Using Simulation Results to Choose a Latent Trait Model

Wendy M. Yen
CTB/McGraw-Hill

A latent trait model goodness-of-fit statistic was defined, and its relationships to several other commonly used fit statistics were described. Simulation data were used to examine the behavior of these fit statistics under conditions similar to those found with real data. The simulation data were generated for 36 pseudo-items and 1,000 simulees using three-, two-, and one-parameter logistic latent trait models. The data were analyzed using three-, two-, and one-parameter models. Between-model comparisons were made of the fit statistics, trait estimates, and item parameter estimates. The three generating models produced clearly different patterns of results. The simulation results were compared to results for real data involving seventh- and eighth-grade students' performance on eight achievement tests. The achievement test results appeared most similar to the simulation results based on data generated with the three-parameter model. Some practical problems that can result from using an inappropriate model with multiple-choice tests are discussed.

There are three latent trait models that are in common use, and they differ in terms of the number of item parameters they estimate to describe item characteristic functions. The three-parameter (3-PAR) model estimates item difficulties, discriminations, and lower asymptotes. The two-parameter (2-PAR) model estimates item difficulties and discriminations but assumes that all lower asymptotes are zero. The one-parameter (1-PAR), or Rasch, model estimates item difficulties but assumes that all item discriminations are constant and all lower asymptotes are zero. (See Allen & Yen, 1979; Lord & Novick, 1968; or *Journal of Educational Measurement,* Summer 1977, for a further discussion of these models.) When items have a multiple-choice format and it is possible for examinees of low ability to get an item correct through lucky guessing, the 2-PAR and 1-PAR models appear a priori to be inappropriate. However, in striving for simplicity, the researcher may be tempted to use the 2-PAR or 1-PAR model and hope that the inaccuracies that are introduced are unimportant. The researcher may turn to a statistical goodness-of-fit test to gauge the degree of these inaccuracies.

This paper examines the behavior of a fit statistic, called Q_1, which is similar to fit statistics that are commonly used for latent trait models. In the following sections (1) Q_1 is specified, (2) informal theoretical justification is given for using a chi-square distribution as an approximation to the distri-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 2, Spring 1981, pp. 245–262

0146-6216/81/020245-18$1.90

bution of Q_1 when the latent trait model being tested is true, and (3) the relationships of Q_1 to some other measures of fit are defined. A simulation study is conducted to examine the behavior of Q_1 and two other fit statistics under conditions similar to those found with real data. The usefulness of Q_1 in choosing the most appropriate latent trait model for a set of data is examined; and an additional graphical method of examining the suitability of the three latent trait models also is presented. Finally, real-data examples are presented of how the fit measure, Q_1, and the graphical comparisons were used in helping choose an appropriate latent trait model for several achievement tests.

Measures of Fit in Latent Trait Theory

Q_1: Definition

Q_1 is a fit statistic that is similar to the fit measures commonly used to examine latent trait models. To calculate Q_1, examinees are rank ordered on the basis of their trait estimates and then divided into 10 cells with approximately equal numbers of examinees per cell. The value of the fit statistic for item i is

$$Q_{1i} = \sum_{j=1}^{10} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$+ \sum_{j=1}^{10} \frac{N_j[(1 - O_{ij}) - (1 - E_{ij})]^2}{1 - E_{ij}}$$

$$= \sum_{j=1}^{10} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})}, \qquad [1]$$

where

N_j is the number of examinees in cell j,
O_{ij} is the observed proportion of examinees in cell j that passes item i, and
E_{ij} is the predicted proportion of examinees in cell j that passes item i.

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

where $\hat{P}_i(\hat{\theta}_k)$ is the item characteristic function for item i, which is evaluated using the trait estimate for examinee k, $\hat{\theta}_k$, and the item parameters estimated for item i. The summation is taken over examinees in cell j.

$$\hat{P}_i(\hat{O}_k) = \hat{c}_i + \frac{1 - \hat{c}_i}{1 + e^{-1.7\hat{a}_i(\hat{O}_k - \hat{b}_i)}}. \qquad [3]$$

The terms $\hat{a}_i$, $\hat{b}_i$, and $\hat{c}_i$ are, respectively, the estimated discrimination, difficulty, and lower asymptote for item i. For the 2-PAR model, $\hat{c}_i$ is fixed at 0 for all items; and for the 1-PAR model, $\hat{c}_i$ is fixed at 0 and $\hat{a}_i$ is fixed at a constant for all items.

The Distribution of Q_1

It is now argued informally that when the model being examined is true, Q_1 should be distributed *approximately* as a chi-square variable with $10-m$ degrees of freedom, where m is the number of item parameters estimated by the latent trait model.

Pearson chi-squares. The general form of the Pearson chi-square statistic is

$$\chi^2 = \sum_{j=1}^{J} \frac{(f_{oj} - f_{ej})^2}{f_{ej}}, \qquad [4]$$

where f_{oj} and f_{ej} are observed and expected frequencies of observations falling in cell j. The chi-square distribution is an approximation to the multinomial distribution, and Hays (1973) states:

> In the use of the Pearson χ^2 statistic to approximate multinomial probabilities, it *must* be true that:
> 1. each and every sample observation falls into one and only one category or class interval [cell];
> 2. the outcomes for the N respective observations in the sample are independent;
> 3. sample N is large. (p. 722)

Q_1: *Observations.* In calculating Q_1, the observations are the item scores for the examinees. Each observation is classified into one of 20 cells based on the value of $\hat{\theta}_k$ and on whether the examinee has passed or failed the item. The assumptions are made that (1) the sample size $\Sigma_{j=1}^{J} N_j = N$ is large and that (2) under the null hypothesis (that the model being tested is true), the observations are locally independent (Allen & Yen, 1979, p. 254).

Q_1: *Degrees of freedom.* The degrees of freedom are calculated as the number of independent data points (i.e., observed proportions) used in Q_1 minus the number of independent parameters estimated from the data points to produce E_{ij}. Because the observed proportion passing an item (O_{ij}) and the observed proportion failing ($1-O_{ij}$) are obviously linearly related, there are at most 10 independent observed proportions in Q_1. The overall observed proportion passing the item is $\Sigma_{j=1}^{10} N_j O_{ij}/N = P_i$; if it is assumed that P_i is a fixed value, then one of the O_{ij} is dependent on the other 9. However, it only makes sense to assume P_i is fixed if it is required that $\Sigma_{j=1}^{10} N_j E_{ij}/N = P_i$. Because this requirement is not made for the latent trait models, it is assumed that there are 10 independent data points in Q_1.

Degrees of freedom are subtracted to reflect the extent to which the E_{ij} values are calculated from or are dependent on the values of O_{ij} (Lancaster, 1969, pp. 136, 142-150). E_{ij} is a function of the estimated item parameters and the estimated trait values. The item parameters and trait values are estimated from the examinees' response vectors and not from the O_{ij} values. This means that the estimation process occurs in a sample space that differs from the sample space in which the goodness-of-fit test is carried out; this can cause a distortion in the chi-square approximation to the distribution of Q_1 (Kendall & Stuart, 1967, pp. 428-430; Lancaster, 1969, p. 171). It can be reasoned that because the values of O_{ij} represent fairly well the complete item characteristic function for most items, the item parameter estimates are highly dependent on the values of O_{ij}, and one degree of freedom should be subtracted for each estimated item parameter. On the other hand, examinees' trait estimates are based on all the test items, and for long tests each item contributes a small proportion of the information about each $\hat{\theta}_k$. Therefore, for any given item there should be a negligible loss of degrees of freedom due to the estimation of the trait values. It should be noted that the determination of the cell boundaries for Q_1 is dependent on the observed distribution of estimated trait values. This fact may affect the distribution of Q_1 and its degrees of freedom.

In any event, the argument is made that when the model being tested is true, Q_1 will have *approximately* a chi-square distribution with $10-m$ degrees of freedom, where there are m item parameters estimated for the latent trait model. The validity of this argument will be examined with simulation data.

The Relationship of Q_1 to Other Measures of Fit

Bock's chi-square. Bock (1972) introduced a latent trait model that estimates item parameters for all the answer choices given with the item, not just the correct answer choice. He used a fit measure that examines the suitability of this model. His model can be restricted to deal with only the characteristic function of the correct answer and his fit measure compared to Q_1. With this restriction, Bock's fit measure equals Q_1 except for two factors:

1. Examinees are grouped into J cells, but J does not have to equal 10. J does happen to equal 10 in an example given in Bock's paper (pp. 44-45).
2. Bock uses $E_{ij} = P_i(\hat{\theta}_{med\text{-}j})$, where $\hat{\theta}_{med\text{-}j}$ is the median of the $\hat{\theta}$ values for examinees in cell j. If the values of $\hat{\theta}_k$ are homogenous within each cell, Bock's method of obtaining E_{ij} will produce a fit measure very similar to Q_1.

To test the significance of an item's fit, Bock used a chi-square distribution with $J-m$ degrees of freedom, where m independent item parameters have been estimated.

Q_2. Wright and Panchapakesan (1969) proposed the following fit measure for the 1-PAR model:

$$Q_{2i} = \sum_{\substack{j=1 \\ N_j \neq 0}}^{J} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij})}, \qquad [5]$$

where each cell contains examinees with the same number-correct score and examinees with zero and perfect scores are excluded. Because the number-correct score is a sufficient statistic for θ for the 1-PAR model, each cell contains examinees with the same $\hat{\theta}_k$. Wright and Panchapakesan stated that the distribution of Q_2 is approximately chi-square with $J-1$ degrees of freedom, where J cells have $N_j \neq 0$. Notice that Q_2 is very similar to Q_1, except that Q_2 involves more cells and constant trait estimates within each cell. Q_2 was used in selecting items for the KeyMath Diagnostic Arithmetic Test (Connolly, Nachtman, & Pritchett, 1976) and the Woodcock Reading Mastery Test (Woodcock, 1973). It appears that Whitely and Dawis (1974) used Q_2 in their evaluation of the 1-PAR model; whereas Rentz and Bashaw (1975) used a "Mean Square Fit," which is Q_2 divided by the number of score groups with $N_j \neq 0$.

Q_3. The BICAL computer program provided by Wright and Mead (1977) produces another fit measure for the 1-PAR model.

$$Q_{3i} = \frac{1}{J} \sum_{j=1}^{J} \frac{N_j^2(O_{ij} - E_{ij})^2}{\sum_{k \in j} \hat{P}_i(\hat{\theta}_k)[1 - \hat{P}_i(\hat{\theta}_k)]}$$

$$= \frac{1}{J} \sum_{j=1}^{J} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij}) - \sigma^2_{P_j}}, \qquad [6]$$

where

$$\sigma^2_{P_j} = \frac{1}{N_j} \sum_{k \in j}^{N_j} [\hat{P}_i(\hat{O}_k) - E_{ij}]^2 \ . \qquad [7]$$

$\sigma^2_{P_j}$ is the variance within cell j of the predicted proportion passing the item. In calculating Q_3, examinees are grouped into cells with as close to a uniform distribution as possible, subject to having each number-correct score appear in only one cell, with the recommendation that $N_j \geq 15$ and with the requirement that $J \leq 6$. Notice that Q_3 is very similar to Q_1 with the important differences being the number of cells, the subtraction of $\sigma^2_{P_j}$ from the denominator, and the constant $1/J$.

The denominator of Q_3 (in the first line of Equation 6) is the variance of the compound binomial distribution (Kendall & Stuart, 1969, pp. 126-127; Lord & Novick, 1968, p. 385). It is the theoretical variance of N_jO_{ij} when the N_j examinees in cell j can have different values of $\hat{P}_i(\hat{\theta}_k)$. If the variance in values of $\hat{\theta}_k$ within cell j is small, and therefore $\sigma^2_{P_j}$ is small, Q_1 and JQ_3 will be very similar.

Q_4. Elliott, Murray, and Saunders (1977) used the following statistic for measuring the fit of the 1-PAR model in choosing items for the British Ability Scales.

$$Q_{4i} = \sum_{\substack{j=1 \\ N_jE_{ij} \geq 5}}^{J} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}} \ . \qquad [8]$$

Each cell contains examinees with the same number-correct score, with the condition that adjacent cells are pooled until $N_jE_{ij} \geq 5$. Examinees with zero and perfect scores are not included. Elliott et al. (1977) stated that Q_4 has a chi-square distribution with $J-1$ degrees of freedom, where there are J cells. Notice that Q_4 will have systematically lower values than Q_2; Q_4 essentially ignores the item scores that are incorrect (see Equation 1) and violates the requirement that "each and every sample observation falls into one and only one category or class interval [cell]."

Fit Measures Examined in this Paper

Thus, there are several fit measures that have been used in latent trait research studies that are fairly similar to Q_1. Q_2 and Q_4 group examinees by number-correct scores, which is suitable for examining the 1-PAR model but not suitable for examining the 2-PAR and 3-PAR models. Because this paper examines all three models, it was decided to modify Q_2 and Q_4, grouping examinees in the manner that was used for Q_1. When this is done, Q_2 becomes essentially equivalent to Q_1 or JQ_3, depending on the denominator that is used. It was also decided to use 10 cells for all the fit measures and to delete the constant $1/J$ from Q_3. Thus, comparisons would be made among three fit measures, Q_1, Q_3^*, and Q_4^*, with

$$Q^*_{3i} = \sum_{j=1}^{10} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}(1 - E_{ij}) - \sigma^2_{P_j}} \ , \qquad [9]$$

and

$$Q^*_{4i} = \sum_{j=1}^{10} \frac{N_j(O_{ij} - E_{ij})^2}{E_{ij}} \ . \qquad [10]$$

Method

Simulation Study

The trait and item parameters that were used in producing simulated response vectors are described as θ_k, a_i, b_i, and c_i. These values were prespecified and were *not* estimated from any data set. Trait and item parameters that were estimated from real or simulated response vectors are identified as $\hat{\theta}_k$, $\hat{a}_i$, $\hat{b}_i$, $\hat{c}_i$.

Item responses were generated to simulate the performance of students from two adjacent grades. Five hundred trait values (θ_k) from each of two normal distributions were generated using the IMSL normal random deviate generator GGNML (IMSL, 1979). One θ_k distribution had an observed mean of −.18 and a standard deviation of .95; the other θ_k distribution had an observed mean of .16 and a standard deviation of 1.02. The difference in means for these two groups is typical of the difference between Grades 7 and 8 in performance on the Comprehensive Tests of Basic Skills, Form S (CTB/McGraw-Hill, 1973). The two trait distributions were pooled to produce one distribution of 1,000 traits. Item discriminations (a_i), item difficulties (b_i), and item lower asymptotes (c_i), were specified to have unimodal, somewhat platykurtic distributions similar to those found with real data. The means and standard deviations of these distributions were $\mu_a = 1.15$, $\sigma_a = .40$, $\mu_b = -.05$, $\sigma_b = 1.15$, $\mu_c = .24$, and $\sigma_c = .04$. (None of the c_i values equaled zero.) Parameter sets for 36 items were created by randomly choosing without replacement from the parameter distributions an a_i, b_i, and c_i for each item. The resulting correlations of the parameters across items were $r_{ab} = .20$, $r_{ac} = -.23$, and $r_{bc} = .08$.

For the 3-PAR model the probability of simulee k passing item i, $P_i(\theta_k)$, was generated by evaluating the 3-PAR item characteristic function using the prespecified θ_k, a_i, b_i, and c_i. For each (i,k) pair, a random number from 0 to 1 was generated from a uniform distribution using IMSL subroutine GGUBS (IMSL, 1979). If the random number was less than $P_i(\theta_k)$, simulee k was said to pass item i; otherwise, simulee k failed item i. This item response generation procedure resulted in pass/fail response arrays based on the 3-PAR model for 36 items for 1,000 simulees.

Item responses also were generated using the 2-PAR and 1-PAR models. For the 2-PAR model, c_i was set equal to 0 for all items, and then the pass/fail responses were generated for the 36 items and 1,000 simulees. For the 1-PAR model, c_i was set equal to 0 and a_i was set equal to 1.0 for all items, and then the pass/fail responses were generated. Thus, there were three sets of item responses, one set generated with the 3-PAR model, another set generated with the 2-PAR model, and a third set generated with the 1-PAR model.

Each of the three sets of item responses was analyzed with each of the three latent trait estimating models, producing nine analyses. The item responses were analyzed with the LOGIST computer program (Wood, Wingersky, & Lord, 1976) to produce estimated trait values ($\hat{\theta}_k$) and item parameters ($\hat{a}_i$, $\hat{b}_i$, $\hat{c}_i$). When default options were used, LOGIST automatically estimated traits and item parameters using the 3-PAR model. To estimate traits and item parameters using the 2-PAR model, LOGIST was constrined to hold $\hat{c}_i = 0$ for all items. To estimate traits and item parameters using the 1-PAR model, LOGIST was constrained to hold $\hat{c}_i = 0$ and to hold $\hat{a}_i$ constant for all items. For all the estimating models, traits were not estimated for simulees with zero or perfect scores.

The fit statistics—Q_1, Q_1^*, and Q_1^*—were calculated for each item for each of the nine combinations of generating models and estimating models.

Achievement Test Study

To examine the extent to which the simulation results were helpful in choosing the most appropriate model for real data, eight tests from the Comprehensive Tests of Basic Skills, Form S, Level 3

(CTB/McGraw-Hill, 1973) were analyzed. The tests were Reading Vocabulary (40 items), Reading Comprehension (45 items), Spelling (30 items), Language Mechanics (20 items), Language Expression (35 items), Mathematics Computation (48 items), Mathematics Concepts (25 items), and Mathematics Applications (25 items). These are multiple-choice tests with four answer choices per item. Data for 999 seventh- and eighth-grade students were analyzed with the three latent trait models.

Results

Simulation Study

Q_1 and Q_3^* had correlations of 1.0 for each of the nine combinations of generating models and estimating models. Linear regression equations were calculated for predicting Q_3^* from Q_1 in each of the nine situations. The regression coefficients varied in value from 1.00 to 1.02 and the additive constants varied from −.06 to +.06. In order to have Q_3^* and Q_1 be so highly related, $\sigma^2_{P_j}$ must have been very small. This implies that Bock's fit measure would also have been very similar to Q_1. Because the results for Q_1 and Q_3^* were so similar, only the results for Q_1 are reported.

Table 1 contains the correlations taken over items of Q_1 and Q_4^*. In general, these two fit statistics led to different conclusions about which items best fit a model. Table 2 contains the means taken over the 36 items of the Q_1 and Q_4^* statistics. Notice that the Q_4^* statistic tended to be about half the size of the Q_1 statistic. This result is consistent with the formulas for Q_1 and Q_4^* (Equations 1 and 10).

The columns of mean fit values have very similar patterns for the data generated by the 3-PAR and by the 2-PAR models. When the data were generated by the 3-PAR model, the 2-PAR estimating model did a surprisingly good job of fitting the data. If a set of real data were being analyzed with all three models, it would be difficult to decide on the basis of the pattern of mean fit values whether the 3-PAR or 2-PAR model was more appropriate. It would be clear that the 1-PAR model was appropriate when the mean fit statistics had similar values for all three estimating models.

Significance tests were conducted to determine whether the fit statistics had chi-square distributions with $10-m$ degrees of freedom. These significance tests were traditional Pearson chi-square tests (Hays, 1973, pp. 725–727) constructed with six cells with equal numbers of fit statistics expected to fall in each cell. The expectations were generated using a chi-square distribution with $10-m$ degrees of freedom, where m item parameters had been estimated by the latent trait model. The significance probabilities of these tests appear in Table 3. (The significance probability is the probability of obtaining a distribution of fit statistics as extreme or more extreme than the observed distribution when the null hypothesis is true, i.e., the fit statistics do have a chi-square distribution with $10-m$ degrees of freedom.)

Given the probabilities in Table 3, it is very unlikely that the Q_4^* statistic has a chi-square distribution. Q_4^* took on values that were too small to be consistent with a chi-square distribution with

Table 1
Correlations of Q_1 and Q_4^* for the Simulation Data

Estimating Model	Generating Model		
	3-PAR	2-PAR	1-PAR
3-PAR	.82	.39	.40
2-PAR	.84	.40	.46
1-PAR	.97	.94	.27

Table 2
Mean Item Fit for the Simulation Data

Estimating Model	Generating Model: 3-PAR		2-PAR		1-PAR	
	Q_1	Q_4*	Q_1	Q_4*	Q_1	Q_4*
3-PAR	7.9	3.5	8.9	3.9	8.7	4.1
2-PAR	10.0	4.4	8.9	4.0	8.5	4.1
1-PAR	35.9	18.8	29.0	14.5	10.0	5.0

$10-m$ degrees of freedom. On the other hand, it appeared likely that Q_1 did have a chi-square distribution with $10-m$ degrees of freedom when the null hypothesis was true (i.e., when the estimating model equaled the generating model).

When an estimating model has fewer parameters than the generating model, the significance probability should be very low. This result occurred except when the data were generated by the 3-PAR model and the 2-PAR model was used for estimation. As mentioned earlier, the 2-PAR model did a surprisingly good job of fitting 3-PAR data, and Q_1 did not detect the fact that the 2-PAR model was not appropriate for the data. Some light can be shed on this finding by examining the relationship between the trait estimates produced by the different models.

When the model used for estimating trait levels and item parameters matched the model used for generating a set of data, the estimated trait levels, item difficulties, and discriminations had strong linear relationships with the trait levels, item difficulties, and discriminating powers in the generating model. In such a situation the lowest correlation of estimated values with the generating values was .89, which was found for the 3-PAR discriminations. The square root of the mean squared difference between estimated and generating lower asymptotes for the 3-PAR model was .03. In other words, when an estimating model was appropriate, it tended to accurately recover the trait levels and parameters of the generating model. Because of this finding, results are presented only for estimated traits and item parameters. Results involving the generating traits and parameters can be inferred from the results involving the estimated traits and parameters for the estimating model that matches the generating model.

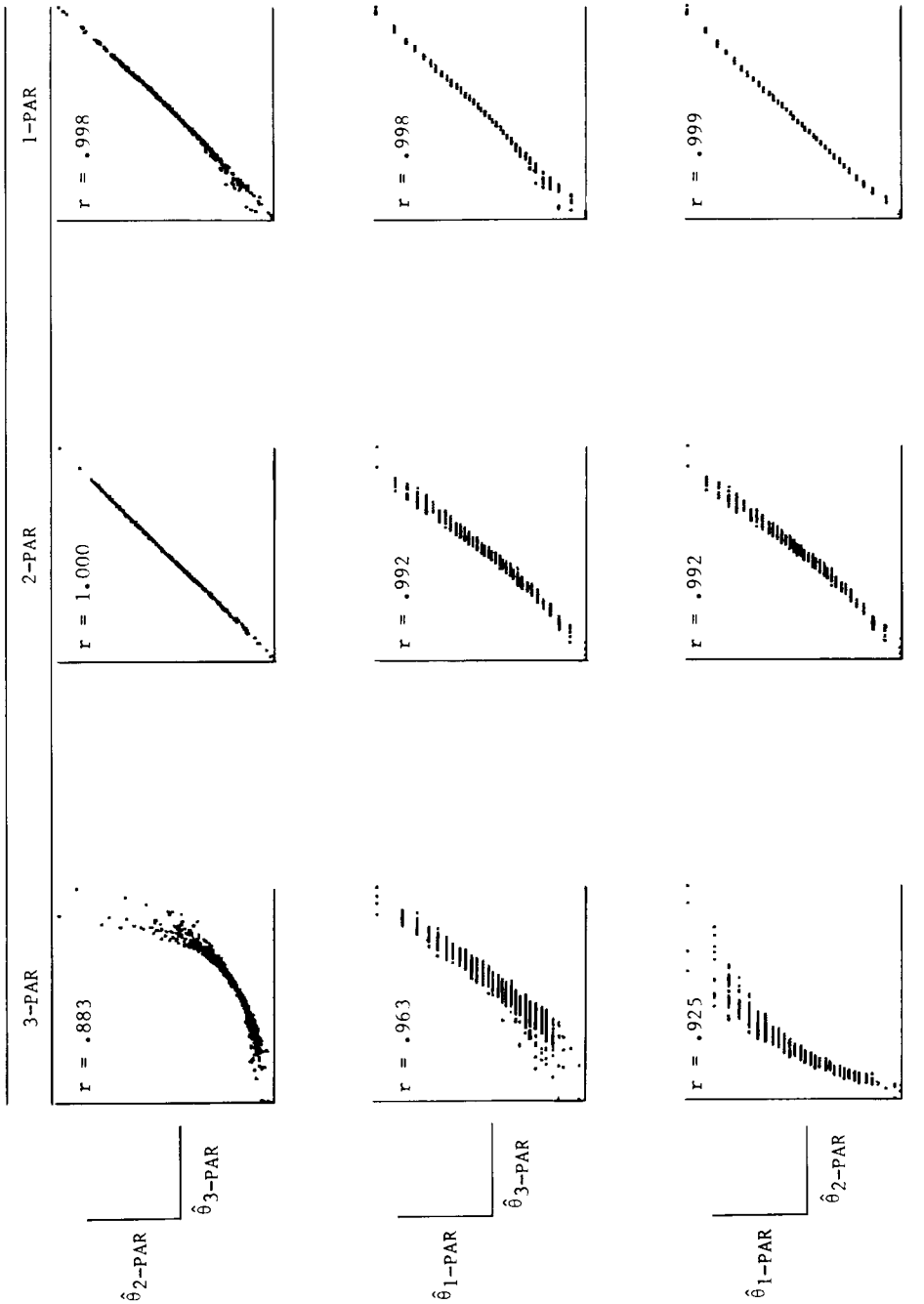
The relationships between trait estimates from the three models are contained in Figure 1. Each column of plots represents the results for one generating model. For each generating model, traits were obtained for three estimating models. The $\binom{3}{2} = 3$ pairs of estimated trait values are related in the

Table 3
Significance Probabilities of the Fit Statistics for the Simulation Data

Estimating Model	Generating Model: 3-PAR		2-PAR		1-PAR	
	Q_1	Q_4*	Q_1	Q_4*	Q_1	Q_4*
3-PAR	.60	.00	.05	.00	.25	.00
2-PAR	.31	.00	.25	.00	.28	.00
1-PAR	.00	.01	.00	.09	.60	.00

Note. Significance probabilities listed as .00 are less than .0005.

Figure 1
Relationships Between Estimated Trait Values



three plots for each generating model, and the plots are labeled in the left margin of the figure. Because the selection of the scale values in the estimation process is arbitrary, numerical labels of the axes of the plots are not important. When data were generated with the 1-PAR or 2-PAR model, every pair of estimated traits was essentially linearly related. When data were generated with the 3-PAR model, pronounced nonlinear relationships between the estimated trait values were created.

An examination of the correlations would not have revealed the curvilinear relationships between the various estimated trait values. Correlation ratios (Hays, 1973, p. 683) were calculated for the $\hat{\theta}$'s in each of the plots in Figure 1. Two correlation ratios were calculated for each plot, using each of the two sets of $\hat{\theta}$'s as the independent variable. The "independent" $\hat{\theta}$'s were grouped in steps of .10 (e.g., $\hat{\theta}$ with values from 1.01 to 1.10 fell in one group). The two correlation ratios for each plot had very similar values, and their average was taken and labelled $\bar{\eta}^2$. The squared correlation was subtracted from the average correlation ratio for each plot. This difference indicates the proportion of variance shared by the two variables beyond the variance shared in a linear relationship. This difference was greater than .05 for two of the plots:

1. When data were generated with the 3-PAR model and traits estimated by the 3-PAR and 2-PAR models were compared, $\bar{\eta}^2 - r^2 = .19$.
2. When data were generated by the 3-PAR model and traits were estimated by the 1-PAR and 2-PAR models, $\bar{\eta}^2 - r^2 = .10$.

These differences between $\bar{\eta}^2$ and r^2 quantitatively verify the existence of the obvious curvilinear relationships in Figure 1.

Figure 2 displays the relationships between the item difficulties estimated by the three models. The pattern of these relationships parallels the pattern found for the traits; but because there are fewer items than examinees, the patterns are more difficult to see with the items than with the trait values.

For data generated with the 2-PAR and 3-PAR models, the relationships between the estimated item discriminations and the relationships between item discriminations and item difficulties are contained in Figure 3. The relationships between the 3-PAR and 2-PAR estimated item discriminations were dramatically different, depending on whether the data were generated with the 3-PAR or 2-PAR model. When the 3-PAR model generated the data, the 3-PAR and 2-PAR estimated discriminations had a weak relationship. When the 2-PAR model generated the data, the 3-PAR and 2-PAR estimated discriminations had a very strong linear relationship.

When the 3-PAR model was used for estimation, there was a low positive correlation between $\hat{a}_i$ and $\hat{b}_i$ regardless of which model generated the data. This low positive correlation was consistent with the low positive correlation between a_i and b_i for the item parameters used in the data generation. When the 2-PAR model was used for estimation, there was a substantial difference in the correlation between $\hat{a}_i$ and $\hat{b}_i$ found when the 2-PAR model generated the data versus when the 3-PAR model generated the data. If the 2-PAR model generated the data, the low positive correlation between $\hat{a}_i$ and $\hat{b}_i$ appeared. However, if the 3-PAR model generated the data, use of the 2-PAR model for estimation produced a strong negative correlation between $\hat{a}_i$ and $\hat{b}_i$.

Achievement Test Study

To examine how similar the simulation data were to the Comprehensive Tests of Basic Skills (CTBS) data, means and standard deviations of estimated item parameters were compared. When the 3-PAR model generated the simulation data and was used in the estimation, the mean $\hat{a}$ = was 1.39, with SD of .57; mean $\hat{b} = -.02$, $SD = 1.08$; and mean $\hat{c} = .23$, $SD = .01$. These estimates were obtained with estimated traits with a mean of zero and a standard deviation of one. When the 3-PAR es-

Figure 2
Relationships Between Estimated Item Difficulties

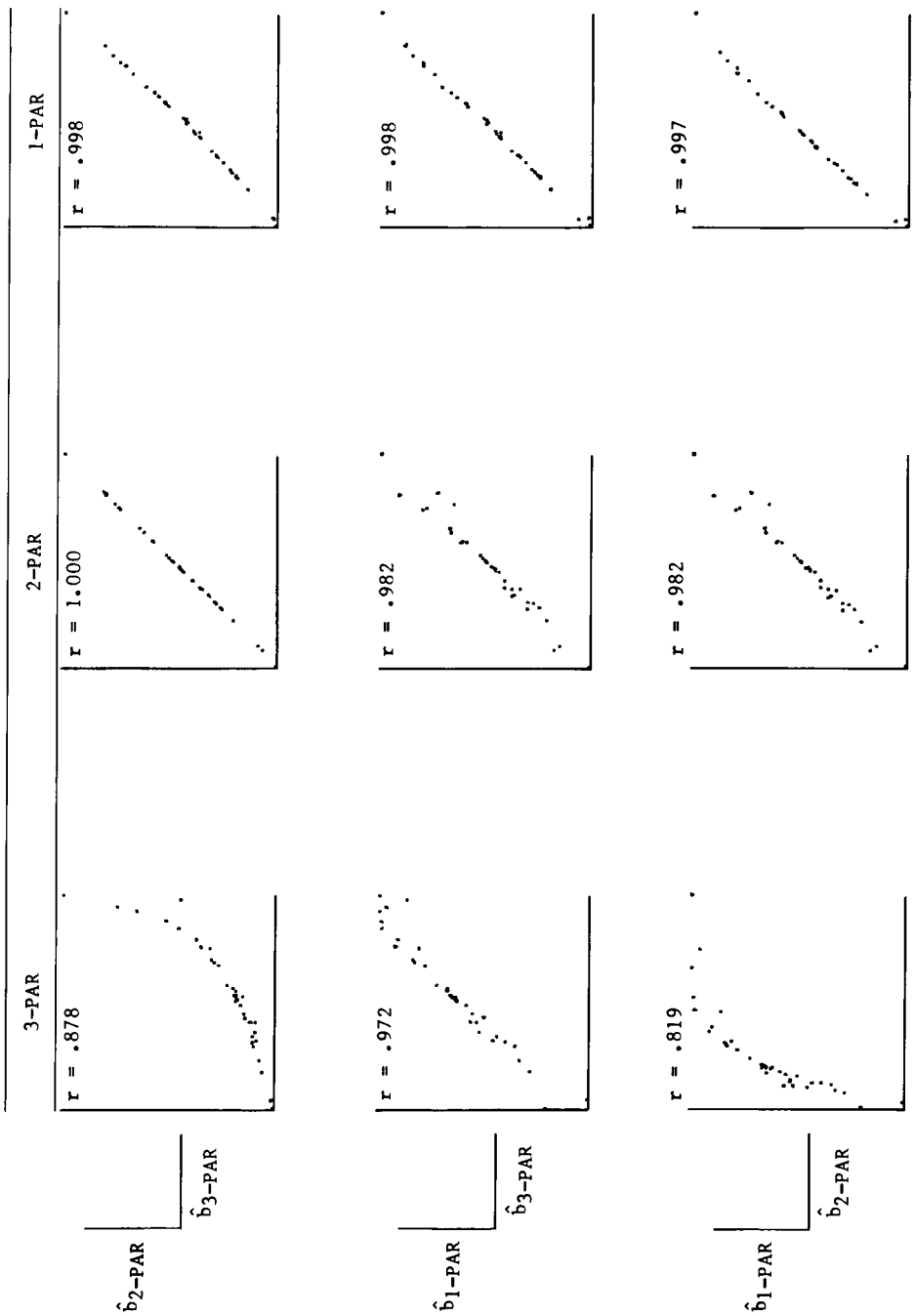
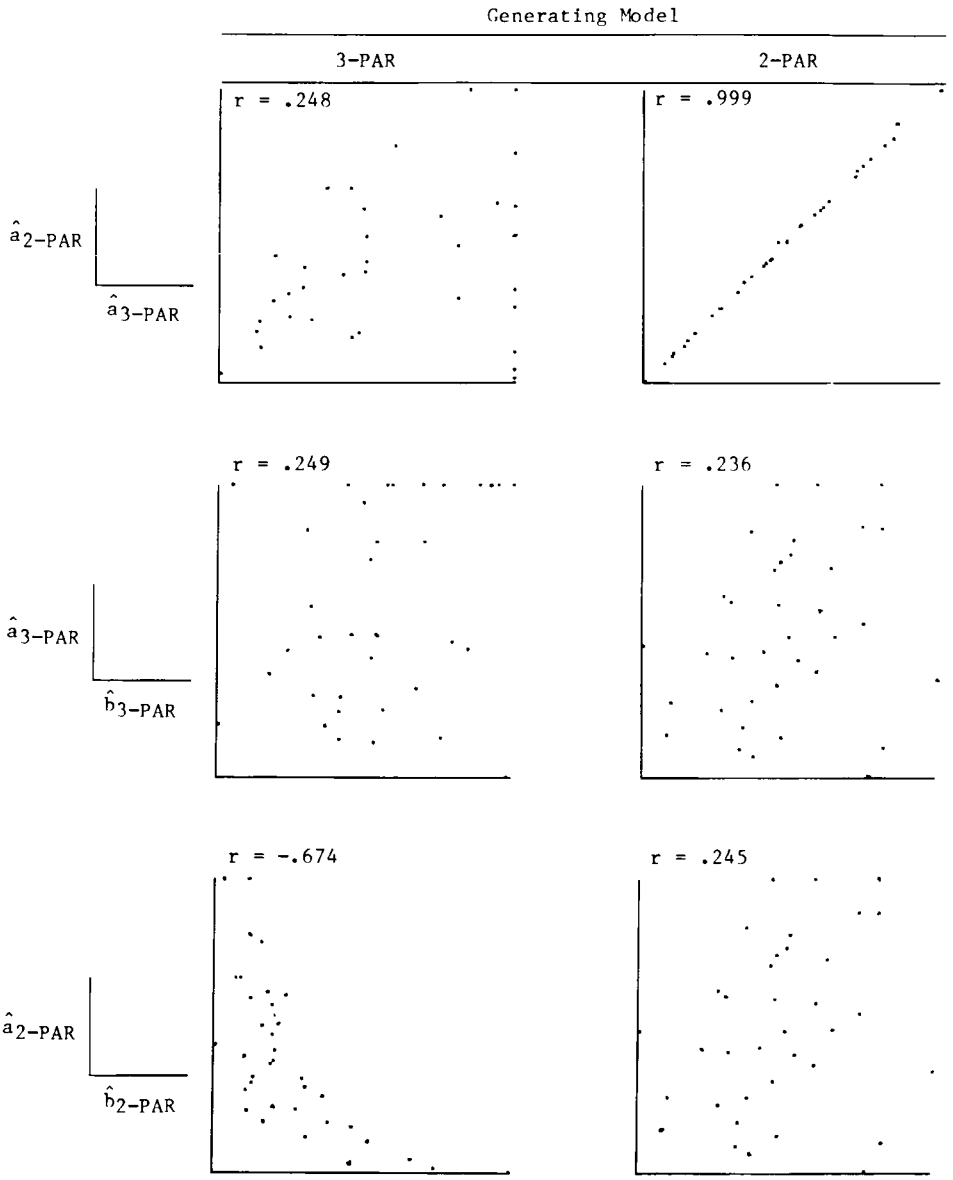


Figure 3
Relationships Between Estimated Item Discriminations
Across Different Estimating Models
and Between Estimated Item Discriminations and Item Difficulties
Within One Estimating Model



timating model was used for the CTBS data and the estimated traits had a mean of zero and a standard deviation of one, the range of item parameter statistics were mean $\hat{a}$ of 1.18 to 1.68, with *SD* of .55 to .80; mean $\hat{b}$ of −.08 to .15, *SD* of .54 to .75; mean $\hat{c}$ of .17 to .23, with *SD* of .01 to .03. The

Table 4
Mean Values of Q_1 for the CTBS Tests

Estimating Model	Read. Vocab.	Read. Comp.	Spell.	Lang. Mech.	Lang. Exp.	Math. Comp.	Math. Conc.	Math. Applic.
3-PAR	11.9	11.4	8.5	14.5	9.5	12.5	11.3	9.8
2-PAR	14.4	13.0	8.9	12.2	10.6	14.7	12.6	12.1
1-PAR	30.7	32.4	24.3	27.2	29.4	26.3	24.6	25.8

simulation data, therefore, appeared to be similar to the CTBS data except that the simulation data had items that tended to have a greater variance in item difficulties than did the CTBS data.

The mean values of Q_1 for the three latent trait models for each of the eight CTBS tests are contained in Table 4. These means have higher values but similar patterns to those in Table 2 for the simulation data generated with the 3-PAR and 2-PAR models. For every test except Spelling and Mathematics Applications, the significance probabilities of the Q_1 statistics were less than .05 for all three models. The estimated trait values and item parameters for each CTBS test were examined, and their relationships appeared very similar to the relationships found for the simulation data generated with the 3-PAR model. For example, Table 5 contains the correlations between the estimated item difficulties and discriminations for the 3-PAR and 2-PAR models. These correlations are similar in pattern to the correlations found with simulation data generated with the 3-PAR model.

All the results for Reading Vocabulary, which were typical of the results for the other tests, are contained in Figure 4. The $\hat{\theta}$ plots in the first column of Figure 4 can be compared with each of the columns of plots in Figure 1; the Reading Vocabulary trait estimates appear most similar to the trait estimates based on simulation data generated with the 3-PAR model. The $\hat{b}_i$ plots in the second column of Figure 4 can be compared with each of the columns of plots in Figure 2; it is not clear which data generation model is most consistent with the results for the Reading Vocabulary item difficulties. The column of $\hat{a}_i$ plots in Figure 4 can be compared to the two columns of plots in Figure 3; the estimated item discriminations for the Reading Vocabulary data clearly are most consistent with the results for the simulation data generated with the 3-PAR model.

Discussion

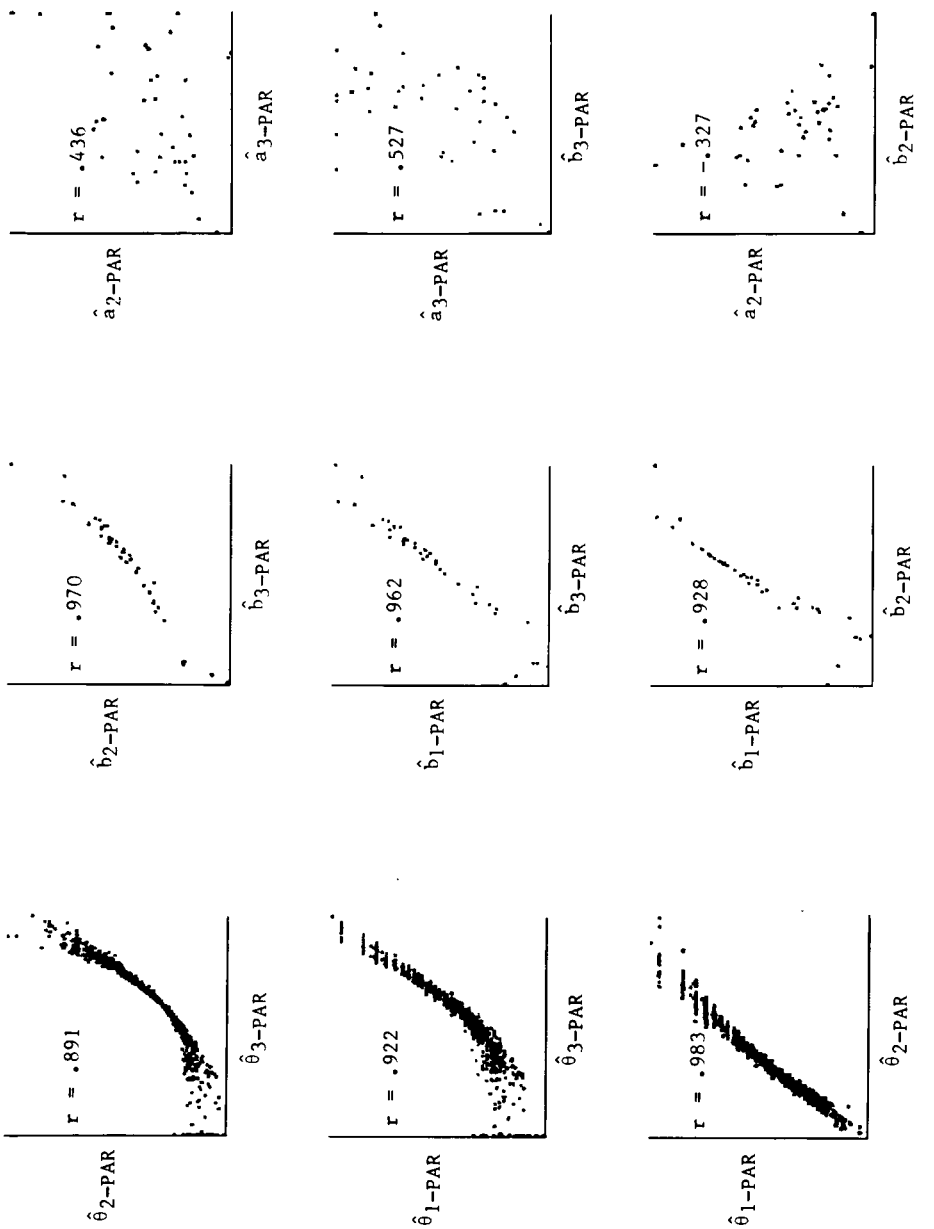
Behavior of Fit Measures with Simulated Data

For most of the sets of simulated data, Q_3^* and Q_1 had only moderate correlations; and for all sets of data, Q_3^* took on values that were about half the size of the values of Q_1. When an estimating model was true, the values of Q_3^* were systematically too small for Q_3^* to be distributed as a chi-square variable with $10-m$ degrees of freedom. The statistics Q_1 and Q_3^* produced nearly identical re-

Table 5
Correlations Between Estimated Item Discriminating Powers and Estimated Item Difficulties for the CTBS Tests

Estimating Model	Read. Vocab.	Read. Comp.	Spell.	Lang. Mech.	Lang. Exp.	Math. Comp.	Math. Conc.	Math. Applic.
3-PAR	.53	.33	-.22	.56	-.18	.58	.41	.40
2-PAR	-.33	-.41	-.57	.23	-.47	.23	-.44	-.45

Figure 4
Relationships Between Trait Estimates and Between Item Parameter Estimates for the Reading Vocabulary Data



sults for all nine sets of simulation data, with the values of Q_3^* tending to be only slightly larger than the values of Q_1. Thus, for the type of data sets examined, which simulated the pooling of examinees from two adjacent grades, the use of 10 cells was sufficient to produce relatively small values of $\sigma^2_{P_j}$ and small within-cell variances in $\hat{\theta}_k$. This result implies that with 10 cells Bock's (1972) fit measure would also be nearly identical to Q_1 for data sets with similar amounts of variance in $\hat{\theta}_k$.

When an estimating model matched the generating model, the hypothesis that Q_1 had a chi-square distribution with $10-m$ degrees of freedom would never have been rejected at the .05 level. However, for these data sets the mean value of Q_1 was always greater than $10-m$, indicating that perhaps too many degrees of freedom were subtracted for the estimated parameters.[1] In any case, it appears reasonable to describe Q_1 as having *approximately* a chi-square distribution with $10-m$ degrees of freedom when (1) the null hypothesis is true, (2) $N \doteq 1000$, and (3) 36 items are used. The exact distribution of Q_1 has not been specified. The most important information about Q_1 that was gained from the simulation study was the pattern of mean values for the nine combinations of generating and estimating models (Table 2). It is this pattern of the mean values of Q_1 that is useful in interpreting results for real data.

It was not a goal of this study to describe the behavior of Q_1 for a wide variety of conditions but to describe its behavior for a set of conditions similar to the conditions existing for the real achievement test data. The number of cells and the number and distribution of examinees are variables that can be expected to influence the value of Q_1, but the simulation was structured to be equal or very similar to the real data with respect to these variables. The number of items and the values of the item parameters can be expected to influence Q_1, but the items in the simulation study were similar to the items analyzed for the real data. The fact that the pattern of results for the simulation data generated with the 3-PAR model so closely paralleled the patterns of results for all eight achievement tests supports the contention that the simulated data were successful in modeling the real data.

Inappropriate Use of a Model

The most surprising simulation results appeared when data that were generated with the 3-PAR model were analyzed with the 2-PAR model:

1. The 2-PAR model fit the data almost as well as the 3-PAR model fit the data (Table 2).
2. The traits (and item difficulties) as estimated by the 2-PAR and 3-PAR models were nonlinearly related, with the 2-PAR trait estimates more stretched out at the high end than the 3-PAR traits (Figure 1).
3. The 2-PAR and 3-PAR estimated item discriminations had a weak relationship (Figure 3).
4. The 2-PAR estimated item difficulties and discriminations displayed a strong negative correlation (Figure 3).

The explanation of these results centers around the fact that the 2-PAR model with $c_i = 0$ attempts to fit or explain item characteristic functions with nonzero asymptotes. It is most awkward for the 2-PAR model to explain a nonzero asymptote when an item is difficult and has a moderate to high discrimination. When an item is easy, there is little data available to establish the item's lower asymptote. When an item has a low discrimination, the nonzero asymptote can be fairly well explained solely on the basis of a low discrimination. The 2-PAR model deals with the difficult, highly discriminating items with nonzero asymptotes by stretching out the trait (and difficulty) scale at the high end. This stretching decreases the slope of the item characteristic curves of difficult items, and the result-

[1]The expected value of a chi-square variable is its degrees of freedom.

ing item characteristic curves can be explained with low discriminations.

Thus, although the 2-PAR model is almost as accurate as the 3-PAR model in explaining pass/fail response arrays generated by a 3-PAR model, the 2-PAR model produces trait estimates and item parameters that are different from the trait levels and parameters used in generating the data. Because an item's discrimination can be affected by the item's difficulty level relative to the ability level of the group of examinees, the 2-PAR estimation of discriminations can be sample dependent for 3-PAR data. Difficult items can have systematically lower discriminations if they happen to be calibrated using a group of examinees with a relatively low ability level than if they happen to be calibrated with relatively able examinees. This sample dependency means that the predictions of the 2-PAR model can be accurate for the sample in which the 2-PAR item parameters are estimated, but the predictions for some items can be inaccurate in cross-validations involving samples of examinees that differ systematically from the sample in which the parameters were estimated. These problems with the 2-PAR estimating model can occur when the data are generated with the 3-PAR model; the 2-PAR estimating model works well with data generated with 2-PAR or 1-PAR models.

When the simulation data were generated by the 3-PAR model, the 1-PAR estimating model produced trait estimates that were fairly linearly related to the 3-PAR trait estimates (Figure 1). Because r_{ab} was fairly close to zero for the 3-PAR generating model, the 1-PAR model (with r_{ab} constrained to equal zero) produced a trait scale similar to the 3-PAR trait scale. However, if r_{ab} had been strongly positive or negative for the 3-PAR generating model, it would be expected that the 1-PAR estimating model would have produced a trait scale with a more pronounced curvilinear relationship with the 3-PAR trait scale. Indeed, such a relationship occurred with the Reading Vocabulary data where $r_{\hat{a}\hat{b}} = .53$ for the 3-PAR model (Figure 4). In a similar fashion, when data were generated by the 2-PAR model, the 1-PAR and 2-PAR trait estimates were linearly related for the most part. However, if $r_{\hat{a}\hat{b}}$ had been strongly positive or negative for the 2-PAR generating model, it would be expected that the 1-PAR and 2-PAR trait estimates would have shown a pronounced curvilinear relationship.

There are several practical problems that can arise when a 2-PAR or 1-PAR model is used inappropriately. One problem is the sample dependency of some item parameter estimates that can occur with the 2-PAR model. If item parameter estimates are sample dependent, one of the major theoretical advantages of the use of the models is lost. Another problem is inaccurate model predictions. When a 1-PAR model is used with multiple-choice items, the model predictions can be expected to be much less accurate (as measured by Q_1) than the 2-PAR or 3-PAR model predictions. This inaccuracy can be a problem if, for example, the model is being used to predict the probability that certain examinees will pass items they did not take.

A third problem is the lack of a perfect correlation between the estimated trait values and actual trait values when the estimating model is inappropriate. The correlations between the different trait estimates are high but not perfect. Examinees will receive somewhat different rank orderings or percentiles when different models are used. The nonlinear relationships between the various trait estimates implies that they can produce different results in correlational studies and different results when comparing, for example, group means and growth.

Another problem arising with the inappropriate use of a 1-PAR model is discussed by Gustafsson (1979):

> The Rasch model can be used successfully to solve the problem of vertical equating, as long as there is no correlation between item discrimination and difficulty. When there is a positive or negative correlation between these two parameters, there will be a bias [in the equating], with the direction of bias being dependent upon the sign of the correlation. . . .In [multiple-choice] tests there tends to be a negative correlation between item discrimination and item difficulty, if a pa-

rameter representing the lower asymptote of the item characteristic curve is not estimated. (p. 157)

An inaccurate vertical equating can be a serious problem in many testing situations.

There is another potential problem with the use of the 1-PAR model with data that are actually consistent with the 2-PAR or 3-PAR model. One of the assumptions of the 1-PAR model is that item discriminations are constant over the items. When the 1-PAR model is fit to a set of data, it chooses a trait scale that tends to minimize the correlation between item discriminations and difficulties. This choice of the scale is not in itself problematical and can actually be seen as desirable. However, a problem can arise when the 1-PAR model is used to equate scores on two tests if in one of these tests the correlation between item difficulty and discrimination (as measured by the 2-PAR model) is positive and in the other test this correlation is zero or negative. In this situation, the 1-PAR model will produce trait scales that are nonlinearly related for the two tests. Such a nonlinear relationship can be detected (e.g., by examining plots of trait values for examinees who take both tests) and the linear equating not used.

This study does not examine the problems involved in the inappropriate use of the 3-PAR model. Using Q_1, the 3-PAR model would have been rejected at the .05 level for all but two of the achievement tests. The rejection of the model is not too surprising because it would not be expected that any model would be strictly true in describing complex test-taking behavior and it would be expected that a sample size of about 1,000 would be large enough to make Q_1 a fairly powerful fit statistic. However, the practical implications of the inaccuracies of the 3-PAR model remain to be defined.

Choosing A Model For Real Data

The simulation results were the foundation for a two-step procedure for choosing the most appropriate of the three latent trait models for a set of real data.

Step 1: Compare mean values of Q_1 for all three models. If these three values are approximately equal, the 1-PAR model is appropriate. If the mean value of Q_1 for the 1-PAR model is substantially greater than the mean values for the other models, then the 1-PAR model is not appropriate. Proceed to Step 2 to determine whether the 2-PAR or 3-PAR model is more appropriate.

Step 2: The simulation results lead to the expectation that the mean values of Q_1 for the 2-PAR and 3-PAR models will tend to be similar and these mean values will not help choose which of these two models is more appropriate. Instead, it is helpful to examine the relationships between the trait estimates and item parameters for the models. If the 2-PAR and 3-PAR trait estimates are nonlinearly related, the 2-PAR and 3-PAR item discriminations have a low correlation, and the correlation between item difficulties and item discriminations is substantially lower for the 2-PAR model than for the 3-PAR model, then the 3-PAR model is more appropriate. If, however, the 2-PAR and 3-PAR trait estimates are highly linearly related, the 2-PAR and 3-PAR item discriminations are highly linearly related, and the correlation between item difficulties and item discriminations is very similar for the 2-PAR and 3-PAR models, then the 2-PAR model can be used.

When this procedure was followed for the eight achievement tests, it was clear that the 3-PAR model was the most appropriate of the three models to use for every test. The two-step procedure was far more helpful than examining the significance probabilities of the Q_1 values for the three models.

A test user may wish to choose one of the three latent trait models a priori and not go to the trouble or expense of examining data with all three models. The results obtained here for the eight achievement tests indicate that it is likely that it will not be appropriate to use the 2-PAR or 1-PAR

models with multiple-choice tests and that the 3-PAR model should be the a priori choice. However, use of the 3-PAR model requires larger sample sizes for accurate item parameter estimates, more computer time to obtain those estimates, and more technical sophistication than the use of the 2-PAR or 1-PAR model, and many multiple-choice test users choose a priori the 1-PAR model. If a multiple-choice test user chooses a 1-PAR or 2-PAR model, the user should be prepared to demonstrate the appropriateness of the model or, failing that, to deal with possible practical problems arising from the use of an inappropriate model, as described in the previous section.

References

Allen, M. J., & Yen, W. M. *Introduction to measurement theory.* Monterey, CA: Brooks/Cole, 1979.

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

Connolly, A. J., Nachtman, W., & Pritchett, E. M. *KeyMath Diagnostic Arithmetic Test manual.* Circle Pines, MN: American Guidance Service, 1976.

CTB/McGraw-Hill. *Comprehensive Tests of Basic Skills, Form S.* Monterey, CA: Author, 1973.

Elliott, C. D., Murray, D. J., & Saunders, R. *Goodness of fit to the Rasch model as a criterion of test unidimensionality.* Manchester: University of Manchester, 1977.

Gustafsson, J. E. The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement,* 1979, *16,* 153-158.

Hays, W. L. *Statistics for the social sciences.* San Francisco: Holt, Rinehart, & Winston, 1973.

IMSL Library (7th ed.). Houston, TX: International Mathematical and Statistical Libraries, 1979.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics* (Vol. 2). London: Griffin, 1967.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics* (Vol. 1). London: Griffin, 1969.

Lancaster, H. O. *The chi-squared distribution.* New York: Wiley, 1969.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Menlo Park, CA: Addison-Wesley, 1968.

Rentz, R. R., & Bashaw, W. L. *Equating reading tests with the Rasch model: Final report.* Athens, GA: University of Georgia, Educational Research Laboratory, September 1975.

Whitely, S. E., & Dawis, R. V. The nature of objectivity with the Rasch model. *Journal of Educational Measurement,* 1974, *11,* 163-178.

Wood, R. L., Wingersky, M. S., & Lord, F. M. *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters* (RM-76-6). Princeton, NJ: Educational Testing Service, 1976.

Woodcock, R. W. *Woodcock Reading Mastery Tests manual.* Circle Pines, MN: American Guidance Service, 1973.

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

Wright, B., & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement,* 1969, *29,* 23-48.

Author's Address

Send requests for reprints or further information to Wendy M. Yen, CTB/McGraw-Hill, Del Monte Research Park, Monterey, CA 93940.