Open access • Journal Article • DOI:10.1177/014662167700100413

Applicability of the Rasch Model with Varying Item Discriminations.

Thomas E. Dinero, Edward H. Haertel

Institutions: Kent State University, University of Chicago

Published on: 01 Sep 1977 - Applied Psychological Measurement (SAGE Publications)

Topics: Item analysis, Rasch model, Polytomous Rasch model, Item response theory and Goodness of fit

Applicability of the Rasch Model with Varying Item Discriminations

Thomas E. Dinero
Kent State University

Edward Haertel
University of Chicago

Among the varieties of logistic models, those attributed to Birnbaum (involving the parameters of item discrimination, item difficulty, and person ability) and Rasch (involving only item difficulty and person ability) have received attention. The present research simulated the responses of 75 subjects responding to 30 items under the Birnbaum model and then attempted a fit to the data using the Rasch model. When item discriminations varied from a variance of .05 to .25 within distributions of different form (uniform, normal, and positively skewed), the poorest overall fit appeared within the uniform distribution. For each distribution there was only a slight increase in the lack of fit as the variances increased.

In 1960, Rasch (cited in Lord & Novick, 1968; Whitely & Dawis, 1974; Wright & Panchapakesan, 1969) presented three models to explain misreadings, number of words read, and general achievement. Each of these is a two-parameter model, incorporating only the ability of each person and the difficulty of each test item to explain the observed data. The most impressive implication of the models is that item calibration and individual measurement are independent both of each other and of the context in which they take place.

In the classical model of item analysis, two principal characteristics of an item merit attention—item difficulty and item discrimination. In many situations, these indices seem to offer the test users important information about their tests. Most champions of the classical model would be careful to admonish users to be sensitive to the interdependency of their results and the subjects who have yielded them.

The suggestion, as in the Rasch model, that the probability of a correct response to an item depends only upon the examinee's ability and the difficulty of the item is an attractive one. Without the complicating effects of item discriminations, individuals are clearly pitted against their criterion, and should supply neatly interpretable data. Of course, whether the picture is as clear as this has yet to be shown.

The present research simulated data from several hypothetical tests for which the effects of item discrimination varied, using a two-parameter logistic model. Fitting of Rasch's model was predicted on the fact that his model may be understood as a one-parameter logistic function. With this bridge to more general models, then, the Rasch assumption of equal item discriminations could be tested.

The ICC Model

The clearest demonstration of the relationship between the person and the item is the item

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 4 Fall 1977 pp. 581–592

characteristic curve. This is a hypothetical plot of the probability of getting a particular item correct as a function of the latent ability of the person (or class of people). At least since Guilford (1936), it has been assumed that, within the ability range of the test, this probability is best described by the normal ogive function.

The logistic test model

$$\Psi = \frac{e^x}{1+e^x} \qquad [1]$$

is much simpler and more efficient computationally. It provides results essentially identical to the normal ogive model since Haley (cited in Birnbaum, 1968) has shown that if $\Phi(X)$ is the cumulative normal distribution function, then

$$\left| \Phi(x) - \Psi(1.7x) \right| < 0.01 \qquad [2]$$

for all X. Hence, the logistic ogive is almost indistinguishable from the normal ogive after a linear transformation on X. Within the context of test theory this model takes on a specific form credited to Birnbaum,

$$P_g(\theta) = \Psi\left[\beta_i(\theta - \alpha_i)\right] \qquad [3]$$

where α is the difficulty of item i, β is the item discrimination, and θ is examinee ability. The probability density function of this model is

$$P_g(Ug=1 \mid \theta) = \frac{e^{\beta_i(\theta-\alpha_i)}}{1+e^{\beta_i(\theta-\alpha_i)}} \qquad [4]$$

$$P_g(Ug=0 \mid \theta) = \frac{1}{1+e^{\beta_i(\theta-\alpha_i)}} \qquad [5]$$

where U_i is 1 if the examinee responds correctly on item i and U_i is 0 if he/she does not (Birnbaum, 1968). This, then, is the most general statement of the two-parameter logistic model, incorporating maximum information about the item and the examinee.

Rasch (1966a,b) has presented a model that can be seen as a simplification of this, in which $X = \beta(\theta - \alpha)$ can be explained in terms of θ and α alone, because the item discrimination (β) has been assumed constant across items (therefore, without loss of generality, equal to one). The implications of this lie in the fact that one can estimate θ independently of α and vice versa. As Wright and Panchapakesan (1969) have indicated, however, the model implies that:

> 1. the model (that is, the latent trait underlying the responses to all items on the test) is unidimensional;
> 2. there are no strong relationships among persons or items other than those specified by the model so that responses of persons to items are stochastically *independent* given their parameters in the model;
> 3. items and persons do not differ substantially with respect to other response factors not represented in the model such as item discrimination, person sensitivity, guessing, or indifference. (p. 2)

Wright and Panchapakesan have also indicated that since few test authors can write items at a predetermined discriminating level, "grossly dissimilar items" should be discarded (p. 4), resulting in a set of items with "similar discrimination and minimal guessing." If item writers were to have a decision rule for doing this, they would then be assured *a fortiori* of building a test in conformity with the Rasch model.

The present paper describes one solution to the establishment of such a criterion. A monte carlo computer program using the Rasch model was designed to input person and item parameters, generate probabilities of success, simulate a test-taking situation, produce the raw item score matrix, and estimate the parameters of the Rasch item characteristic curve. All four subsections may be used independently of each other, parameters can be read in or generated internally, and link-ups with other subsections are determined *only* by the intent of the user. The subsection which estimates the parameters of the Rasch item characteristic curve will accept as input either a raw item score matrix or a matrix of probabilities of success. In addition, the data-generating function follows Birnbaum's (1968)

two-parameter model, and the data calibration follows Wright and Panchapakesan (1969). This allowed the present methodology: generation of data using Birnbaum's model for simulation, and analysis of this data using Wright and Panchapakesan's calibration based on Rasch's model. Poor calibration would then suggest lack of robustness of the Rasch calibration to departures from homogeneity of item discriminations.

Method

The Simulation Program

There were three general sections in the present simulation. The first phase read item difficulties, discriminations, and person abilities, or generated them according to user specifications. Following this, the parameters were combined according to the Birnbaum formulation into a person × item matrix of probabilities. In the second (and actual simulation) phase, a series of uniformly distributed random numbers between 0 and 1 was generated. These numbers were compared with the probabilities generated in phase one and the "raw data" matrix was generated according to the rule:

$a_{ji} = 1$ if $P(a_{ji} = 1) >$ random number

$a_{ji} = 0$ if $P(a_{ji} = 1) <$ random number

The matrix of a_{ji}'s could have been read in at this point instead of being generated.

The third phase involved item calibration based on either the matrix of raw item scores or the person × item matrix of probabilities generated in the first phase of the simulation. Rasch (1966a,b) has shown that, assuming the one-parameter model, unweighted total scores (that is the sum of the a_{ji} for person or score group j) are sufficient statistics for estimating latent ability, which is denoted by θ. Wright and Panchapakesan (1969) have elaborated Rasch's original least squares approach; in addition, they have presented a maximum likelihood estimation procedure which is more precise.

Several points need to be made about this estimation procedure. First, there is one and only one ability level for any one score (or score group). Second, item calibration (that is, determining the alpha or item difficulties) generally precedes person measurement (determining the thetas or person abilities). Third, if any row or column of the data matrix contains all 1's or 0's, the corresponding score group or item cannot be used in the calibration. In other words, items with difficulties of 0.0 or 1.0 do not discriminate among examinees and must be eliminated; and persons with either perfect scores or scores of zero must also be eliminated.

The maximum likelihood method of estimation is treated briefly here; the discussion closely follows Wright and Panchapakesan (1969). The estimations of the item difficulty and person ability are based upon the assumption that, within any score group (i.e., that set of examinees who received identical raw scores), the probability of success on item i is approximately the proportion of examinees with the total score for that group who responded correctly to item i. The parameters of the model are estimated in such a way that the predicted probabilities of success for each score group on each item approximate these proportions. Since only the arithmetic differences of person ability parameters and item difficulty parameters appear in the model, adding a constant to all item difficulty and person ability parameters would not affect the model at all. This indeterminacy is typically resolved by setting the mean item difficulty equal to zero.

The standard error of the item difficulties is derived from the variances of the probabilities predicted by the model, under the assumption that the actual responses to a given item within a given score group are binomially distributed. Wright and Panchapakesan (1969) showed these values to be

$$\sigma_{\alpha_i} = \left(\sum_{j=1}^{k-1} r_j \frac{e^{\theta_j - \alpha_i}}{(1+e^{\theta_j - \alpha_i})^2} \right)^{1/2} \quad [6]$$

where r_j is the number of persons in score group j. The standard error of the person ability parameters depends in part on the uncertainty of the item difficulty parameters and is somewhat

more complex. The interested reader is referred to Wright and Panchapakesan (1969, Equation 29).

The maximum likelihood estimation procedure is necessary only for item calibration; once generated, the item estimates can be used to calculate person abilities directly. Initially, the implicit equations for item difficulties and person abilities are solved simultaneously using an iterative procedure. Once the items have been calibrated, the ability estimate for any examinee depends only upon his/her total raw score. Moreover, any set of calibrated items may be combined to form a new test; and a similar set of implicit equations may be solved iteratively to determine the ability estimate corresponding to any possible raw score on the new test. Additionally, these estimates of α may be used to calculate the standard error of the estimate of ability (θ) corresponding to each raw score.

For each item, its goodness-of-fit to the Rasch model is computed by forming a standard deviate

$$y_{ji} = \frac{aji - \mathcal{E}(aji)}{V(aji)^{\frac{1}{2}}} \qquad [7]$$

where a_{ji} is the obtained item score for person j on item i, $\varepsilon(a_{ji})$ is the expectation of a_{ji} based on item difficulty (α_i) and person ability (θ_j), and $V(a_{ji})^{1/2}$ is the standard deviation of the a_{ji}. The squares of the standard deviates summed over people yield an approximate χ^2 with $N - 1$ degrees of freedom which can be used to test the fit of item i to the model, where N is the number of standard deviates entering the sum.

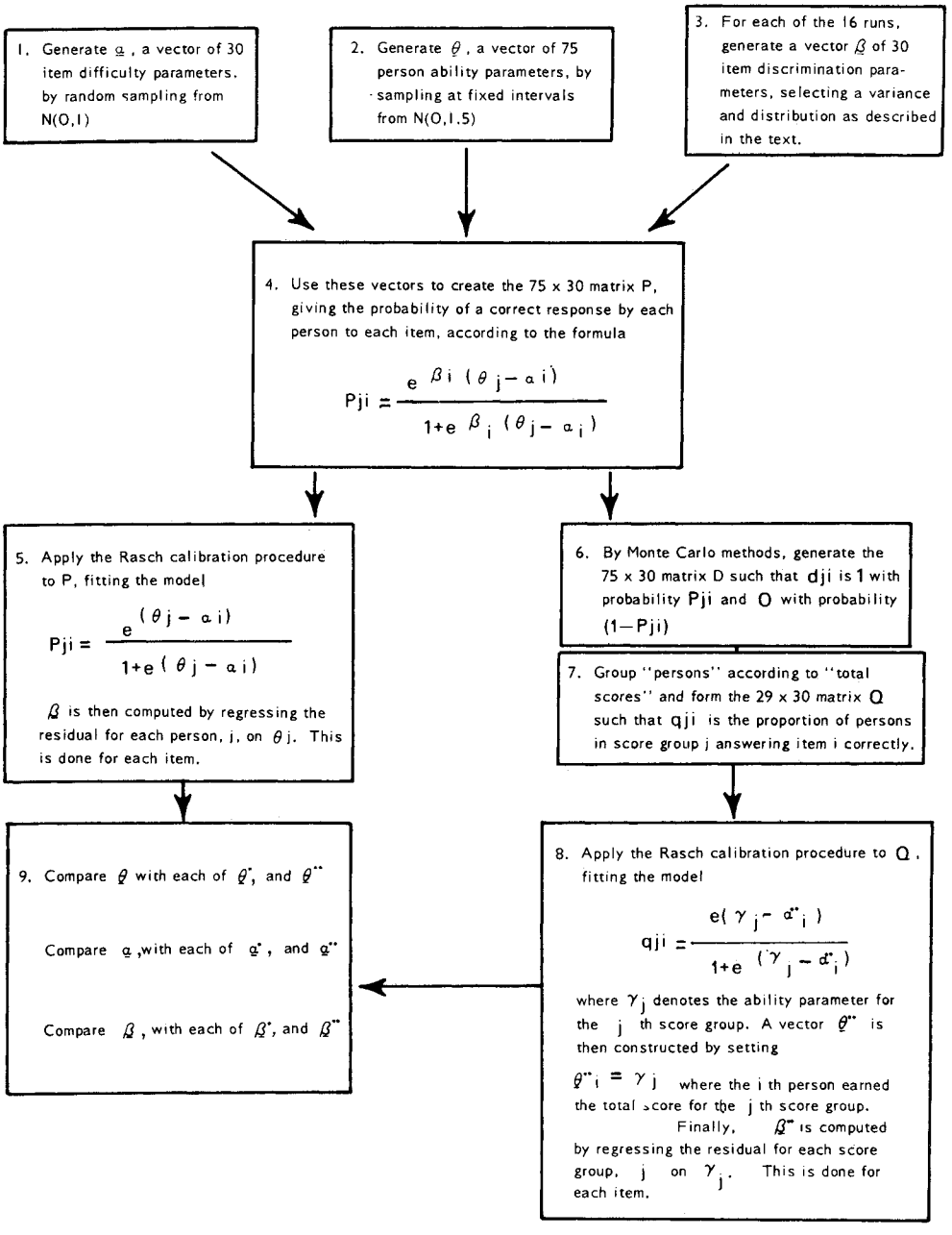
Procedure

The plan of the study, described immediately below, is also summarized in Figure 1. The central concern of the present research was the effect of item discriminations on fit to the Rasch model. It was believed that a certain tolerance is allowed in the application of the theory, although the exact amount was unknown. Goodness-of-fit should vary as a function of the degree to which item discriminations are the same, that is, the degree to which they do not vary among themselves. This degree of fit was therefore examined as a function of the variance of the item discriminations. For the present simulations, variances were assumed to be .05, .10, .15, .20, and .25. One run was also made at $\sigma_\beta^2 = 0$ to examine the degree of accuracy of the item calibration. As there was also some question about the shape of the distribution of these values, this quality was also varied. Three distributions were used: normal, uniform, and positively skewed. This latter form is thought to be the most reasonable for a well-constructed test, since discrimination values should never be negative; with a mean of one, the distribution would skew right. The actual shape was operationalized as approximately a chi-square distribution with one degree of freedom. Since the Rasch procedure automatically scales the person and item parameters in such a way as to make the average item discrimination equal to one, this value was taken as the mean of all distributions studied.

There were thus 16 simulation runs, one for the pure Rasch model and three at each degree of discrimination variability; the distributions of item discrimination all had means equal to one. All item and person parameters except item discrimination were the same for all 16 runs. For each run a test length of 30 items was employed, with item difficulties randomly sampled from a normal distribution with mean 0 and standard deviation 1. The obtained random sample which was used for all runs had a mean of .113 and a sample standard deviation of .940, with a range from −1.553 to 2.070.

For each of the 16 simulation runs (see Table 1 for an outline), two calibrations were performed. First, the person by item matrix of probabilities was calculated directly from the known parameters. This hypothesized matrix of probabilities was calibrated using the maximum likelihood procedure; it is referred to here as the P matrix calibration. This approximated the result of administering the test an infinite number

FIGURE 1
PLAN OF THE SIMULATIONS



Note: Regressions to estimate β' and β'' are performed in the (linear) logistic metric. These are inconsistent estimates, but provide a useful first order check for trends in the residuals.

Table 1
General Outline of Procedure[a]

Variability of Item Disciminations (σ^2_β)	Distribution of Item Discriminations		
	Uniform	Normal	Positive Skew
.05	2	7	12
.10	3	8	13
.15	4	9	14
.20	5	10	15
.25	6	11	16

[a] Table entries are run numbers. Additionally, one simulation was run using constant discriminations (Run 1) to simulate perfect fit to the Rasch model assumptions.

of times to each of the 75 persons and calibrating the data obtained in the conventional manner, with observations pooled by examinee rather than by total score. The only difference between this calibration and one based on an infinite sample of distinct people pooled according to total score lies in the fact that when this a priori probability matrix is calculated directly, the ability parameters for each examinee may be different. In calibrating actual data, however, only $k-1$ distinct ability estimates are possible, corresponding to each possible total raw score on a k-item test.

The second calibration of each of the 16 simulation runs was performed on a data matrix obtained by simulating an administration of the test to 75 persons and analyzing the obtained raw data matrix. The abilities of these persons were sampled at fixed intervals from a normal distribution with mean 0 and variance 1.5. The obtained sample had a mean of 0.0 and a sample variance of 1.475. While it is known that the best item calibration is done with a good deal of replicability within each score group, hence large N (Whitely & Dawis, 1974), the computer time and cost were prohibitive for this. The number 75 was considered sufficient because (1) an additional calibration was obtained on an "infinite" sample, and (2) the 75 "persons" used were "centered on the test" (i.e., the test was of exactly the right difficulty for them), resulting in a very efficient administration with respect to amount of information obtained.

For each simulation, after the θ and α parameters were estimated for each item, i, the residual $a_{ji}-\varepsilon(a_{ji})$ were regressed on θ_{ji}. The fitted slope provides an (inconsistent) estimate of item discrimination, which may be compared with the known β_i used in generating the simulated data. For the purposes of this investigation, the extent to which these estimates reproduce the β values input was not of primary importance.

In the computer model, all simulations allow item difficulties and discriminations and person abilities either to be read in or generated internally. For the present research, all three were generated randomly with the following characteristics. For all runs the set of item difficulties was the same, having been randomly selected from the unit normal distribution. The person abilities were sampled at fixed intervals from a normal distribution with a mean of zero and a variance of 1.5. With these data fixed, 16 simulation runs were attempted. The first used a standard default option built into the program and generated a unit vector of item discriminations; this, then, was the run in which the precision of the item calibration routine could be tested, since the input conformed exactly to the Rasch model. Each of the remaining simula-

tions, however, deviated from the Rasch assumption of similar discriminations in two ways, since differing variances and distribution shapes were used. Five of the runs had discriminations uniformly distributed with mean equal to one and variance equal respectively to .05, .10, .15, .20, and .25, with each run showing increasingly stronger deviation from Rasch's assumption. For each of these runs, discriminations were sampled at fixed intervals from the appropriate uniform distribution. The next five simulations had discriminations normally distributed around a mean of one and variances respectively .05, .10, .15, .20, and .25. Values were once more sampled at fixed intervals.

The remaining five analyses were based on the chi-square distribution with one degree of freedom. This distribution has a mean of one and a variance of two. Data points were selected in the following manner. Thirty values were first sampled at fixed intervals from the chi-square distribution with one degree of freedom. These thirty points, with their mean of one and variance of two, were then converted to a data set having a mean of one and a variance of .25, using a linear transformation. This set of points was adjusted slightly to obtain the desired range of discriminations while holding the first two moments constant; and, finally, the obtained set was linearly transformed to each of a set of points having a mean of one and the variances used above (.05, .10, .15, .20, .25).

Results

Table 2 presents results for the degree of misfit for items in the item calibration process; Table 3 presents similar results from person calibration. Both item fit and person fit are described for calibration based on the probability

Table 2
Degree of Misfit of Items for Item Calibrations

Run	σ^2_β	Dist.	P Matrix MS Error	P Matrix Maximum Misfit	Data Matrix[a] MS Error	Data Matrix[a] Maximum Misfit
1	.00	-	.120	-3.000	.156	1.388
2	.05	U	1.492	3.034	.391	-1.202
3	.10	U	1.490	-3.024	.501	3.559
4	.15	U	1.486	-3.025	.345	1.956
5	.20	U	1.483	-3.026	.324	1.964
6	.25	U	1.480	-3.027	.389	-1.212
7	.05	N	.093	-3.031	.002	-0.165
8	.10	N	.171	3.564	.045	.467
9	.15	N	.010	.433	.087	-0.841
10	.20	N	.018	.571	.053	.845
11	.25	N	.027	.695	.070	-0.818
12	.05	S	.008	.749	.086	-1.153
13	.10	S	.005	.301	.107	.876
14	.15	S	.010	-0.277	.065	.932
15	.20	S	.020	-.430	.050	.778
16	.25	S	.024	.457	.314	2.713

[a] Computed by score group

Table 3
Degree of Misfit of Items for Person Calibrations

Run	σ^2_β	Dist.	P Matrix MS Error	P Matrix Maximum Misfit	Data Matrix[a] MS Error	Data Matrix[a] Maximum Misfit
1	.00	-	0.0	±.0005	.064	-0.509
2	.05	U	1.116	2.472	1.132	-2.342
3	.10	U	1.184	-2.611	1.240	-2.612
4	.15	U	1.234	-2.716	1.368	-2.772
5	.20	U	1.288	-2.801	1.355	-2.802
6	.25	U	1.331	-2.875	1.449	-2.953
7	.05	N	.021	-0.519	.150	-1.237
8	.10	N	.040	-0.879	.119	-0.860
9	.15	N	.075	+1.049	.194	-1.632
10	.20	N	.106	-.1262	.241	-1.979
11	.25	N	.138	-1.450	.694	3.589
12	.05	S	.009	-0.319	.075	.676
13	.10	S	.009	.233	.091	.771
14	.15	S	.059	-.947	.185	-1.279
15	.20	S	.075	-.883	.214	-1.689
16	.25	S	.366	3.215	.450	-3.186

[a] Computed by score group

matrix (*P*) and the raw data matrix. The criteria of interest are the mean square error and the most extreme point of lack of fit. While they are both self-explanatory, the latter deserves some explication. The extreme instance of misfit may be misinterpreted unless it is taken into account that (1) it is a single score and, by its nature, an extreme one and (2) it is believed to be an outlier for many of the simulations. As a last point, the standard errors of estimate of person and item parameters are rarely less than .1 logit and can be quite a bit greater.

Several patterns were noted in the results, although there are no statistical tests to confirm them. First, however, it should be noted the average item fit for both the *P* matrix and the raw data matrix was zero. This is, of course, due to the constraint imposed upon the item difficulties, as discussed above; and the information was, therefore, not included in Tables 2 and 3.

The poorest fit seemed to be for the uniform distribution, where the maximum misfits were considerably larger than for either of the other distributions and the error variance larger. The MS errors across increasing discrimination variability were generally lower in the theoretical (*P* matrix) calibration; as can be expected, the random error introduced in the simulation of test-taking clouded the issue. This was true for both person and item fit based on either the theoretical or simulated data.

Supplementary Analyses

In addition to looking at the variance of the difference between the parameters and their estimates (the MS error), the correlations between these values were also calculated. Table 4 includes these correlations for all runs. Again, the uniform distributions yielded very low correla-

Table 4
Correlations between α and θ Parameters and Their Estimates

Run Number	σ^2_β	Dist.	Correlations: Item Difficulty(α) P Matrix	Item Difficulty(α) D Matrix	Ability Estimate (θ) P Matrix	Ability Estimate (θ) D Matrix
1	.00	-	1.0000	.9753	.9638	.9774
2	.05	U	.2133	-.0709	.5949	-.2363
3	.10	U	.2147	-.1396	.6164	-.1183
4	.15	U	-.2137	-.2049	.6349	-.2126
5	.20	U	-.2152	-.1581	.6499	-.0415
6	.25	U	-.2156	-.1883	.6630	-.0552
7	.05	N	.9899	.9260	.9999	.9727
8	.10	N	.9781	.9530	.9434	.9792
9	.15	N	.9624	.9017	.9995	.9720
10	.20	N	.9459	.8754	.9992	.9837
11	.25	N	.9284	.8501	.9988	.9742
12	.05	S	.9953	.9617	.9598	.9718
13	.10	S	.9971	.9529	.9994	.9646
14	.15	S	.9723	.9027	.9979	.9774
15	.20	S	.9623	.8908	.9984	.9830
16	.25	S	.9773	.9409	.9996	.8940

tions for the alphas and low-moderate correlations for the thetas. For either the normal or skewed distributions, there was no evidence that variance of the distribution of the discriminations has any effect at all on the accuracy of the ability or difficulty estimates.

An argument for the use of item discrimination may still be made in terms of the extraction of maximal information from the test. In order to assess the degree to which unweighted total score approximates the mathematically correct scoring (in which each response is weighted by its items' discrimination), the mathematically correct scoring was correlated with the unweighted number of items correct for each simulation. These results are shown in Table 5.

The pattern across distribution forms was consistent with the other analyses. The minimum of these correlations across all 16 simulation runs was .8069. The magnitude of these correlations suggests that a slight increase in test length could compensate for whatever loss of information the use of unweighted raw scores might entail. This conclusion, unfortunately, cannot be generalized to the case in which a test is of inappropriate difficulty for the examinees. In such a case, a correlation may be induced between item difficulty and item discrimination, because items at one end of the continuum of item difficulties represented in the test will function better, and hence appear more discriminating, than items at the other end of the difficulty continuum.

Conclusions

The present research suggests that the lack of an item discrimination parameter in the Rasch model does not result in poor calibration in the presence of varying item discriminations. While the robustness of the model to other departures from assumptions remains to be investigated,

Table 5
Correlations Between the Unweighted Total Score Approximations and the Weighted Total Scores

Run Number	σ^2_β	Distribution	$r_{x.x(c)}$
1	.00	--	.9847
2	.05	U	.8123
3	.10	U	.8195
4	.15	U	.8069
5	.20	U	.8547
6	.25	U	.8183
7	.05	N	.9921
8	.10	N	.9875
9	.15	N	.9877
10	.20	N	.9885
11	.25	N	.9887
12	.05	S	.9888
13	.10	S	.9851
14	.15	S	.9827
15	.20	S	.9872
16	.25	S	.9786

such studies are also indicated for the normal ogive model and the more general logistic models. Until it is shown to be either inadequate or inferior to some other model, the use of the simplest model is to be recommended, if only on the basis of mathematical elegance and the sufficiency of total number of items correct as a statistic for testee ability.

There is, in addition, a secondary benefit to be gained from use of the Rasch model: if dichotomously scored items were used, the attenuation paradox (Lord & Novick, 1968) may be avoided. As Samejima (1969) has shown, information loss with high discriminating items is greater if the items are scored dichotomously than if they are scored trichotomously or tetrachotomously. If dichotomous items are used with lower, but fairly homogeneous, discriminations, the attenuation paradox may be avoided.

The substitution of equal item discriminations, rather than maximum item discriminations, as a goal in item writing appears counter-intuitive to the test construction expert steeped in classical test theory. While it is true that a highly discriminating item is capable of providing more information concerning the placement of an individual on the continuum of some latent trait, the highly discriminating item functions over a narrower range of abilities than a less discriminating item. An item with perfect discrimination would provide complete information about a single point on the ability continuum and no information about any other point. Therefore, for any given test, an optimal range of discrimination will exist. If the test characteristic curve is to rise steeply through a narrow range of abilities, more highly discriminating items will be desirable than if the test is to function over a broad range of abilities.

In the context of classical test theory, discrimination is an attribute of an item measured with respect to some population of examinees. In the context of the Rasch model, however, the estimated item discrimination is just one check

on the fit of the model to the data, to determine if any linear trend appears in the residuals across levels of ability. Stated differently, the discrimination is the slope of the regression of item difficulty on person ability. If this slope is 1, there is no evidence of misfit according to this particular criterion. In assessing the fit of items, a more useful criterion appears to be the approximate χ^2 suggested by Wright and Panchapakesan (1969). If this value is large, the estimate of item discrimination may be useful in explaining the observed lack of fit. Many common perturbations of the measurement process (e.g., guessing, inappropriate item difficulty relative to a particular set of examinees, or the relevance of special information possessed by some well-defined subset of examinees) will be reflected in either low or high discriminations. In such cases, the Rasch model should be regarded not as requiring items to be more highly constrained than other models, but rather as providing a mechanism for detecting problems which threaten the validity of *any* measurement, regardless of the model employed. In other words, the Rasch model appears more restrictive in part because it forces the user to take cognizance of more information concerning potential perturbations in the measurement.

The present results might be seen to parallel Wainer's (1976) discussion of regression weights: within the constraints of positive inter-predictor and predictor-criterion correlation, information loss is minimal using equal regression weights. In the present case, with varying discriminations, these values might have been used to weight item responses (Lord & Novick, 1968). However, they were ignored in the Rasch estimation; an acceptable fit was, nevertheless, obtained. It may be added that this fit was in spite of varying item difficulties leading to varying item means and variances.

No guidelines are provided regarding minimum and maximum permissible β values. The item discriminations should be examined in conjunction with other fit statistics, and any outliers (exceptionally large or small β's, χ^2's, or even item difficulties, for that matter) should be examined to try to determine whether a problem exists.

The Rasch model appears to be highly robust to differing discriminations, except in the case of uniformly distributed discrimination parameters. Because in the actual application of the model the true values of the discriminations are unknown, item difficulties are estimated following calibration, by regressing probability of success on ability in the (linear) logistic metric. The poorer the fit of an item, the larger the standard error of estimate of its discrimination may be. In the light of these considerations, the authors suggest Wright's (1969) approximate χ^2 statistic for the evaluation of fit.

Limitations of the Present Research and Suggestions for Future Research

In this study, the only source of misfit which was introduced into the data was nonhomogeneity of item discriminations. The calibration procedure proved quite robust to perturbations of this kind. Actual data, however, are influenced by a wide variety of effects, e.g., guessing, carelessness when items are too easy, practice effects which distort the shape of the item characteristic curve and/or induce violations of the assumption of local independence of persons and items.

These additional sources of misfit raise several questions:

1. If more than one parameter is to be estimated for each item, is discrimination the best choice to accompany difficulty, or would more variance be accounted for by a parameter representing, say, level of asymptote of the item characteristic curve (sensitivity to guessing)?
2. Would the Rasch calibration procedure be less robust to variation in item discriminations if those variations occurred in the context of other sources of misfit?
3. If variations in item discriminations alone do not preclude the use of the Rasch model,

what evidence is there that models incorporating more parameters are superior to the Rasch model in fitting actual data?

These question should be addressed by future research on the Rasch model.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical Theory of Mental Test Scores* (Part V). Reading, MA: Addison-Wesley, 1968.

Guilford, J. P. *Psychometric Methods.* New York: McGraw-Hill, 1936.

Lord, F. M. and Novick, M. R. *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley, 1938.

Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology,* 1966, *19,* 49–57. (a)

Rasch, G. An individualistic approach to item analysis. In P. F. Lazarsfeld and N. W. Henry (Eds.), *Readings in Mathematical Social Science.* Chicago: Science Research Associates, 1966, 89–108. (b)

Wainer, H. Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin,* 1976, *83,* 213–217.

Whitely, S. E. and Dawis, R. V. The nature of objectivity with the Rasch model. *Journal of Educational Measurement,* 1974, *11,* 3, 163–178.

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

Author's Address

Thomas E. Dinero, 406 White Hall, Kent State University, Kent, OH 44242.