Small *N* Does Not Always Justify Rasch Model

Dato N. M. de Gruijter
University of Leyden

In many applications of item response theory, it is of little consequence whether the Rasch model or a more accurate, but more complicated item response model is used. With small sample sizes, it might be advantageous to employ the Rasch model. A clear counterexample is the case of optimal item selection under guessing.

The choice of an appropriate item response model—frequently restricted to the choice between the one-parameter Rasch model and the three-parameter logistic model—is an intricate one. The Rasch model is simple and has attractive features; however, it might be too simple to be adequate.

Preferably the choice between models should not be made on the basis of statistical fit only, as a model might be rejected for statistical reasons in large samples, even when discrepancies between model and data are small. Further, the possibility should be considered that the model is adequate for the intended application despite deviations from the model (Gustafsson, 1980). Recently Lord (1983) added another argument for choosing the Rasch model even when it is not the true model. In a paper with the provocative title ''Small *N* Justifies Rasch Model'', he suggested that the additional parameters of the three-parameter model cannot be determined with reasonable accuracy for small sample sizes. Under these circumstances the Rasch model might give more accurate results in some applications.

The present author argues that there are applications in which the distinction between the Rasch model and an alternative model which includes a guessing parameter is very important. One such application is the selection of items by means of the information function in order to obtain maximum discrimination between abilities near a particular latent trait value. This paper demonstrates that with the Rasch model, items that are too easy are selected when guessing is involved.

The Information Function and Optimal Item Selection

In the three-parameter logistic model, the probability of a correct answer to item i given ability θ is expressed by

$$P_i(\theta) = c_i + (1 - c_i)\Psi[Da_i(\theta - b_i)] \quad , \tag{1}$$

where

$$\Psi[Da_i(\theta - b_i)] = \{1 + \exp[-Da_i(\theta - b_i)]\}^{-1} \quad , \tag{2}$$

c_i is the lower asymptote or pseudo-guessing parameter,
a_i is the slope parameter,

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 2, June 1986, pp. 187–194
© Copyright 1986 Applied Psychological Measurement Inc.
0146-6216/86/020187-08$1.65

b_i is the item difficulty parameter, and

D is a constant, set equal to 1.7 in order to obtain correspondence with the normal ogive model. The probability as a function of θ (Equation 1) is called the item characteristic curve (ICC).

The Rasch model can be regarded as a special case of the three-parameter model with c_i equal to zero and with equal slopes for all items. The factor $D\bar{a}$, where $\bar{a}$ is the common slope, can be eliminated through the transformation $\theta^* = D\bar{a}\theta$ and $b_i^* = D\bar{a}b_i$.

Given the item parameters of the items in a test, the maximum likelihood estimator $\tilde{\theta}$ of θ is asymptotically normally distributed with mean θ and variance $I^{-1}(\theta)$, the inverse of the test information function. Samejima (1977) demonstrated that this function can also be used with relatively small N. The test information function equals the sum of the item information functions

$$I_i(\theta) = P_i'(\theta)^2/[P_i(\theta)Q_i(\theta)] \quad , \tag{3}$$

where $P_i'(\theta)$ is the first derivative of $P_i(\theta)$ and $Q_i(\theta) = 1 - P_i(\theta)$. When, in maximum likelihood estimation of person parameters, θ, $I_1(\theta_0)$ exceeds $I_2(\theta_0)$, the first item has a higher relative efficiency for discriminating abilities near θ_0. For optimal discrimination near θ_0, items should be chosen with a high information value for $\theta = \theta_0$. Given items with equal cs and as, this amounts to finding items with an optimal difficulty level b_i.

The item information function for the three-parameter logistic model can be written as

$$I_i(\theta) = D^2a_i^2\psi[Da_i(\theta - b_i)]\{[P_i(\theta) - c_i]/P_i(\theta)\} \quad , \tag{4}$$

where

$$\psi(x) = \Psi(x)[1 - \Psi(x)] \quad . \tag{5}$$

For given values a_i and c_i, this function is maximized at the value θ_0 when

$$b_i = \theta_0 - D^{-1}a^{-1} \log \{½[1 + (1 + 8c_i)^{½}]\} \quad , \tag{6}$$

where log designates the natural logarithm (Birnbaum, 1968, p. 463). When guessing plays no role, the optimal b-value equals θ_0. Otherwise, the optimal b-value is lower than θ_0.

Remarkably, the optimal b_i for the closely related three-parameter normal ogive differs systematically from the outcome of Equation 6 when c_i exceeds zero (Wolfe, 1981); the information function does not seem to be robust. Wolfe argued that in tailored/adaptive testing, different items might be selected under the two different models.

The P-value corresponding to θ_0 for the optimal item difficulty depends on c_i only; $P_i(\theta_0)$ for optimal b_i equals

$$1 - 2(1 - c_i)/[3 + (1 + 8c_i)^{½}] \tag{7}$$

(De Gruijter & van der Kamp, 1984). This term equals .50 for $c_i = 0.0$ and .68 for $c_i = .25$. When the three-parameter model with $c = .25$ for all items applies, items with $P_i(\theta_0)$ close to .68 should be selected. The Rasch approach, in which it is assumed that c equals zero, would lead to a very different item selection with $P_i(\theta_0)$ close to .5.

Rasch Model Fit Under Guessing

The well-known facts, restated above, are clear: With the Rasch model, the wrong items are selected in optimal item selection when guessing plays a role. One way to avoid this problem is to verify whether the Rasch model fits. Unfortunately, the use of a statistical test does not guarantee rejection of the Rasch model for small and medium-sized samples, even when guessing is substantial. When the ICCs are close together, the latent scale is undetermined and a nonlinear transformation of the original scale gives a Rasch-conform scale. For this reason, Gustafsson (1980) concluded that a statistical test on a peaked test does not make sense when guessing is suspected. Even when item difficulties differ, the Rasch model may appear adequate. Meredith and Kearns (1973) demonstrated for items with equal slopes that a nonlinear transformation of the latent scale gives the Rasch model if $b_i + \log(c_i)$ is constant over items. When this relationship is true for all items from the relevant item domain, nothing is wrong and the Rasch model results—including those with respect to the optimal P-value—can be used. However, when the relationship is an accidental one which is not valid in the whole domain, analyses with the Rasch model can be highly misleading.

Results from a simulation study by Gustafsson (1980) for equal slopes and lower asymptotes demonstrate that the power of statistical tests is low and that large samples are needed in order to obtain an acceptable rejection rate for the Rasch model. Apparently the model is quite pliable. A demonstration is given below.

Five hypothetical items were used, with $c = .25$, $a = 1.0$, and b-values equal to -1.0, $-.5$, 0.0, .5, and 1.0. The θs were assumed to have an approximately standard normal distribution, a discrete distribution with lower limit $\theta_l = -2.5$ and upper limit $\theta_u = 2.5$. The ICCs of the items were assumed to be known. They were approximated by Rasch curves, $P_i^*(\theta)$, with a lower asymptote equal to zero and with equal slopes. The bs and the common a-value of the $P_i^*(\theta)$ were obtained from the minimization of

$$F = \sum_i \sum_k f(\theta_k) \, [P_i(\theta_k) - P_i^*(\theta_k)]^2 \quad , \tag{8}$$

where k is the kth latent class of θ and $f(\theta_k)$ is the relative frequency of $\theta = \theta_k$. In actual estimation of item parameters from item scores, obtained according to probabilities $P_i(\theta)$, random fluctuations naturally play a role.

The ICC of the third item (the item with $b = 0.0$) and its approximation are given in Figure 1. The approximation fails for the lower θs where the ICC approaches the value .25 and the approximation continues to drop.

The minimum of the weighted squared distances between the model and Rasch curves is not obtained by minimizing Equation 8, but by minimizing

$$F' = \sum_i \sum_k f(\theta_k) \, [P_i(\theta_k) - P_i^*(\theta_k^*)]^2 \quad , \tag{9}$$

where θ* is a nonlinear transformation of θ, with respect to the item parameters for the Rasch approximation and the parameters θ_k^*. It was decided to approximate this goal by fixing the Rasch item parameters from the minimization of Equation 8 and minimizing F' in a second stage only with respect to the θ*s.

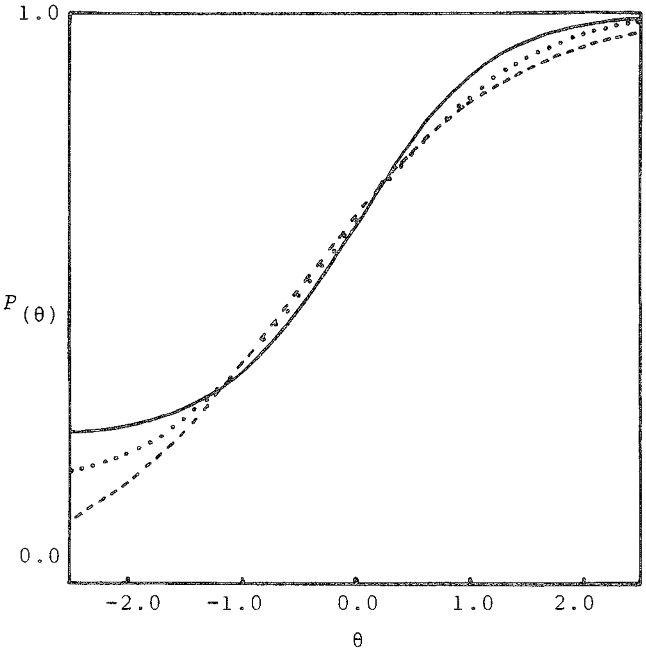
At low θs, the resulting θ*s exceeded the earlier ones; as a result, the Rasch curves approximated the true ICCs better at low θs. The effect of the scale transformation θ* is given for the third item in Figure 1. In this figure the final values $P_i^*(\theta^*)$ are plotted against the original θs.

From the foregoing it can be stated that the Rasch model is flexible due to a malleable latent scale. Of course, when tests are analyzed according to the Rasch model for samples with different ability levels, the various Rasch scales will not be compatible (see also Gustafsson, 1980). It is important to note that the outcome of statistical tests might be less revealing than expected, when guessing is involved.

The Accuracy of the Three-Parameter Model

The Rasch model is not an acceptable model for item selection if the lower asymptotes of the items exceed zero, but the alternative three-parameter model has its own problems. Lord (1983) stated that the parameters in this model are inaccurately estimated in small and medium-sized samples; this conclusion seems to be supported by the values of the standard errors for item parameter estimates computed by Thissen and Wainer (1982) for the situation where the θs can be considered known.

Figure 1
Item Characteristic Curve (solid line),
Rasch Approximation With θ Fixed (dashed line),
and Rasch Approximation Based on a New Latent Scale
(dotted line) for an Item With $b = 0.0$



The results of Thissen and Wainer give an overly negative picture of the possibilities of the three-parameter model. It is possible to restrict the variation in the estimate of the lower asymptote in a meaningful way. When a prior distribution for c is introduced, the standard errors for all item parameter estimates drop to more acceptable levels (De Gruijter, 1984). Further, the item parameter estimates are correlated to the effect that the ICC itself is fairly accurately estimated in the ability range into which most of the person parameters fall. When confronted with a choice between setting c equal to zero and trying to obtain a reasonable estimate of c, the latter approach clearly is to be preferred when guessing is suspected. In the next section the Rasch model is compared with the alternative model

$$P_i(\theta) = c + (1 - c)\Psi[Da(\theta - b_i)] \quad , \tag{10}$$

the three-parameter model with equal cs and equal slopes, that is, the Rasch model with a common pseudo-guessing parameter.

An Empirical Demonstration of the Bias in Item Selection

For the demonstration, a hypothetical test with $n = 40$ items was chosen. The a-values of these items had a common value of 1.0; the c-values were set equal to .25. Ten different b-values were taken, each occurring four times: $-.8$, $-.7$, $-.6$, $-.5$, $-.4$, $-.3$, $-.2$, $-.1$, 0.0 and .1. Items were to be selected on the basis of the value of the information function at the θ level corresponding to a relative true score on the 40-item test equal to $\tau_0 = .608$. The corresponding value of θ, θ_0, which can be obtained by solving

$$\tau_0 = n^{-1} \sum_i [c + (1 - c)] \{1 + \exp[-Da(\theta - b_i)]^{-1}\} \tag{11}$$

for θ, equals $-.408$. The values τ_0 and θ_0 were chosen in such a way that $P_i(\theta_0)$ became equal to .5 (the optimal value according to the Rasch model) for $b = 0.0$.

The highest information at $\theta = \theta_0$—given $a = 1.0$ and $c = .25$—was obtained for b equal to $-.591$. The highest information in the test was obtained for $b = -.6$. The relative efficiency of the items with respect to an item with $b = -.6$ for discriminating near θ_0,

$$\mathrm{RE}(i,\ b = -.6) = I_i(\theta_0)/I_{b=-.6}(\theta_0) \quad , \tag{12}$$

is given in the second column of Table 1. From the entries in the table, it can be concluded that items with b-values close to zero are considerably less efficient for discrimination near θ than more optimal items with b-values close to $-.6$.

In order to demonstrate the inadequacy of the Rasch analysis for item selection (even with small sample size), a simulation study was done with the hypothetical 40-item test and 400 hypothetical examinees, with θ randomly sampled from the standard normal distribution. Notice that with this distribution, 34% of the examinee population lies below θ_0, or τ_0. The mean of the population distribution equals zero, that is, the mean is identical to the b-value for which $P_i(\theta_0)$ equals .5.

The data were first analyzed with a least squares estimation procedure for the three-parameter model, in which item parameters and population distribution parameters are estimated from marginal proportions correct and marginal proportions of item pairs (De Gruijter & Mooijaart, 1983). The model from Equation 10—the true model—was fitted by constraining the a-values to be equal, just as the c-values were constrained. A discrete population distribution for θ was specified, with five equally sized latent classes. The θ values of two latent classes were fixed in order to fix the latent interval scale for θ; the three

Table 1

Estimated Relative Efficiency (RE) for Various Difficulty (b) Levels

		Estimated RE	
b	RE	$\hat{c}_i=\hat{c}$	$\hat{c}_i=0$
-0.8	.96	.96	.90
-0.7	.99	.98	.94
-0.6	1.00	1.00	1.00
-0.5	.99	1.02	1.04
-0.4	.97	1.01	1.08
-0.3	.92	0.95	1.13
-0.2	.87	0.93	1.13
-0.1	.80	0.94	1.13
0.0	.72	0.84	1.15
0.1	.64	0.73	1.15

remaining θ values were free to vary. With this specification of the latent distribution, no advantage was taken of the knowledge that the true population had a normal distribution. As a starting value for $\hat{c}$, the value .20 was chosen.

The final estimate of c was .21, which is much closer to the true value than $c = 0$ in the Rasch model. Equation 11 was used with estimated item parameters in order to obtain an estimate of θ_0. Next, the item informations at $\hat{\theta}_0$ were computed, also using the item parameter estimates. For each true b-value the estimated information values were averaged over the four items with that b-value. These averages were divided by the average for true b $(-.6)$. The resulting estimated average relative efficiencies are given in the third column of Table 1 (Estimated RE, $\hat{c}_i = \hat{c}$). For $b = -.3$ the outcome is .95, which means that the items with true b equal to $-.3$ had an average estimated information at $\hat{\theta}_0$ 95% of the value of the average for true b $(-.6)$. The maximum value in the third column is shifted somewhat with respect to the maximum in the second column.

A second analysis was done with the Rasch model. For this analysis unconditional maximum likelihood estimation (the CALFIT program) was used. One item had a large, but unsystematic misfit; this item was retained for further analyses. Using the estimated test characteristic curve (Equation 11), $\hat{\theta}_0^*$, the latent ability in the Rasch model corresponding to τ_0, was obtained. With this value and the estimated Rasch parameters average relative efficiencies were obtained, which are given in Table 1 in the fourth column. Observe the large shift in the maximum relative efficiency: The maximum is obtained for more difficult items with a true b-value close to 0.0, which is as expected for the Rasch analysis.

If only 5 test items were to be selected from the 40 available items on the basis of the estimated RE, the Rasch analysis would have resulted in a selection with true b-values $-.2$, $-.1$, 0.0 and .1 (twice). The first analysis with $\hat{c} = .21$ would have resulted in the more adequate selection $-.6$, $-.5$ (twice) and $-.4$ (twice). Similar results were obtained with an even smaller sample size. When the procedure was repeated for a sample size of 100, taking the first 100 hypothetical examinees, the five optimal items according to the Rasch analysis had true b-values equal to $-.2$, $-.1$ (twice) and .1 (twice). The analysis with the model from Equation 10 resulted in an estimated $\hat{c}$ of .23 and an item selection with true b-values equal to $-.7$, $-.5$ (twice), $-.4$ and $-.3$.

The Effects on Measurement Errors

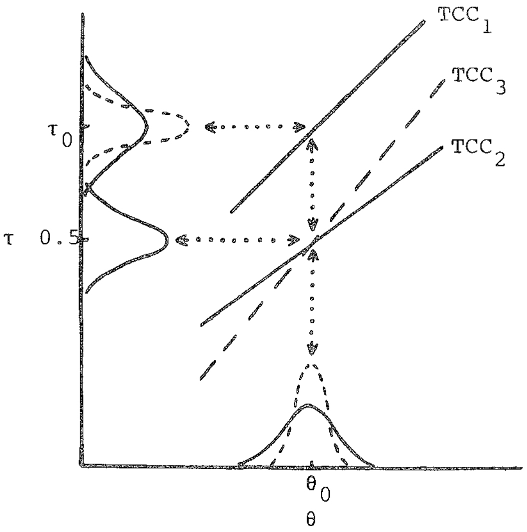
The fact that the wrong items were selected in the Rasch analysis suggests that the accuracy of these items was overestimated in the analysis. In other words, it seems plausible that the measurement at θs near θ_0 with an item selection based on the Rasch model is less accurate than the Rasch results suggest. This conjecture cannot be verified directly, by comparing the true information at θ_0 for the five selected items in the sample of 400 examinees, $I(\theta_0)$, with the Rasch information, $\hat{I}(\hat{\theta}_0^*)$, due to differences between the two latent scales. First a transformation to a common scale (e.g., the relative true score scale of the original 40-item test) is needed. The true information of the five selected items for the relative true score scale can be written as

$$I(\tau) = I(\theta)\ (d\theta/d\tau)^2 \quad . \tag{13}$$

A similar equation is obtained for the estimated information function on the basis of the Rasch model, $\hat{I}(\tau)$. The ratio $\hat{I}(\tau_0)/I(\tau_0)$ equals 1.3, that is, the estimated accuracy at τ_0 exceeds the maximum attainable accuracy. The Rasch analysis overestimates the accuracy of the Rasch item selection.

In Figure 2 it is schematically shown how this can happen. The relative true score on a test is shown as a function of θ near θ_0. This test characteristic curve (TCC), designated as TCC_1, is assumed to coincide with the estimated TCC near θ_0, which is not too unreasonable even with relatively small sample sizes. The TCC of an item selection based on a Rasch analysis is also given (TCC_2). The relative true score for

Figure 2
Error Variances Associated With θ_0 for the Total Test (TCC$_1$), Rasch Selection (TCC$_2$), and the Estimated Curve of the Selection Under the Rasch Model (TCC$_3$)



this selection equals .5 for θ_0, the θ level of interest. The Rasch approximation to the true TCC is given by TCC$_3$. TCC$_3$ can be constructed by computing, for each relative true score on the total test, the relative true score of the selection on the basis of a Rasch analysis and θ. In the example, TCC$_2$ and TCC$_3$ cross for θ equal to θ_0; the Rasch approximation is assumed to be unbiased for $\theta = \theta_0$. Notice further that all TCCs have been linearly approximated in the neighborhood of θ_0.

The error distribution for θ_0 (i.e., the distribution of relative observed scores x given θ_0) is indicated on the left side of the figure. This distribution is relevant under number-correct scoring. Number-correct scoring seems appropriate in this application, because (1) the need for differential item weights is reduced due to the item selection, and (2) complications due to inaccurately estimated item parameters—this was Lord's problem (Lord, 1983)—are avoided. Using TCC$_2$, the error distribution can be transformed into an error distribution for θ, and next, using TCC$_1$, into an error distribution for τ, the relative true score on the total test. These error distributions are also displayed in Figure 2.

In the Rasch analysis, TCC$_2$ is replaced by TCC$_3$. The resulting error distributions, indicated with dashes, have a *smaller* variance. The gain obtained with the wrong model is an illusion, however. When an ability $\tau_c \neq \tau_0$ is chosen, the Rasch analysis results in a biased estimate. Its expected value is as far from τ_0, in terms of the error distribution, as τ_c is in terms of the error distribution for the correct model. The ratio of the local slope of TCC$_2$ and the standard deviation of errors, $\sigma(x \mid \theta_0)$, has remained invariant through all scale transformations. In other words, the information function on the basis of number-correct scoring (Birnbaum, 1968, p. 453) has remained constant.

Due to this invariance property, the efficiency of different item selections near θ_0 can also be compared on the basis of their number-correct information functions, at least when the item selections are large enough in view of the asymptotic properties of information functions. The number-correct information functions can be approximated by the test information functions (i.e., the sums of the item information functions) because the selection tends to create relatively homogeneous tests. From this it is clear that item selections based on the guessing model are more efficient at $\theta = \theta_0$ than those based on a Rasch analysis. The relative efficiency of selections similar in composition to the five-item Rasch selection, based on a sample size of 400, is about .75 the efficiency of the guessing model selections, as can be seen in the second column of Table 1.

Discussion

The attractive Rasch model has great potential. It should not, however, be used for optimal item selection when guessing is a serious possibility. When the Rasch model is used for optimally selecting items, and guessing plays an important role in item responses, overly difficult items tend to be selected with an attendant overestimation of the efficiency of the items. This selection bias is especially troublesome when these more difficult items are marginal with respect to difficulty, content, and other qualities.

A logistic model with a pseudo-guessing parameter seems to be the correct choice for the situation with guessing. Even this choice is not without its problems, however. A small deviation from this model can lead to different conclusions, as was noticed by Wolfe (1981). Still more problems are to be expected when, for example, the true model is Samejima's guessing model (Samejima, 1979). Nonparametric methods for the determination of the relative efficiencies of items could be developed using item-item curves (De Gruijter, 1982; Levine, 1982), but it is doubtful whether such methods will be accurate. The results presented above suggest that it is unwise to emphasize optimal information in item selection. A more modest approach to item selection would be to define some minimal information value above which items would be acceptable for selection.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading MA: Addison-Wesley.

De Gruijter, D. N. M. (1982). *Tentamineren en beslissen* (SVO series no. 63). Harlingen, The Netherlands: Flevodruk.

De Gruijter, D. N. M. (1984). A comment on "Some standard errors in item response theory." *Psychometrika, 49*, 269–272.

De Gruijter, D. N. M., & Mooijaart, A. (1983). Least squares estimation of the item parameters in the three-parameter logistic model. *Tijdschrift voor Onderwijsresearch, 8*, 218–223.

De Gruijter, D. N. M., & van der Kamp, L. J. Th. (1984). *Statistical models in psychological and educational testing*. Lisse, The Netherlands: Swets & Zeitlinger.

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

Levine, M. V. (1982). Fundamental measurement of the difficulty of test items. *Journal of Mathematical Psychology, 25*, 243–268.

Lord, F. M. (1983). Small *N* justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing*. New York: Academic Press.

Meredith, W., & Kearns, J. (1973). Empirical Bayes point estimates of latent trait scores without knowledge of the trait distribution. *Psychometrika, 38*, 533–554.

Samejima, F. (1977). Effects of individual optimization in setting the boundaries of dichotomous items on accuracy of estimation. *Applied Psychological Measurement, 1*, 77–94.

Samejima, F. (1979). *A new family of models for the multiple-choice model* (Research Report No. 79-4). Knoxville TN: University of Tennessee.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47*, 397–412.

Wolfe, J. H. (1981). Optimal item difficulty for the three-parameter normal ogive response model. *Psychometrika, 46*, 461–464.

Author's Address

Send requests for reprints or further information to Dato N. M. de Gruijter, Educational Research Center, University of Leyden, Boerhaavelaan 2, 2334 EN Leyden, The Netherlands.