A Monte Carlo Investigation of Several Person and Item Fit Statistics for Item Response Models

H. Jane Rogers
University of Massachusetts

John A. Hattie
University of New England, Australia

This study investigated the behavior of several person and item fit statistics commonly used to test and obtain fit to the one-parameter item response model. Using simulated data for 500 persons and 15 items, the sensitivity of the total-t, mean-square residual, and between-t fit statistics to guessing, heterogeneity in discrimination parameters, and multidimensionality was examined. Additionally, 25 misfitting persons and a misfitting item were generated to test the power of the three fit statistics to detect deviations in a subset of observations. Neither the total-t nor the mean-square residual were able to detect deviation from any of the models fitted. Use of these statistics appears to be unwarranted. The between-t was a useful indicator of guessing and heterogeneity in discrimination parameters, but was unable to detect multidimensionality. These results show that use of person and item fit statistics to test and obtain overall fit to the one-parameter model can lead to acceptance of the model even when it is grossly inappropriate. Assessments of model fit based on this strategy are inadequate. Alternative methods must be sought.

Item response models entail strong assumptions about the data to which they are applied, and the often-cited advantages of their use accrue only when these assumptions are fulfilled—that is, when the data fit the model, at least to a moderate degree. Without acceptable fit, the validity of results cannot be ensured. Since the 1960s, when the use of item response models was first proposed, many goodness-of-fit tests have been suggested, but the assessment of model fit has remained at issue (Traub & Wolfe, 1981).

To a large extent, this problem has arisen out of the necessity of using approximate tests of fit: test statistics with distributions which only approximate a recognized distribution. Statistics with asymptotically known distributions can be derived, but until recently computational problems have prevented their application and test makers have relied on the readily calculated approximate tests. Even given solutions to these computational problems, difficulties remain. When sample sizes are small, the asymptotic tests are of dubious validity; when sample sizes are large, the statistics gain sufficient power to detect practically insignificant deviation from the model. Moreover, although approximate methods can be used to assess the fit of persons and items, there are no precise statistical tests which allow the identification of subsets of persons and items which are extreme in some way. The use of approximate tests has thus continued despite frequent criticisms of their validity. "Rough but useful" (Wright & Stone, 1979) has been their justification.

A common strategy for testing and obtaining overall fit is to delete persons and items found to be misfitting on the basis of these approximate tests. The widely used BICAL program (Wright, Mead, & Bell, 1979) uses person and item fit statistics in

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 11, No. 1, March 1987, pp. 47–57

0146-6216/87/010047-11$1.80

lieu of a test of overall fit under the assumption that when persons and items found to be misfitting are deleted, the remaining data fit the model. This practice of deleting items and persons in order to obtain fit of the remaining data has been strongly questioned. Traub and Wolfe (1981) warned that once person and item fit has been tested using results predicted by the model, it is no longer appropriate to test overall fit. Yet it cannot be assumed that the data indeed fit the model after deletion of misfitting persons and items. Gustafsson (1980) believed that this practice is likely to result in spurious fit, and gave a number of reasons why it should be discouraged. Primary among these is that the statistics may fail to detect deviation from the model, either through lack of sensitivity to certain violations of the model assumptions or through the effect of "trading relationships" between different violations.

Despite warnings from Gustafsson and others, there appears to have been little empirical investigation of the consequences of using person and item fit statistics to test and obtain overall fit. Few monte carlo studies of the behavior of these statistics have been reported. The present paper reports an investigation of some commonly used person and item fit statistics for the one-parameter model, examining their sensitivity to guessing, heterogeneity in discrimination parameters, and multidimensionality.

The fit statistics investigated were the mean-square residual for persons and items (Wright & Stone, 1979), the total-t for persons and items (Wright et al., 1979), and the between-t for items (Wright et al., 1979). Rogers (1983), in a review of fit statistics for item response models, pointed out the theoretical problems associated with each. Because of their weak theoretical bases, justification for the use of these statistics can be found only by examination of their behavior under known conditions.

Method

Description of the Statistics

The mean-square residual. The mean-square residual is based on the standardized residual

$$z_{ij} = \frac{(x_{ij} - p_{ij})}{[p_{ij}(1 - p_{ij})]} , \quad (1)$$

where x_{ij} is the dichotomous response of person i to item j, and

p_{ij} is the probability of a correct response for person i to item j, obtained from the model equation using estimates of the person and item parameters.

This statistic may be squared and summed over persons to give a measure of item fit, or summed over items to measure person fit. The simple z uses the normal approximation to the binomial; it thus follows that the sum of squared z scores has an approximate chi-square distribution. Dividing this χ^2 statistic by the degrees of freedom associated with its distribution gives the mean-square residual

$$\sum z_{ij}^2/(k-1) \quad (2)$$

for person fit, where k is the number of items, and

$$\sum z_{ij}^2/(N-1) \quad (3)$$

for item fit, where N is the number of persons. It is claimed that these statistics have F distributions with $(k-1, \infty)$ and $(N-1, \infty)$ degrees of freedom, respectively (Wright & Stone, 1979). A log transformation is then applied to produce a t statistic which is said to have an approximately standard normal distribution (Wright & Stone, 1979).

The major flaw of the mean-square residual is that it uses an approximation—the normal to the binomial—which is valid only for large sample sizes. In this application the residual is based on a single observation which can only take values of 0 or 1; its probability distribution cannot be reasonably approximated by a normal curve. Given this fundamental flaw, it is unlikely that the transformed statistic is normally distributed. Use of tabulated z values as critical values for the identification of misfitting persons and items must therefore be considered suspect.

The total-t. Wright et al. (1979) claimed to fortify the standardized residual against off-target data by weighting each squared residual by the information it contains. The statistic produced is

$$v = \frac{\sum(x_{ij} - p_{ij})^2}{\sum p_{ij}(1 - p_{ij})} , \quad (4)$$

with summation over j for person fit and i for item

fit. This statistic is called the total-t fit statistic. The total-t is incorporated into the BICAL program by Wright et al. (1979). The authors' description of their procedure is that "the sum of squared differences $(x-p)^2$ is . . . divided by its model expectation $p(1-p)$ to form a mean square statistic" (p. 12). Its distribution, however, was not discussed. In fact, the statistic appears to be a ratio of variance estimates; if the estimates were independent, this ratio would have an F distribution. But the estimates are not independent, and consequently the distribution of the statistic is not known. Its derivation and justification are unclear. Wright et al. converted this v statistic, by means of a cube-root transformation, to a statistic which is said to have an asymptotic standard normal distribution, at least theoretically. They remarked that in practice this is not always the case, and recommended that items and persons with a t value exceeding 2.0 be examined for response irregularities.

Like the mean-square residual, the total-t is based on dubious distributional assumptions. That the statistic has a distribution at least approximated by the one claimed and that the cutoff value of 2.0 yields a valid indicator of person or item misfit must be demonstrated empirically, since it cannot be proven on theoretical grounds.

The between-t. Wright et al. (1979) arranged persons into score groups before computing resid uals, producing the statistic called the between-t fit statistic, given by the formula

$$v_{Bj} = \sum_{r=1}^{m} \frac{\left(x_{rj} - \sum_{i \in r} n_i p_{ij}\right)^2}{\sum_{i \in r} n_i p_{ij}(1-p_{ij})} [k/(m-1)(k-1)] \quad , \quad (5)$$

where n_i is the number of persons with total score i in score group r,
p_{ij} is the probability of persons with total score i answering item j correctly, and
m is the number of score groups.

This statistic is said to be a mean-square; a cube-root transformation is applied and a t statistic, claimed to have an asymptotic standard normal distribution, is obtained.

The rationale provided for the between-t develops from the parameter invariance which is characteristic of item response models. For the one-parameter model, estimates obtained in different score groups should be the same as those obtained in the whole sample; consequently, the observed number of successes in each score group can be compared with the number predicted using the total sample parameter estimates.

The between-t is also calculated in the BICAL program, using six score groups. However, its actual formula in the program omits the factor $k/(k-1)$. Because of this omission, the degrees of freedom associated with the chi-square statistic are $(m-1)$.

Research Design

To investigate the sensitivity of the above person and item fit statistics to violations of the one-parameter model assumptions, a monte carlo study was designed. Each statistic was examined for sensitivity to guessing, heterogeneity in discrimination parameters, and multidimensionality.

Data generation. Data were generated using the multidimensional logistic item response model described by Hattie (1984) and McDonald (1979) and represented by the equation

$$p_j(\theta_1, \ldots, \theta_k) = c_j + (1-c_j) \frac{1}{1+\exp[-Da_j(\sum a_{j\ell}\theta_\ell - b_j)]} \quad , \quad (6)$$

where $\theta_1, \ldots, \theta_k$ are the k latent traits,
$\sum a_{j\ell}\theta_\ell$ is the θ parameter across the ℓ dimensions,
$a_{j\ell}$ is the discrimination parameter for item j on dimension ℓ ($\ell = 1, \ldots, k$),
b_j is the difficulty parameter for item j, and
c_j is the guessing parameter for item j.

The test length was 15 or 16 items in each data simulation. The 16th item was added to provide an item known to misfit the one-parameter model: All responses to this item were randomly generated. Sample size was 500 or 525. The larger sample contained 25 persons simulated to be misfitting; these persons were assigned incorrect responses to

the five easiest items of the 15-item test and correct responses to the five most difficult, with random responses for the five items of intermediate difficulty. Samples of this size are large enough to permit reasonably stable parameter estimates for the one-parameter model (Swaminathan & Gifford, 1979).

Sensitivity to multidimensionality was assessed by generating data to reflect one or two dimensions. For the one-dimensional case, two sets of discrimination parameters were used. In the first set, all discriminations were set equal to 1, in accordance with the Rasch model. In the second set, values of .6, 1.0, and 1.4 were chosen, with equal numbers of items at each value. This choice of discrimination parameters reflects empirical values, which typically fall between .5 and 2.0 (Ree, 1979). For the two-dimensional case, correlations between the dimensions of .1 and .5 were chosen to represent differing "degrees" of multidimensionality. Orthogonal dimensions were not considered because they generally do not occur in ability tests. Discrimination parameters were then calculated in the manner described by Hattie (1984) to produce the desired correlation between dimensions.

Difficulty parameters were in the range $[-2,2]$, with three items each at -2.0, -1.0, 0.0, 1.0, and 2.0. Two levels of guessing were chosen: (1) all zero, as required by the one- and two-parameter models, and (2) all .15, simulating the occurrence of guessing on a five-option multiple-choice test with reasonably attractive distractors. θ parameters were drawn randomly from a standard normal distribution. For the two-dimensional datasets, two θ parameters were drawn.

The three factors (dimensionality, discrimination, and guessing) were completely crossed to give eight combinations. For each of these combinations, which will be referred to as generating models, data matrices of three sizes were generated: 500 persons $\times$ 15 items; 525 persons $\times$ 15 items; and 500 persons $\times$ 16 items. Finally, to ensure stability of results, 75 replications were made of each model, resulting in the production of 600 datasets in all.

Two programs, BICAL (Wright et al., 1979) and NOHARM (Fraser, 1980), were used to fit the one-parameter model to the data. BICAL fits the one-parameter model by means of unconditional maximum likelihood estimation procedures in which person and item parameters are estimated simultaneously. NOHARM fits the unidimensional normal ogive model by finding the best approximation to it in terms of the Hermite-Tchebycheff polynomial series, using harmonic analysis (see McDonald, 1980, for details). NOHARM requires the assumption of a normal distribution of ability.

The NOHARM procedure was designed only to estimate item parameters, and makes no attempt to estimate θ. For the purposes of the present study, it was necessary to modify the program to obtain θ estimates so that the fit statistics could be calculated. θ estimates were obtained by converting number-correct scores to standardized (z) scores. (This approach seemed reasonable because the average correlation between number-correct scores and one-parameter θ estimates across 75 datasets exceeded .99.)

Both programs were used to fit the one-parameter model to each dataset so that differences in the rates of item and person rejection could be observed. This comparison enabled an assessment of the validity of the recommended cutoff values independent of the program under which they were determined.

Critical values for each statistic were chosen on the bases of theoretical expectation and common usage. Items or persons with values of the statistics exceeding the critical values were deemed to be misfitting. For both the total-t for persons (PTF) and items (ITF), critical values of 1.5 and 2.0 were used, in accordance with the recommendations of Wright et al. (1979). If the distributions of the statistics are normal as claimed, a value of 2.0 should yield a type I error rate of around 5%. For the mean-square residual for persons (PMSQ) and items (IMSQ), a cutoff value of 3.0 was used, as suggested by Wright and Stone (1979). If the distribution is approximately normal, this value should give a type I error rate of less than 1%. The value of 3.0 was apparently chosen, however, because a smaller z value results in too many rejections. Additionally, critical values from the F distribution were applied to the PMSQ and IMSQ before the log transformation to the normal distribution was per-

formed. This was done as a check on the distribution claimed for these statistics and on the validity of the transformation. For the untransformed PMSQ (RPMSQ), the critical value used was the 95th percentile of the F distribution with 14 and ∞ degrees of freedom, a value of 1.7. For the untransformed IMSQ (RIMSQ), the theoretical 5% cutoff point was $F_{.05}$ (499, ∞) = 1.1.

In the case of the between-t for items (IBF), critical values of 1.5 and 2.0, again as suggested by Wright et al. (1979), were used, along with chi-square distribution critical values for the statistic before transformation (RIBF). Here the critical values were $\chi^2(5,.05) = 11.1$ and $\chi^2(5,.01) = 15.1$. Degrees of freedom were 5 because the IBF was calculated, following Wright et al. (1979), on the basis of six groups.

Under BICAL, only the PTF, ITF, and IBF were calculated; under NOHARM, all fit statistics were calculated. Table 1 provides a summary of the fit statistics used, their acronyms, cutoff values applied, and the estimation procedures under which they were calculated.

Analysis

For each dataset of the 500 × 15 size, the number of persons and the number of items exceeding the critical values of the statistics were calculated. These numbers became the dependent variables of analysis. Under BICAL, 6 variables, corresponding to the critical values described above, were obtained. Under NOHARM, 12 dependent variables were obtained (see Table 1).

Multivariate analysis of variance (MANOVA) was then performed to examine the main effects of guessing, discrimination, and multidimensionality (and interaction of these factors) on the 6 variables obtained from BICAL and, separately, on the 12 variables from NOHARM. The MANOVA results were not expected to be of interest in themselves, but were used to isolate significant effects for closer study. (For this reason, the dimensionality effect was separated into 2 two-level comparisons rather than taken as a single three-level factor.) Where significant differences were found, univariate F values were examined to determine which of the statistics were sensitive to violations of the model assumptions. A significance level of .0001 was chosen for the univariate analyses to ensure that the overall error rate resulting from the large number of comparisons was no larger than .01. This more stringent alpha level was also expected to reduce the likelihood of unimportant differences

Table 1
Statistics Used, Their Acronyms and Cutoff Values, and the Estimation Procedures Under Which They Were Calculated

Statistic	Acronym	Cutoff Value	BICAL	NOHARM
Total-t person fit	PTF	1.5	X	X
		2.0	X	X
Mean square residual person fit:				
transformed to normal	PMSQ	3.0		X
untransformed, F distribution	RPMSQ	1.8		X
Total-t item fit	ITF	1.5	X	X
		2.0	X	X
Mean-square residual item fit:				
transformed to normal	IMSQ	3.0		X
untransformed, F distribution	RIMSQ	1.0		X
Between-t item fit:				
transformed to normal	IBF	1.5	X	X
		2.0	X	X
untransformed, χ^2 distribution	RIBF	11.1		X
		15.1		X

reaching significance due to the large number of observations per cell. Mean rejection rates were then considered to assess the meaningfulness of the statistical significance and the consequent utility of the statistic.

For each dataset of the 525×15 size, person fit values for the 25 misfitting persons were calculated. Under BICAL, only the PTF was computed; under NOHARM, the PTF and PMSQ were computed. The proportion of these 25 persons with values of the statistics exceeding the critical values was obtained in order to assess the consistency with which these misfitting persons were detected.

For each dataset of the 500×16 size, values of the item fit statistics for the misfitting item were obtained. Under BICAL, values of the ITF and IBF were calculated; under NOHARM, values of the ITF, IBF, and IMSQ were obtained. For each statistic with each generating model, the percentage of replications in which the value of the statistic exceeded the critical value was calculated, in order to determine the consistency with which the misfitting item was detected by the statistic.

Results

Results of the multivariate analysis of variance are presented in Table 2. As the table shows, significant differences were found for all contrasts specified. Univariate F values for each fit statistic were then examined; results are reported by statistic below. Results within each statistic are prefaced by the size of the data matrix on which they are based.

In each analysis, it was expected that if the statistic were doing its job—detecting misfit—many more persons or items would be identified as misfitting under violations of the model assumptions than would be identified when the model fitted the data. In the following sections, sensitivity of the statistic to a violation of the model assumptions is defined as a statistically significant increase in the number of misfitting persons or items.

The Total-t for Persons

500×15. Examination of univariate F values showed that under the conditions simulated, the PTF was insensitive to heterogeneity in discrimination parameters and to multidimensionality when the dimensions were moderately correlated. It was, however, sensitive to guessing and to multidimensionality when the correlation between dimensions was slight. Interaction between guessing and dimensionality makes interpretation of these main effects difficult: When there was no guessing, more persons were rejected if the data were two-dimensional than if one-dimensional, but when guessing was present, the number of persons rejected increased for one-dimensional data and de-

Table 2
Multivariate F Values for the Effects of Guessing, Unequal Discriminations, and Multidimensionality Under BICAL and NOHARM

Effect	BICAL (df=6,585)	NOHARM (df=12,585)
Guessing	115.42*	177.49*
Discriminations	42.95*	27.98*
Dimensionality I [1 vs 2(.1)]	67.03*	89.92*
Dimensionality II [1 vs 2(.5)]	20.56*	9.26*
Guessing x Discriminations	37.52*	16.56*
Guessing x Dimensionality I	31.78*	41.71*
Guessing x Dimensionality II	8.58*	4.21*

*Statistically significant at $p < .001$.

Table 3
Mean Number of Persons Rejected by PTF and PMSQ, Under BICAL and NOHARM, for Each Generating Model

Generating Model				Mean Number of Rejections	
Number of Dimensions	Correlation	Discrimination	Guessing	BICAL	NOHARM
PTF with a Cutoff of 1.5					
1		all 1	.00	26	26
1		all 1	.15	31	30
1		mixed	.00	26	22
1		mixed	.15	29	28
2	.1		.00	31	29
2	.1		.15	30	29
2	.5		.00	28	23
2	.5		.15	30	30
PMSQ with a Cutoff of 3.0					
1		all 1	.00		31
1		all 1	.15		33
1		mixed	.00		30
1		mixed	.15		27
2	.1		.00		29
2	.1		.15		18
2	.5		.00		32
2	.5		.15		32

creased for two-dimensional data, resulting in nearly equal numbers of rejected persons.

Examination of mean numbers of persons rejected, presented in Table 3, showed that when the model fit the data, approximately 5% of examinees were rejected, as would be expected if the claim of a normal distribution for the PTF were true. (In Tables 3 and 4, figures are rounded to the nearest integer.) However, even under the most extreme violation of the one-parameter model assumptions, the rejection rate increased by no more than 2%. The statistic was not useful as an indicator of overall misfit of the model.

525 × 15. Although it did not warn of model inappropriateness, the PTF did detect extreme deviation within datasets under the one-parameter model. Regardless of the nature of the remaining data, all 25 persons whose responses were generated to be extreme were consistently detected.

The Mean-Square Residual for Persons

500 × 15. The PMSQ was insensitive both to guessing and to heterogeneity in discrimination parameters, rejecting fewer persons when these conditions occurred than when they did not. It was also insensitive to multidimensionality. Table 3 reports the mean numbers of persons rejected under each condition. Rejection rates were similar to those produced by the total-*t* for persons; approximately 6% were rejected when the model fit the data, but only 1% more were rejected under the most extreme violation of the model. Like the total-*t*, the mean-square residual for persons does not provide any indication of model misfit.

525 × 15. Despite its failure to detect overall misfit, the PMSQ did consistently detect the 25 misfitting persons in all datasets.

The Total-*t* for Items

500 × 15. As a fit statistic for the one-parameter model, the ITF was found to be sensitive to guessing and inequality of discrimination parameters but not to multidimensionality. Interaction between guessing and discrimination occurred, such that the pres-

ence of guessing and unequal discrimination parameters in combination produced fewer misfitting items than produced by the presence of guessing alone.

Examination of mean rejection rates (Table 4) showed that the sensitivity of the ITF was insufficient to make it a useful indicator of either guessing or heterogeneity in discrimination parameters. Over the 75 replications of datasets which fit the one-parameter model, an average of 0 items was detected as misfitting—an acceptable result if the statistic were normally distributed. However, at most 2 items were rejected when discriminations varied, and an average of 1 item was rejected when guessing was present.

500 × 16. The randomly answered item was detected by the ITF in over 95% of replications for each generating model, except where both guessing and multidimensionality with slightly correlated dimensions were present in the remaining data; in that case the detection rate was reduced to 80%.

The Mean-Square Residual for Items

500 × 15. The IMSQ was insensitive to all violations of the model assumptions—variation in item discrimination parameters, guessing, and multidimensionality. Table 4 reports mean rejection rates. As shown, when the model was appropriate for the data the IMSQ rejected many more items than expected, but rejection rates decreased under violations of the model assumptions. The RIMSQ yielded a similar result.

500 × 16. The randomly answered item was consistently detected when found in one- or two-parameter datasets or in datasets which were multidimensional with moderately correlated dimensions. When there was guessing or multidimensionality with only slight correlation between dimensions, the item was detected in 50% to 70% of replications. When both occurred, the item was not detected.

The Between-*t* for Items

500 × 15. Examination of univariate *F* values showed the IBF to be sensitive to guessing and heterogeneity in discrimination parameters, but not to multidimensionality. Interaction between guessing and discrimination occurred, with the combined presence of guessing and unequal discriminations resulting in the rejection of fewer items than the presence of guessing alone.

Consideration of rejection rates (Table 4) shows that with a cutoff of 2.0, the IBF rejected consistently more items under NOHARM than under BICAL. Use of the RIBF with an *F*-distribution cutoff at the 95th percentile produced slightly better results for NOHARM; it was seen that the RIBF, though still rejecting more items than expected when the model fit the data, was a more useful indicator of guessing and, to a lesser extent, heterogeneity in discrimination parameters than either of the other two item fit statistics examined. For BICAL, a similar result was observed for the IBF with a cutoff of 2.0.

500 × 16. The presence of a randomly answered item was detected by the IBF more than 95% of the time except when both guessing and multidimensionality with nearly orthogonal dimensions were present in the remaining data. For this case, the detection rate was 80%. Similar, but slightly smaller, detection rates were found for the RIBF.

Discussion

It is clear from these results that the person fit statistics examined lack power to detect deviation from the one-parameter model. Although the mean-square residual and the total-*t* yielded rejection rates of approximately 5% for one-parameter model data, as expected under the assumption of a normal distribution, rejection rates increased by no more than 2% under any violation of the model assumptions. It appears that deviation of the degree resulting when the model is inadequate for the data as a whole is not extreme enough to cause large-scale rejection of persons. Misfitting persons were consistently detected only when these persons were extreme relative to the remaining data. It is therefore unlikely that "sleepers," "fumblers," and "plodders" (Wright & Stone, 1979, p. 171) will be consistently identified by the person fit statistics.

Table 4
Mean Number of Items Rejected by ITF, IMSQ, and IBF, Under BICAL and NOHARM, for Each Generating Model

Generating Model				Mean Number of Rejections	
Number of Dimensions	Correlation	Discrimination	Guessing	BICAL	NOHARM
ITF with a Cutoff of 1.5					
1		all 1	.00	0	0
1		all 1	.15	1	1
1		mixed	.00	2	1
1		mixed	.15	1	1
2	.1		.00	0	0
2	.1		.15	1	1
2	.5		.00	0	0
2	.5		.15	1	1
IMSQ with a Cutoff of 3.0					
1		all 1	.00		4
1		all 1	.15		3
1		mixed	.00		4
1		mixed	.15		2
2	.1		.00		1
2	.1		.15		0
2	.5		.00		4
2	.5		.15		3
IBF with a Cutoff of 2.0					
1		all 1	.00	1	4
1		all 1	.15	4	8
1		mixed	.00	4	6
1		mixed	.15	4	7
2	.1		.00	1	4
2	.1		.15	2	6
2	.5		.00	1	4
2	.5		.15	4	8

Such persons will seldom be as aberrant as is necessary for detection. The claims of Wright and Stone (1979) are thus not substantiated.

The corresponding item fit statistics showed a similar pattern of results. For the total-*t* for items, the 5% rejection rate expected under the null hypothesis was not exceeded when the model fit the data, but for the mean-square residual for items, the assumption of a normal distribution did not appear to be reasonable. For both statistics, increases in rejection rates due to overall misfit were not large enough to suggest model inappropriateness. Misfitting items were detected by the ITF and IMSQ only when extreme in their context.

In the present study, the between-*t* for items proved to be more sensitive to departures from the model. It was observed, however, that when both NOHARM and BICAL were used to fit the one-parameter model, consistently more items were rejected under NOHARM than under BICAL. This is not an indication that NOHARM fits the model less accurately than does BICAL, but rather is a result of the method used in BICAL to obtain the maximum likelihood parameter estimates. An iterative procedure is used in which the criterion for convergence is, in fact, the minimization of a function of the IBF. As a result, no other model-fitting technique will produce values of the IBF as low as does

BICAL. The critical value of 2.0 suggested by Wright et al. (1979) is therefore dependent on the use of the BICAL routine. The results also showed that the transformation used to produce a supposedly normally distributed form in fact reduced the usefulness of the between-*t*; better results were obtained for the statistic in its chi-square form. In spite of this improvement, however, use of a critical value from the chi-square distribution does not appear to be appropriate.

These findings are not unexpected, given the warnings of George (1979), Hambleton, Swaminathan, Cook, Eignor, and Gifford (1979), and others about the theoretical problems associated with these fit statistics. That the between-*t* has some value is supported by other simulation studies of its performance. Yen (1981) found that the between-*t* performed very similarly to the Yen Q_1 statistic, which is derived from contingency table data of score group by dichotomous response. Although the between-*t* differs in derivation, it can be shown to differ mathematically from Q_1 only by a constant and by the subtraction of a term in the denominator, this term being the variance of predicted probability of success within score groups. When all members of a group have the same ability, the variance will be zero. In Yen's study, it was very small. Yen concluded from her results that Q_1 (and thus the between-*t*) has an approximately chi-square distribution, although the mean value of the statistic under the null hypothesis was consistently higher than expectation. The results reported are, however, based on only one replication of the data.

McKinley and Mills (1985), in a more extensive simulation study, also found that the between-*t* performed similarly to Q_1, and was in addition able to detect model misfit due to guessing and unequal discriminations and, to a lesser extent, multidimensionality. As in the present study, the between-*t* identified too many items as misfitting under the null hypothesis. These results suggest that the between-*t* will be of use in investigations of model fit, although strict adherence to a tabulated chi-square critical value seems unwarranted. As Wright and Panchapakesan (1969) advised, instead of deleting items exceeding a critical value, it is better to examine items for which the value of the statistic is large. With the caveat that dimensionality be further investigated by other means, Yen's suggestion that the statistic be used to compare the fit of the one-, two-, and three-parameter models is endorsed.

Conclusions

The results reported in this study suggest that the mean-square residual and total-*t* person and item fit statistics will contribute very little to an investigation of one-parameter model fit and, indeed, if relied on solely, will provide incorrect information about the appropriateness of the model. Given these results, based on 75 replications of 8 different types of data, it must be concluded that these fit statistics as they are currently used are inadequate measures of fit. The between-*t* item fit statistic appears to have some value, but the critical value commonly used may need reconsideration.

The present study clearly demonstrates that use of person and item fit statistics to assess and obtain overall fit to the one-parameter model can lead to acceptance of the model when it is grossly inappropriate. The exclusion of persons and items identified as misfitting by the statistics provides no assurance of fit of the remaining data to the model.

References

Fraser, C. (1980). *NOHARM: A program for estimating parameters of the normal ogive by harmonic analysis robust method*. Unpublished manuscript, University of Toronto.

George, A. A. (1979, April). *Theoretical and practical consequences of the use of standardised residuals as Rasch model fit statistics*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Gustafsson, J. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 33*, 205–233.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1979). Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research, 48*, 467–510.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49–78.

McDonald, R. P. (1979). The structural analysis of multivariate data: A sketch of general theory. *Multivariate Behavioral Research, 14*, 21–38.

McDonald, R. P. (1980, October). *Fitting a latent trait model by the analysis of covariance structures*. Paper presented at the annual convention of the Northeastern Educational Research Association.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9*, 49–58.

Ree, M. (1979). Estimating item characteristic curves. *Applied Psychological Measurement, 3*, 371–375.

Rogers, H. J. (1983). *An investigation of some fit statistics for latent trait models*. Unpublished master's thesis, University of New England, Australia.

Swaminathan, H., & Gifford, J. A. (1979). *Estimation of parameters in the three-parameter latent trait model* (Laboratory of Psychometric and Evaluative Research Report No. 90). Amherst MA: University of Massachusetts, School of Education.

Traub, R. E., & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. *Review of Research in Education, 9*, 377–435.

Wright, B. D., Mead, R. J., & Bell, S. R. (1979). *BICAL: Calibrating items with the Rasch model* (Research Memorandum 23-B). 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.

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.

Author's Address

Send requests for reprints or further information to H. Jane Rogers, School of Education, University of Massachusetts, Amherst MA 01003, U.S.A.