The Robustness of Rasch Estimates

Fons J. R. van de Vijver
Tilburg University

The small scale applicability of Rasch estimates was investigated under simulated conditions of guessing and heterogeneity in item discrimination. The accuracy of the Rasch estimates was evaluated by means of the correlation between the item/person parameters and their estimates, the standard deviations of the estimates, and the difference as well as the root mean squared difference between parameters and estimates. Within the range of the present investigation (from 10 to 50 items and from 25 to 500 persons) these criteria yielded favorable results under conditions of heterogeneous item discrimination. Under conditions of guessing, robustness could only be demonstrated for the correlational criterion. Guessing affects the difference measures between the parameter values and estimates quite strongly in a systematic way. It is argued that, notwithstanding these estimation errors, the Rasch model is to be preferred over nonstandard estimation procedures, from which the validity is unclear, or the use of the three-parameter model with its computational problems in small samples.

This article is concerned with one particular item response model, namely, the Rasch model. The choice for the Rasch model rather than the two- or three-parameter model was motivated first and foremost by the small scale applicability of the Rasch model. The often reported lack of convergence of the estimation procedures for small data sets (cf. Hulin, Lissak, & Drasgow, 1982; Lord, 1968) does not occur in the Rasch model. Even if in more-parameter models the iterative procedure converges, not all parameters may be estimated accurately. In particular, the accuracy of the guessing parameter is often poor (cf. Hulin et al., 1982; Thissen & Wainer, 1982).

A second reason to focus on the Rasch model was that the few comparisons between the different methods carried out so far showed converging results, or as Traub and Lam (1985) stated: "When a model has been fit to a set of data by different methods, any difference in estimates associated with a difference in methods has been too small to be of practical consequence" (p. 26). This implicitly favors the use of the computationally simpler and less computer time-consuming model, that is, the Rasch model.

In this article the robustness of Rasch estimates was investigated against violations of the assumptions of no guessing and homogeneity of item discrimination. A few related monte carlo studies on the Rasch model have been reported. Hambleton and Traub (1971) compared the information and efficiency of three scoring systems, namely, of the one-, two-, and three-parameter models, with respect to various ability levels. Using 15-item tests, they found that person estimates were rather robust against heterogeneity of item discriminations. Under conditions of guessing, however, their dependent measure showed less favorable results for the one- and two-parameter models, in particular for examinees with low scores. Although their con-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 10, No. 1, March 1986, pp. 45–57

0146-6216/86/010045-13$1.90

clusions will generalize over various test lengths, it should be noted that the authors used a test length of 15 items, which is rather small, if not too small, for an application of the three-parameter model.

In another monte carlo study, Dinero and Haertel (1977) investigated the impact of variation in item discriminations on the correlation between parameter values (difficulties as well as abilities) and their estimates. They found that the form of the distribution of the item discrimination parameter critically affected the robustness of the estimates. For normal distributions of the item discrimination parameters the authors observed high correlations. The correlations decreased considerably when the discrimination parameters were distributed uniformly. Since it seems difficult to believe that the shape of the item discrimination parameter distribution influences the robustness of Rasch estimates so dramatically, major parts of their study were replicated in the present study.

Wainer and Wright (1980) investigated the robustness of ability estimates under conditions of guessing. They found that jackknife procedures outperformed the standard estimation procedure of the Rasch model in tests of up to 40 items.

In the present study the accuracy of the estimates was evaluated in terms of four measures (cf. Hulin et al., 1982). First, the correlations between the item parameters and their estimates as well as between the person parameters and their estimates were calculated. Second, the standard deviations of the item and person estimates were computed. Third, the logit scale was split in 20 intervals and for each interval the bias was calculated, being the mean difference between the item or person parameters and their respective estimates. Finally, the same was done for the root of the mean squared difference between parameters and estimates.

Procedure

The program structure was as follows:

1. A number of vectors of specified length were generated: (1) uniformly distributed item discrimination parameters drawn from a uniform distribution and then rescaled in order to make the product of all parameters equal to one, (2) standard normally distributed item difficulties, (3) uniformly distributed lower asymptotes, and (4) standard normally distributed person parameters.
2. On the basis of these vectors, the matrix $\mathbf{P}$ with elements p_{vi} was computed, containing the theoretical probabilities of correct answers by means of

$$p_{vi} = c_i + (1 - c_i) \frac{\exp[a_i(\theta_v - b_i)]}{1 + \exp[a_i(\theta_v - b_i)]}, \quad (1)$$

where a_i is the item discrimination parameter of item i,
b_i is the difficulty of the item,
c_i is the lower asymptote, and
θ_v is the ability of person v.
Under the Rasch model $a_i = 1$ and $c_i = 0$ for each item.
3. A matrix $\mathbf{R}$ with elements r_{vi} was filled with uniformly distributed random numbers within the interval (0,1).
4. The dichotomous data matrix $\mathbf{D}$ was computed with elements d_{vi} so that: $d_{vi} = 1$ if $p_{vi} \geq r_{vi}$ and $d_{vi} = 0$ if $p_{vi} < r_{vi}$.
5. All persons and items with only zeros or ones were removed.
6. It was checked whether the maximum likelihood estimates based on these data were unique. If not, the program returned to the beginning (this situation did not occur during the simulations). This procedure stems from Fischer (1981) who has derived necessary and sufficient conditions for the existence of unique maximum likelihood estimates in the Rasch model.
7. Conditional item and person parameters were estimated. The procedure for the estimation of item parameters used here does not lead to computational inaccuracy and can be applied even for large numbers of items (Verhelst, Glas, & van der Sluis, 1984).
8. The following dependent variables were computed:
(a) The correlation between (item and person) parameter values and their estimates, (b) The standard deviations of these estimates, (c) The

bias, that is, for each interval (smaller than -2.70, between -2.70 and -2.40, ... , larger than 2.70), the average difference between parameters and estimates, and (d). The root mean squared error (RMSE), the square root of the average squared difference per interval between parameter values and their estimates. To reduce the number of points in the graphs the results of the bias and RMSE statistics were averaged pairwise (mean of the first and second interval, mean of the third and fourth interval, etc.).

Simulation runs were performed for the following number of items (k) and persons (n): $k = 10$ and $n = 25$; $k = 25$ and $n = 50$; $k = 25$ and $n = 100$; $k = 25$ and $n = 500$; and $k = 50$ and $n = 500$. The levels of items and persons were not crossed completely because little information would be gained by a complete crossing at the cost of much computer time.

The simulations consisted of four parts; in each part all (n,k)-combinations were used (see Table 1). First, data were generated under the Rasch model to provide a criterion against which subsequent findings could be evaluated; for each (n,k)-combination 50 runs were performed. As these data were generated under the (null) hypothesis of the validity of the Rasch model, they are referred to as "null data."

Then, data sets were generated in which the items had heterogeneous discrimination parameters. Three different levels of variability of the discrimination parameter were used, with lower and upper limits ranging from .90 to 1.10, from .50 to 1.50, and from .00 to 2.00, respectively, with 50 runs for each level per (n,k)-combination. Thereafter, data were generated in which guessing occurred. Three different average guessing levels were used, ranging from .10 to .30, from .25 to .45, and from .40 to .60, with 50 runs for each level. In the final run of simulations, data were generated in which both assumptions were violated. The three levels of the two factors, item discrimination and guessing, were crossed, with each crossing containing 25 runs.

Table 1
The Design of the Study

Item Discr.		Guessing		Number of
Lower Limit	Upper Limit	Lower Limit	Upper Limit	Replications per (n,k)-Combination
No Assumptions Violated				
1.00	1.00	.00	.00	50
Heterogeneous Item Discrimination				
.90	1.10	.00	.00	50
.50	1.50	.00	.00	50
.00	2.00	.00	.00	50
Guessing				
1.00	1.00	.10	.30	50
1.00	1.00	.25	.45	50
1.00	1.00	.40	.60	50
Heterogeneous Item Discrimination and Guessing				
.90	1.10	.10	.30	25
.90	1.10	.25	.45	25
.90	1.10	.40	.60	25
.50	1.50	.10	.30	25
.50	1.50	.25	.45	25
.50	1.50	.40	.60	25
.00	2.00	.10	.30	25
.00	2.00	.25	.45	25
.00	2.00	.40	.60	25

Results

Correlations

In Table 2 the correlations between the parameters and their estimates are presented. In all data sets it was observed that the correlation between item difficulties and their estimates increases with the sample size, and that the correlation between person parameters and estimates increases with test length. This is fairly obvious as the influence of "calibration error" decreases. For instance, when the sample size increases, with the number of items remaining constant, the sum of the estimated probabilities of a correct response for each person, equal to the sufficient statistic, will better approximate the sum of the probabilities expected under the Rasch model. The influence of improbable item responses will diminish in large data sets.

It appears from these simulations that in the null case 25 items/persons is sufficient to produce correlations of over .90 between person/item parameters and their estimates. In the two-parameter model a similar figure has been found (Hulin et al., 1982).

The correlations in Table 2 between item parameters and their estimates are typically higher than between person parameters and their estimates. This is a consequence of the fact that the number of persons in the simulations was always larger than the number of items. Since this is generally the case in empirical data sets, it follows that item estimates as a rule are more accurate than person estimates.

From Table 3 it can be gathered that the correlations are not very sensitive to heterogeneity of item discriminations. A decrement in the correlation could only be observed for discrimination parameters with extreme variation. However, this largest dispersion used, ranging from .00 to 2.00, is unlikely to be present in empirical data. The correlation between the person parameters and their estimates was unaffected by variation in item discriminations, even when these values range from .00 to 2.00. Thus, with the correlation between parameters and estimates as the main criterion, the Rasch model appears to be robust against heterogeneity of item discriminations. This confirms the results of Hambleton and Traub (1971).

The reason for this robustness can easily be understood. Suppose that two tests are administered to the same people, one test meeting the assumptions of the Rasch model and the other test having the same item difficulties, but heterogeneous item discrimination values. The vectors with the person totals derived from these two tests will

Table 2
Correlations and Standard Deviations for the Null Data

		Statistic*			
Items	Persons	$r_{b\hat{b}}$	$r_{\theta\hat{\theta}}$	$\hat{s}_b$	$\hat{s}_\theta$
10	25	.92	.79	1.22	1.21
25	50	.95	.90	1.09	1.15
25	100	.98	.90	1.09	1.13
25	500	.99	.90	1.02	1.11
50	500	.99	.95	1.01	1.06
Average		.97	.89	1.09	1.13

*$r_{b\hat{b}}$ = correlation between item parameters and estimates;
$r_{\theta\hat{\theta}}$ = correlation between person parameters and estimates;
$\hat{s}_b$ = estimated standard deviation of item parameters (population value = 1.00);
$\hat{s}_\theta$ = estimated standard deviation of person parameters (population value = 1.00).

not differ much from each other because the underestimations and overestimations in the test, which can be expected under the Rasch model, will cancel out when summed over the items. Suppose, in addition, that all items have the same difficulty parameter b_i. The item characteristic curves (ICCs) of the second test will all intersect at point b_i. The item totals will not differ for the various items of both tests, if the average of the person parameters does not differ too much from b_i, that is, as long as the number of persons observed at either side of b_i is approximately the same. Only those items with highly uneven numbers of observations at either side, the very easy or difficult items, will be affected.

When guessing is introduced, the correlations are more strongly affected, with higher guessing rates giving rise to lower correlations, though the impact remains fairly small (see Table 4). The correlation between the item parameters and their estimates, averaged over all combinations of sample size and test length, was .90, whereas a value of .97 was observed in the null case. For the correlation between the person parameters and their estimates, these values were .89 for the null case and .78 under heterogeneity of the item discriminations. The reason for this robustness is rather clear; guessing does not tend to alter the original distances between either the items or the persons or both, except for a simple linear transformation, thereby maintaining high correlations. Only for very high guessing rates (e.g., yes/no alternatives) may the relative distance not be well preserved, and by consequence, the correlation may decrease.

When both assumptions are violated simultaneously, the average correlations drop from .97 in the null case to .85 for the item difficulties, and from .89 to .75 for the person abilities (see Table 5). Apparently, the two violations do not amplify each other.

Table 3
Correlations and Standard Deviations under Heterogeneous Item Discrimination

Items	Persons	Discr. Par. Lower Limit	Discr. Par. Upper Limit	Statistic* $r_{b\hat{b}}$	$r_{\theta\hat{\theta}}$	$\hat{s}_b$	$\hat{s}_\theta$
10	25	.90	1.10	.90	.79	1.19	1.21
		.50	1.50	.91	.81	1.21	1.20
		.00	2.00	.85	.76	1.19	1.14
25	50	.90	1.10	.95	.90	1.10	1.18
		.50	1.50	.93	.90	1.06	1.10
		.00	2.00	.90	.89	1.05	1.02
25	100	.90	1.10	.97	.89	1.05	1.10
		.50	1.50	.97	.89	1.01	1.08
		.00	2.00	.89	.88	1.01	.98
25	500	.90	1.10	.99	.90	1.02	1.11
		.50	1.50	.98	.90	1.07	1.11
		.00	2.00	.91	.88	1.03	.98
50	500	.90	1.10	.99	.94	1.01	1.05
		.50	1.50	.98	.94	1.01	1.04
		.00	2.00	.90	.93	.95	.98
Average		.90	1.10	.96	.88	1.07	1.13
		.50	1.50	.95	.89	1.07	1.11
		.00	2.00	.89	.87	1.05	1.02

*See Table 2 for an explanation of the symbols.

Table 4
Correlations and Standard Deviations under Conditions of Guessing

Items	Persons	Guess. Par. Lower Limit	Guess. Par. Upper Limit	Statistic* $r_{b\hat{b}}$	$r_{\theta\hat{\theta}}$	$\hat{s}_b$	$\hat{s}_\theta$
10	25	.10	.30	.86	.74	.94	1.01
		.25	.45	.84	.63	.83	.90
		.40	.60	.75	.53	.80	.76
25	50	.10	.30	.90	.86	.83	.92
		.25	.45	.86	.80	.77	.82
		.40	.60	.83	.73	.79	.78
25	100	.10	.30	.94	.84	.82	.90
		.25	.45	.92	.80	.72	.81
		.40	.60	.84	.73	.62	.78
25	500	.10	.30	.97	.85	.75	.89
		.25	.45	.96	.79	.70	.80
		.40	.60	.94	.73	.63	.74
50	500	.10	.30	.97	.91	.78	.84
		.25	.45	.95	.88	.68	.74
		.40	.60	.93	.83	.62	.69
Average		.10	.30	.93	.84	.82	.91
		.25	.45	.91	.78	.74	.81
		.40	.60	.86	.71	.69	.75

*See Table 2 for an explanation of the symbols.

Standard Deviations

Under the assumption of a normal distribution of abilities in the latent distribution, Andersen and Madsen (1977) have described a procedure to estimate the mean and standard deviation of this distribution. Such a distributional assumption, however, is not strictly needed to estimate parameters of the latent distribution of item difficulties and person abilities, since the standard deviations of the latent distributions can be simply estimated on the basis of the estimates. In the present study the standard deviations of both the item and person distributions were estimated and compared with the parameter value, which was always equal to 1.00. The results are presented in Tables 2 through 5. It can be seen that, even in the null case, the dispersion of the latent distribution is usually somewhat overestimated, particularly in small data sets. Quite a number of persons and items seem to be required for accurate estimation. Remarkably, the standard deviations are more accurately estimated when the items have varying item discriminations (see Table 3). Again, it appears that heterogeneity in item discrimination does not invalidate the use of the Rasch model.

Under conditions of guessing, there is a dramatic shrinkage of the estimated standard deviation of both item and person estimates (see Table 4). Higher guessing rates lead to smaller standard deviations of the estimated latent distribution. This finding is not surprising, since in this shrinkage the standard deviation of the Rasch estimates simply behaves like the number of items correct, which also shows smaller standard deviations under guessing. When the two assumptions are violated simultaneously, the standard deviations are mainly determined by the guessing level (see Table 5). Thus, it appears to be quite possible to estimate the standard deviation of the latent distributions without any distributional assumption, provided a sufficiently large number of observations are available and guessing does not occur.

Table 5
Correlations and Standard Deviations under Violation of Both Assumptions

Discr. Par. Lower Limit	Discr. Par. Upper Limit	Guess. Par. Lower Limit	Guess. Par. Upper Limit	Statistic* $r_{b\hat{b}}$	Statistic* $r_{\theta\hat{\theta}}$	Statistic* $\hat{s}_b$	Statistic* $\hat{s}_\theta$
.90	1.10	.10	.30	.93	.83	.84	.90
		.25	.45	.89	.77	.75	.81
		.40	.60	.87	.70	.70	.76
.50	1.50	.10	.30	.91	.81	.84	.87
		.25	.45	.89	.78	.75	.81
		.40	.60	.85	.69	.68	.73
.00	2.00	.10	.30	.84	.78	.83	.80
		.25	.45	.82	.75	.71	.73
		.40	.60	.77	.66	.68	.69

*See Table 2 for an explanation of the symbols.

Bias and RMSE: The Null Data

In Figure 1 the bias of the item estimates is presented. It can be seen that the bias of the estimates is small, except for the extremes of the logit scale where the bias can become larger. Figure 2 shows that the size of the RMSE is mainly determined by the sample size, with larger samples giving rise to a smaller RMSE. The finding of a higher accuracy of the item estimates in larger samples replicates the finding of the previous section. Furthermore,

Figure 1
Bias of the Item Estimates

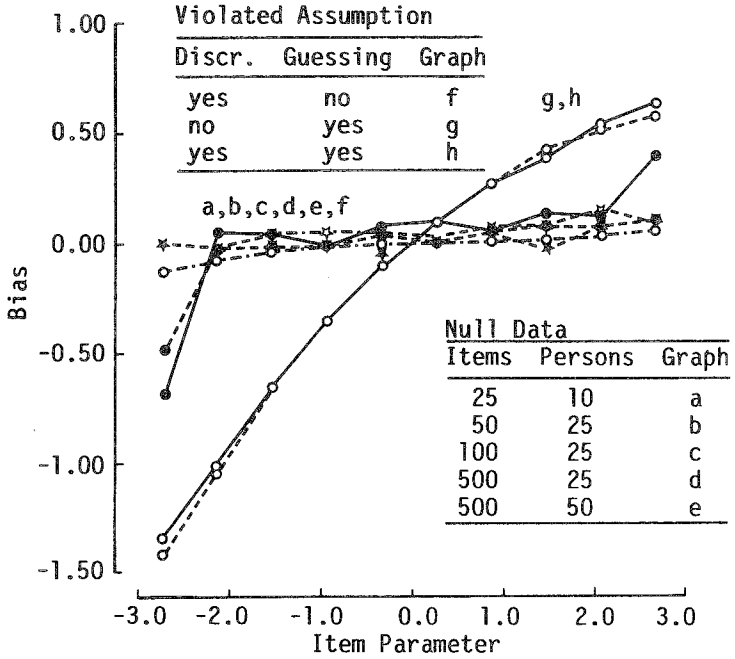


Figure 2
RMSE of the Item Estimates
(See Figure 1 for Explanation of Symbols)

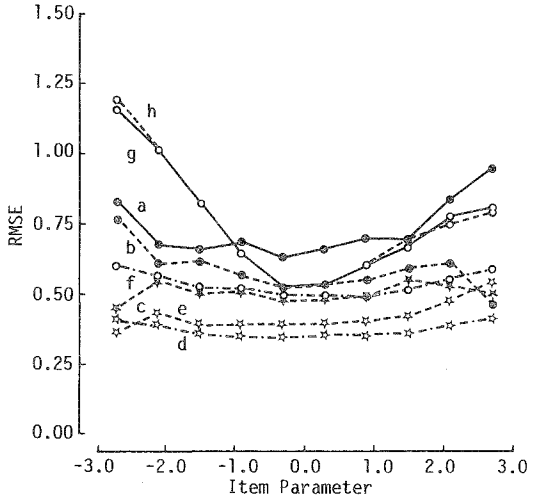
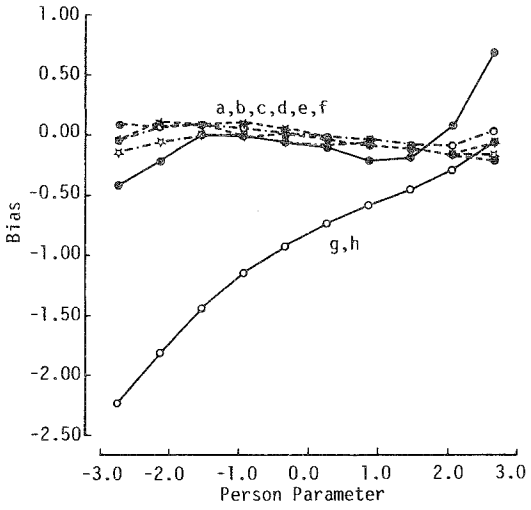


Figure 3
Bias of the Person Estimates
(See Figure 1 for Explanation of Symbols)



it illustrates the consistency of conditional estimates in the Rasch model. At the same time, it has to be noted that the difference between a parameter and its estimate is usually substantial as exemplified by the present RMSE values.

Analogous results were observed for the bias and RMSE of the person estimates, as can be seen in Figure 3. In general, the bias was low; the RMSE of the person estimates decreases with test length. From a comparison of Figure 3 and Figure 4, it can be gathered that the RMSE is smaller for the item estimates than for the person estimates, because the number of persons in the samples was larger than the number of items in the tests.

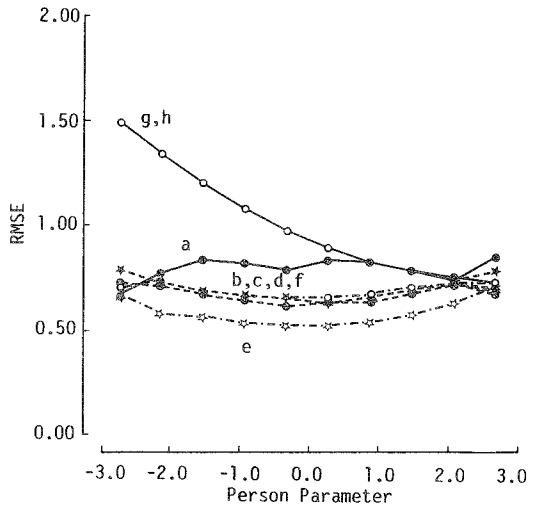
Bias and RMSE Under Violation of the Assumptions

An inspection of Figures 1 through 4 shows that heterogeneity of item discriminations hardly affects the bias and RMSE, replicating the previous findings of robustness of the item estimates in this respect. Guessing, on the other hand, influences both the bias and RMSE measures dramatically. For difficult items the bias graphs are below zero and for easy item values the reverse holds. Thus, the item estimates are systematic underestimations and overestimations of their parameter values.

The effect of this distortion is a reduction in the dispersion of the estimates. It appears that the parameter values of difficult items are systematically underestimated and the parameter values of easy items are overestimated when guessing occurs. This may seem counterintuitive. It may seem more likely to expect the bias and RMSE to decrease with item difficulty, that is, to expect the smallest bias in the easiest items, since guessing is less likely to occur in these items. It should be borne in mind, however, that the logit scales are determined up to a constant. The (arbitrary) midpoint of the item scale was set at zero, as usually done. This leads to a larger bias and RMSE at the extremes. The indeterminacy implies that it is perfectly legitimate to move the bias graphs along the vertical axis. Anchoring the items at the easiest item (e.g., by fixing this item at -3.00) would have yielded a picture with probably a higher intuitive appeal.

Although not further documented here, it was found that both the bias and RMSE increase with higher guessing levels and that the systematic estimation errors do not decrease with an increase in sample size or test length. Thus, guessing leaves

Figure 4
RMSE of the Person Estimates
(See Figure 1 for Explanation of Symbols)



the rank order of the item difficulties largely intact, thereby maintaining high correlations between the item parameters and their estimates, but also introduces large differences between the two, as demonstrated in the shrinkage of the variance of the item estimates.

The results of the person estimates are comparable to those from the previous section. Both bias and RMSE have low values under heterogeneous item discriminations and high values under guessing. Once again, there is a very strong negative correlation between ability and bias.

On the basis of Figure 2, a correction for bias seems straightforward; a procedure similar to the traditional correction for guessing would be indicated. The feasibility of such a guessing correction, however, depends on the data generating algorithm used in which guessing, an item characteristic, occurs whenever the person does not know the correct answer. In more complicated guessing models, for example, models in which guessing is considered as a function of both the person and the item, the traditional correction may turn out to be inferior to other procedures (cf. Wainer & Wright, 1980). At the same time, it should be noted that the validity of different guessing models is difficult to evaluate.

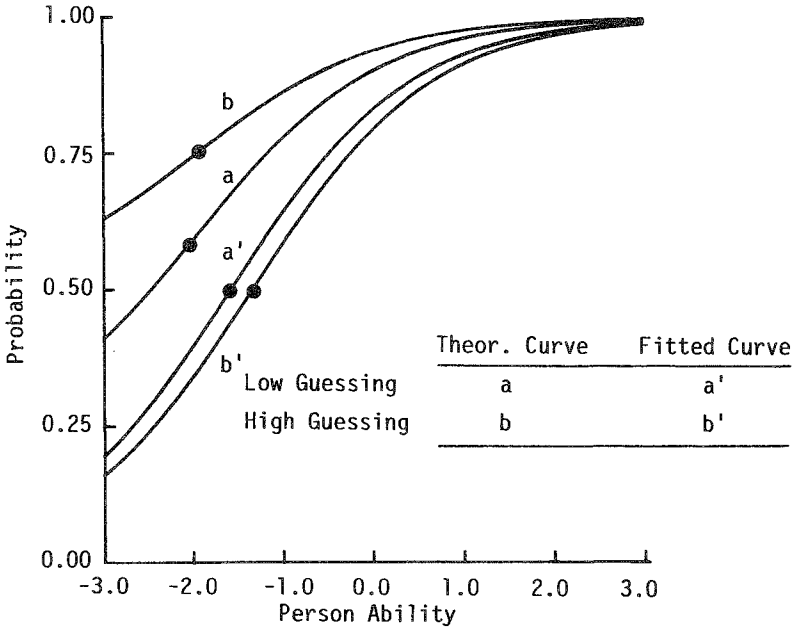
A closer examination of the results with respect to guessing is presented in Figures 5a and 5b. In the first of these figures the theoretical and empirically fitted ICCs are given for two easy items ($b_i = -1.95$ for both items), one with a low guessing parameter ($c_i = .20$) and one with a high guessing parameter ($c_i = .50$). For both items the estimated item difficulties ($\hat{b}_i = -1.58$ and $\hat{b}_i = -1.34$, respectively) were larger than the parameter values, with the larger difference for the item with the higher guessing parameter. The item difficulty, the point of inflection of a curve, is marked in the graphs of Figures 5a and 5b. Apparently, in both cases, the item difficulty is overestimated. In Figure 5b the same procedure has been followed for two difficult items ($b_i = 1.95$ for both items). In this case the estimated item difficulties were less than the parameter values ($\hat{b}_i = 1.41$ for $c_i = .20$, and $\hat{b}_i = .85$ for $c_i = .50$); thus, the item difficulties were underestimated here. Again, the discrepancy was larger for the item with the higher guessing rate. As the estimated values of both easy and difficult items are closer to their average than their corresponding parameter values, it can be concluded that guessing reduces the dispersion of the item difficulties by overestimating the parameters of easy items and underestimating the parameters of difficult items.

Replication of Dinero and Haertel's (1977) Study

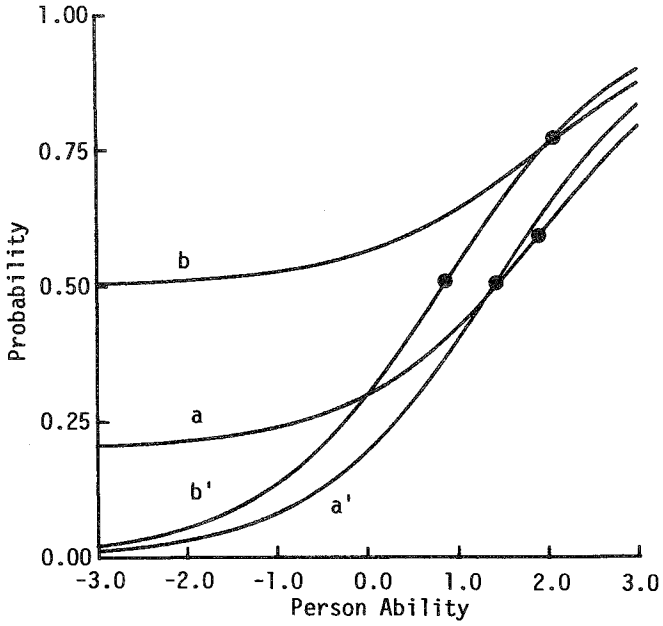
Since the results of the present simulations sharply contrast with those of Dinero and Haertel's (1977) investigation, major parts of their study were replicated. Dinero and Haertel generated two types of data matrices. The first was the so-called "**P** matrix" (matrix **P** in Step 2 of the program structure described above) in which the cells of the data matrix with the probabilities are summed and rounded to the nearest integer in order to obtain the sufficient statistics. The second data matrix, the so-called "**D** matrix," is generated in the same way as the data matrix **D** described in Step 4 of the program structure. In the present replication these **P** and **D** matrices were analyzed for two types of population distributions of the item discrimination parameters,

Figure 5
Theoretical and Fitted Item Characteristic Curves for Two Easy Items and Two Difficult Items under Conditions of Guessing
(The marks at the curves indicate the item difficulties)

(a) Easy Items



(b) Difficult Items



namely, uniform and normal, each combined with five different levels of variance of the item discrimination parameter for 75 persons and 30 items.

The results of this replication are presented in Table 6. When item discrimination parameter values were sampled from a normal distribution, Dinero and Haertel's (1977) results could reasonably be replicated for both **P** and **D** matrices (note that each cell in Table 6 is based on only one simulated data set). No noteworthy differences between the statistics for the two population distributions of the item discriminations emerged. The lowest correlation for uniformly distributed discrimination parameters obtained was .80, which is very far from the low negative values reported by Dinero and Haertel.

Discussion

In this monte carlo study, it was shown that, even in small samples and for short tests, heterogeneity of the item discriminations hardly affects the accuracy of Rasch estimates. The robustness of the model in this respect somewhat downplays the importance of detecting heterogeneity by means of fit tests (e.g., Gustafsson, 1980; Molenaar, 1983; van den Wollenberg, 1982). For an adequate interpretation of the present simulations, it should be noted that all the ICCs in these simulations were monotonically increasing. Consequently, the robustness reported here may apply to ability and achievement tests rather than to attitude questionnaires.

The question of robustness of Rasch estimates is more complicated when guessing occurs, since it depends on the criterion used. When the difference between the item or person parameter and its estimate is used for this purpose, Rasch estimates are definitely not robust. Guessing reduces the dispersion in both person and item estimates. This problem cannot be eliminated by increasing the test length or sample size; the difference between the parameter and its estimate almost exclusively depends on the value of the lower asymptote. In the present simulations it was found that this difference increases with item difficulty and decreases with ability. A similar finding has been reported by Hambleton and Traub (1971). It should be noted that this result is a consequence of the anchoring used. On the logit scales, item difficulties and person abilities are determined up to a constant; adding a constant to all difficulties and abilities does not influence the likelihood of the data matrix. In the present simulations the anchoring was done by fixing the average of the estimated difficulties at zero, thereby equating the average of the item parameters

Table 6
Results of the Replication of the Dinero and Haertel (D & H) Study

Item Discrimination Parameter		Item Difficulty				Ability			
		P Matrix		D Matrix		P Matrix		D Matrix	
Distribution	Variance	D & H	van de Vijver	D & H	van de Vijver	D & H	van de Vijver	D & H	van de Vijver
—	.00	1.00	.99	.98	.98	.96	.99	.98	.92
Uniform	.05	.21	.99	−.07	.95	.59	.99	.24	.96
Uniform	.10	.21	.99	−.14	.95	.62	.99	−.12	.96
Uniform	.15	−.21	.99	−.20	.94	.63	.80	−.21	.95
Uniform	.20	−.22	.99	−.16	.94	.65	.99	−.04	.96
Uniform	.25	−.22	.98	−.19	.94	.66	.99	−.06	.94
Normal	.05	.99	.99	.93	.96	.99	.73	.97	.95
Normal	.10	.98	.99	.95	.96	.94	.94	.98	.96
Normal	.15	.96	.96	.90	.97	.99	.99	.97	.96
Normal	.20	.95	.99	.88	.97	.99	.99	.98	.96
Normal	.25	.93	.99	.85	.97	.99	.99	.97	.96

and their estimates. However, other anchorings are perfectly legitimate. Needless to say, the (squared) difference between parameter and estimates can be arbitrarily eliminated at any point of the logit scale.

The measures used here in the evaluation of robustness, the correlation between parameters and estimates, the standard deviations of the estimates, and the bias and RMSE differ considerably in their degree of precision. Both the correlations and the standard deviations are rather crude measures in comparison with bias and RMSE. For applications of the Rasch model in which the exact values rather than, for example, the rank order of the estimates are needed, the bias and RMSE measures are more relevant than the other measures.

Different approaches are available in item response theory to deal with guessing. One involves the use of the three-parameter model. However, this model cannot be used in small data sets as demonstrated by Hulin et al. (1982) and Thissen and Wainer (1982), among others. The accuracy of the estimates of the guessing parameter is often poor.

Another approach to deal with guessing involves the use of nonstandard estimation procedures (cf. Molenaar, 1983). Basically, nonstandard procedures will weigh responses differentially; the relative contribution of a correct response to the estimated ability is considered to be low (cf. Mislevy & Bock, 1980; Wainer & Wright, 1980) or even absent (cf. Waller, 1974) when the response is very unlikely. The practical value of these approaches will depend on the validity of the underlying model. Obviously, guessing may lead to extremely unlikely correct answers, which should be given little or no weight in the estimation of ability, but these answers can also be indicative of a poor "person fit," that is, of the fact that the test is not an adequate instrument for the examinee due to motivational problems, fatigue, poor instructions, and so forth (cf. Tatsuoka, 1984). The attribution problem, whether difficult items were correctly guessed or easy items were missed for some reason other than lack of ability, is far more complex than these correction procedures suggest.

The present study has some interesting consequences for assessment in applied settings. When dealing with small data sets, there does not seem to be an urgent need to investigate the homogeneity of the item discriminations, a condition which is not easily checked in small data sets. On the other hand, guessing can invalidate the use of the Rasch model, but the presence of guessing is inherent to the response format used and its detection is not dependent on psychometric analysis. In either case there is a need to use formal fit statistics.

References

Andersen, E. B., & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika, 42,* 357–374.

Dinero, T. E., & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement, 4,* 581–592.

Fischer, G. H. (1981). On the existence and uniqueness of maximum likelihood estimators in the Rasch model. *Psychometrika, 46,* 59–77.

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

Hambleton, R. K., & Traub, R. E. (1971). Information curves and efficiency of three logistic test models. *British Journal of Mathematical and Statistical Psychology, 24,* 273–281.

Hulin, C. L., Lissak, R. L., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement, 6,* 249–260.

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic test model. *Educational and Psychological Measurement, 28,* 989–1020.

Mislevy, R., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement, 42,* 725–737.

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika, 48,* 49–72.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95–110.

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

Traub, R. E., & Lam, Y. R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology, 36,* 19–48.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123–140.

Verhelst, N. D., Glas, C., & van der Sluis, A. (1984). Estimation problems in the Rasch model: The basic symmetric functions. *Computational Statistics Quarterly, 1,* 245–262.

Wainer, H., & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika, 45,* 373–391.

Waller, M. I. (1974). *Removing the effects of random guessing from latent trait ability estimates* (Research Bulletin 74–32). Princeton NJ: Educational Testing Service.

Author's Address

Send requests for reprints or further information to Fons J. R. van de Vijver, Department of Psychology, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands.