# An Application of Item Response Theory to Language Testing: Inadequacy of the Rasch Model

Inn-Chull Choi

## 1. Introduction

### 1.1. Problem

Many item response theory (IRT) models have been developed, based on their different assumptions and needs. Although the Rasch model was developed by Rasch (1966) independently of the other item response models and along quite different lines, Rasch's model can be viewed as an item response model in which the item characteristic curve (ICC) is a one-parameter logistic function with the assumptions of equal discriminating power and no guessing.

The Rasch one-parameter model offers several practical benefits. It is the simplest among the logistic models with many statistically refined properties, and is the least expensive to employ in practice. The simple form of this model, however, can constitute its chief weakness when applied to empirical data. Empirical ICCs representing real data do not necessarily have the same slope, nor do they necessarily have lower asymptotes of zero. Given a wealth of empirical evidence accumulated in the measurement field and the fact that even the instructions of the Test of English as a Foreign Language (TOEFL) specifically encourage test takers to guess when in doubt, it may be difficult to accept the notion of no-guessing assumed by the Rasch model. Substantial empirical evidence also disconfirms the assumption of uniform item discrimination index in that items differ in the extent to which they correlate with the underlying trait (Goldstein & Blinkhorn, 1977; Whitely, 1977; Wood, 1978; Gustafsson, 1980). Traub (1983: 65) concludes that "..., *the Rasch model is least likely to fit educa-*

*tional achievement data.*"

Despite these limitations, the Rasch model has been predominant in language testing.[1] There has been no systematic study of the appropriateness or adequacy of the Rasch model and the rationale for its dominance and for the absence of use of other models in the field of language testing. However, when the practical limitations (e.g., sample size, availability of computation tools) no longer apply, it may be well worth the effort to investigate the adequacy of the Rasch model and to compare the appropriateness of the one-parameter Rasch model, the two-parameter model, and the three-parameter model.

### 1.2. Purpose

The present study was intended to check the model data fit in order to investigate the adequacy of the one-parameter Rasch model which has been questioned for its restrictive assumptions but is still prevalent in language testing.

Two major procedures were used to check goodness-of-fit of the Rasch, the two-parameter model, and the three-parameter model with the data: $-2$ log-linear likelihood ratio (produced by BILOG), and residual analyses (Hambleton & Swaminathan, 1985).

### 1.3. Limitations

As there are no universally accepted indices available for checking model assumptions and goodness-of-fit, the findings of the study are only suggestive and the validity of the results is confined within the appropriateness of the statistical measures and procedures employed for this study. Since BILOG is employed to estimate the parameters, the results should be interpreted within the power and limitation of this computer program.[2]

---

[1] Choi (1989) points out that the Rasch model is virtually the one and only model that has been employed in the application of IRT to current research in language testing.

[2] LOGIST is recommended for longer tests and larger samples and when some items are omitted or not reached, whereas BILOG is suggested for short tests and/or small examinee samples (Mislevy & Stocking, 1989). Therefore, considering the short subtests and the relatively small sample size employed for the current study, BILOG was chosen for the present study.

### 1.4. Hypothesis

Because of its strong (restrictive) assumptions, it is hypothesized that the Rasch one-parameter model may not fit the present data as well as the two- and the three-parameter models. Depending on the extent to which test takers resort to guessing, the two-parameter model may work as well as the three-parameter model.

## 2. Procedures

The data for this paper were obtained as part of a larger study that investigated the comparability of two batteries of EFL proficiency tests (Bachman et al., 1989). Analyses were conducted with a variety of statistical packages (SPSSX ver. 3, 1988; SAS ver. 5, 1985; LISREL PRELIS ver. 7, 1988; BILOG ver. 2, 1986).

### 2.1. Measures & Data Collection

Measurement Instruments

The present study focused on the reading comprehension test of the two test batteries, i.e., section 3 of the TOEFL developed and administered by Educational Testing Service, and paper 1 of the First Certificate of English (FCE) developed by University of Cambridge Local Examinations Syndicate. The TOEFL section 3 has 60 question items (two of which are experimental items), consisting of two parts: vocabulary and reading comprehension. Excluding the experimental item, each part has 29 items, all of which are of the four-choice multiple-choice type. The FCE paper 1 has 40 items comprising two parts: vocabulary and reading comprehension. The first part has 25 items, the latter 15 items. All the items are of the four multiple-choice type.

Subjects & Administration

For the sake of the pre-planned quantitative analyses for the project, all subjects were asked to take both the TOEFL and the FCE. 1,613 subjects took the TOEFL at eight different sites: 171 in Thailand, 141 in Egypt, 197 in Japan, 196 in Hong Kong, 196 in Spain, 305 in Brazil, 199 in France, and 208 in Switzerland. 1,400 subjects took the FCE at the same sites: 171 in Thailand, 89 in Egypt, 150 in Japan, 195 in Hong Kong, 190 in Spain, 203 in Brazil, 196 in France, and 206 in Switzerland. The subjects are all

adult students who are learning English as a foreign language.

The tests were administered from November 16 through December 18, 1988 under strict supervision. Both tests were scored by optical scanning machines for minimal data entry error.

## 2.2. Research Design

The Unidimensionality Assumption

Demonstrating that a model's assumptions are largely unsatisfied is neither a necessary nor a sufficient condition, but only suggestive of poor fit. Even though no satisfactory procedures to solve the dimensionality assessment problem have yet been found, there are numerous procedures that are currently advocated or being investigated (Hattie, 1985).

The traditional approach to determining test dimensionality is through factor analysis of the inter-item correlations matrices. Lord and Novick (1968) propose that a sufficient condition for unidimensionality exists if a matrix of tetrachoric inter-item correlations can be reproduced by a single common factor. This condition is not necessary, however, and thus allows only the affirmation of the unidimensionality and not its rejection (Hulin, Drasgow, & Parson, 1983).

Another widely-used means of testing the unidimensionality assumption may be a nonparametric conditional approach involving contingency table based on conditional association being a necessary condition for local independence (Holland, 1981; van den Wollenberg, 1982; Molenaar, 1983; Rosenbaum, 1984; Yen, 1984; Holland & Rosenbaum, 1985).

Stout's(1987) nonparametric model yields the statistical index T used to assess departure from unidimensionality. His method is based on the fundamental principle that local independence should hold approximately when sampling from a subpopulation of examinees of approximately equal ability. Based on two disjoint subtests-the ***assessment subtest*** (chosen to be as unidimensional as possible) and the ***partitioning subtest*** (the remaining subtest), T is computed to test the null hypothesis of unidimensionality.

The dimensionality of the datasets was first checked and the essentially unidimensional datasets were selected to ensure the appropriate and valid use of IRT modeling applied to the model-data fit study. Two nonparametric methods: the nonparametric conditional Stout's approach (1987) and the

factor analytic approach based on tetrachoric correlation matrices (Lord & Novick, 1968) were employed to check the dimensionality of the data.[3] The factor analytic approach rather than content analysis was employed to produce the assessment subtest for Stout's method.

The Equal Discrimination Assumption

This applies only to the Rasch model assuming uniform discriminability. The assumption was checked as suggested by Hambleton & Swaminathan (1985) through identification of percent of item-test score correlations (outliers) falling outside some acceptable range (i.e., average item-test score correlation $\pm .15$).

Model Fit

Following the procedures for checking model assumptions, the essentially unidimensional data sets were used to explore goodness-of-fit among the models of interest.

Fit Statistics

After each estimation cycle, BILOG provides the value of $-2$ times the log likelihood function (Mislevy & Bock, 1984). Since, with large numbers of observation, the null hypothesis that the model fits the data is nearly always rejected using the chi-square test, the $-2$ log likelihood statistics does not provide a useful check of the extent of model fit provided by each of the IRT models examined (Hambleton & Swaminathan, 1985). Rather, these statistics were used to investigate relative model fit between pairs of different IRT models. A rather simplistic analysis of frequencies and ICC of misfit items with each IRT model may also be revealing.

Residual Analyses

The goodness of fit of a latent trait model is most directly evaluated through *residual analysis*. The rationale for the residual analyses is that when an IRT model fits a given dataset, all of the model assumptions must be met to a reasonable degree; if not, larger residuals are expected. There are many studies of residual analysis differing somewhat in how the various steps of the algorithm are implemented (Wright & Panchapakesan, 1969; Wright & Mead, 1976; Bock, 1972; Yen, 1981).

The procedure involved fitting the one-, the two-, and the three-parameter

[3] I would like to thank Dr. William Stout for generously providing the software for me to apply to my study.

models to the datasets, using the model parameter estimates to predict the item performance data. The difference between the actual and predicted proportion-correct score in each of 12 equal interval categories of theta scale and for each item was divided by the corresponding standard error of the proportion-correct estimate to obtain a standardized residual (Hambleton & Swaminathan, 1985; Hambleton & Murray, 1983). Graphic presentation was used to aid the overall understanding of the results.

## 3. Results and Discussion

### 3.1. The Unidimensionality Assumption

The interpretation of the factor analysis of tetrachoric correlation matrices does not seem to indicate any clear evidence that the tests are essentially unidimensional for adequate IRT modeling. The screen plots appear to show that none of the data sets have a single factor solution.

Stout(1987) has proposed a nonparametric model to produce a statistical index T, which indicates the extent of departure from unidimensionality. The summary results are given in Table 1.

N.B.: Under H0 : d=1, T should be N(0, S), S<=1.
Under H1 : d>1, T will have a positive mean.
The following notations will be used hereafter.
VOC: Vocabulary Dataset
RDG: Reading Dataset
TOT: Total Dataset

**Table 1.** Summary Results of Stout's Approach

| Test | T | p | $H_0$(alpha=.01) | Dimensionality |
|---|---|---|---|---|
| | | | TOEFL | |
| VOC | .51774 | .30232 | accept: | unidimensional |
| RDG | 3.97816 | .00003 | reject: | multidimensional |
| TOT | 5.23227 | .00000 | reject: | multidimensional |
| | | | FCE | |
| VOC | −1.41120 | .48006 | accept: | unidimensional |
| RDG | 1.70854 | .04377 | accept: | unidimensional |
| TOT | −.16721 | .48006 | accept: | unidimensional |

Although the unidimensionality hypothesis of the FCE reading subtest could be rejected at alpha of .05, the results for this subtest may be invalid because of its short length of only 15 items. Stout suggests that at least 20 items are needed for valid interpretation of the results (personal communication). Thus, the three essentially unidimensional datasets, i.e., TOEFL vocabulary, FCE vocabulary, and FCE total data sets were chosen for the further model-data fit studies.

### 3.2. The Equal Discrimination Assumption

To check the equal discrimination assumption of the Rasch model, Hambleton & Swaminathan(1985) have suggested identifying the percent of item-test score correlations that fall outside some acceptable range, i.e., $r_{bis} \pm .15$.

The results are provided in Table 2 below:

**Table 2.** *Percentage of Outliers of $r_{bis}$ (%)*

| | VOC | RDG | TOT |
|---|---|---|---|
| TOEFL | 31 | 27 | 31 |
| FCE | 4 | 26 | 17 |

Although no absolute criterion as to the variance of item-test score correlations is available, these results given in Tables 3 and 4 nevertheless suggest that the assumption of equal item discrimination indices is violated to a considerable degree. It may not be possible to assume equal discrimination for all the data sets except for FCE Vocabulary. This untenability of equal discrimination assumption raises a question as to the appropriateness of the Rasch model for these data sets.

### 3.3. Goodness-of-Fit among Models

Fit Statistics

−2 Log Likelihood Ratio

After each estimation cycle, BILOG provides a statistic of −2 times the log likelihood ratio which can be used as a rough estimate of the relative fit between a pair of different IRT models. The three IRT models were compared as to their model fit on the basis of the corresponding fit

statistics.

The −2 log likelihood ratios for each data set produced with each IRT model are given in Table 3.

**Table 3.** BILOG Fit Statistics

| | 1 PL | 2 PL | 3 PL |
|---|---|---|---|
| TOEFL | | | |
| VOC | 45580.557 | 44947.119 | 44865.184 |
| FCE | | | |
| VOC | 29436.182 | 29326.189 | 29306.212 |
| TOT | 46381.678 | 45917.253^ | 45920.886^ |

PL stands for Parameter Logistic Model.

"^" indicates the unusual pattern, i.e., that the fit statistic of the lower parameter model is smaller (meaning better fit) than that of the higher parameter model. But the negligible discrepancy may be attributed to relatively small sample size (N=1,088) for the FCE data set.

The differences in −2 log likelihood ratios between pairs of models are provided in Table 4.

**Table 4.** Differences in Fit Statistics between Pairs of Models

| | 1PL~2PL | 2PL~3PL | 1PL~3PL |
|---|---|---|---|
| TOEFL | | | |
| VOC | 633.438** | 81.935** | 715.373** |
| FCE | | | |
| VOC | 109.993** | 19.977 | 129.970** |
| TOT | 464.425** | −3.633 | 460.792** |

** denotes the statistical significance at a=.001

Table 4 demonstrates that, with all the data sets, there were statistically significant differences in model fit between the one- and the two-parameter models, and between the one- and the three-parameter models. But only with the TOEFL data set were there significant differences in model fit between the two- and the three-parameter models. This may be for two reasons: (a) the instructions of the TOEFL encourage students to guess, when in doubt of a possible answer, or (b) item selection for the TOEFL is based on the three-parameter model.

It should be noted that with FCE Reading and Total data sets, the two-

parameter model provides a negligibly better (nearly equal) fit than the three-parameter model. This may be due to the fact that the test takers are not told anything about guessing in the instructions for the FCE tests, and they may thus not resort to guessing as much as in taking the TOEFL.

Frequencies of Misfit Items

Misfit items are identified by BILOG at alpha of .01. Table 5 demonstrates a clear pattern that the higher (more powerful) parameter model has far fewer misfit items than the lower (more restrictive) model. Thus, it may provide an additional evidence for the claim that higher parameter models provide much better fits than does the one-parameter model.

**Table 5.** Numbers of Misfit Items

| TEST | 1 PL | 2 PL | 3 PL |
|---|---|---|---|
| TOEFL | | | |
| VOC | 14 | 0 | 0 |
| FCE | | | |
| VOC | 4 | 0 | 0 |
| TOT | 9 | 1 | 1 |

Residual Analyses[4]

Standardized Residual(SR)

Standardized residuals(SRs) can be interpreted as a global indicator of model-data fit in that the magnitude of SR represents the discrepancy between the observed proportion correct and the predicted proportion correct based on item parameters and person parameter. When an adequate model-fit is obtained, the SRs are expected to be small and distributed randomly around 0.

Table 6 provides a complete summary of frequencies (i.e., numbers of items) and percentages of absolute value of four SR categories for three data sets based on three IRT models.

As shown in Table 6, the SRs are substantially smaller for the two- and the three-parameter model indicating that these fit the data strikingly better than does the Rasch model. As for the two- and the three-parameter

[4] Due to limited space, the illustrations of only one dataset were presented for each section. The patterns of the figures are essentially consistent with the rest of the datasets.

**Table 6.** Frequencies & Percentages of Standardized Residuals

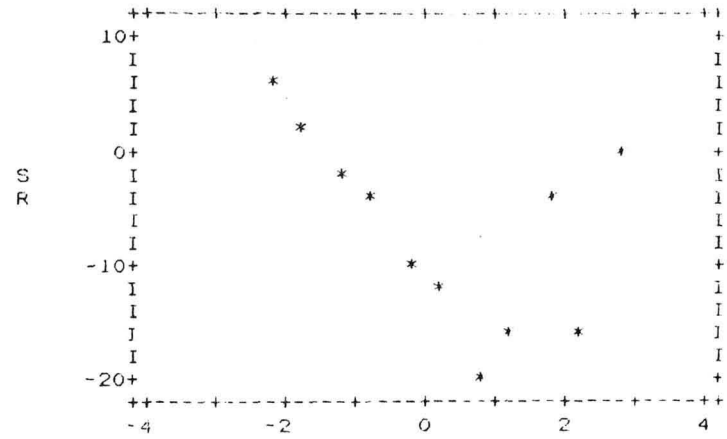
| | 1 PL | | 2 PL | | 3 PL | |
|---|---|---|---|---|---|---|
| | FREQ | % | FREQ | % | FREQ | % |
| TOEFL VOC | | | | | | |
| 0<=\|SR\|<1 | 42 | 12.1 | 211 | 60.6 | 220 | 63.2 |
| 1<=\|SR\|<2 | 31 | 8.9 | 105 | 30.2 | 102 | 29.3 |
| 2<=\|SR\|<3 | 21 | 6.0 | 31 | 8.9 | 24 | 6.9 |
| 3<=\|SR\| | 254 | 73.0 | 1 | 0.3 | 2 | 0.6 |
| FCE VOC | | | | | | |
| 0<=\|SR\|<1 | 23 | 8.4 | 184 | 61.1 | 183 | 61.0 |
| 1<=\|SR\|<2 | 25 | 9.1 | 92 | 30.0 | 90 | 30.0 |
| 2<=\|SR\|<3 | 25 | 9.1 | 22 | 7.3 | 22 | 7.3 |
| 3<=\|SR\| | 202 | 73.4 | 2 | 0.6 | 5 | 1.7 |
| FCE TOT | | | | | | |
| 0<=\|SR\|<1 | 44 | 10.0 | 314 | 65.4 | 323 | 67.3 |
| 1<=\|SR\|<2 | 27 | 6.1 | 129 | 26.9 | 114 | 23.8 |
| 2<=\|SR\|<3 | 22 | 5.0 | 31 | 6.5 | 37 | 7.7 |
| 3<=\|SR\| | 347 | 78.9 | 6 | 1.2 | 6 | 1.2 |

models, the percentage of SR whose absolute value is less than 1 is much larger than that of SR whose absolute value is equal to or greater than 3. The opposite is obviously true of the Rasch model.

Given the fact that all of the model assumptions must be met to a reasonable degree when an IRT model fits a data set, Table 6 demonstrates that the Rasch model is clearly inferior to the two- and the three-parameter models in providing model fit. It should also be noted that the two-parameter model in general provides nearly as good a fit as the three-parameter model.

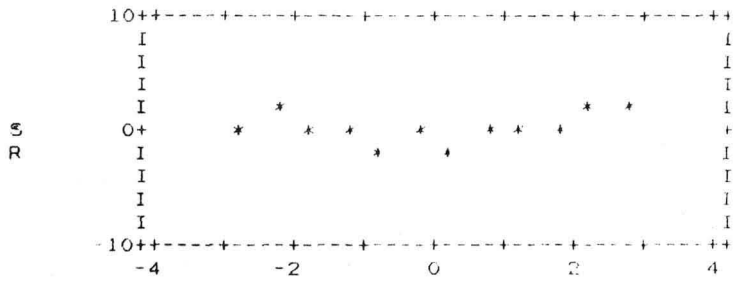
Plot of SRs

A perfect model fit would lead to 0 SRs, which will produce a series of dots in a straight line parallel to the theta scale axis in graphic display. The following Figures 1, 2, 3 of the plots of SRs for a typical sample item #1 of FCE total dataset illustrate considerable differences in model fit among three IRT models. That is, the Rasch model fails while the two-parameter model in general provides as good a model fit as the three-parameter model. It should be noted that the graphic displays of SRs reveal



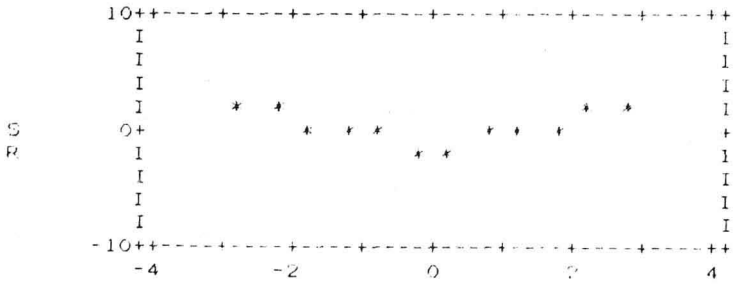

THETA SCALE

**Fig. 1.** Plot of SRs based on 1PL model for FCE Item #1




THETA SCALE

**Fig. 2.** Plot of SRs based on 2PL model for FCE Item #1



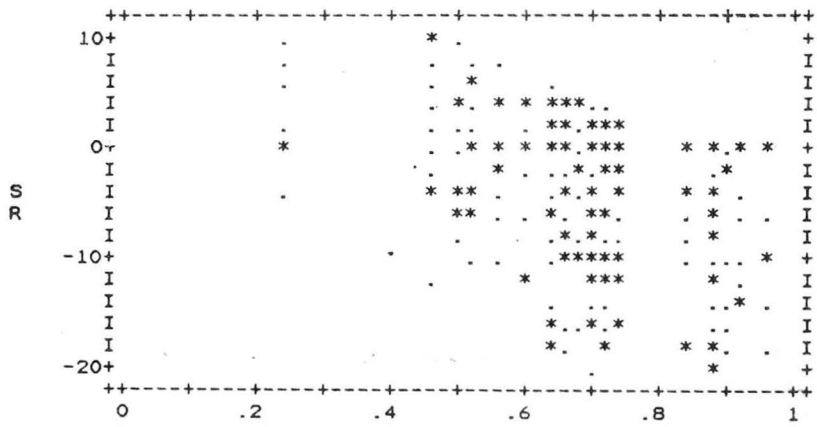

THETA SCALE

**Fig. 3.** Plot of SRs based on 3PL model for FCE Item #1

that around the middle range of the ability scale, the Rasch model yields the largest negative SRs whose patterins strikingly differ from those of the two- and the three-parameter models. That indicates that the Rasch model tends to overestimate the proportion correct for the group of middle ability level. At extreme ability ends, however, the Rasch model provides SRs somewhat close to those of the two- and the three-parameter models. These findings are consistent with those of Hambleton & Swaminathan (1985).

Scatterplot of SRs & Item Difficulty Indices

The following Figures 4, 5, 6 are the illustrations of SRs plotted against classical item difficulty indices with the three IRT models for the FCE vocabulary dataset. The scatterplots of SRs and classical difficulty indices for all the data sets indicate that the Rasch model tends to underestimate the proportion correct with more difficult items and overestimate the proportion correct with less difficult items.

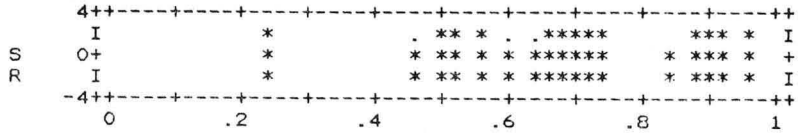



CLASSICAL ITEM DIFFICULTY INDEX (p)
(*: Multiple occurrence: ·: Single occurrence)

**Fig. 4.** Scatterplot of 1PL Model SRs and Classical Item Difficulty Indices for FCE Vocabulary

Scatterplot of SRs & Item Discrimination Indices

The following Figures 7, 8, 9 show the scatterplots of SRs and classical item discrimination indices with the three IRT models for the TOEFL vocabulary dataset. These figures illustrate a clear pattern indicating that the Rasch model produces an interaction effect between SRs and classical



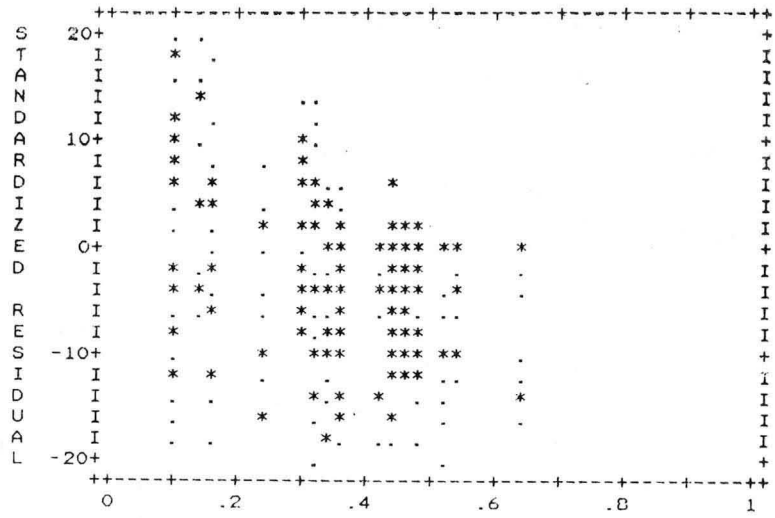

CLASSICAL ITEM DIFFICULTY INDEX (p)

**Fig.** 5. Scatterplot of 2PL Model SRs and p for FCE Vocabulary


CLASSICAL ITEM DIFFICULTY INDEX (p)

**Fig.** 6. Scatterplot of 3PL Model SRs and p for FCE Vocabulary

discrimination indices. Since the Rasch model is based on an equal discrimination assumption, it seriously fails to fit the data especially with items of lower discriminability. This finding suggests that the assumption of equal




CLASSICAL ITEM DISCRIMINATION INDEX ($r_{bis}$)

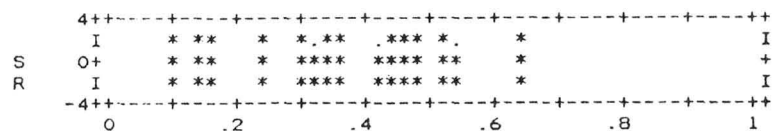
(*: Multiple occurrence: ·: Single occurrence)

**Fig.** 7. Scatterplot of 1PL Model SRs and Classical Discrimination Indices for TOEFL Vocabulary

discrimination is problematic enough to make the Rasch model inadequate in fitting the data.


CLASSICAL ITEM DISCRIMINATION INDEX ($r_{bis}$)

**Fig. 8.** Scatterplot of 2PL Model SRs & $r_{bis}$ for TOEFL Vocabulary




CLASSICAL ITEM DISCRIMINATION INDEX ($r_{bis}$)

**Fig. 9.** Scatterplot of 3PL Model SRs & $r_{bis}$ for TOEFL Vocabulary

## 4. Conclusions and Implications

### 4.1. The Unidimensionality Assumption

The factor analysis reveals that none of the six data sets have a single factor solution. However, based on the results of Stout's method, the TOEFL vocabulary, the FCE vocabulary and total datasets prove to be essentially unidimensional at alpha of .01.

### 4.2. Model Fit

Fit Statistics

The fit statistic seems to be a very useful indicator to compare the goodness-of-fit among the three IRT models. With the FCE reading and total data sets, the two-parameter model provided a negligibly better (statistically non-significant) fit than did the three-parameter model. This suggests that the application of the three-parameter model to data sets which are basically free from guessing may not result in as good a fit as that obtained with the two-parameter model.

Residual Analyses

The residual analyses produced results in a strikingly consistent and clear

pattern as was hypothesized, i.e., the Rasch model provided a less adequate fit than the two- and the three-parameter models. The graphic displays of standardized residuals are very revealing concerning the model-fit provided by three IRT models. The plot of standardized residuals against classical item difficulty and classical discrimination indices also support the above findings. With the Rasch model, there is a clear interaction effect between SRs and discrimination indices–larger SRs were obtained with items of lower discrimination. Also, the Rasch model tends to underestimate the proportion correct with more difficult items, and overestimate the proportion correct with less difficult items.

### 4.3. Inadequacy of the Rasch Model

The results of the fit statistics and the residual analyses turned out to be compatible with each other in terms of inadequacy of the Rasch model. As was hypothesized, the results provide strong evidence that there is a striking difference between the Rasch model and the other two models in providing model fit. It has been clearly demonstrated that the restrictive Rasch model fails to provide model–fit with the language test data sets used for the present study. This finding suggests that the general finding in the psychometrics that the two- and the three-parameter models provide better fit than the restrictive Rasch model, may also be true of the language test data (Goldestein & Blinkhorn, 1977; Whitely, 1977; Wood, 1978; Gustafsson, 1980).

The exploration of classical item discrimination indices ($r_{bis}$) of all the data sets may also provide the basis for rejecting the equal discrimination assumption of the Rasch model. Even though the use of the Rasch model has been predominant in language testing, the findings of the present study do suggest that IRT models more powerful than the Rasch model are more appropriate for the purpose of applying IRT to data from language tests.

### 4.4. Two-Parameter Model vs. Three-Parameter Model

It may also be important to note that the two-parameter model works nearly as well as the three-parameter model. The fit statistics of BILOG indicate that the two-parameter model works negligibly better than the three-parameter model with the FCE data sets. And, the residual analyses

suggest that, even with the TOEFL, which encourages guessing, the two-parameter model provides as good a model fit as the three-parameter model. This finding may have significant implications for practical limits on sample size required for legitimate use of IRT modeling. The justifiable use of the two-parameter model instead of the three-parameter model may relieve measurement researchers of the burden of obtaining the large sample sizes required for the three-parameter model estimation.

## 5. Summary

In summary, applying IRT to a model-data fit study with two widely used EFL tests, the present study investigated the adequacy of the Rasch model which is still dominant in language testing. Based on the results of the unidimensionality checks, three data sets were selected for further investigation of the model fit.

The hypothesis that the Rasch model may be inadequate for language test data was supported by the results of several statistical procedures. Given the finding that the Rasch model clearly fails to provide an adequate model fit, the prevailing use of the Rasch model in language testing should be re-evaluated. The relatively small differences in model fit between the two- and the three-parameter models suggest that the effect of differential discrimination of the test items is more significant than the guessing factor in the application of IRT modeling tests.

## REFERENCES

Bachman, Lyle, F., Fred Davidson, Kathy Ryan, & Inn-Chull Choi. (1989) *The Cambridge-TOEFL Comparability Study: Final Report.* Urbana, IL: University of Illinois at Urbana-Champaign.

Bock, R.D. (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika* 37, 29-51.

Choi, Inn-Chull. (1989) *An Application of Item Response Theory to Language Testing: Model-Data Fit Studies.* Unpublished Ph.D. Thesis. Urbana: University of Illinois at Urbana-Champaign.

Goldestein, H., & Blinkhorn, S. (1977) Monitoring educational standards-an inappropriate model. *Bulletin of the British Psychological Society* 30, 309-311.

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

Hambleton, R.K., & Hariharan Swaminathan. (1985) *Item Response Theory: Principles and Applications.* Boston: Kluwer-Nijhoff.

Hambleton, R.K., & Linda Murray. (1983) Some goodness of fit investigations for item response models. In R.K. Hambleton, editor, *Applications of item response theory*, Vancouver, BC: Educational Research Institute of British Columbia.

Hatties, J.A. (1985) Methodology review: Assessing unidimensionality of test and items. *Applied Psychological Measurement* 9, 139-64.

Hicks, Marilyn, M. (1984) *A Comparative Study of Methods of Equating TOEFL Test Scores.* Princeton: Educational Testing Service.

Holland, P.W. (1981) When are item response models consistent with observed data? *Psychometrika* 46, 79-92.

Holland, P.W., & Rosenbaum, P.R. (1985) Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics* 14, 1523-1543.

Hulin, Charles, L., Fritz Drasgow, & Charles K. Parsons. (1983) *Item Response Theory: Application to Psychological Measurement.* Homewood, IL: Dow Jones-Irwin.

Lord, Frederic, M. & Melvin, R. Novick. (1968) *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Mislevy, Robert J., & R. Darrell Bock. (1984) *BILOG: Item Analyses and Test Scoring with Binary Logistic Models*, Version 2. 2. Mooresville, In: Scientific Software, Inc.

Mislevy, Robert J., & M.L. Stocking. (1989) A consumer guide to LOGIST and BILOG. *Applied Psychological Measurement* 13, 57-75.

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

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

19, 49-57.

Rosenbaum, P.R. (1984) Testing the conditional indepen ence and monc-tonicity assumptions of item response theory. *Psychometrika* 49, 425-435.

Stout, William. (1987) A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika* 52, 589-617.

Traub, R.E. (1983) A priori considerations in choosing an item response model. In R.K. Hambleton, editor, *Appolications of item response theory*, Vancouver, BC: Educational Research Institute of British Columbia.

Whitely, S.E. (1977) Models, meanings and misunderstandings: Some issues in applying Rasch's theory. *Journal of Educational Measurement* 14, 227-235.

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

Wood, R. (1978) Fitting the Rasch model-A heady tale. *British Journal of Mathamatical and Statistical Psychology* 31, 27-32.

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

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

Yen, W. M. (1981) Using simulating retults to choose latent trait model. *Applied Psychological Measurement* 5, 245-262.

Yen, W.M. (1984) Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement* 8, 125-145.

## ABSTRACT


### An Application of Item Response Theory to Language Testing: Inadequacy of the Rasch Model

Inn-Chul Choi

Even though the application of IRT to language testing has recently attracted much attention, no model-data fit research has been conducted to explore the appropriateness of IRT modeling in language testing. No study has examined the adequacy of the Rasch model which has been predominant in language testing.

Employing fit statistics, and the residual analyses, the current paper conducted model-data fit studies on two EFL reading tests with the 1-, 2-, and 3- IRT models to investigate the adequacy of the Rasch model. The results clearly indicate that the Rasch model fails to provide an adequate fit for these data, thus suggesting that the prevailing use of the Rasch model in language testing needs to be re-evaluated. The comparison of model-fit between the 3-parameter model and the 2-parameter model also suggests that for language tests, the discrimination parameter is more significant than is the guessing parameter.



136-053
서울시 성북구 동선동 3가
성신여자대학교 영어영문학과