Some Empirical Results Related to the Robustness of the Rasch Model

Robert Forsyth, Upatham Saisangjan, and Jerry Gilmer
University of Iowa

The primary purpose of this study was to investigate the invariance properties of the Rasch model using data from standardized achievement tests that were not constructed to conform to the Rasch model. The item responses of approximately 3,400 examinees (Grades 9, 10, 11, and 12) to four separately timed sections of the *Iowa Tests of Educational Development* were analyzed. The results indicated that the Rasch model does yield reasonably invariant item parameter and ability parameter estimates for different tests and different examinee groups, even though the assumptions of the model are not met.

As noted by many measurement experts, the use of latent trait models may help solve some of the measurement problems frequently encountered by practitioners (e.g., linking and equating of tests and building item banks). However, these models usually require rather strong assumptions about the nature of the data. The major assumption of all latent trait models is unidimensionality or, equivalently, local independence of items. Additional assumptions are required for each specific latent trait model.

For dichotomously scored items, the one-parameter logistic, or Rasch, model, is the simplest latent trait model and consequently makes the most restrictive assumptions. In addition to the unidimensionality assumption, the model assumes that all items exhibit the same degree of discrimination and that the item characteristic curve (ICC) has a lower asymptote of zero.

For many tests (e.g., standardized achievement tests) it may be assumed that the response data for a group of examinees will not conform exactly to the Rasch model. Standardized achievement tests are frequently built according to a content-by-process table of specifications. For such tests it is reasonable to conclude that the assumption of unidimensionality is not met in any absolute way, although one common factor usually accounts for a large proportion of the observed variance. Also, some tests of this type (e.g., reading comprehension and mathematics problem-solving) frequently employ clusters of items based on the same stimulus material. Consequently, the local independence assumption may be violated to some degree. In addition, it is highly probable that the items in such tests vary in discriminating power. Finally, multiple-choice items are frequently used in these tests and some students may employ a test-taking strategy that includes guessing. If guessing is prevalent, it is probable that the ICC will have a nonzero lower asymptote.

In view of the above comments, it would seem that the latent trait models, and particularly the Rasch model, would not be very useful in mea-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 5, No. 2, Spring 1981, pp. 175–186
© Copyright 1981 Applied Psychological Measurement Inc.
0146-6216/81/020175-12$1.60

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

surement situations involving standardized achievement tests. However, the potential benefits of the latent trait models are such that it seems important to investigate the robustness of these models with data that do not conform in all respects to the assumptions of the models. Since the Rasch model has several advantages relative to the other latent trait models (e.g., the total score is a sufficient statistic for estimating ability), investigations of the robustness of this model are of special interest.

During the last 10 years, many investigations of the robustness of the Rasch model have been conducted. The frequency of such investigations has increased markedly within the last few years. The results of these studies have not been conclusive. Some have found the Rasch model very robust under a variety of circumstances (see, e.g., Dinero & Haertel, 1977; Forster, 1976; Hutten, 1980; Rentz & Bashaw, 1977; Slinde, 1978; Tinsley & Dawis, 1977). Others have reported less encouraging results, at least in certain situations (see, e.g., Loyd & Hoover, 1980; Slinde & Linn, 1978, 1979a).

The benefits derived from using the Rasch model accrue primarily because the estimates of the item parameters are invariant across different groups of examinees and because the estimates of the person parameters are invariant across different sets of items, if the data fit the model. The major purpose of the present study was to investigate these invariance properties for the Rasch model using data for existing standardized achievement tests that were not constructed to conform to the model.

Method

Sample and Instrumentation

The data used in this study were part of a larger study that was conducted in the fall of 1978 to equate the seventh edition of the *Iowa Tests of Educational Development* (ITED; Iowa Testing Programs, 1979) to the sixth edition. Approximately 11,000 students in Grades 9 through 12 in 34 Iowa high schools participated in this equating study. Based on data from previous years, the schools were chosen to provide distributions of examinee scores that were representative of the state of Iowa distributions. Within each school and within each grade, random thirds of the students took one of three forms of the ITED (two new forms and one old form). The data for the present study were obtained from the responses of approximately 3,400 examinees to one of the new forms of the ITED.

Four of the seven subtests of the ITED were selected for analysis:

1. Correctness of Expression (Test E; 81 items);
2. Ability to Do Quantitative Thinking (Test Q: 54 items);
3. Ability to Interpret Literary Materials (Test L; 73 items); and
4. Vocabulary (Test V; 60 items).

These tests were chosen because they represent a variety of test formats and test content and, therefore, should provide a reasonable examination of the robustness of the model.

Each subtest of the ITED consists of three "blocks" of items and is divided into two overlapping levels, designated I and II. Level I includes Blocks 1 and 2 and was taken by 944 students in Grade 9 and 927 students in Grade 10. Level II includes Blocks 2 and 3 and was taken by 899 students in Grade 11 and 650 students in Grade 12. Note that all examinees took the Block 2 items. In general, the items in Block 1 are less difficult than the items in Block 2, and the Block 2 items are less difficult than the Block 3 items.

Data Analysis

Violations of assumptions. For each of the four subtests, some assessment of the degree to which the response data violated the unidimensionality assumption and the equal discrimination assumption was made. An assessment of the extent of guessing on the part of the examinees was not possible.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

To examine the unidimensionality assumption, a principal components factoring procedure was employed with the matrix of tetrachoric inter-item correlations for both the 9th and 11th grade data. The percent of common variance accounted for by the first factor in the unrotated factor solution was used as an index of unidimensionality.[1]

To investigate the equal discrimination assumption, the LOGIST program developed by Wood, Wingersky, and Lord (1976) was used to estimate the item discrimination parameters for the two-parameter logistic model. The variation of these discrimination estimates was used as an indication of the extent of the violation of the equal discrimination assumption. The means, standard deviations, and ranges of these discrimination values were found for each subtest for Grades 9 and 11.

Invariance of item parameter estimates. The LOGIST program was used to estimate the person and item parameters for the one-parameter logistic model. Two main sets of data analysis procedures were undertaken to investigate the robustness of the one-parameter logistic model. The first was concerned with the invariance of the item parameter estimates across different samples of examinees. The second focused on the invariance of the ability estimates across different sets of items.

For the Block 2 items in each subtest, four estimates of each item parameter were obtained. These were derived from the following four data matrices:

1. The responses of examinees in Grade 9 to the items in Blocks 1 and 2.
2. The responses of examinees in Grade 10 to the items in Blocks 1 and 2.
3. The responses of examinees in Grade 11 to the items in Blocks 2 and 3.
4. The responses of examinees in Grade 12 to the items in Blocks 2 and 3.

The Grade 10, Grade 11, and Grade 12 estimates were transformed to the Grade 9 scale using a procedure described by Wright and Stone (1979). This procedure yielded four sets of item parameter estimates, each set having the same mean value. If the model is robust, the four estimates for each item should be very similar.

To gain some insight into the similarity of the item parameter estimates, correlations between the 9th grade estimates and each of the other sets of estimates were computed. Although these correlations provide some help in evaluating the robustness of the model, the implications of the variations of the item parameter estimates are probably best considered by examining the four raw score to ability score conversions derived from the four sets of item parameter estimates.[2] Therefore, for each subtest, four ability estimates were computed for each possible raw score for the Block 2 items. These four ability estimates were obtained using the estimates of the item parameters derived from the four response matrices described above.

Invariance of ability parameter estimates. For each of the four subtests, two ability estimates were obtained for each examinee. For examinees in Grades 9 and 10, these two estimates were based on the Block 1 and Block 2 items, respectively. For examinees in Grades 11 and 12, items in Blocks 2 and 3 were used to obtain the two ability estimates. These two ability estimates were compared using the "standardized difference scores" procedure employed by Wright (1968), Whitely and Dawis (1974), and Slinde and Linn (1978, 1979a). The formula

[1] As noted by Urry (1977) and Hambleton, Swaminathan, Cook, Eignor, & Gifford, (1978) there are several problems with the use of a factor analytic approach to assess the degree of unidimensionality. However, since this procedure has been used frequently by other researchers and since more viable procedures do not seem to be widely available, the factor analytic approach was used in this study.

[2] Several inferential procedures have been proposed to test the invariance of the item parameter estimates (see, e.g., Gustafsson, 1979; Hashway, 1978; Wright, 1977). However, none of these procedures was employed in this study. It was accepted that the assumptions of the Rasch model were violated to some extent.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

used to obtain a standardized difference score for an individual, say D_i, is

$$D_i = \frac{\hat{\theta}_{i1} - \hat{\theta}_{i2}}{\sqrt{S_{i1}^2 + S_{i2}^2}} , \qquad [1]$$

where $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$ are the ability estimates for individual i on the items for Blocks 1 and 2, respectively, and S_{i1}^2 and S_{i2}^2 are the estimated error variances associated with these ability estimates.

Results And Discussion

Table 1 presents the means, standard deviations, and Kuder-Richardson 20 reliability coefficients by grade for each block of items. The difference between the Grade 12 mean for the Block 2 items and the Grade 9 mean for the same items is approximately three-fourths of a standard deviation unit for each of the four subtests. The increase in the average item difficulty from Grade 9 to Grade 12 seems to be fairly typical of such increases for standardized achievement tests (see, e.g., *Comprehensive Tests of Basic Skills, Form S*, CTB/McGraw Hill, 1974, p. 33). However, the differences between grade means are certainly less (in standard deviation units) than the differences between the means of the examinee groups used in some investigations of the robustness of the Rasch model (see, e.g., Slinde & Linn, 1978, 1979a).

The difference between the average p-values for the two blocks of items for each grade and each test is generally less than such differences reported in other studies that have investigated the item-free characteristics of the Rasch model. Typically, these other studies have divided a particular test into an "easy" and a "difficult" test on the basis of individual item difficulty indices. Consequently, the difference between the average p-values for the two tests is fairly large. For example, Slinde and Linn (1979a) used this procedure to form an easy and a difficult reading test. For high-, middle-, and low-ability groups, the differences between the average p-values for the two tests were .24, .31, and .15, respectively. In the present study, only the two blocks of items for the vocabulary test show differences in average p-values of this magnitude (see Table 1).

Violations of Assumptions

Table 2 presents the first eight eigenvalues associated with the matrix of inter-item correlations for each subtest.[3] The percentage of variance accounted for by each factor is also shown in Table 2. The percentage of common variance accounted for by the first factor ranged from about 65% (Test L, Grade 9) to approximately 74% (Test V, Grade 11). These values are typical of those found in other robustness studies involving standardized achievement tests (see, e.g., Slinde & Linn, 1979b). In general, it seems reasonable to assume that one very dominant factor and several minor factors were associated with each subtest.

Table 3 shows the means, standard deviations, and ranges of the item discrimination indices obtained by fitting a two-parameter logistic model to the response data.[4] It is difficult to compare the results in Table 3 with previous studies, since different definitions of the discrimination index have been used. However, these results seem to indicate that the equal discrimination assumption was violated most for Tests L and V and least for Tests E and Q.

Invariance of Item Parameter Estimates

Table 4 shows the correlations between the item difficulty parameter estimates for Grade 9 and each of the other sets of item parameter esti-

[3] Tetrachoric correlations were used. Also, the diagonal elements of each matrix were communality estimates.

[4] These discrimination indices were estimated given the difficulty parameter estimates derived by fitting a one-parameter logistic model to the response data.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 1
Means, Standard Deviations, and Reliability Coefficients by Grade and Test

Grade and Statistic	Test											
	E (27 Items)			Q (18 Items)			V (20 Items)			L		
	Block 1	Block 2	Block 3	Block 1	Block 2	Block 3	Block 1	Block 2	Block 3	Block 1 (27 Items)	Block 2 (19 Items)	Block 3 (27 Items)
9 (N=944)												
M	15.16	13.82	-	10.12	7.81	-	12.39	7.51	-	14.66	8.86	-
SD	5.12	5.65	-	4.09	3.92	-	4.70	4.43	-	5.72	4.01	-
KR20	.79	.83	-	.80	.76	-	.83	.80	-	.84	.75	-
10 (N=927)												
M	16.21	15.08	-	11.05	8.96	-	13.38	8.86	-	15.60	9.32	-
SD	5.34	5.98	-	4.10	4.11	-	4.69	4.73	-	5.98	4.29	-
KR20	.81	.85	-	.81	.79	-	.84	.82	-	.86	.78	-
11 (N=899)												
M	-	16.78	15.69	-	9.78	7.55	-	10.34	7.60	-	10.64	14.57
SD	-	5.38	5.95	-	4.23	4.27	-	5.06	4.61	-	4.43	5.74
KR20	-	.82	.85	-	.79	.79	-	.84	.81	-	.80	.84
12 (N=650)												
M	-	17.83	16.69	-	10.75	8.51	-	11.39	8.43	-	11.80	15.70
SD	-	5.23	6.15	-	4.07	4.58	-	5.03	4.89	-	4.01	6.01
KR20	-	.83	.86	-	.78	.82	-	.84	.83	-	.77	.86

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 2
First Eight Eigenvalues and Associated Percentage of Common Variance

Grade 9		Grade 11	
Eigen-value	% Common Variance	Eigen-value	% Common Variance
	Test V		
12.89	72.2	13.99	73.8
1.40	7.8	1.50	7.9
.85	4.7	.75	4.0
.68	3.8	.71	3.8
.59	3.3	.61	3.2
.53	3.0	.51	2.7
.49	2.7	.45	2.4
.44	2.5	.43	2.3
	Test E		
12.82	65.4	15.81	64.5
1.46	7.4	2.30	9.4
1.18	6.0	1.63	6.7
1.05	5.4	1.28	5.2
.95	4.9	1.08	4.4
.84	4.3	.97	4.0
.66	3.4	.75	3.1
.65	3.3	.69	2.8
	Test L		
11.70	64.5	13.84	71.7
1.69	9.3	1.20	6.2
1.17	6.5	.95	4.9
.88	4.9	.83	4.3
.77	4.2	.74	3.8
.72	4.0	.63	3.3
.64	3.5	.59	3.1
.57	3.1	.52	2.7
	Test Q		
10.53	69.4	11.09	71.0
1.28	8.5	1.15	7.4
.73	4.8	.89	5.7
.65	4.3	.66	4.2
.60	4.0	.55	3.5
.52	3.1	.48	3.0
.44	2.9	.41	2.7
.41	2.7	.39	2.5

Table 3
Means, Standard Deviations, and Ranges of the Estimates of the Discrimination Parameters

Test and Statistic	Grade 9	Grade 11
Test E		
Mean	.54	.59
SD	.17	.16
Range	.15- .88	.22- .88
Test Q		
Mean	.69	.63
SD	.21	.21
Range	.36-1.19	.24-1.28
Test L		
Mean	.59	.67
SD	.23	.26
Range	.24-1.11	.20-1.13
Test V		
Mean	.75	.76
SD	.23	.31
Range	.35-1.37	.38-1.75

mates. For each set of four correlations, the correlation associated with Test L was the lowest. However, all of the correlations were .90 or greater. These results are consistent with the results reported by Tinsley and Dawis (1975), Slinde (1978), and Forster (1976) and indicate that the rank order of the item parameter estimates was very similar for each grade.

As noted previously, the implications of the differences among item difficulty parameter estimates are probably best considered by examining the four raw score to ability score conversions defined by the four sets of item parameter estimates. Table 5 gives these four conversions for each possible raw score value for Test E. It seems obvious that the variations in ability estimates shown in Table 5 are of little practical consequence. The raw score to ability score conversions for the other three subtests exhibited even less variation than the Test E conversions.

In summary, the Rasch model seems to be extremely robust with respect to the person-free

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 4
Pearson Correlations between the Item Parameter Estimates for Grade 9 and Each of the Other Sets of Item Parameter Estimates.

	Test			
Estimates From	E	Q	V	L
Grades 9 and 10	.99	.97	.97	.95
Grades 9 and 11	.96	.96	.96	.93
Grades 9 and 12	.97	.92	.93	.90

Table 5
Raw Score to Ability Score Conversions for Four Sets of Item Parameter Estimates: Test E

	Item Parameter Estimates Based On			
Raw Score	Grade 9	Grade 10	Grade 11	Grade 12
1	-3.62	-3.61	-3.77	-3.83
2	-2.85	-2.84	-2.99	-3.04
3	-2.37	-2.37	-2.37	-2.55
4	-2.01	-2.01	-2.13	-2.17
5	-1.72	-1.72	-1.83	-1.86
6	-1.47	-1.46	-1.56	-1.59
7	-1.24	-1.24	-1.33	-1.35
8	-1.04	-1.03	-1.11	-1.13
9	-0.84	-0.84	-0.91	-0.92
10	-0.66	-0.66	-0.71	-0.72
11	-0.48	-0.48	-0.52	-0.53
12	-0.31	-0.31	-0.34	-0.34
13	-0.14	-0.14	-0.16	-0.16
14	0.02	0.02	0.02	0.02
15	0.19	0.19	0.20	0.21
16	0.36	0.36	0.38	0.39
17	0.54	0.53	0.57	0.58
18	0.72	0.71	0.77	0.78
19	0.91	0.91	0.98	0.99
20	1.11	1.11	1.19	1.21
21	1.34	1.33	1.43	1.46
22	1.59	1.58	1.70	1.73
23	1.88	1.87	2.00	2.04
24	2.23	2.22	2.37	2.41
25	2.70	2.69	2.86	2.90
26	3.45	3.45	3.63	3.69

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 6
Comparison of Ability Estimates Derived from Different Item Blocks: Test E

Grade and Statistic	Block 1	Block 2	Standardized Difference
9 (N=937)			
Mean	-.025	.009	-.068
SD	.9497	1.0959	1.1411
Std. Error	.031	.036	.046
10 (N=917)			
Mean	.020	.047	-.038
SD	1.0429	1.1637	1.0898
Std. Error	.034	.038	.036
	Block 2	Block 3	Standardized Difference
11 (N=882)			
Mean	.052	.026	.037
SD	1.0853	1.1608	1.1612
Std. Error	.037	.039	.039
12 (N=638)			
Mean	.075	.103	-.029
SD	1.1342	1.2244	1.1885
Std. Error	.045	.048	.047

property, at least for tests similar to those in the ITED battery and for examinee groups similar to those used in this study. These results are very consistent with the results reported by other investigators who employed different standardized achievement tests and used examinees at other educational levels (see, e.g., Rentz & Bashaw, 1977; Slinde, 1978).

Invariance of Ability Parameter Estimates

The means, standard deviations, and standard errors for the distributions of ability estimates and the distributions of standardized difference scores for each of the four ITED subtests are given in Tables 6 through 9. As Wright (1968) has noted, these standardized difference scores would be expected to have a mean of zero and a standard deviation of one if the two blocks of items were statistically equivalent.

In general, the results shown in Tables 6 through 9 provide considerable evidence to support the contention that the Rasch model is robust with respect to the item-free measurement property. The means of the standardized difference scores for Tests E and L (Tables 6 and 8) were very close to zero.[5] These results tend to agree with results reported by Slinde (1978), Tinsley and Dawis (1977), Whitely and Dawis (1974), and Wright (1968).

For Grades 10 and 12, the means of the standardized difference scores for Test Q show statistically significant departures from zero (Table 7). The practical significance of these departures may be questioned, however. The maximum departure from zero occurred for the

[5] The variances of these difference scores are statistically greater than one; however, the practical implications of the observed differences seem minimal.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 7
Comparison of Ability Estimates Derived from Different Item Blocks: Test Q

Grade and Statistic	Block 1	Block 2	Standardized Difference
9 (N=927)			
Mean	.030	-.0245	.036
SD	1.2596	1.1182	1.0415
Std. Error	.041	.037	.034
10 (N=902)			
Mean	.056	-.049	.094
SD	1.2671	1.1534	.9654
Std. Error	.042	.038	.032
	Block 2	Block 3	Standardized Difference
11 (N=863)			
Mean	-.006	-.047	.042
SD	1.1857	1.1609	1.0575
Std. Error	.040	.040	.036
12 (N=618)			
Mean	.009	-.088	.112
SD	1.1585	1.2601	1.0221
Std. Error	.047	.051	.041

Grade 12 data. The observed difference between the two ability estimate means for Grade 12 was approximately .10. This difference represents a difference of approximately one-half of a raw score point in the middle of the raw score distributions and is approximately one-fourth of the standard error of measurement for scores in the middle of the ability score distributions. As Slinde and Linn (1979b) noted, differences of this magnitude could "reasonably be tolerated in relation to the types of decisions likely to be made with equated tests" (p. 449).

The means of the standardized difference scores for the vocabulary test (Table 9) showed the most marked departure from the expected value of zero. The Grade 10 data exhibited the most extreme results. The observed difference between the two ability estimate means for Grade 10 was approximately .28. This difference represents a difference of more than one raw score point in the middle of the raw score distributions and is approximately two-thirds of the estimated standard error of measurement for scores in the middle of the ability score distribution. Differences of this magnitude would probably be considered unacceptable in many measurement situations (e.g., test equating). The means of the standardized difference scores for the other grades showed less marked departures from zero, however.

These vocabulary results are similar to the results reported by Slinde (1978) for a group of Grade 5 examinees who received ability estimates on each of two different standardized vocabulary tests and for a group of Grade 4 examinees who obtained ability estimates on each of two different standardized reading tests.

It is interesting to note that the unidimensionality assumption seems more plausible for Test V than for any of the other tests (see Table

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 8
Comparison of Ability Estimates Derived from
Different Item Blocks: Test L

Grade and Statistic	Block 1	Block 2	Standardized Difference
9 (N=935)			
Mean	.005	.012	-.036
SD	1.1190	1.0425	1.1589
Std. Error	.037	.034	.038
10 (N=914)			
Mean	.036	.012	.008
SD	1.2055	1.1304	1.1685
Std. Error	.040	.037	.039
	Block 2	Block 3	Standardized Difference
11 (N=884)			
Mean	.038	.012	.014
SD	1.1922	1.0784	1.0851
Std. Error	.040	.036	.036
12 (N=636)			
Mean	.028	.026	-.013
SD	1.1123	1.1798	1.1511
Std. Error	.044	.047	.046

2). Thus, the observed results may indicate that the unidimensionality assumption is not as critical as the other assumptions. However, in a fairly extensive study that compared the fit of empirical data to the Rasch model, Hutten (1980) found that unidimensionality was an important indicator of model fit. For 20 data sets, the degree of unidimensionality correlated −.55 with a model-fit statistic. For these same data sets, the degree of variability of the discrimination indices and the extent of guessing were not significantly related to the model-fit statistic.[6]

Two factors do seem to be related to the results reported in Tables 6 through 9: (1) the difference between average *p*-values (see Table 1) for the two blocks of items used to obtain the ability estimates and (2) the difference between average discrimination values for these two blocks of items. To investigate the relationship between these two factors and the degree of departure from expectation, three statistics were calculated for each of the 16 data sets. The first statistic was related to the departure of the data from expectation (assuming the Rasch model was appropriate). This statistic was the ratio of the absolute value of the standardized difference mean to the standard error of the standardized difference mean. This ratio ranged from .21 (Test L, Grade 10) to 8.53 (Test V, Grade 10). The second statistic, the absolute difference between average *p*-values for the two blocks of items used to compute the standardized difference scores, ranged from .02 (Test V, Grade 10) to .24 (Test V, Grade 9). The absolute difference between the average discrimination values for the two blocks of items was the third statistic computed. The range of this statistic was from .01 (Test Q, Grade 12) to .23 (Test V, Grade 9).

[6] See Hutten (1980) for the definitions of unidimensionality, degree of variability, and extent of guessing used.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

Table 9
Comparison of Ability Estimates Derived from Different Item Blocks: Test V

Grade and Statistic	Block 1	Block 2	Standardized Difference
9 (N=903)			
Mean	.096	-.090	.179
SD	1.3112	1.1083	1.0557
Std. Error	.044	.037	.035
10 (N=868)			
Mean	.079	-.203	.308
SD	1.3049	1.1209	1.0497
Std. Error	.044	.038	.036
	Block 2	Block 3	Standardized Difference
11 (N=856)			
Mean	.009	-.078	.103
SD	1.2836	1.1412	1.0149
Std. Error	.044	.039	.035
12 (N=611)			
Mean	.024	-.101	.151
SD	1.3033	1.1919	.9944
Std. Error	.053	.048	.040

For these 16 sets of values, the rank-order correlations between the ratio statistic and the other two statistics were .82 (absolute difference in average *p*-values) and .55 (absolute difference in average discrimination indices). Both of these correlations are statistically significant ($p < .05$). These findings, if replicated with other data sets, may have important implications regarding the use of the Rasch model in the vertical equating of certain tests.

One final observation should be made. It would have been possible to derive additional ability estimates for each examinee. For example, the item parameter estimates for Block 2 based on the Grade 12 response data could have been used to obtain ability estimates for Grade 9 examinees. However, given the results reported in the section related to the invariance of item parameters (see Table 5), it is clear that such estimates would be very similar to the ability estimates for the Block 2 items based on the Grade 9 response data.

Conclusions

The major purpose of this study was to investigate the robustness of the Rasch model with respect to the item-free person measurement property and the person-free test calibration property, using standardized achievement tests that were not constructed to conform to the Rasch model. Within the limitations imposed by the particular set of standardized tests used and by the nature of the examinee groups used, the results of this study indicate that the Rasch model does yield reasonably invariant item parameter and ability estimates for different tests and different examinee groups, even though the assumptions of the model are not met.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/

References

CTB/McGraw Hill. *Comprehensive Tests of Basic Skills* (Technical Bulletin No. 1). Monterey, CA: Author, 1974.

Dinero, T. E., & Haertel, E. Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement*, 1977, *1*, 581-592.

Forster, F. *Sample size and stable calibrations*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Gustafsson, J.-E. *Testing and obtaining fit of data to the Rasch model*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, B. R., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research*, 1978, *48*, 467-510.

Hashway, R. M. *Objective mental measurement*. New York: Praeger Publishers, 1978.

Hutten, L. R. *Some empirical evidence for latent trait model selections*. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Iowa Testing Programs. *Iowa Tests of Educational Development, Form X-7*. Iowa City, IA: Author, 1979.

Loyd, B. H., & Hoover, H. D. Vertical equating using the Rasch model. *Journal of Educational Measurement*, 1980, *17*, 179-193.

Rentz, R. R., & Bashaw, W. L. The national reference scale for reading: An application of the Rasch model. *Journal of Educational Measurement*, 1977, *14*, 161-180.

Slinde, J. A. *An evaluation of the Rasch model in providing objective measurement when vertically equating tests*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign, 1978.

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

Slinde, J. A., & Linn, R. L. A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement*, 1979, *16*, 159-165. (a)

Slinde, J. A., & Linn, R. L. The Rasch model, objective measurement, equating, and robustness. *Applied Psychological Measurement*, 1979, *4*, 437-452. (b)

Tinsley, H. E. A., & Dawis, R. V. An investigation of the Rasch sample-free item and test calibration. *Educational and Psychological Measurement*, 1975, *35*, 325-329.

Tinsley, H. E. A., & Dawis, R. V. Test-free person measurement with the Rasch simple logistic model. *Applied Psychological Measurement*, 1977, *1*, 483-488.

Urry, V. W. Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement*, 1977, *14*, 181-196.

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

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

Wright, B. D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service, 1968.

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

Wright, B. D., & Stone, M. H. *Best test design*. Chicago: Mesa Press, 1979.

Author's Address

Send requests for reprints or further information to Robert A. Forsyth, 318 Lindquist Center, University of Iowa, Iowa City IA 52242.

Downloaded from the Digital Conservancy at the University of Minnesota, http://purl.umn.edu/93227.
May be reproduced with no cost by students and faculty for academic use. Non-academic reproduction
requires payment of royalties through the Copyright Clearance Center, http://www.copyright.com/