DOCUMENT RESUME

ED 191 915 TM 800 539

AUTHOR	George, Archie A.
TITLE	Theoretical and Practical Consequences of the Use of Standardized Residuals as Rasch Model Fit Statistics.
INSTITUTION	Texas Univ., Austin. Research and Development Center for Teacher Education.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	Apr 79
NOTE	56p.: Paper presented at the Annual Meeting of the American Educational Research Association (63rd, San Francisco, CA, April 8-12, 1979).
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Academic Ability: Computer Programs: Difficulty Level: *Goodness of Fit: *Item Analysis: Item Banks: *Latent Trait Theory: *Mathematical Models: Test Items
IDENTIFIERS	Item Discrimination (Tests): *Rasch Model: *Standardized Residuals

ABSTRACT

The appropriateness of the use of the standardized residual (SR) to assess congruence between sample test item responses and the one parameter latent trait (Rasch) item characteristic curve is investigated. Latent trait theory is reviewed, as well as theory of the SR, the apparent error in calculating the expected distribution of the SR, and implications of using the SR for item analysis. Empirical results using actual data are presented to support the theoretical anlaysis, as well as a demonstration of the practical implications of the failure to reject items which do not fit the model. Conclusions based on the findings include: (1) discriminations of all the items in a test must be very similar in order for Rasch model analyses to work in practice: (2) the SR mean square fit statistic does not detect unacceptable variation in discrimination: and (3) item discrimination needs to be monitored and controlled using more exact tests of fit than the residual mean square. Finally, an alternate linear model is described which may provide a practical solution to problems encountered in the construction of item banks and tailored testing. (RL)

THEORETICAL AND PRACTICAL CONSEQUENCES OF THE USE OF STANDARDIZED RESIDUALS AS RASCH MODEL FIT STATISTICS

Archie A. George

Procedures for Adopting Educational Innovations Program
Research and Development Center for Teacher Education
The University of Texas at Austin

Paper presented at the annual meeting of the
American Educational Research Association,
San Francisco, April 9, 1979, Session 6.17

THEORETICAL AND PRACTICAL CONSEQUENCES OF THE USE OF STANDARDIZED RESIDUALS AS RASCH MODEL FIT STATISTICS[1]

Archie A. George
Research and Development Center for Teacher Education
The University of Texas at Austin

The standardized residual is a statistic which has been used to assess the congruence between a sample of test item responses and the one parameter latent trait (Rasch) item characteristic curve. However, the central thesis of this paper is that the statistic is not appropriate for this purpose. Its theoretical distribution is based on the central limit theorem which, of course, requires a large-sample size and yet its calculation involves very small sample sizes. The practical consequences of the use of this statistic are described and illustrated.

More specifically, this paper consists of five main sections. The first section contains a very brief review of latent trait theory and the models receiving greatest attention at this time. The second section contains a description of the standardized residual statistic and an explanation of its theoretical basis. The apparent error in the calculation of the expected distribution of this statistic is pointed out, and the implications of the use of the statistic for item analysis are described and justified on a theoretical basis. In the third section, empirical results using actual data are presented which support the theoretical analysis. In the fourth section, the practical

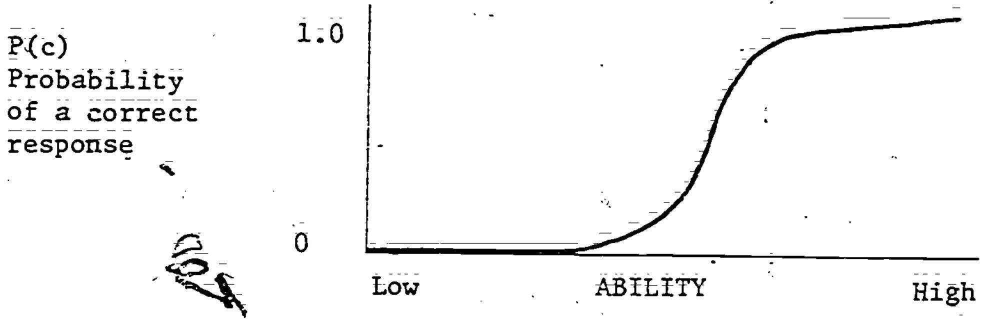
[1] The research described herein was conducted under contract with the National Institute of Education. The opinions expressed are those of the author and do not necessarily reflect the position or policy of the National Institute of Education, and no endorsement by the National Institute of Education should be inferred.

implications of the failure to reject items which do not fit the model are demonstrated. Finally, an alternate model is described which may provide a practical solution to problems encountered in the construction of item banks and tailored testing.

Part I. Latent Trait Theory

Latent trait theory has been proposed as an alternative to classical test theory for the assessment of ability and educational achievement. A latent trait model specifies a relationship between observable performance and unobservable traits (abilities) which are assumed to underlie performance. The latent trait models currently under study in educational measurement specify mathematical formulae which relate ability to probability of a correct response on specific test items. Figure 1 shows a hypothetical relationship, between ability and correct response, which is called the item characteristic curve.

Figure 1.



Lord (1952, 1953) investigated the use of the normal ogive as a model for performance on mental test items. The mathematical complexity of this model did not encourage its full development, but the groundwork was laid for later work using simpler models. Currently, three latent trait models seem to be receiving the most intense research. These models are very closely related mathematically even though they are the result of different lines of development.

The simplest is the Rasch model:

$$P(c) = \frac{\lambda}{1 + \lambda},$$

where $\lambda = e^{\beta-\delta}$.

In this model, δ is the difficulty of the item and β is the ability of the person. This model is often referred to as the "one parameter" model, because the item characteristic curve is completely determined by only one parameter, the difficulty of the item.

A model of intermediate complexity has been proposed:

$$P(c) = \frac{\lambda}{1 + \lambda},$$

where $\lambda = e^{\alpha(\beta-\delta)}$.

This model is the same as the Rasch model, except for the parameter α, which is called the "item discrimination" parameter. This parameter describes the slope of the item characteristic curve. Thus, the two parameter model, as this model is frequently referred to, has more flexibility. That is, a wider variety of item characteristic curves can be described using this model.

A three parameter model has also been proposed:

$$P(c) = \gamma + (1-\gamma) \frac{\lambda}{1 + \lambda},$$

where $\lambda = e^{\alpha(\beta-\delta)}$.

In this model, γ is referred to as a "guessing" parameter. It functions to modify the lower asymptote of the item characteristic curve so that the probability of a correct response never reaches zero, no matter how low the ability of the test taker.

The Rasch model is currently receiving a lot of attention from test users, school districts, and others. This is primarily due to the efforts of Ben Wright and his colleagues at the University of Chicago. The simplicity of the one parameter model has enabled statisticians to develop techniques of estimating the item difficulty and person ability. What could be described as cookbook procedures have been developed for the use of the Rasch model in many testing situations (Cohen, 1976, in Wright, 1977). Estimation problems are more difficult for the two and three parameter models.

One advantage offered by use of latent trait models is that different tests can be used to measure students of different ability while maintaining comparability of scores. That is, each student could receive a different test and the abilities of the students could still be placed on a single scale. Low ability students can be tested using easy tests and high ability students tested using more difficult tests, and the measurements reported on a single scale.

The Rasch model offers one advantage over the other latent trait models; the number of items correct on a test is all that is necessary to estimate a person's ability. Correct responses to difficult items do not count any more than correct responses to easy items. The Rasch model has been shown to be the only logistic latent trait model which has this property (Andersen, 1977). According to the other logistic models, a person's ability is estimated as a function of the difficulty of the items that person marked correctly, not merely how many.

Part II. A Theoretical Analysis of the Standardized Residual Fit Statistic

According to Wright (1977), the two major advantages of the Rasch model are sample-free item calibration and test-free person measurement. Sample-free item calibration refers to the concept that the difficulty of test items can be estimated regardless of the abilities of the persons who respond to the items. Test-free person measurement refers to the concept that the ability of persons can be estimated regardless of the particular items in the test to which they they respond. Together, these properties allow for the construction of a personalized testing program which measures high ability students using tests containing difficult items and low ability students using tests containing easy items. Scores on all tests can be converted (vertically equated) to measures of ability on a common scale, and new test items can be calibrated without controlling for the ability of the sample.

Several investigators have examined the ability of the Rasch model to live up to these promises. Anderson, Kearney and Everett (1968) and Tinsley and Dawis (1975) found Rasch item difficulties to be fairly invariant for particular sets of items when based on different samples of test takers. Whitely and Dawis (1974) and Slinde and Linn (1978) attempted to replicate Wright's (1968) results for the problem of vertical equating. According to Slinde and Linn, Whitely and Dawis' results were not as good as those of Wright, but were judged to "lend some support for the item-free person measurement claim of the Rasch model" (1978, p. 26). However, Slinde and Linn found that for the math data analyzed in their study, "the Rasch model did not provide a satisfactory means of vertical equating" (1978, p. 34). They went on to say that it may be necessary to more carefully select items that fit the model. [This is also the recommendation of Keats and Boldt, as reported by Angoff (1971, pp. 529-530).] Slinde and Linn did not test items for fit to the Rasch model. They acknowledged that this may have been responsible for the inadequate vertical equating results they obtained.

While working with the Rasch model, Dr. Donald Veldman and I noted that the test of item fit recommended by Wright (1977, p. 102) indicated that a very large number of items fit the Rasch model better than could be expected by chance. That is, the standardized residual fit statistics were very low for many items. I found this to be true in published articles, also (Perline, Wright and Wainer, 1977; Wright and Mead, 1977), and initiated this study of the fit statistic.

The Standardized Residual statistic is computed as follows. For each person who attempts an item, a z-square is computed:

$$z^2 = \frac{(X-P)^2}{P(1-P)} ,$$

where $P = e^x/(1+e^x)$, $x = b-d$

and $X = 1$ if correct response

0 if incorrect.

where b = estimate of the person's ability

and d = estimate of the item's difficulty.

This z^2 conveniently reduces to e^{-x} for a correct response and e^x for an incorrect response. Clearly, when x is large (i.e., a person's ability is much greater than the item's difficulty), the predicted probability of a correct response is very high. For instance, if $x = 4$,

$$P(c) = e^x/(1+e^x) = 54.60/55.60 = .98.$$

If the person gets the item correct, a z^2 of e^{-4} (.02) is assigned. If the person misses the item, a z^2 of e^4 (54.6) is assigned. These z^2 are summed over persons and divided by the number of persons to obtain the value of the standardized residual. This statistic has also been called the mean squared error, or fit mean square. Wright, Mead and Draba (1976) claim that the statistic will be high for items with both high and low discriminations. This statistic is supposedly distributed as a chi square with N-1 degrees of freedom (Panchapakesan, 1969; Wright and Panchapakesan, 1969), or as a mean square with expected value 1.0 and variance $2L/(L-1)(N-1)$, L = number of items, N =

number of persons (Perline, Wright and Wainer, 1977). Wright and his colleagues have recommended several modifications of this simple $\sum z^2/N$ formula. These adjustments were made because the distribution of the statistic does not conform to theoretical expectation. However, there are fundamental problems with the statistic which cannot be corrected by such adjustment.

The basic idea in the formulation of the mean z^2 is that deviations of obtained scores from expected values can be converted to z-scores by dividing the deviations by the standard deviation, squaring, and averaging across persons. Lower mean z^2 should indicate small deviation, in general, while large mean z^2 should indicate greater deviations. Table 1 shows how heavily dependent the standard deviation of the binomial distribution is upon the sample size for several values of the probability of a correct response. In computing the z^2 fit statistic, the sample size is always 1, since it is calculated for each person-item encounter and averaged over persons to assess item fit, and over items to assess person fit. From this point of view, the test should be very conservative (i.e., rarely reject items), because the standard deviations are large, which makes z^2 values small.

A more serious problem is the use of a normal approximation to the binomial, which is implicit in the expectation that the sum or mean z^2 is chi squared distributed. Table 2 shows the magnitude of errors which are introduced when the z^2 fit statistic is assumed to be sampled from a normal distribution. Table 2 shows several actual probabilities of a correct response to an item [P(c)], and the probabilities which would be inferred to exist if the z^2 were a normal deviate.

When the actual probability is .50, the inferred probability for both correct and incorrect responses (1 and 0) is .32. When the actual probability of a correct response is .80, a correct response is inferred to have a

	Sample Size			
P(c)	1	5	10	30
.50	.50	.22	.16	.09
.60	.49	.22	.15	.09
.70	.46	.20	.14	.08
.80	.40	.18	.13	.07
.90	.30	.13	.09	.05
.95	.22	.10	.07	.04
.98	.14	.06	.04	.03
.99	.10	.04	.03	.02

Table 1.

SD of binomial Distribution ($\sqrt{\frac{pq}{N}}$)

P(c)	Inferred Probability of Correct Response	Inferred Probability of Incorrect Response
.50	.32	.32
.60	.41	.22
.70	.52	.13
.80	.61	.05
.90	.74	.0027
.95	.82	.0000
.98	.89	.0000
.99	.91	.0000

Table 2.

Actual and inferred probabilities assuming the standardized residual is a z-score sampled from a normal distribution.

probability of .61, and an incorrect response a probability of .05. These errors are introduced because the calculated z is assumed to have been sampled from a normal distribution, which would be appropriate only for large sample sizes.

The values in Table 2 were calculated as follows. Since

$$P(c) = e^{x}/1+e^{x},$$

$$x = ln\left[\frac{P(c)}{1-P(c)}\right]$$

for any specified P(c).

The z^2 for a correct answer is e^{-x}, and for an incorrect answer e^{x}, according to the formula presented earlier:

$$z^2 = \frac{[X - P(c)]^2}{P(c)[1-P(c)]}$$, which reduces

algebraically to e^{-x} when X = 1 and e^{x} when X = 0. The square root of each z^2 (i.e., e^{-x}, e^{x}) was calculated and the corresponding value [P(c)] found in a normal probability table. The inferred probabilities in Table 2 are 2[1-P(c)], which represents the probability of obtaining a value as deviant or more deviant than one with the z score with which the table was entered.

With a little reflection, these considerations reveal how very small fit statistics come about for some items. Items which are easy are answered correctly more often than the Rasch model predicts. Each time this happens, a very small z^2 is added to the "sum of squared residuals," and the result is a small mean squared residual -- which supposedly reflects good fit to the Rasch curve! It is also possible to see how greatly this statistic can be affected by a few students of low ability obtaining correct responses, perhaps by guessing. For each unexpected correct answer, a very large z^2 is added to the sum of squared residuals, producing a larger mean squared residual. Table 3 shows the z^2 values for correct and incorrect responses at several values of P(c).

P(c)	z^2 for Correct Response	z^2 for Incorrect Response
.50	1.00	1.00
.60	.67	1.50
.70	.43	2.33
.80	.25	4.00
.90	.11	9.00
.95	.05	19.00
.98	.02	49.00
.99	.01	99.00

Table 3.

Values of the squared standardized residual for several probabilities of a correct response.

Notice the extremely high and low values of z^2 for values of P(c) above .90.

A more direct illustration might be to consider the implications of the statistic for a number of persons of similar ability whose proportion of correct responses deviates from the theoretical proportion in certain ways. For example, if 40 persons obtained a total score on a test to which the Rasch model assigns an ability rating of 1.386, the model predicts that 80 percent of these should get an item with difficulty 0.0 correct. If exactly 32 (80%) do indeed perform as predicted, the sum of squared residuals would be:

$$\Sigma z^2 = 32\ (e^{-1.39}) + 8\ (e^{1.39}) = 40.00$$

and the mean squared residual = 1.00. There is, of course, no difference between obtained and expected proportion for this sample.

Let us suppose that 36 of the 40 individuals (90%) provided correct responses:

$\Sigma z^2 = 36\ (.250) + 4\ (4.00) = 25.00$, and the mean squared residual = .625.

Thus, it can be seen that a deviation in this direction might lead one to infer that the data fit the model "better than" when no deviation was present!

On the other hand, let us suppose that 28 of the 40 individuals (70%) provided correct responses:

$\Sigma z^2 = 28\ (.250) + 12\ (4.00) = 55.00$, and the mean squared residual = 1.375.

In this case, a person would be led to the inference that the data did not fit the model as well as the previous data, even though exactly the same deviation is present.

It might be appropriate, since there is a fairly large group involved, to use the normal approximation to the binomial to test for significance of the differences between hypothesized and expected frequencies of correct responses. That is, for each case:

1) $z = \frac{.8 - .8}{\sqrt{\frac{(.8)(.2)}{40}}} = \frac{0}{.063} = 0$

2) $z = \frac{.9 - .8}{\sqrt{\frac{(.8)(.2)}{40}}} = \frac{.1}{.063} = 1.58$

3) $z = \frac{.7 - .8}{\sqrt{\frac{(.8)(.2)}{40}}} = \frac{-.1}{.063} = -1.58$

What a difference grouping of scores makes!

Items which are very difficult for the sample of students on which fit statistics are based often appear to fit the Rasch curve poorly because a few students do answer these correctly. Items which are very easy are usually inferred to fit the Rasch curve very well, because even more students answer them correctly than are predicted by the model. Items which have difficulties near the ability level of the sample usually appear to be a good fit because of another factor -- the Rasch curve is flatter than most actual item characteristic curves. This results in more incorrect responses than predicted when P(c) is less than .50 and more correct responses when P(c) is greater than .50. Most z^2 are less than 1.00; so is the mean z^2. Figure 2 shows three item characteristic curves: the Rasch curve, one steeper curve and one flatter curve. The steeper and flatter curves have exactly the same deviation from the Rasch. The mean squared deviations were calculated for each curve and are shown in Table 4. These values indicate that the steeper curve fits the model best — even better than the Rasch curve itself! The flatter curve fits the least well. It is for this reason, in my opinion, that the mean squared residual has become a widely used index of fit of data to the Rasch model. The steeper the item characteristic curve, the better the item is inferred to

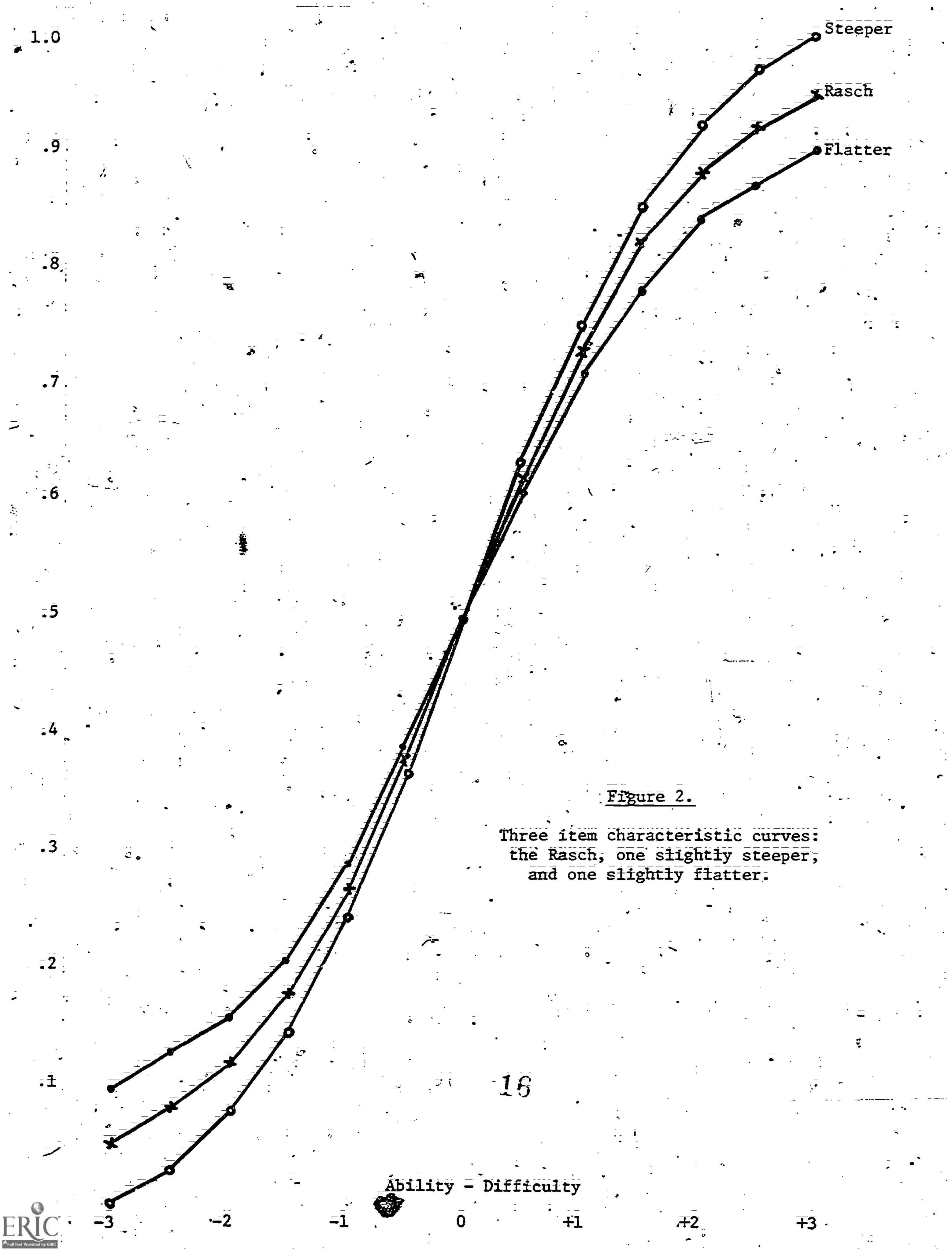


Figure 2.

Three item characteristic curves: the Rasch, one slightly steeper, and one slightly flatter.

x	Correct e^{-x}	Incorrect e^{x}	Steeper Curve	Rasch Curve	Flatter Curve
-3.0	20.09	.05	0/100 *	5/95	10/90
-2.5	12.18	.08	3/97	8/92	13/87
-2.0	7.39	.14	8/92	12/88	16/84
-1.5	4.48	.22	15/85	18/82	21/79
-1.0	2.72	.37	25/75	27/73	29/71
-.5	1.65	.61	37/63	38/62	39/61
0.0	1.00	1.00	50/50	50/50	50/50
.5	.61	1.65	63/37	62/38	61/39
1.0	.37	2.72	75/25	73/27	71/29
1.5	.22	4.48	85/15	82/18	79/21
2.0	.14	7.39	92/8	88/12	84/16
2.5	.08	12.18	97/3	92/8	87/13
3.0	.05	20.09	100/0	95/5	90/10
		Σ	855.56	1316.55	1737.74
		Mean Squared Residual :	.66	1.01	1.34

Table 4.

Calculation of mean squared residual for three curves in Figure 2.

*These ratios show the proportion correct to incorrect. To calculate the mean squared residual, multiply the numerator by the value in the e^{-x} column, multiply the denominator by the value in the e^{x} column and add these two values, then sum this result across the 13 intervals shown and divide by 1300.

fit the Rasch model. Thus, tests built using the mean squared residual have, for the most part, been as good as tests built using traditional test statistics. However, such tests are only believed to have been constructed by selecting items which fit the Rasch model. The selected items actually were those with the highest discrimination, just as in traditional analyses.

Part III. Examination of a Set of Actual Test Data Which Illustrates the Problems Encountered When Using the Standardized Residual as a Fit Statistic

Because of the inadequacies in the standardized residual fit statistic, it was necessary to construct another procedure for testing the fit of the Rasch model to actual data. The procedure I devised is as follows. Veldman's (1978) version of the PROX procedure (Cohen, 1976, in Wright, 1977) was used to estimate the item difficulties and person abilities for each of the 1664 students whc responded to a 237 item English achievement test. All 1664 students had scores greater than 0 and less than 237, enabling them to have their abilities estimated according to the model. Fit of each item to the Rasch curve was assessed by fitting a least squares line to the data in the range of ±2 units of the estimated ability-difficulty (b-d). A t-test was used to determine the probability that the data were sampled from a population in which the slope of the line was the same as the slope of the Rasch model between these two values.

Figure 3 shows the Rasch item characteristic curve between the values ±2.0 $\beta-\delta$ and a line fitted by a least squares to a uniformly distributed set of data points on the Rasch curve. The slope of this line is .2094, and seems to approximate the Rasch curve quite nicely. Table 5 shows the formulae for the t-test which determined the significance of the difference between the Rasch slope (.2094) and the slope of each item's data in this range.

I have taken some care to investigate the adequacy of the linear model as a substitute for the Rasch curve in the range ±2 (b-d). The error introduced was calculated by computing the sum of squared deviations from the Rasch curve of data which conform exactly to the Rasch model. That is, $\sum [P(c)]\,[1-P(c)]^2 + [P(nc)]\,[0-P(c)]^2$ where P(c) is the proportion of correct scores predicted by the Rasch model,

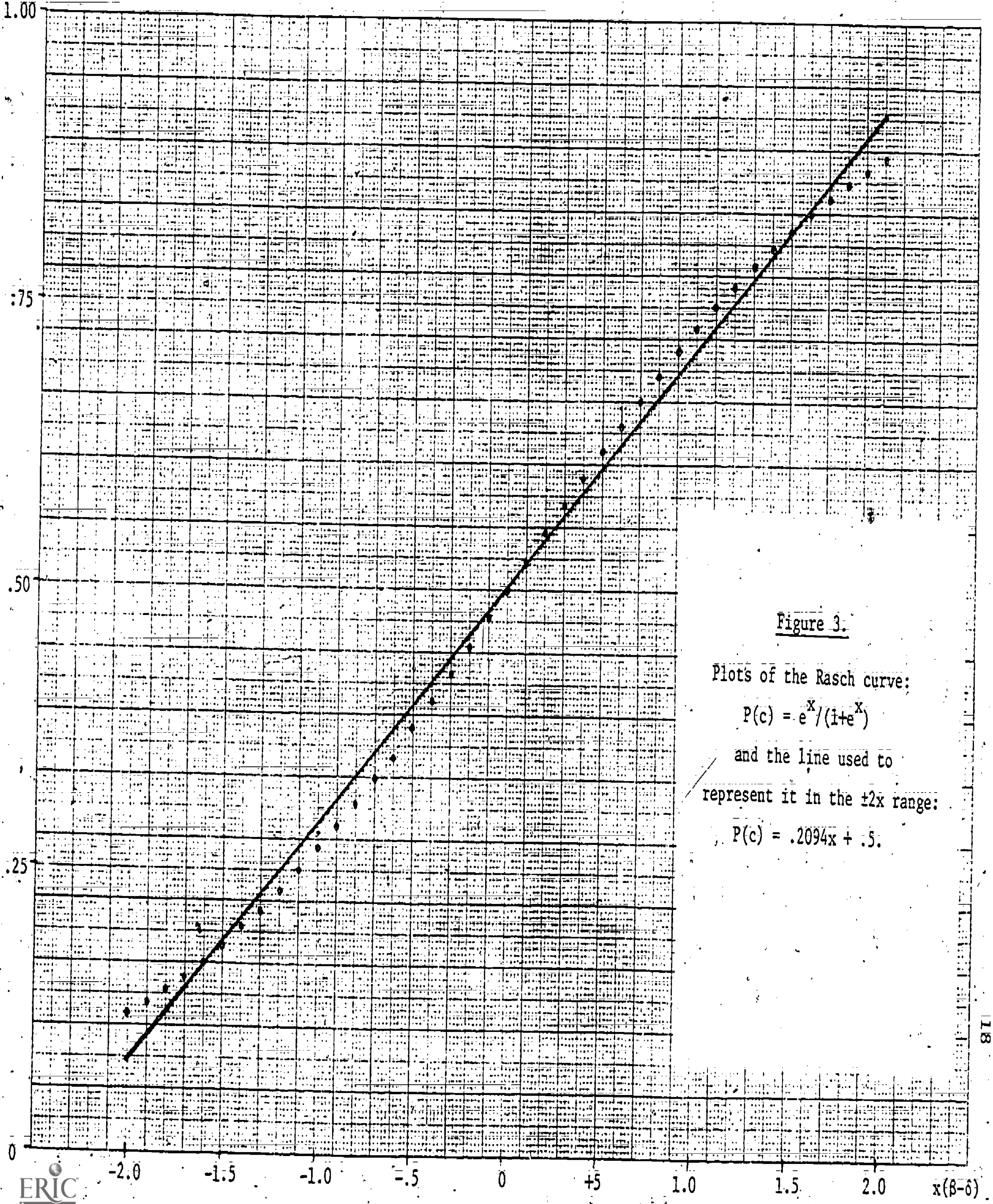


Figure 3.

Plots of the Rasch curve:

$P(c) = e^{x}/(1+e^{x})$

and the line used to

represent it in the ±2x range:

$P(c) = .2094x + .5.$

$$SSx = \sum_{i=1}^{N} (b_i - d)^2 - \frac{\left[\sum_{i=1}^{N} (b_i - d)\right]^2}{N}$$

$$SSy = \sum_{i=1}^{N} X^2{}_i - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}$$

$$SSxy = \sum_{i=1}^{N} (b_i - d)\ X_i - \frac{\left[\sum_{i=1}^{N} (b_i - d)\right]\left[\sum_{i=1}^{N} X_i\right]}{N}$$

$$\hat{\beta}_1 = SSxy/SSx$$

$$SSE = SSy - (SSxy)^2/SSx$$

$$s = \sqrt{\frac{SSE}{N-2}}$$

$$t = \left(\frac{\hat{\beta}_1 - .20940}{s}\right)\left(\sqrt{SSx}\right),\ df = N-2$$

b_i = estimated ability of person i

d = estimated difficulty of item

X_i = 1 if person i answered item correctly

= 0 otherwise

Table 5.

Formulae used to assess the significance of the difference between the slope of actual item characteristic curves and the slope of the Rasch curve.

1-P(c) is the distance from the Rasch curve to the correct score,

P(nc) is the proportion of incorrect scores, and

0-P(c) is the distance from the Rasch curve to the incorrect score.

Summing these values across 41 intervals (-2.0 to +2.0 in .1 increments, see Figure 3), assuming a uniform distribution of data throughout this interval, results in an error sum of squares (ESS) of 7.719602. Using the linear model, the formula is $\Sigma [P(c)]\ [1-\hat{P}(c)]^2 + [P(nc)]\ [0-\hat{P}(c)]^2$, where $\hat{P}(c)$ is the proportion of correct scores predicted by the linear model. The error sum of squares in this case is 7.733142, or an error increase of only .2%!

If one were to test for the adequacy of the linear model using an F-test comparing the two models (see Ward and Jennings, 1973),

$$Y = a_1X^{(1)} + a_2X^{(2)} + \ldots + a_{41}X^{(41)} + E^1 \qquad (1)$$

where the $X^{(i)}$ are binary (0,1) vectors specifying membership in one of the 41 levels of b-d and Y is a vector containing observed scores (1 = correct, 0 = incorrect) and

$$Y = b_oU + b_1X + E^2, \qquad (2)$$

where U is a unit vector (all 1's) and X contains the value of b-d for each of the observed scores in Y,

the F would be:

$$F = \frac{(ESS\ \ - ESS_1)/(41-2)}{ESS_1\ /\ (41N - 41)},$$

where N is the number of data points at each level of b-d. In order to obtain a significant (p = .05) F with 39 and 41N-41 degrees of freedom (F = 1.51), approximately 33,841 data points are necessary!

For those who like to think in terms of multiple correlation coefficients (R^2), these are

$$R^2 = 1 - \frac{(ESS)}{N\sigma^2}$$

$$= 1 - \frac{(7.7196)\left(\frac{N}{41}\right)}{(N)\ (.25)} = .2469$$

for the Rasch model and

$$R^2 = 1 - \frac{(7.7331)\left(\frac{N}{41}\right)}{(N)\ (.25)} = .2455$$

for the linear model.

A Monte Carlo test of these calculations was performed. This involved creation of 33,825 data points which conformed exactly to the Rasch curve (825 in each of the 41 levels of b-d), and analysis of these data using the two models specified above using Veldman's PRIME subroutine REGRAN. The computed R^2 were .2470 and .2457. The F was 1.513. Thus, for all practical purposes the linear model seems to be an acceptable approximation of the Rasch curve in this range.

Table 6 shows the item numbers, standardized mean square residuals, proportion of the sample obtaining correct responses, Pearson product moment correlations of item-total test scores, least squares slope of the actual item in the ±2 b-d range, the value of the t statistic testing for slope of .2094, and the number of cases on which the t was calculated (i.e., the number of data points in the ±2 b-d range) for a selected set of items.

Table 6 is ordered according to decreasing slope of the least squares lines. Some interesting relationships among the statistics in this table are apparent. Items with steep slopes tend to have low residual mean squares and also have high item-test correlations (CORR). In general, when the difficulty of an item is about .5 (50% correct) and the slope was high (.24+), the item-total correlation was .45 or greater, and the slope was significantly greater than the Rasch model. According to the slope test, only about one-third of

Item #	Mean Square Residual	Proportion Correct	Rasch Difficulty	Correlation	Slope	t	Test-N
156	8.70	.03	5.33	.00	.42	.95	27
48	.75	.98	-3.38	.31	.40	1.73	32
72	.62	.97	-2.78	.37	.39	2.96	65
53	.69	.97	-2.94	.35	.36	2.05	52
138*	.74	.33	2.24	.51	.36	10.45	1428
31	.62	.95	-2.16	.39	.35	3.45	141
136*	.70	.28	2.50	.48	.35	8.98	1351
29	.95	.97	-2.88	.33	.35	1.91	57
10	.54	.84	-.64	.61	.34	7.11	759
206	.64	.87	-.88	.56	.34	6.16	612
17*	.73	.38	1.97	.53	.34	9.63	1495
18*	.74	.45	1.61	.57	.34	10.58	1558
107	.66	.12	3.75	.33	.33	3.62	659
34	.97	.97	-2.68	.30	.33	2.13	71
133*	.73	.65	.64	.60	.33	10.07	1480
96*	.76	.37	2.02	.50	.32	8.42	1484
98	.63	.89	-1.14	.55	.32	4.67	481
97	.64	.90	-1.23	.54	.32	4.23	411
202*	.74	.65	.66	.59	.32	9.33	1479
12*	.78	.48	1.48	.54	.32	9.07	1562
232*	.70	.70	.36	.60	.32	8.39	1370
131	.62	.84	-.63	.59	.32	5.59	759
226	.59	.82	-.43	.61	.32	6.09	878
207	.68	.87	-.92	.54	.32	4.97	599
208	.67	.80	-.24	.60	.32	6.62	993
59	.40	.97	-2.97	.38	.31	1.34	52
14*	.65	.77	-.04	.60	.31	7.15	1130
215*	.85	.51	1.33	.53	.31	8.64	1566
205	.59	.87	-.91	.57	.31	4.67	599
66	.78	.96	-2.40	.34	.31	1.91	97
13*	.79	.57	1.03	.56	.31	8.07	1556
75	.42	.97	-2.94	.42	.31	1.20	52

Item #	Mean Square Residual	Proportion Correct	Rasch Difficulty	Correlation	Slope	t	Test-N
5*	.80	.50	1.38	.55	.30	7.78	1568
90*	.74	.77	-.03	.57	.30	6.29	1130
77	.24	.98	-3.00	.44	.30	1.14	51
15*	.78	.67	.54	.57	.30	7.28	1448
152*	.83	.38	1.99	.47	.30	6.60	1495
16*	.77	.78	-.14	.55	.30	5.95	1066
197*	.69	.75	.08	.59	.30	6.30	1218
166*	.88	.54	1.17	.53	.30	7.13	1567
3*	.76	.68	.51	.58	.30	6.76	1428
101	.67	.80	-.27	.57	.30	5.29	972
236	.75	.68	-1.34	.56	.30	6.73	1428
137	.84	.49	1.41	.51	.30	6.89	1564
135	.85	.44	1.66	.49	.29	6.66	1551
20	.77	.24	2.76	.41	.29	4.79	1249
228	.71	.75	.05	.57	.29	5.78	1200
37	1.03	.57	1.04	.52	.29	6.81	1556
227	.71	.74	.16	.58	.29	5.95	1260
198	.70	.74	.16	.58	.29	5.91	1260
			(79 items omitted)				
92	.74	.94	-1.88	.41	.23	.44	201
83*	.97	.62	.77	.46	.23	1.26	1506
114*	1.07	.58	.99	.40	.22	1.17	1550
87*	1.04	.52	1.31	.41	.22	1.12	1566
7*	.98	.67	.53	.44	.22	1.03	1428
108*	1.11	.42	1.74	.38	.22	.93	1536
28	.56	.92	-1.59	.46	.22	.38	285
210*	1.03	.56	2.14	.42	.22	.94	1564
32	.62	.91	-1.35	.47	.22	.38	366
165*	1.01	.79	-.23	.43	.22	.58	1025
192*	.98	.64	.68	.45	.22	.65	1479

Item #	Mean Square Residual	Proportion Correct	Rasch Difficulty	Corre-lation	Slope	t	Test-N
229*	1.04	.61	.87	.43	.22	.63	1526
62	.87	.85	-.65	.44	.22	.36	759
214	1.22	.13	3.62	.24	.21	.18	753
231*	.95	.64	.67	.44	.21	.37	1479
82	.87	.91	-1.41	.40	.21	.13	343
79	.69	.93	-1.64	.42	.21	.09	266
11	.80	.83	-.52	.48	.21	.13	823
19*	.91	.22	2.9	.33	.21	-.06	1167
27	.66	.94	-1.96	.38	.21	-.05	182
163*	1.13	.75	.09	.39	.21	-.23	1218
167*	.98	.67	.51	.43	.21	-.26	1428
150	.76	.94	-1.93	.36	.20	-.19	186
42	.24	.98	-3.42	.38	.20	-.06	31
151	.78	.89	-1.11	.44	.20	-.44	481
217	1.19	.96	-4.55	.18	.20	-.18	185
38	.80	.92	-1.52	.40	.19	-.48	311
172*	1.22	.27	2.59	.26	.19	-.98	1318
55	.33	.98	-3.03	.37	.19	-.18	44
140	.57	.94	-1.88	.41	.19	-.47	201
190*	1.13	.75	.09	.39	.19	-1.25	1218
47*	.93	.74	-.54	1.04	.19	-1.55	1260
58	1.03	.92	-1.51	.32	.19	-.84	311
132*	1.07	.56	1.10	.37	.18	-1.87	1564
157	.82	.88	-1.03	.42	.18	-1.16	548
143*	1.40	.47	-.01	.27	.18	-2.07	1557
159*	1.06	.72	.28	.39	.18	-2.05	1317
168*	1.35	.27	2.58	.21	.18	-1.83	1318
153	.94	.82	-.39	.41	.17	-1.95	901
			(15 items omitted)				
124*	1.29	.72	.24	.29	.15	-4.12	1302
171*	1.17	.53	1.24	.31	.15	-4.51	1568

Item #	Mean Square Residual	Proportion Correct	Rasch Difficulty	Corre-lation	Slope	t	Test-N
182	1.46	.86	-.77	.25	.15	-3.15	677
187*	1.12	.30	2.37	.27	.15	-3.70	1387
112	1.67	.85	-.68	.20	.15	-3.48	742
158	1.07	.81	-.36	.35	.15	-3.45	917
173*	1.14	.54	1.20	.32	.15	-4.55	1563
44	.77	.85	-.67	.43	.15	-2.94	742
120*	1.05	.78	-.10	.35	.14	-3.90	1083
141	.91	.86	-.82	.36	.14	-3.07	644
105	1.47	.17	3.30	.15	.14	-3.02	940
142	.89	.96	-2.34	.35	.14	-1.15	104
130*	1.09	.73	.19	.35	.14	-4.68	1284
123*	1.27	.75	.04	.31	.14	-4.58	1200
23	.73	.90	-1.25	.42	.14	-2.46	411
117	.97	.84	-.56	.37	.14	-3.60	801
26	.64	.94	-1.96	.37	.14	-1.72	182
116	.54	.98	-3.60	.25	.14	-.58	28
179	1.64	.13	3.64	.13	.13	-2.67	737
71*	1.15	.64	.72	.30	.13	-5.32	1492
40	.69	.89	-1.14	.42	.13	-2.82	481
39*	1.03	.75	.07	.36	.13	-4.85	1218
115	.87	.93	-1.75	.33	.13	-2.23	233
33*	1.30	.73	.18	.30	.13	-5.03	1259
194*	1.15	.59	.95	.32	.13	-5.69	1538
36*	1.31	.75	.07	.27	.13	-5.12	1218
175*	1.61	.70	.34	.21	.13	-5.58	1341
149*	1.46	.78	-.10	.25	.13	-5.02	1083
119	1.08	.87	-.93	.31	.13	-3.60	585
113	.66	.98	-3.19	.24	.13	-.94	38
161*	1.25	.70	.38	.29	.12	-5.59	1370
65*	1.44	.54	1.20	.20	.12	-6.20	1567
84*	1.30	.50	1.37	.25	.12	-6.38	1568
63	.97	.89	-1.10	.32	.11	-3.94	492

Item #	Mean Square Residual	Proportion Correct	Rasch Difficulty	Corre-lation	Slope	t	Test-N
147	.98	.83	-.54	.36	.11	-4.95	801
70*	2.08	.21	3.00	.06	.10	-5.56	1110
178*	2.01	.36	2.08	.06	.10	-7.20	1474
139*	1.55	.36	2.08	.16	.10	-7.32	1475
181	1.50	.79	-.21	.19	.09	-7.65	1025
176*	1.62	.66	.57	.17	.08	-9.06	1448
25	.36	.98	-3.55	.31	.08	-1.00	30
185	3.51	.15	3.46	.01	.08	-5.73	855
169	1.30	.48	1.45	.19	.06	-10.02	1564
109	1.24	.16	3.39	.18	.06	-5.65	880
180	2.31	.33	2.21	.02	.05	-9.82	1445
177	1.45	.69	.44	.16	.05	-10.67	1414
155	1.51	.70	.36	.15	.04	-11.26	1370
209	1.37	.22	2.89	.17	.03	-8.60	1167
52	1.51	.61	.87	.10	.02	-12.96	1526
184	2.24	.31	2.34	.01	-.00	-12.64	1395
160	1.10	.03	5.42	.14	-.02	-.97	22
73	1.77	.53	1.22	-.10	-.04	-17.18	1568
148	2.19	.41	1.82	-.01	-.06	-17.88	1521
146	2.13	.45	1.59	-.14	-.10	-21.75	1553

Table 6.

Statistics on selected achievement test items based on a sample of 1664 elementary students.

*Items which were selected for further analyses -- see Section IV.

the items (78/237) fit the model ($|t|>2.0$ rejection value), and for many of these (25) the test was based on less than 100 cases. Thus, it could be argued that only 53 items fit the Rasch model. A common rule of thumb for rejection of an item is RMSQ>3.0 (Wright and Mead, 1977, p. 51), which would lead to rejection of only 2 items!

Some items in Table 6 are clearly very poor test items from a traditional point of view, and yet appear to be acceptable when the mean square residual is examined. For example, item #73, near the bottom of the list, has a mean square residual of 1.77 and an item-total correlation of -.01! Figure 4 shows the Rasch curve and a plot of item #73. It would be absurd to include this item in a test, at least from any point of view but the mean square residual. Table 7 shows the calculations of the RMSQ for item #73.

From the data in Table 6, an interesting correlation can be calculated. The residual mean square correlates .50 with the Rasch difficulty estimate. This seems strange -- why should a fit statistic be biased toward easy items? This degree of correlation between the mean square residual and the proportion of correct responses is also seen in data which has been published. For the parole data in Perline, Wright and Wainer (1977, p. 13), the correlation is .55. The exam data in Rasch (1960, p. 106) gives rise to a .26 correlation, using Rasch's estimates of the difficulty and mean square residuals calculated by myself.

It might be pointed out that correlations between statistics can be misleading. For example, Perline, Wright and Wainer (1977) attempt to demonstrate that Rasch ability estimates are essentially identical with ability estimates obtained using additive conjoint measurement. The correlations between ability estimates based on two data sets are .997 and .990. However, correlations between Rasch ability estimates and the raw test scores to which they correspond are .994 and .993 on these same data. From this point of view, there is no

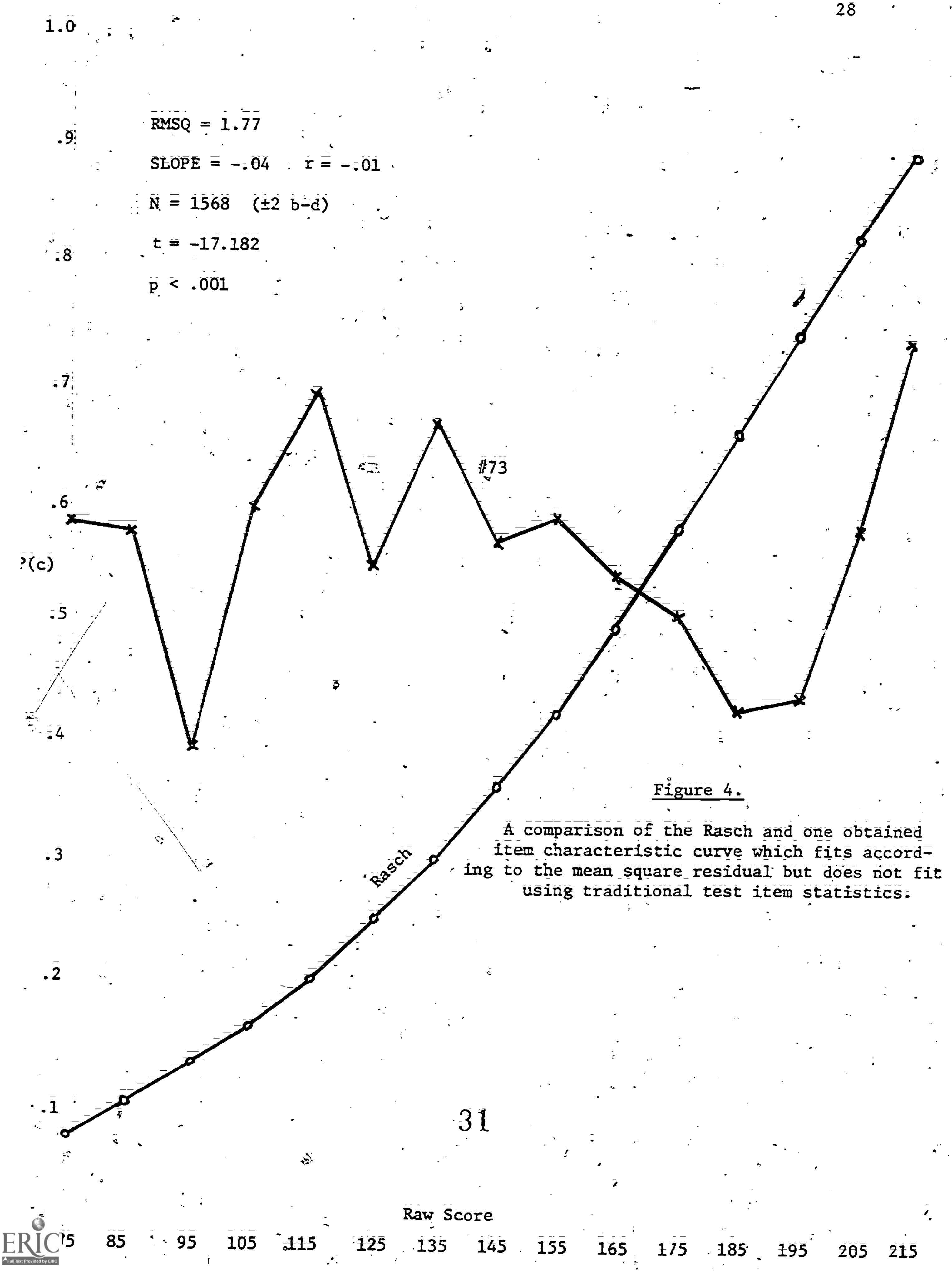


Figure 4.

A comparison of the Rasch and one obtained item characteristic curve which fits according to the mean square residual but does not fit using traditional test item statistics.

	P(c)	N	N(c)	e^{-x}	N($\bar{c}$)	e^{x}	z^2
	0	0					
	0	1	0	180.33	1	.0055	.01
	.20	5	1	79.33	4	.0126	79.38
	.40	5	2	44.97	3	.02	90.00
	.38	8	3	28.81	5	.03	86.58
	.45	11	5	19.80	6	.05	99.30
	.40	5	2	14.24	3	.07	28.69
Portion of curve in Figure 4.	.59	17	10	10.56	7	.09	106.23
	.58	19	11	7.99	8	.13	88.93
	.40	25	10	6.13	15	.16	63.70
	.60	42	25	4.75	17	.21	122.32
	.69	54	37	3.70	17	.27	141.49
	.55	65	36	2.88	29	.35	113.63
	.67	84	56	2.24	28	.45	138.04
	.57	122	70	1.73	52	.58	151.26
	.59	147	87	1.33	60	.75	160.71
	.54	181	98	1.00	83	1.00	180.00
	.51	195	99	.73	96	1.36	202.83
	.43	252	108	.52	144	1.92	332.64
	.44	222	98	.35	124	2.84	386.46
	.58	156	90	.22	66	4.56	320.76
	.75	44	33	.12	11	8.50	97.46
	1.00	4	4	.04	0	22.23	.16
		1664					z^2=2991.78

$$\text{RMSQ} = \frac{\Sigma z^2}{1664} = 1.80$$

Table 7.

Calculation of the fit statistic for item #73 in Figure 4, using data grouped into 10 point raw score intervals.

difference between raw scores and Rasch ability estimates. The correlation coefficient simply is not sensitive to the scale changes in the distance between ability estimates which are introduced by Rasch ability estimates and multidimensional scaling. This scaling is an essential attribute of latent trait modeling because it allows for test-free person measurement, at least in theory.

The practical significance of the .55 correlation can be seen in Table 8, which shows a cross-tabulation of mean square residuals with Rasch difficulty estimates. Notice that virtually all the items which have difficulty estimates of -1.0 or lower (very easy items) also have a mean square residual (RMSQ) of .9 or less which would indicate very good fit to the Rasch model. Items of medium difficulty tend to have low RMSQ, while items of moderate to high difficulty tend to have high RMSQ. These figures indicate that the fit statistic is very sample dependent, which Wright and Mead (1977, p. 50) also point out.

The reason for this relationship between the RMSQ and item difficulty is that many items are answered correctly more often than the Rasch model predicts. This results in low RMSQ for easy items and high RMSQ for difficult items.

How selective must one be in order to construct tests which can be used in test-free person measurement? This is an empirical question that this paper only begins to address. For example, Figure 5 shows close correspondence between the responses to item #220 and the Rasch curve. The slope of this item characteristic curve is significantly steeper than the Rasch curve, but it may not be deviant enough to warrant rejecting the item. That is, test-free person measurement may not be impaired by retaining item #220 as if it fit the model. However, more careful item selection is necessary than can be done using the mean square residual. For example, Figure 6 shows an item which has an item

		Rasch difficulty estimate		
		−∞ to −1.0	−1.0 to 1.0	1.0 to ∞
mean square residual	.00 to .90	60	53	19
	.90 to 1.1	6	31	15
	1.1 to ∞	1	22	30

N = 237

$\chi^2 = 60.34$

$p < .01$

Table 8.

Cross-tabulation of the mean square residuals with Rasch difficulty estimates.

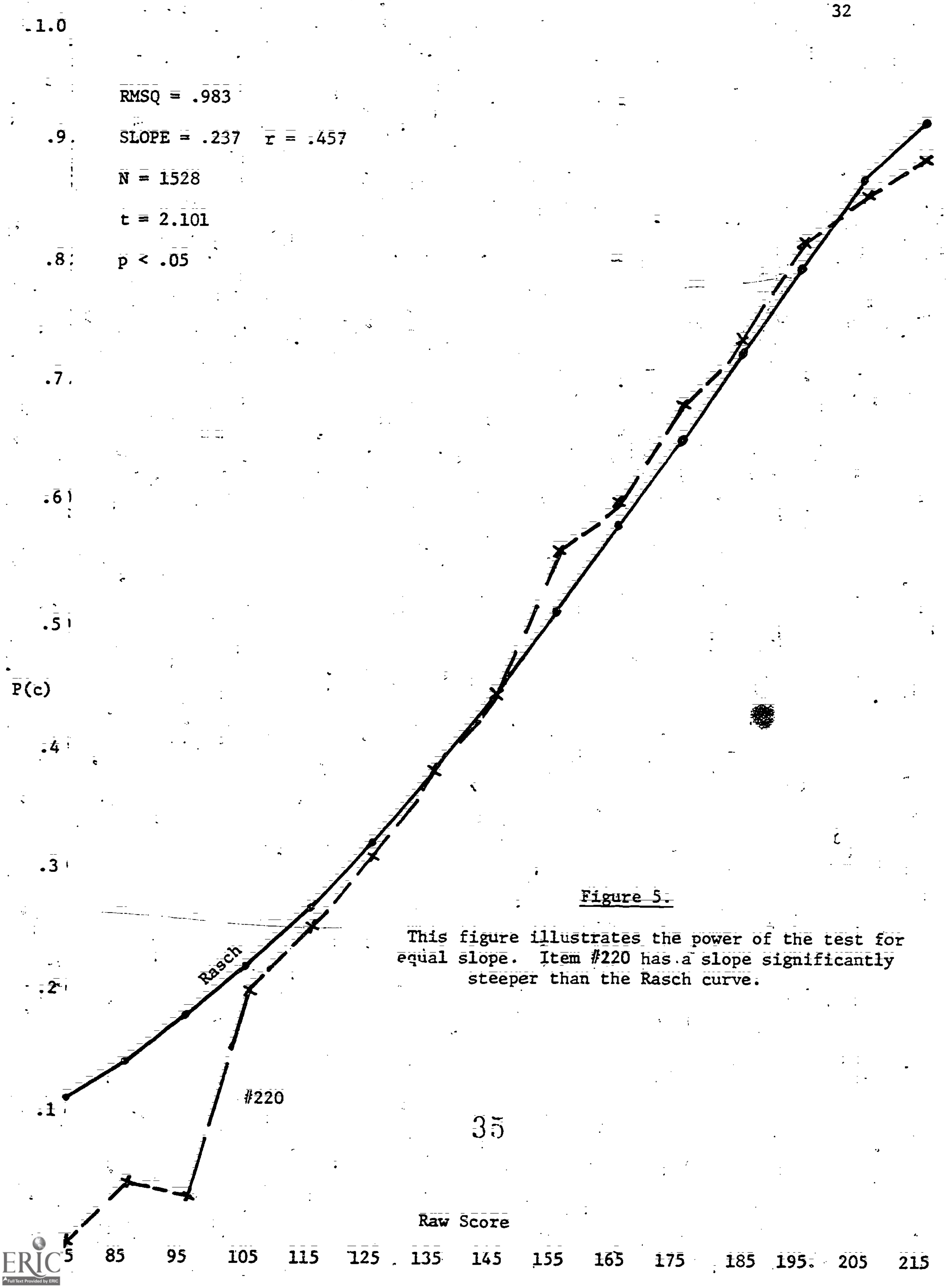


Figure 5.

This figure illustrates the power of the test for equal slope. Item #220 has a slope significantly steeper than the Rasch curve.

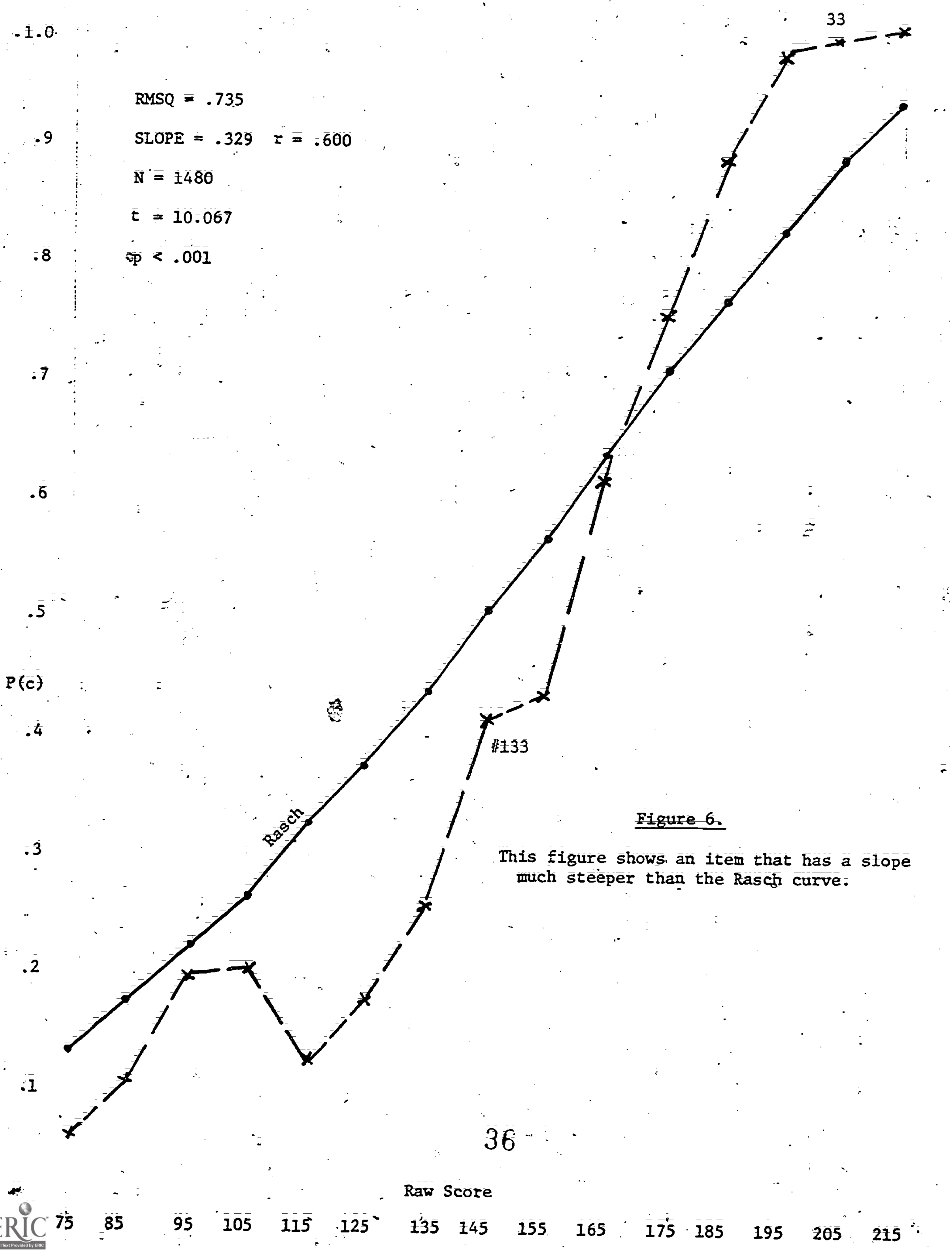


Figure 6.

This figure shows an item that has a slope much steeper than the Rasch curve.

which has an item characteristic curve much steeper than the Rasch curve. Inclusion of items with widely varying slopes in a test violates the basic assumption of Rasch modeling, that all items have the same discriminations (Wright, 1977, p. 103).

Part IV. The Practical Implications of Failure to Reject Items Which Do Not Fit the Rasch Model

From Table 6, it appears that the items on this test battery differ greatly with respect to the slope of the item characteristic curves. Do these differences actually affect the properties of tests which might be built using subsets of the items? Are these differences great enough to impair the use of such tests in a testing program based on Rasch model theory? The slope of the item characteristic curve is cne index of item discrimination, and the Rasch model assumes all items in a test have the same discrimination. If one were to base judgements of acceptability of the items on the mean square standardized residuals, only a few of the items would be unacceptable. What practical consequences might arise if items with slopes as different as are found here are included in tests?

The Estimation of Item Difficulty and Person Ability

In order to answer these questions, four tests of 20 items each were constructed. The first, or "Steep" test contained only items which had slopes of .29 or greater based on a sample size of 1000 or more data points between ±2 b-d. The second, or "Rasch" test contained only items which had slopes not significantly different than .2094, based on 1000 or more data points between ±2 b-d. The third, or "Flat" test contained only items which had slopes less than .15, based on 1000 or more data points between ±2 b-d. A fourth, or "Mixed" test was composed of seven items from the Steep test, six from the Rasch test, and seven from the Flat test.

A sample of 1480 respondents was selected from the 1664, each of which had a score between 0 and 20 on each of the four tests. When each of these tests was analyzed separately, very intersting results were obtained. In both the Steep and Flat tests, the slopes of the item characteristic curves suddenly seemed to be very close to the slope of the Rasch curve. The Steep test no longer seemed to contain items with steep slopes and the Flat test no longer appeared to contain items with flat slopes. The Rasch test still contained items which fit the Rasch curve. Only the Mixed test seemed to contain items with divergent slopes.

Table 9 shows the slopes of the 20 items which were in the Mixed test as they were computed in two contexts. First, in the context of the Mixed test and, secondly, in the context of the tests containing only items with similar slopes. These statistics were based on the same sample of 1480 students.

These results were very disturbing. The Rasch model program apparently adjusts the slope of the items in any given test to the best fit with the Rasch curve. Looking at the output from program BICAL (Wright and Mead, 1977), there is little if any indication that the Flat test is any better or worse than the Rasch and Steep tests. This adjusting of discriminations is alluded to in the manual for use of the BICAL computer program. The program provides a "discrimination index," which "is in fact the linear trend across score groups. Values larger than one indicate that the observed characteristic curve for an item is steeper than the average best fitting logistic curve for all items; values less than one indicate the curve is flatter" (Wright and Mead, 1977, p. 53).

However, there are very substantial differences in the quality of the

Item #		Context	
		Mixed Test	Homogeneous Test
133		.30	.22
96	Steep items	.27	.20
202		.28	.22
12		.30	.25
232		.31	.22
14		.31	.27
215		.27	.19
192		.21	.25
229	Rasch items	.22	.20
231		.21	.23
19		.18	.15
163		.19	.25
167		.22	.25
71		.16	.32
39		.18	.29
33	Flat items	.17	.26
194		.13	.20
36		.14	.21
175		.14	.18
149		.14	.18

Table 9.

Slopes of item characteristic curves for 20 items contained in the Mixed test as computed in that context and in the context of items with similar slopes.

Steep, Rasch, Flat, and Mixed tests. Table 10 shows some traditional test statistics on these four tests. The alpha coefficients readily indicate t the tests which contain items with steeper slopes have higher internal reliability. Item-total correlations, shown in the lower portion of Table 10, also indicate dramatic differences between the items in these tests. The changes in slope observed in Table 9 between the mixed and homogeneous context do not appear in Table 10. That is, the item-total correlations in Table 10 remain fairly constant regardless of the heterogeneity of the discriminations of the items in the test.

In order to verify these results, which were computed using Veldman's PRIME library program RASCH, and to look for further information which might be provided by Wright's program BICAL, similar runs were made using BICAL. Several initial comparison runs verified that all but one of the statistics which were computed by Veldman's RASCH program agreed to the third decimal with those provided by Wright's BICAL when the PROX procedure was specified. The single exception was the fit statistic, which was adjusted by very complex factors in BICAL for reasons noted above. The BICAL program provides much more output, however, and also allows for a supposedly more accurate calibration procedure, referred to as UCON, for "unconditional maximum likelihood estimation." [The "unconditional" refers to the notion that the exact probability of each response vector given a particular total raw score is not computed in the process of estimating item parameters. The unconditional procedure approximates the conditional procedure, and is reported to be less expensive and less subject to round off errors (Wright and Mead, 1977, p. 23).]

It was apparent from the output of these analyses that the procedure for adjusting the average slope of the item characteristic curves to that specified by the Rasch model (regardless of the data fed into the program) involves

	Steep	Rasch	Flat	Mixed
Mean	31.68	31.45	31.95	32.59
Sigma	5.57	4.13	3.17	4.32
Alpha	.91	.78	.61	.81

Item-Total Correlations Within Different Contexts

Item #		20 Item Mixed Test	20 item Homogeneous Test
133		.63	.64
96	Steep test	.54	.57
202		.60	.64
12		.59	.64
232		.63	.63
14		.61	.65
215		.56	.59
192		.49	.50
229	Rasch test	.48	.43
231		.48	.49
19		.38	.37
163		.42	.46
167		.45	.49
71		.38	.47
39	Flat test	.40	.45
33		.37	.41
194		.37	.35
36		.34	.36
175		.25	.26
149		.30	.30

Table 10.

Some traditional test statistics for the items in the 20-item Mixed test computed in that context and the context of the homogeneous tests.

adjustments of the estimated abilities of the students and difficulties of the items. Recall that Rasch model item characteristic curves use ability minus difficulty (b-d) as the abscissa (X axis). Even minor tranformations of the values of the abscissa can easily affect the calculated slope of a line. To reduce the slope of a line by one-half, simply double the values of the abscissa.

The procedure used to adjust the slope of an item characteristic curve to that of the Rasch model does this by transforming the scale of the ability and difficulty estimates. For example, the range of ability in the sample of 1480 students was estimated to be 7.43 units based on the Steep test and 6.51 units based on the Mixed test. Item difficulty estimates also vary systematically. The distance between item #96 (the most difficult item common to the Steep and Mixed tests) and item #14 (the easiest item in common) was estimated to be 2.66 units when difficulty estimates were obtained in the 20-item Steep test and 2.22 units when based on the 20-item Mixed test, using the same sample of 1480 students' data. The difference in the scale of ability estimates is combined with the difference in the scale of difficulty estimates when the ability - minus - difficulty (b-d) scale is formed to produce a change in the apparent slope of an item characteristic curve. Notice in Table 9 that the slope of the item characteristic curve for each item in the Steep test is less when based on the homogeneous (Steep) test than when based on the Mixed test.

It has been stated that "item discrimination cannot be estimated directly or efficiently in the way Rasch item difficulty and person ability can" (Wright, 1977, p. 104). Wright criticized Lord (1975) for imposing arbitrary constraints on his procedure for estimating discrimination, but it seems that Wright's procedure also imposes a very rigid constraint -- that the "average best fitting logistic curve for all items" (Wright and Mead, 1977, p. 53) must be the same as the Rasch model. The Rasch model does not have a parameter for item discrimination; this does not mean that the Rasch model does not specify what the discrimination is -- it merely asserts that every item has the same discrimination. In order to demonstrate that the items in a particular test fit the Rasch model, it is necessary to show that they have the particular discrimination specified by that model. This is done by adjusting the ability and difficulty estimates and using these as the abscissa when plotting the item characteristic curves.

Person-Free Test Calibration

What effect do variations in the slope of the item characteristic curve have on person-free test calibration? In order to investigate this, three ability groups were defined out of the sample of 1480. Wright and Mead (1977, p. 46) suggest that extremely high and low scores should not be included in item difficulty calibration attempts. Therefore, students with more than 49 or less than 17 correct responses to the 60 items in the Steep, Rasch, and Flat tests were excluded. The low ability group contained 494 students with scores of 17 through 31. The high ability group contained 494 students with scores of 40 through 49. A medium ability group contained 367 students with scores of 32 through 39. A total of 1355 students thus remained, divided into these three ability groups.

A new Mixed test was also defined, containing 10 items from the Steep test and 10 from the Flat test. This Mixed test was composed of items which had very nearly the same level of difficulty in order to illustrate as clearly as possible the potential hazards of including items with different discriminations in a single test. Between 25% and 68% of the 1355 students marked each of these items correctly.

Difficulty estimates for the 20 items in this Mixed test were calculated using the high and low ability groups. Table 11 shows the estimated difficulties

Difficulty Estimates

Item #	Ability Group High	 Low
17	-.45	.23
18	.37	1.58
133	-.08	1.18
202	-1.75	-.16
12	-1.45	-.39
215	-.04	.79
13	-.14	.39
5	-.74	.17
152	-.24	.64
166	.70	1.16
171	.33	-.33
173	.38	-.48
71	-.26	-.74
194	.16	-.78
175	-.14	-1.59
161	-.61	-1.03
65	.39	-.44
84	.68	-.44
178	1.53	.05
139	1.37	.18

Table 11.

Difficulty estimates for a set of 20 items based on high and low ability groups.

of each item based on two ability groups. These difficulty estimates were computed using Wright and Mead's program BICAL. The correlation between difficulty estimates based on the high and low ability groups is -.42. "If gross variation in item discrimination is tolerated in the final pool of test items, then the possibility of person-free test calibration is lost (Wright, 1968, p. 100).

Table 12 shows the fit statistics for each of these 20 items as reported by program BICAL for the high and low ability groups and the total sample. Notice that only the "Between Group Fit Mean Square" based on the total sample gives any indication that these items do not fit the Rasch model. According to Wright and Mead, "the between group mean square tests the agreement between the observed item characteristic curve and the best fitting Rasch characteristic curve as estimated by the groups selected" (1977, p. 52). A non-significant between groups mean square "indicates that statistically equivalent estimates of difficulty would result from using either the low scores or the high scores for calibration (Wright and Mead, 1977, p. 51). Even so, they do not recommend dropping items from a test if they have high between group fit mean squares. This statistic is apparently the only indication on the BICAL printouts that there may be something amiss with the Mixed test. Even so, it only indicates this when these items have been calibrated on a sample containing a wide distribution of ability. When the calibrations and fit statistics are based only on the high or low ability groups, there is no way to tell from the BICAL printouts that the difficulty estimates would not be stable across ability groups.

The reason these difficulty estimates are so different when based on the high and low ability groups is that the item characteristic curves cross. Items from the Flat test are "more difficult" within the low ability group, but "less difficult" within the high ability group. Rasch model proponents have clearly recognized that items which have crossing characteristic curves

	Between Group Fit Mean Squares			Total Fit Mean Squares		
Item #	High	Low	Total	High	Low	Total
17	1.37	1.08	7.41	1.03	.95	.92
18	2.35	2.75	15.72	1.02	.82	.83
133	1.58	2.15	20.69	.99	.84	.82
202	1.68	3.67	15.55	.84	.99	.82
12	.33	1.81	9.62	.98	1.01	.87
215	.90	3.50	14.89	.98	.85	.86
13	2.39	2.94	4.48	.95	1.01	.94
5	1.82	2.43	13.86	.89	.94	.85
152	.97	2.16	10.30	1.02	.91	.90
166	.43	1.57	3.50	1.07	.86	.93
171	2.02	1.12	2.13	.98	1.01	1.08
173	.86	1.49	6.87	1.08	1.05	1.18
71	.86	.32	1.83	.99	1.11	1.10
194	.35	1.34	6.80	1.05	1.08	1.18
175	1.55	1.03	19.02	1.08	1.08	1.34
161	.82	1.18	2.13	.98	1.12	1.09
65	.15	2.65	7.18	1.05	1.13	1.19
84	2.53	.67	10.42	1.05	1.08	1.22
178	3.31	4.90	28.54	1.12	1.15	1.40
139	.94	1.91	11.78	1.06	1.06	1.22

Table 12.

Residual mean square fit statistics for the 20 items in the Mixed test computed by program BICAL using three sa., .es: high ability, low ability, and total group.

cannot be allowed to remain in a test if Rasch model procedures are to be used (Wright, 1968). What I have demonstrated is that the standardized residual mean square fit statistic does not provide the information necessary to select items with similar slopes.

Vertical Equating

These differences in difficulty estimates affect the results of attempts to link tests, also. For example, no items are common to both the Steep and Flat tests, but they can be linked through the Mixed test. Ten items in the Mixed test are common to the Flat test and ten to the Steep test. Following the procedure specified by Wright (1977, p. 107), we can link the Flat and Steep tests using the calibration of the Mixed test.

The constant necessary to translate all item difficulties in the calibration of the Mixed test onto the scale of the Flat test would be

$$t_{mf} = \sum_{i=1}^{10} (d_{im}-d_{if})/10.$$

The constant of the translation between the Steep test and the Mixed test would be

$$t_{sm} = \sum_{i=1}^{10} (d_{is}-d_{if})/10.$$

Table 13 shows the difficulties estimated for the linking items on these three tests, with the high and low ability estimates for the Mixed test. Using the difficulty estimates based on the low ability group, the two constants are

$t_{mf} = .830$ and

$t_{sm} = .404.$

Using the difficulty estimates based on the high ability group, the

Item # N	Flat 1355	Mixed low(494)	 high(494)	Steep 1355
17		.233	-.453	1.199
18		1.580	.371	.735
133		1.180	-.077	-.526
202		-.158	-1.750	-.549
12		-.393	-1.451	.590
215		.787	-.035	.393
13		.393	-.140	-.058
5		.172	-.745	.389
152		.642	-.239	1.238
166		1.162	.696	.159
171	.306	-.333	.327	
173	.312	-.478	.380	
71	-.074	-.735	-.261	
194	.026	-.785	.161	
175	-.660	-1.593	-.140	
161	-.450	-1.034	-.613	
65	.377	-.436	.389	
84	.476	-.444	.680	
178	1.225	.055	1.533	
139	1.160	.182	1.368	

Table 13.

Linking the Flat and Steep tests through the Mixed test, using either high or low ability groups to calibrate the difficulties of items in the Mixed test.

two constants are

$$t_{mf} = -.117 \quad \text{and}$$

$$t_{sm} = -.739.$$

Thus, the difference between the difficulties of the Flat and Steep test is estimated to be 1.234 units using the link based on low ability students and -.856 units using the link based on the high ability students. The average scores of the 1355 students to these tests were 11.19 (Steep test) and 12.02 (Flat test), indicating that the Flat test is slightly easier than the Steep test.

The ability estimate for a person with a score of 11 on the Steep test is .24 according to the BICAL printout based on the 1355 students in the sample and the 20 items in the test. However, translating to the scale of the Flat test, a score of 11 on the Steep test would reflect an ability of

$$b_r = 1.234 + \left[\sqrt{1 + \frac{1.960}{2.89}}\right] ln\left(\frac{11}{20-11}\right) = 1.49$$

using the low ability link and

$$b_r = -.856 + \left[\sqrt{1 + \frac{1.960}{2.89}}\right] ln\left(\frac{11}{20-11}\right) = -.60$$

using the high ability link. Such a great difference (2.09 units) would be intolerable in a testing situation. (In terms of a more familiar metric, -.60 is roughly equivalent to the 25th percentile and 1.49 to the 76th percentile on the Steep test.)

To my knowledge, no other study has reported such unacceptable results using the Rasch model. The selection of items for this test of sample-free item calibration and vertical equating was intended to produce as dramatic results as possible. The fact that many other studies have obtained at least marginally acceptable results using the Rasch model indicates that the Rasch model is fairly robust. That is, even though variation in item discrimination can have dramatic, adverse effects on item calibration and test linking; in many situations such effects have not been observed. Future studies need to monitor item discriminations more closely and thus determine the extent to which the assumption of equal discrimination can be violated in practice.

Part V. Conclusions and Suggestions for Further Research

This investigation has demonstrated two things. First, the discriminations of all the items in a test must be very similar in order for Rasch model analyses to work in practice. Second, the standardized residual mean square fit statistic does not detect unacceptable variation in discrimination. These findings were used to construct tests in which item discriminations were dissimilar enough to produce discrepancies in difficulty estimates based on high and low ability groups. These discrepancies were then shown to lead to serious errors in test linking. The major implication is that item discrimination needs to be monitored and controlled using more exact tests of fit than the residual mean square. If item discrimination is carefully controlled, then vertical equating, sample-free test calibration and test-free person measurement might be possible.

The successful use of the linear model to determine the slope of item characteristic curves may be indicative of its potential as a latent trait model. If one were to restrict the range of specification of the model to, say, $.1 < P(c) \leq .9$, then the linear model is practically equivalent to the logistic model (see Figure 4).

The linear model has two parameters: the slope (or discrimination) of the item and the intercept (or difficulty) of the item. One might define the difficulty of an item as the point at which the item characteristic curve indicates a .5 probability of success. Previously in this paper, the linear model was specified as $P(c) = mx + b$, where m is the slope of the line, x equals ability minus difficulty and b equals .5, the expected value of P(c) when ability minus difficulty equals zero. The model could be reformulated by defining x as the ability of the person and b equal to $(.5-m\delta)$, where m is the slope of the line and δ is the difficulty of the item. Least squares estimators of m and b are available which are unbiased, efficient, and consistent. These estimators are very well suited to the problem at hand because tests of hypotheses concerning m and b are very robust to violations of assumptions of equal variance and normality of the distributions of P(c) given x (George, 1977). These are exactly the type of distributions that occur in test item data.

Perhaps the major difficulty with estimating fit of items to any model is that few models do well outside of the $.1 < P(c) < .9$ range. I suggest that one does not need to model performance outside of this range. Indeed, a fully consistent program of individualized testing would not use items which are too easy or difficult to estimate ability. Nor would data from persons with abilities too far above or below an item's difficulty be used to estimate that difficulty. It is likely that, regardless of which model one wishes to use, data outside the .1 to .9 P(c) range adversely affects the estimates of the parameters of the model. Thus, it should be possible to devise algorithms which improve the estimates of ability and difficulty and discrimination by eliminating items and persons which do not provide reliable information.

One example of the use of the linear model would be in selection of a set of items which satisfies the assumption of the Rasch model that all items in a test have the same slope. If we define this to mean that no two items in a test have item characteristic curves which cross in the $.1 < P(c) < .9$ range, the linear model can be used to select such items. In order to determine whether two item characteristic curves cross in this range, we specify their equations to be $P(c) = m_1\beta + .5-m_1\delta_1$ and $P(c) = m_2\beta + .5-m_2\delta_2$. The point of intersection is

$$pi = \frac{m_1(.5-m_2d_2) - m_2(.5-m_1d_1)}{m_1 - m_2},$$

where d_1 and d_2 are the estimated difficulties of the items. If pi is less than .1 or greater than .9, then both items would be included in the test. If pi is greater than .1 and less than .9, then the item with the steeper slope should be retained and the item wi-h the flatter slope dropped from the test.

This selection procedure was applied to the 237 item test discussed previously using Rasch difficulty estimates provided by the PROX procedure and slope estimates based on the responses of individuals within ±2 (b-d) units (see Table 6). Thirty items were retained, each of which did not cross any of the others retained. Every other item in the test (i.e., 207 items) crossed one or more of the retained items and had a flatter slope. Table 14 shows the item numbers, slope and difficulty estimates of these 30 items. Other statistics for most of these items are included in Table 6.

This procedure should provide a very good method of building tests that have all the properties desired by Rasch model proponents. The range of difficulty estimated by the PROX procedure for the 30 items is -2.16 to +4.81 in the context of the 237 items. It may simply not be necessary to use any other items from the 237 to obtain as accurate an estimate of ability as is possible. The alpha coefficient for this 30 item test is .92, based on the 1664 persons described earlier. (Rasch model item difficulty estimates in the context of the 30 item test range from -2.54 to 5.61, further illustrating the context specific nature of these estimates.)

Another aspect of latent trait modeling that is very promising is the

Item #	Slope of Linear Model	Rasch Difficulty Estimate
10	.342	-.639
12	.320	1.482
13	.306	1.033
14	.315	-.042
15	.301	.539
17	.337	1.974
18	.336	1.608
20	.295	2.764
31	.348	-2.159
35	.286	-1.907
96	.324	2.020
97	.322	-1.232
98	.323	-1.139
107	.333	3.755
126	.277	-1.338
127	.290	.223
133	.329	.644
136	.347	2.502
138	.356	2.241
162	.264	4.813
166	.296	1.172
197	.299	.083
198	.293	.157
206	.339	-.886
207	.317	-.918
208	.316	-.244
215	.314	1.327
226	.317	-.434
227	.293	.157
232	.318	.356

Table 14.

Items retained by testing for non-intersection of item characteristic curves in the range of .1 to .9 probability of correct response.

hope of determining when items and tests are biased for or against specific individuals or groups of individuals. The Rasch model fit statistic discussed earlier has been proposed as an index of such bias (Wright, Mead and Draba, 1976). However, it is clear from this paper how inadequate that statistic is. The linear model might provide a practical means of detecting such bias by letting x equal item difficulty instead of person ability. The slope (m) might then be a good index of the extent to which each person brought the target ability to bear on the items in the test. In addition, the standard error of estimate can be calculated for each parameter in this model. This statistic provides an index of the accuracy of the estimate of each person's ability.

Ultimately, the future of latent trait models depends upon the acceptance they receive from those who do testing on a daily basis. It is important that these people understand the models we propose and how they should be used. Most practitioners have worked with linear models in other contexts. This is one more reason I advocate a linear model over the logistic models currently receiving the most intense analyses.

References

Andersen, E. B. Sufficient statistics and latent trait models. Psychometrika, 1977, 42, 69-81.

Anderson, J., Kearney, G. E., & Everett, A. V. An evaluation of Rasch's structural model for test items. The British Journal of Mathematical Psychology, 1968, 21, 231-283.

Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D. C.: American Council on Education, 1971, 508-600.

George, A. Effects of violations of assumptions on the power and robustness of the F-test for equality of means using fixed effects linear regression. Dissertation, University of Texas, 1977.

Lord, F. M. A theory of test scores. Psychometric Monograph, 1952, No. 7.

Lord, F. M. The relation of test score to the trait underlying the test. Educational and Psychological Measurement, 1953, 13, 517-548.

Lord, F. M. Relative efficiency of number-right and formula scores. British Journal of Mathematical and Statistical Psychology, 1975, 28, 46-50.

Panchapakesan, N. The simple logistic model and mental measurement. Doctoral dissertation, University of Chicago, 1969.

Perline, R., Wright, B. D., & Wainer, H. The Rasch model as additive conjoint measurement. Research Memorandum No. 24, Statistical Laboratory, Department of Education, The University of Chicago, 1977.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Slinde, J. A., & Linn, R. L.. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Tinsley, H. E., & Dawis, R. V. An investigation of the Rasch simple logistic model: Sample free item and test calibration. Educational and Psychological Measurement, 1975, 35, 325-339.

Veldman, D. The PRIME system: Computer programs for statistical analysis. Austin: Research and Development Center for Teacher Education, The University of Texas, 1978.

Ward, J. H., & Jennings, E. Introduction to linear models. Englewood Cliffs, New Jersey: Prentice-Hall, 1973.

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

Wright, B. D. Sample free test calibration and person measurement. Proceedings of the 1967 invitational conference on testing problems. Princeton, New Jersey: Educational Testing Service, 1968, 85-101.

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

Wright, B. D., & Mead, R. J. BICAL: Calibrating rating scales with the Rasch model. *Research Memorandum No. 23*. Chicago: Statistical Laboratory, Department of Education, The University of Chicago, 1976.

Wright, B. D., Mead, R. J., & Draba, R. E. Detecting and correcting test item bias with a logistic response model. *Research Memorandum No. 22*, Statistical Laboratory, Department of Education, The University of Chicago, 1976.

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