The Analysis of Item-Ability Regressions: An Exploratory IRT Model Fit Tool

Neal M. Kingston and Neil J. Dorans
Educational Testing Service

The use of item-ability regressions (the comparison of the regression of the observed proportion of people answering an item correctly on estimated θ with the estimated item response function) to investigate the psychometric properties of particular item types in a given population was explored using data from four administrations of 10 item types (a total of 806 items) from the Graduate Record Examinations General Test. Although the method does not allow an absolute determination of fit for a latent trait model (in this case, for the three-parameter logistic model), it does show that certain item types consistently fit the model worse than other item types, and it led to and supported a specific hypothesis as to why the model probably did not fit these item types.

Item response theory (IRT), when the assumptions of the chosen model (local independence and form of model) are met, provides powerful tools for the analysis of items, tests, and people (Hambleton & Swaminathan, 1984; Lord, 1980). Numerous attempts have been made to assess the fit of item response models to observed data (Holland, 1981; Rosenbaum, 1984; Yen, 1981, 1984), but existing techniques have not been used in an exploratory fashion. The thrust of this article is to provide insight as to why items do not fit the model.

To explore the fit of the three-parameter logistic model to Graduate Record Examinations (GRE) data, a heuristic graphical technique and some quantitative summaries of that technique were used in a roughly normative manner. This exploratory technique, which will be referred to as analysis of item-ability regressions, compares the regression of the observed proportion of people answering an item correctly on estimated θ (empirical regression) with the item response function based on the estimated item parameters (estimated regression; Hambleton, 1980; M. Stocking, personal communication, 1980).

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 9, No. 3, September 1985, pp. 281–288

0146-6216/85/030281-08$1.65

The Analysis of Item-Ability Regressions Technique

The untransformed ability scale (θ estimated on the metric for which the trimmed calibration sample, examinees with estimated θ between −3 and 3, has a mean of 0 and a standard deviation of 1) is split into 15 intervals of width .4 in the range for −3.0 to +3.0. P_i, the proportion of people in interval i answering the item correctly, adjusted for omits, is computed for each interval. That is,

$$P_i = \frac{n_i^+ + (n_i^o/A)}{n_i} \quad , \tag{1}$$

where n_i^+ is the number of examinees in the ith interval who answered the item correctly,
n_i^o is the number of examinees in the ith interval who omitted the item,
A is the number of alternatives per item,

n_i is the number of examinees in interval i who answered the item or any item subsequent to that item.

The 15 P_i are plotted as squares whose areas are proportional to n_i. For each interval, a line of length $4(PQ/n_i)^{1/2}$ is plotted, where P and Q are computed from the estimated item response function. The line is centered on the estimated response function.

It should be noted that although this line is a rough estimate of the .95 confidence interval around the item response function, it is not being used as a statistical test. The reasons why this line does not represent an actual .95 confidence interval include: (1) the use of 2 instead of 1.96 as a coefficient, (2) the use of the inappropriate symmetric normal approximation to the binomial confidence interval around the response function (particularly a problem for extreme values of P), and (3) the use of an interval based on estimated item parameters.

Figures 1a through 1f show six examples of item-ability regressions. The vertical scale in each is the probability of a correct response and ranges from 0 to 1. The horizontal scale is the ability metric and ranges from -3 to $+3$. Various attributes of these item-ability regressions relate to model fit. After looking at more than 1,000 of these plots, it was decided that a useful summary statistic would be the number of times the proportion of the examinees in an interval responding correctly to the item fell outside the $\pm 2(PQ/n_i)^{1/2}$ interval centered on the response function, that is, the number of times the midpoints of the boxes fell off the vertical lines. Thus, the item-ability regressions in Figures 1a and 1b would each be scored 0, those in Figures 1c and 1d would be scored 2 and 3, respectively, and those in Figures 1e and 1f would be scored 5 and 9, respectively.

The GRE Database

This analysis was based on 395 verbal, 275 quantitative, and 136 analytical items. The verbal and quantitative items consisted of all such operational items from administrations of four GRE General Test editions. The analytical items consisted of all operational items from two of these four editions.

The verbal measure consists of four item types: analogies, antonyms, sentence completion, and reading comprehension. The quantitative measure consists of three item types: regular mathematics, data interpretation, and quantitative comparison. The analytical measure (at the time of this research) also consisted of three item types: analysis of explanations, logical diagrams, and analytical reasoning. Examples of each item type can be found in the GRE Bulletin (Educational Testing Service, 1985). Item and person parameters were estimated separately for the verbal, quantitative, and analytical measures using LOGIST (Wingersky, 1983). Sample sizes ranged from approximately 3,000 to 7,250 examinees per item.

Results

Table 1 presents cumulative distributions of item scores on the model fit statistic described above. Data are presented for the three major item classifications and their constituent item types. To aid interpretation of these data, frequencies of model fit score were collapsed into two categories (0, 1, 2+), and compared across item types with a chi-square test of independence. Table 2 presents these results for the three major item classifications.

The high chi-square for Table 2 shows a relationship between broad item classification and model fit. Whether or not the three-parameter logistic model fits data for any of the item types in an absolute sense, Table 2 shows that some item types are fit more closely than others. In particular, the order of fit seems to be (from best to worst): verbal, analytical, quantitative. Since these differences might be due to specific item types, each broad classification was separately analyzed by specific item type. Table 3 presents these results for the verbal, quantitative, and analytical item types.

The four verbal item types presented in Table 3 show no significant difference in model fit. Of the three quantitative item types, the model fits the quantitative comparison items least well.

Figure 1
Some Examples of Item-Ability Regressions

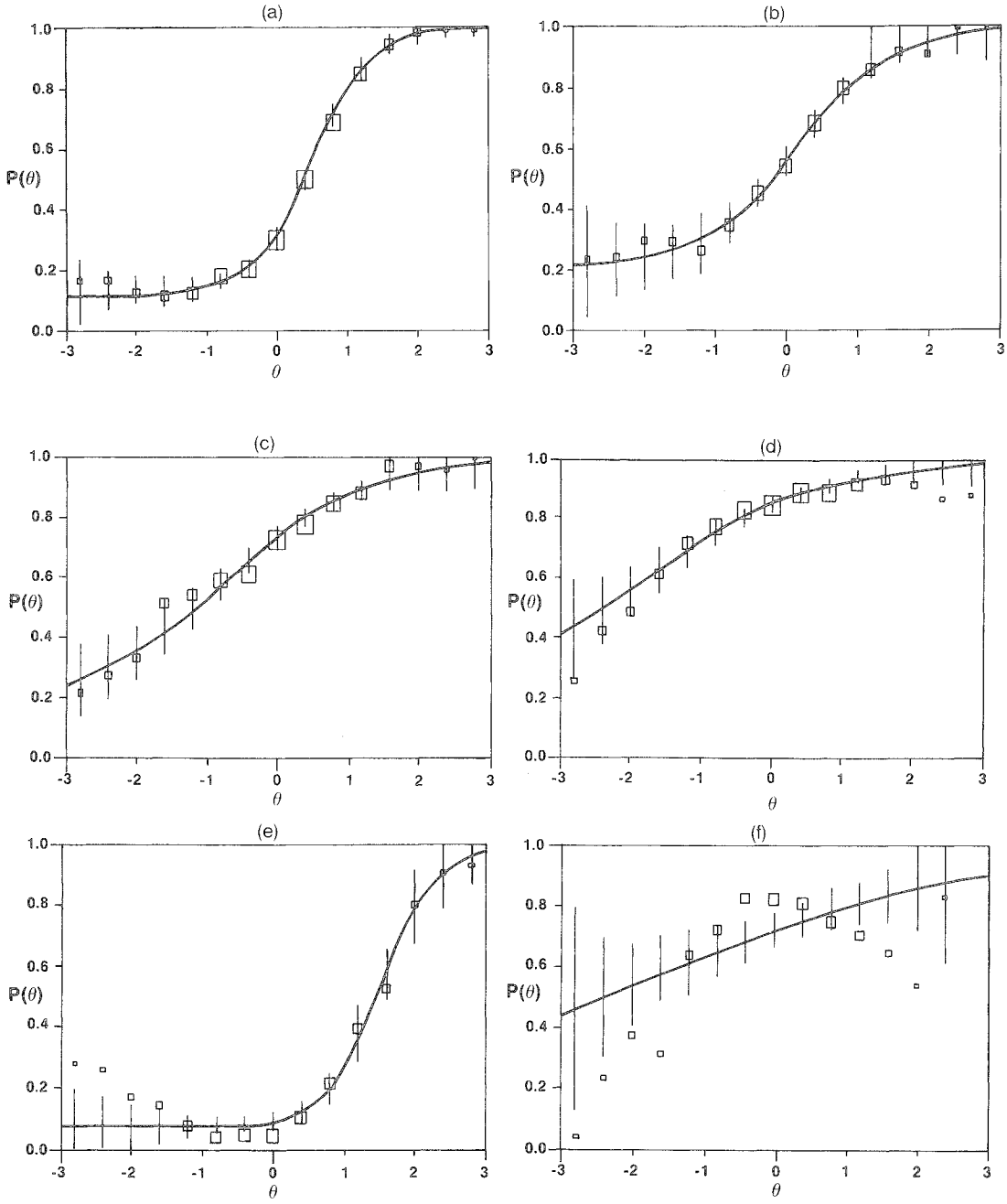

Table 1
Assessment of Model Fit

Item Type	Number of Items	Cumulative Proportion of Items With Model Fit Score Less Than or Equal to: 0	1	2	3	4	5	6	7	8
All Verbal	395	.63	.87	.96	.99	.99+	1.00			
Analogies	90	.62	.84	.93	.98	1.00				
Antonyms	102	.67	.91	.97	.99	.99	1.00			
Sentence Completion	81	.56	.88	.95	1.00					
Reading Comprehension	122	.66	.86	.97	.99	1.00				
All Quantitative	275	.45	.69	.82	.89	.94	.96	.98	.99	1.00
Regular Mathematics	75	.45	.80	.91	.95	.96	.96	.97	.97	1.00
Data Interpretation	55	.56	.80	.90	.94	.98	.98	.98	.98	1.00
Quantitative Comparison	150	.41	.60	.75	.85	.91	.96	.99	.99	1.00
All Analytical	136	.59	.82	.95	.98	.99	.99	.99	.99	1.00
Analysis of Explanations	76	.54	.76	.93	.96	.97	.97	.97	.99	1.00
Logical Diagrams	30	.70	.97	.97	1.00					
Analytical Reasoning	30	.60	.83	.97	1.00					
All Items	806	.56	.80	.91	.96	.98	.99	.99	.99	1.00

Table 2
Comparison of Model Fit for Three Major Item Classifications

Item Classification	Model Fit Score 0-1	2+	Total
Verbal	345	50	395
Quantitative	190	85	275
Analytical	112	24	136
Total	647	159	806

$\chi^2 = 34.55$ with 2 df, $p \leq .0001$

One feature of quantitative comparison items is that they all share the same response options and instructions:

> Directions: Each question in this part consists of two quantities, one in Column A and one in Column B. You are to compare the two quantities and on the answer sheet blacken space:
> A if the quantity in Column A is the greater;
> B if the quantity in Column B is the greater;
> C if the two quantities are equal;
> D if the relationship cannot be determined from the information given.

This might lead to multidimensionality due to the particular correct response of the item. To investigate this, a chi-square test of independence between the keyed response and model fit score (collapsed into two categories) was performed. Results are presented in Table 4. There is no evidence for any response option factors.

Alternatively, it could be argued that another type of multidimensionality caused the model fit problem. Perhaps quantitative comparison items themselves are unidimensional, but are tapping a different dimension from the rest of the quantitative items. Factor analytic results do not indicate this is the case (see Kingston & Dorans, 1982, for a

Table 3
Comparison of Model Fit For Various Verbal, Quantitative, and Analytical Item Types

Item Classification	Model Fit Score 0-1	2+	Total	χ^2	df	p
Verbal Item Types				2.33	3	.5267
Analogies	76	14	90			
Antonyms	93	9	102			
Sentence Completion	71	10	81			
Reading Comprehension	105	17	122			
Total	345	50	395			
Quantitative Item Types				12.77	2	.0017
Regular Mathematics	60	15	75			
Data Interpretation	40	10	50			
Quantitative Comparison	90	60	150			
Total	190	85	275			
Analytical Item Types				6.16	2	.0461
Analysis of Explanations	58	18	76			
Logical Diagrams	29	1	30			
Analytical Reasoning	25	5	30			
Total	112	24	136			

Table 4
Comparison of Model Fit for Different Keyed Responses of Quantitative Comparisons Items

Keyed Response	Model Fit Score 0-1	2+	Total
A	23	15	38
B	21	19	40
C	27	12	39
D	19	60	150

$X^2 = 2.41$ with 3 df, $p \leq .4823$

review of past GRE factor analytic studies), but existing studies used linear models and IRT is based on a nonlinear model. A separate quantitative comparison factor could not be ruled out.

To further investigate this, the quantitative comparison items for one form were calibrated separately. Item-ability regressions for items in this calibration could not be affected by multidimensionality inherent across the three quantitative item types. Table 5 compares the model fit for the 30 quantitative comparison items calibrated with the entire quantitative section against that for the items calibrated as an homogeneous subset.

Since different calibrations of identical items are represented in the two rows of Table 5, a test of independence was not performed. Nonetheless, it seems obvious that any multidimensionality occurs within the item type and not across the three quantitative item types.

Further examination of the items and their directions led to the hypothesis of another type of dimensionality problem. Due to a problem-solving response set, some examinees who did not know the answer to a quantitative comparison item might be more likely to answer D, "the relationship cannot be determined from the information given," than others of equal quantitative ability, in which case the poor model fit of these items might be explained. This problem-solving response set would contribute to a lack of model fit regardless of the keyed response. If the correct answer were A, B, or C, some examinees with a given ability would be less likely to select the correct answer than others because of their propensity for response D. If D were the correct answer, these same examinees would be more likely to select the correct answer than the model predicted. There is no way to test this hypothesis with available data.

Table 3 indicates that the three-parameter logistic model fit analysis of explanations items less well than the other analytical item types. Like quantitative comparison items, these items all share a single response format:

> Directions: For each set of questions, a fact situation and a result are presented. Several numbered statements follow the result. Each statement is to be evaluated in relation to the fact situation and result.
>
> Consider each statement separately from the other statements. For each one, examine the following sequence of decisions, in the order A, B, C, D, E. Each decision results in selecting or eliminating a choice. The first choice that cannot be eliminated is the correct answer.

A Is the statement inconsistent with, or contradictory to, something in the fact situation, the

Table 5
Comparison of Model Fit for Homogeneous and Heterogeneous Calibrations of Quantitative Comparison Items

Calibration	Model Fit Score 0-1	2+	Total
Quantitative Comparison Only	18	12	30
All Quantitative Items	19	11	30
Total	37	23	60

result, or both together? If so, choose A.
If not,

B Does the statement present a possible adequate explanation of the result? If so, choose B.
If not,

C Does the statement have to be true if the fact situation and result are as stated?
If so, the statement is deducible from something in the fact situation, the result, or both together; choose C.
If not,

D Does the statement either support or weaken a possible explanation of the result?
If so, the statement is relevant to an explanation; choose D.

E If not, the statement is irrelevant to an explanation of the result; choose E.

Table 6 presents a test of independence between keyed response and model fit. Analysis of explanations items keyed B or E were not fit well by the model. In fact, some of the B-keyed items are not monotonically increasing; more able students frequently chose the D response. Figure 1f presents the most extreme example, found in the present study, of such an item. Factor analysis (Swinton & Powers, 1980) has provided additional evidence of keyed response-specific factors for analysis of explanations items.

Table 6
Comparison of Model Fit for Different Keyed Responses of Analysis of Explanations Items

Keyed Response	Model Fit Score 0-1	2+	Total
A	10	1	11
B	7	10	17
C	18	1	19
D	16	0	16
E	7	6	13
Total	58	18	76

$\chi^2 = 25.07$ with 4 df, $p \leq .0001$

Discussion

The analysis of item-ability regressions has provided insight into the fit of the three-parameter logistic model to GRE data. The question remains: how likely is this technique to clarify model fit for other test data? In this research, each of more than 800 items were administered to at least 3,000 examinees. Most researchers will not have the luxury of these sample sizes.

Since this is a heuristic method that has not been widely used, it is difficult to know the minimum number of examinees and items required for useful analysis. Certainly, the number of examinees could be reduced to whatever size sample a researcher thought sufficient for accurate parameter estimation. For the three-parameter logistic model with joint maximum likelihood estimation (e.g., LOGIST) and reasonable constraints on the estimation of the lower asymptotes, sample sizes of about 1,000 should be more than adequate. With marginal maximum likelihood estimation (e.g., BILOG), considerably smaller sample sizes should be acceptable.

The power of this analytical technique depends on the number of items and the degree of misfit. Relative misfit was clear within the 76 analysis of explanations items, but not within the 395 verbal items. Although it is suggested that other researchers using this technique try to use samples of at least 100 items, even an analysis of about 50 items should indicate misfit, if the degree of misfit is sufficiently large.

It should also be noted that the identification of misfitting subgroups of items from a larger set of items can be confounded with the proportion of items in that subgroup. For example, if a test administered to high school students consisted of 90 vocabulary items and 10 geometry items, the vocabulary items would dominate the definition of the latent trait. The geometry items would probably appear not to be fit by a unidimensional latent trait model, even though, if they were calibrated separately, they would be fit. In fact, it is not the geometry items that are at fault, but the researcher who believes that vocabulary and geometry measure the same latent trait. Nonetheless, if the test had consisted of 10 vocabulary and 90 geometry

items, the analysis of item-ability regressions would produce a different result. Using approximately equal numbers of items within each subgroup should alleviate this confounding. Having examined a number of factor analyses in addition to the data reported herein, it is believed that no such confounding affects the results of this study.

Conclusion

The analysis of item-ability regressions appears to be a useful exploratory tool for the investigation of factors influencing item-response behavior and the fit of item response models to observed data. In this research, the three-parameter logistic model seems to fit all of the verbal item types and two of the analytical item types, logical diagrams and analytical reasoning, better than the three quantitative item types and the analysis of explanations items. Of the latter four item types, regular mathematics and data interpretation items seem to be fit almost as well as some of the "good fitting" item types. Analysis of explanations items keyed other than B or E were fit by the model quite well (less than 5% of the items keyed A, C, or D had a model fit score of 2 or greater), but those keyed B or E had the highest proportion of model fit scores of 2 or greater of any of the item classifications considered (53%). Quantitative comparison items were difficult to fit the three-parameter logistic model.

References

Educational Testing Service. (1985). *Graduate Record Examinations information bulletin*. Princeton NJ: Educational Testing Service.

Hambleton, R. (1980). Latent ability scales: Interpretations and uses. In S. Mayo (Ed.), *New directions for testing and measurement: Interpreting test performance* (No. 6, pp. 73–97). San Francisco: Jossey-Bass.

Hambleton, R., & Swaminathan, H. (1984). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79–82.

Kingston, N., & Dorans, N.J. (1982). *The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test* (GRE Board Professional Report 79–12P). Princeton NJ: Educational Testing Service.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika, 49,* 425–435.

Swinton, S., & Powers, D. (1980). *A factor analytic study of the restructured aptitude test* (GRE Board Professional Report 77–6). Princeton NJ: Educational Testing Service.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 45–56). Vancouver, BC: Educational Research Institute of British Columbia.

Yen, W. M. (1981). Using simulation results to choose a 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.

Acknowledgments

Financial support for this research was provided jointly by the Graduate Record Examination Board and Educational Testing Service. The opinions expressed herein, however, are those of the authors and are not necessarily endorsed by either organization. The review and advice of our professional colleagues at ETS and two anonymous reviewers is gratefully acknowledged, as is the programming of Marilyn Wingersky and Christopher Constantini.

Authors' Addresses

Send requests for reprints (first author) or further information to Neal M. Kingston, Educational Testing Service, 20-P, Princeton NJ 08541, U.S.A., or Neil Dorans, Educational Testing Service, 32-E, Princeton NJ 08541, U.S.A.