specify the ability group used to generate item responses (LO=low, CE=centered, HI-high). The next three or four digits indicate the number of examinees.

[2] All tests were 75 items in length.

Table 2
Description of Item Parameters for the Simulated Unidimensional Tests

	Model								
Stat	1PL			2PL			3PL		
	a	b	c	a	b	c	a	b	c
Mean	1.00	0.45	0.00	1.16	0.45	0.00	1.16	0.45	0.17
S.D.	0.00	1.68	0.00	0.55	1.68	0.00	0.55	1.68	0.04
Min.	1.00	-2.77	0.00	0.21	-2.77	0.00	0.21	-2.77	0.11
Max.	1.00	2.98	0.00	1.98	2.98	0.00	1.98	2.98	0.25
Range	0.00	5.75	0.00	1.77	5.75	0.00	1.77	5.75	0.14

Table 3
Simulated Ability Distributions Used to Generate
Unidimensional Response Vectors

Ability	N	Mean	S.D.	Min	Max	Range
Low	500	-0.979	1.027	-3.893	2.327	6.220
Centered		0.021	1.027	-2.893	3.327	6.220
High		1.021	1.027	-1.893	4.327	6.220
Low	1000	-1.029	1.003	-3.893	2.327	6.220
Centered		-0.030	1.003	-2.893	3.327	6.220
High		0.971	1.003	-1.893	4.327	6.220
Low	2000	-1.048	0.999	-4.924	2.327	7.251
Centered		-0.048	0.999	-3.924	3.327	7.251
High		0.952	0.999	-2.924	4.327	7.251

Table 4
Target Factor Loadings and Proportion-Correct Scores
for Multidimensional Simulations

Item	Loadings		Proportion-correct		
	I	II	Low	Centered	High
1	0.70	0.20	0.010	0.135	0.260
2	0.70	0.20	0.020	0.145	0.270
3	0.70	0.20	0.030	0.155	0.280
4	0.70	0.20	0.040	0.165	0.290
5	0.70	0.20	0.050	0.175	0.300
6	0.70	0.20	0.060	0.185	0.310
7	0.70	0.20	0.070	0.195	0.320
8	0.70	0.20	0.080	0.205	0.330
9	0.70	0.20	0.090	0.215	0.340
10	0.70	0.20	0.100	0.225	0.350
11	0.70	0.20	0.110	0.235	0.360
12	0.70	0.20	0.120	0.245	0.370
13	0.70	0.20	0.130	0.255	0.380
14	0.70	0.20	0.140	0.265	0.390
15	0.70	0.20	0.150	0.275	0.400
16	0.70	0.20	0.160	0.285	0.410
17	0.70	0.20	0.170	0.295	0.420
18	0.70	0.20	0.180	0.305	0.430
19	0.70	0.20	0.190	0.315	0.440
20	0.70	0.20	0.200	0.325	0.450
21	0.70	0.20	0.210	0.335	0.460
22	0.70	0.20	0.220	0.345	0.470
23	0.70	0.20	0.230	0.355	0.480
24	0.70	0.20	0.240	0.365	0.480
25	0.70	0.20	0.250	0.375	0.500
26	0.70	0.20	0.260	0.385	0.510
27	0.70	0.20	0.270	0.395	0.520
28	0.70	0.20	0.280	0.405	0.530
29	0.70	0.20	0.290	0.415	0.540
30	0.70	0.20	0.300	0.425	0.550
31	0.70	0.20	0.310	0.435	0.560
32	0.70	0.20	0.320	0.445	0.570
33	0.70	0.20	0.330	0.455	0.580
34	0.70	0.20	0.340	0.465	0.590
35	0.70	0.20	0.350	0.475	0.600
36	0.70	0.20	0.360	0.485	0.610
37	0.70	0.20	0.370	0.495	0.620
38	0.20	0.70	0.380	0.505	0.630

Table 4(Continued)
Target Factor Loadings and Proportion-Correct Scores
for Multidimensional Simulations

Item	Loadings		Proportion-correct		
	I	II	Low	Centered	High
39	0.20	0.70	0.390	0.515	0.640
40	0.20	0.70	0.400	0.525	0.650
41	0.20	0.70	0.410	0.535	0.660
42	0.20	0.70	0.420	0.545	0.670
43	0.20	0.70	0.430	0.555	0.680
44	0.20	0.70	0.440	0.565	0.690
45	0.20	0.70	0.450	0.575	0.700
46	0.20	0.70	0.460	0.585	0.710
47	0.20	0.70	0.470	0.595	0.720
48	0.20	0.70	0.480	0.605	0.730
49	0.20	0.70	0.490	0.615	0.740
50	0.20	0.70	0.500	0.625	0.750
51	0.20	0.70	0.510	0.635	0.760
52	0.20	0.70	0.520	0.645	0.770
53	0.20	0.70	0.530	0.655	0.780
54	0.20	0.70	0.540	0.665	0.790
55	0.20	0.70	0.550	0.675	0.800
56	0.20	0.70	0.560	0.685	0.810
57	0.20	0.70	0.570	0.695	0.820
58	0.20	0.70	0.580	0.705	0.830
59	0.20	0.70	0.590	0.715	0.840
60	0.20	0.70	0.600	0.725	0.850
61	0.20	0.70	0.610	0.735	0.860
62	0.20	0.70	0.620	0.745	0.870
63	0.20	0.70	0.630	0.755	0.880
64	0.20	0.70	0.640	0.765	0.890
65	0.20	0.70	0.650	0.775	0.900
66	0.20	0.70	0.660	0.785	0.910
67	0.20	0.70	0.670	0.795	0.920
68	0.20	0.70	0.680	0.805	0.930
69	0.20	0.70	0.690	0.815	0.940
70	0.20	0.70	0.700	0.825	0.950
71	0.20	0.70	0.710	0.835	0.960
72	0.20	0.70	0.720	0.845	0.970
73	0.20	0.70	0.730	0.855	0.980
74	0.20	0.70	0.740	0.865	0.990
75	0.20	0.70	0.750	0.875	0.990

Table 5
Proportions of Items Identified as Misfitting
One-Parameter Data

Sample Size	Statistic	Distribution of Ability/Calibration Model								
		LO			CE			HI		
		1PL	2PL	3PL	1PL	2PL	3PL	1PL	2PL	3PL
500	BCHI	0.00	0.01	0.03	0.07	0.07	0.04	0.00	0.03	0.00
	WCHI	0.00	0.00	0.03	0.05	0.04	0.04	0.03	0.01	0.00
	LCHI	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
	YCHI	0.00	0.01	0.00	0.04	0.07	0.04	0.01	0.03	0.00
1000	BCHI	0.00	0.01	0.07	0.07	0.05	0.04	0.03	0.03	0.01
	WCHI	0.00	0.00	0.05	0.04	0.03	0.04	0.01	0.03	0.01
	LCHI	0.00	0.00	0.04	0.03	0.01	0.03	0.00	0.00	0.00
	YCHI	0.00	0.01	0.07	0.05	0.04	0.04	0.03	0.03	0.01
2000	BCHI	0.03	0.03	0.07	0.08	0.08	0.12	0.03	0.03	0.03
	WCHI	0.01	0.01	0.08	0.07	0.08	0.09	0.03	0.04	0.04
	LCHI	0.01	0.00	0.08	0.00	0.00	0.04	0.00	0.00	0.00
	YCHI	0.04	0.04	0.05	0.08	0.08	0.09	0.03	0.03	0.03

Table 6
Proportions of Items Identified as Misfitting
Two-Parameter Data

Sample Size	Statistic	Distribution of Ability/Calibration Model								
		LO			CE			HI		
		1PL	2PL	3PL	1PL	2PL	3PL	1PL	2PL	3PL
500	BCHI	0.36	0.01	0.11	0.37	0.03	0.04	0.42	0.05	0.01
	WCHI	0.28	0.04	0.08	0.36	0.04	0.04	0.42	0.05	0.04
	LCHI	0.35	0.00	0.07	0.40	0.00	0.01	0.47	0.00	0.00
	YCHI	0.34	0.01	0.03	0.37	0.03	0.04	0.40	0.04	0.01
1000	BCHI	0.47	0.03	0.05	0.51	0.01	0.03	0.65	0.03	0.03
	WCHI	0.45	0.01	0.07	0.53	0.03	0.08	0.64	0.01	0.05
	LCHI	0.47	0.01	0.07	0.51	0.00	0.05	0.69	0.01	0.01
	YCHI	0.45	0.01	0.05	0.49	0.01	0.03	0.63	0.03	0.03
2000	BCHI	0.65	0.03	0.07	0.65	0.05	0.12	0.77	0.04	0.03
	WCHI	0.61	0.01	0.13	0.72	0.07	0.09	0.77	0.05	0.03
	LCHI	0.64	0.01	0.07	0.65	0.00	0.08	0.75	0.01	0.01
	YCHI	0.61	0.01	0.07	0.63	0.05	0.11	0.77	0.03	0.01

Table 7
Proportions of Items Identified as Misfitting
Three-Parameter Data

Sample Size	Statistic	Distribution of Ability/Calibration Model								
		LO			CE			HI		
		1PL	2PL	3PL	1PL	2PL	3PL	1PL	2PL	3PL
500	BCHI	0.43	0.12	0.04	0.53	0.12	0.01	0.55	0.11	0.03
	WCHI	0.53	0.09	0.03	0.57	0.11	0.03	0.47	0.11	0.01
	LCHI	0.43	0.09	0.00	0.53	0.03	0.00	0.56	0.10	0.00
	YCHI	0.44	0.15	0.03	0.53	0.12	0.01	0.53	0.12	0.03
1000	BCHI	0.65	0.09	0.05	0.67	0.17	0.04	0.73	0.15	0.07
	WCHI	0.67	0.15	0.05	0.75	0.19	0.01	0.72	0.16	0.01
	LCHI	0.69	0.12	0.03	0.68	0.07	0.01	0.72	0.19	0.01
	YCHI	0.65	0.09	0.04	0.68	0.17	0.04	0.70	0.19	0.04
2000	BCHI	0.84	0.20	0.05	0.[illegible]	0.33	0.05	0.83	0.39	0.04
	WCHI	0.85	0.29	0.05	0.89	0.28	0.04	0.83	0.29	0.01
	LCHI	0.85	0.13	0.00	0.85	0.20	0.01	0.80	0.28	0.00
	YCHI	0.85	0.20	0.04	0.89	0.33	0.05	0.81	0.37	0.03

Table 8
Proportions of Items Identified as Misfitting
Multidimensional Data

Sample Size	Statistic	Distribution of Ability/Calibration Model								
		LO			CE			HI		
		1PL	2PL	3PL	1PL	2PL	3PL	1PL	2PL	3PL
500	BCHI	0.92	0.86	0.91	0.25	0.29	0.35	0.72	0.73	0.83
	WCHI	0.70	0.63	0.76	0.16	0.13	0.25	0.77	0.61	0.59
	LCHI	0.93	0.85	0.89	0.21	0.19	0.32	0.76	0.71	0.77
	YCHI	0.92	0.81	0.84	0.27	0.27	0.36	0.72	0.69	0.79
1000	BCHI	1.00	0.97	0.98	0.68	0.63	0.76	0.95	0.87	0.92
	WCHI	0.85	0.82	0.86	0.53	0.45	0.55	0.93	0.77	0.65
	LCHI	0.99	0.96	0.95	0.73	0.67	0.76	0.93	0.87	0.87
	YCHI	0.97	0.93	0.92	0.68	0.60	0.68	0.93	0.85	0.87
2000	BCHI	1.00	0.99	1.00	0.92	0.95	0.99	0.99	0.95	0.99
	WCHI	0.92	0.93	0.82	0.75	0.77	0.92	0.97	0.87	0.82
	LCHI	0.99	0.97	0.97	0.97	0.97	1.00	0.96	0.95	0.96
	YCHI	1.00	0.96	0.97	0.92	0.93	0.99	0.99	0.95	0.99

Table 9
Means of MAD Statistics

		Distribution of Ability/Sample Size								
Data	Calibration Model	LO			CE			HI		
		500	1000	2000	500	1000	2000	500	1000	2000
1PL	1PL	0.16	0.15	0.15	0.19	0.19	0.19	0.20	0.20	0.20
	2PL	0.15	0.15	0.15	0.19	0.19	0.19	0.20	0.20	0.20
	3PL	0.15	0.15	0.15	0.19	0.19	0.19	0.20	0.20	0.20
2PL	1PL	0.18	0.18	0.18	0.21	0.21	0.21	0.22	0.22	0.22
	2PL	0.18	0.17	0.17	0.20	0.20	0.20	0.21	0.21	0.21
	3PL	0.18	0.18	0.17	0.20	0.20	0.20	0.21	0.21	0.21
3PL	1PL	0.34	0.33	0.33	0.32	0.32	0.32	0.28	0.28	0.27
	2PL	0.32	0.32	0.32	0.30	0.30	0.30	0.27	0.26	0.26
	3PL	0.32	0.32	0.32	0.30	0.30	0.30	0.26	0.26	0.26
Multidimensional	1PL	0.28	0.28	0.27	0.29	0.29	0.29	0.26	0.26	0.26
	2PL	0.27	0.26	0.26	0.29	0.29	0.29	0.26	0.25	0.25
	3PL	0.26	0.26	0.25	0.29	0.29	0.29	0.24	0.24	0.24

Table 10

Summary of ANOVA on MAD Statistics
for the 1PL Data

Source	df	MS	F	p
Ability	2	0.47	13.88	0.00
Sample Size	2	0.00	0.03	0.97
Ability x Sample	4	0.00	0.01	0.99
Error	666	0.03		
Calibration Model	2	0.00	48.57	0.00
Model x Ability	4	0.00	2.49	0.04
Model x Sample	4	0.00	2.76	0.03
Model x Ability x Sample	8	0.00	0.46	0.89
Error	1332	0.00		

Table 11

Summary of ANOVA on MAD Statistics
for the 2PL Data

Source	df	MS	F	p
Ability	2	0.24	4.38	0.01
Sample Size	2	0.00	0.01	0.99
Ability x Sample	4	0.00	0.01	0.99
Error	666	0.05		
Calibration Model	2	0.02	66.08	0.00
Model x Ability	4	0.00	2.46	0.04
Model x Sample	4	0.00	0.03	0.99
Model x Ability x Sample	8	0.00	0.01	0.99
Error	1332	0.00		

Table 12

Summary of ANOVA on MAD Statistics
for the 3PL Data

Source	df	MS	F	p
Ability	2	0.63	15.51	0.00
Sample Size	2	0.00	0.00	0.99
Ability x Sample	4	0.00	0.01	0.99
Error	666	0.04		
Calibration Model	2	0.05	110.09	0.00
Model x Ability	4	0.00	2.47	0.04
Model x Sample	4	0.00	0.03	0.99
Model x Ability x Sample	8	0.00	0.02	0.99
Error	1332	0.00		

Table 13

Summary of ANOVA on MAD Statistics
for the Multidimensional Data

Source	df	MS	F	p
Ability	2	0.34	11.37	0.00
Sample Size	2	0.01	0.18	0.83
Ability x Sample	4	0.00	0.10	0.98
Error	666	0.03		
Calibration Model	2	0.08	89.55	0.00
Model x Ability	4	0.02	9.53	0.00
Model x Sample	4	0.00	1.79	0.13
Model x Ability x Sample	8	0.00	1.03	0.41
Error	1332	0.00		

Table 14
Means of MAD Statistics

Data	Model	Distribution of Ability/Sample Size								
		LO			CE			HI		
		500	1000	2000	500	1000	2000	500	1000	2000
1PL	1PL	0.02	0.02	0.01	0.02	0.02	0.02	0.03	0.02	0.02
	2PL	0.02	0.02	0.01	0.03	0.02	0.01	0.02	0.02	0.01
	3PL	0.02	0.02	0.01	0.03	0.02	0.02	0.03	0.02	0.01
2PL	1PL	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.05
	2PL	0.02	0.02	0.01	0.03	0.02	0.01	0.02	0.02	0.01
	3PL	0.02	0.02	0.01	0.03	0.02	0.02	0.03	0.02	0.01
3PL	1PL	0.06	0.06	0.05	0.06	0.06	0.06	0.06	0.06	0.06
	2PL	0.04	0.04	0.03	0.04	0.03	0.03	0.03	0.03	0.02
	3PL	0.04	0.03	0.02	0.03	0.03	0.02	0.03	0.02	0.02
Multidimensional	1PL	0.15	0.16	0.17	0.06	0.06	0.07	0.10	0.11	0.12
	2PL	0.15	0.16	0.17	0.05	0.06	0.08	0.09	0.10	0.11
	3PL	0.15	0.15	0.15	0.06	0.07	0.08	0.08	0.10	0.11

Table 15

Summary of ANOVA on WAAD Statistics
for the 1PL Data

Source	df	MS	F	p
Ability	2	0.00	7.27	0.00
Sample Size	2	0.02	64.12	0.00
Ability x Sample	4	0.00	0.13	0.97
Error	666	0.00		
Calibration Model	2	0.00	17.20	0.00
Model x Ability	4	0.00	3.75	0.00
Model x Sample	4	0.00	3.08	0.02
Model x Ability x Sample	8	0.00	1.66	0.10
Error	1332	0.00		

Table 16

Summary of ANOVA on WAAD Statistics
for the 2PL Data

Source	df	MS	F	p
Ability	2	0.01	7.52	0.00
Sample Size	2	0.01	7.83	0.00
Ability x Sample	4	0.00	0.04	0.99
Error	666	0.00		
Calibration Model	2	0.16	540.64	0.00
Model x Ability	4	0.00	11.32	0.00
Model x Sample	4	0.00	7.13	0.00
Model x Ability x Sample	8	0.00	0.09	0.99
Error	1332	0.00		

Table 17

Summary of ANOVA on WAAD Statistics
for the 3PL Data

Source	df	MS	F	p
Ability	2	0.00	6.67	0.00
Sample Size	2	0.02	28.09	0.00
Ability x Sample	4	0.00	0.70	0.59
Error	666	0.00		
Calibration Model	2	0.18	696.41	0.00
Model x Ability	4	0.00	5.34	0.00
Model x Sample	4	0.00	6.76	0.00
Model x Ability x Sample	8	0.00	0.21	0.99
Error	1332	0.00		

Table 18

Summary of ANOVA on WAAD Statistics
for the Multidimensional Data

Source	df	MS	F	p
Ability	2	1.18	170.95	0.00
Sample Size	2	0.05	6.86	0.00
Ability x Sample	4	0.00	0.35	0.84
Error	666	0.01		
Calibration Model	2	0.02	26.97	0.00
Model x Ability	4	0.02	21.93	0.00
Model x Sample	4	0.00	0.77	0.54
Model x Ability x Sample	8	0.00	4.14	0.00
Error	1332	0.00		

An Investigation of the Adequacy of Several Goodness of Fit Statistics

Abstract

A study was conducted to evaluate six goodness of fit procedures using data simulation techniques. The procedures evaluated included the weighted average absolute deviation, the mean absolute deviation, Bock's chi-square, Yen's chi-square, Wright and Mead's chi-square, and the likelihood ratio chi-square statistics. These 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.

Based on the results of this study, the following conclusions seem appropriate. First, sample sizes of 500-1000 seemed to yield the best results. The largest sample size (2000) seemed to make the fit procedures too sensitive. Second, shifts in the mean of the ability distribution caused minor fluctuations, but did not appear to be a major concern. Third, the chi-square statistics performed better than did the two non-chi-square statistics. Finally, the likelihood ratio chi-square procedure appeared to yield the best results.