# Item Location Effects and Their Implications for IRT Equating and Adaptive Testing

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

A context effect occurs when examinees' item responding behavior is affected by the location of an item within a test. Recent advances in testing practice, most notably adaptive testing and certain innovative equating schemes, require items to be more invariant across intended usages than earlier methods. In this paper, location effects are identified as a form of multidimensionality, and examples of testing situations where location effects are important are described. Then, the susceptibility to item location effects of 10 item types from the Graduate Record Examination General Test is investigated by comparing the item difficulty parameters of sets of items across intended usages. Results are replicated using a second form of the test. Two of the 10 item types, analysis of explanations and logical diagrams, are clearly affected by item location in the population tested. One common item type, reading comprehension, appears to be affected somewhat by item context in this population. It is strongly advised that these item types not be used in an adaptive testing program without first assessing their susceptibility to location effects within the population (and subpopulations) of interest.

The three-parameter logistic item response model (Lord, 1980) assumes that the set of items under study is unidimensional; that is, that the probability of successful response by examinees to a set of items can be mathematically modeled by using only one ability parameter, and that performance on an item can be adequately described by a logistic function with no more than three item parameters. Considerable research has been done on the use of unidimensional item response theory (IRT) models with multidimensional tests (Bejar, 1980; Lord, 1980; Reckase, 1979; Traub & Wolfe, 1981). Most of this research has involved differences among nominal item types or content areas within an achievement domain. Kingston and Dorans (1982a) showed that, although certain item types used in the Graduate Record Examination (GRE) General Test violate the unidimensionality assumptions to some extent, for the most part IRT-based methods seemed to be robust to this violation.

Whitley and Dawis (1976) and Yen (1980) have provided evidence that item context or position within a test might be another source of violation of the unidimensionality assumption. This might present problems for certain IRT applications, most notably IRT true-score equating based on precalibration, and computerized adaptive testing.

Precalibration is a variant of IRT true-score equating that uses a data collection design that allows the calibration of item statistics before the test form is operationally administered. The appropriateness of IRT equating based on precalibration requires that changes in position of items in a test between the preoperational (calibration) and operational administrations of the test have no effect on item parameter estimates. The precalibration equating design will be further described later in this paper and will be used to provide an example

*APPLIED PSYCHOLOGICAL MEASUREMENT*
*Vol. 8, No. 2, Spring 1984, pp. 147–154*

*0146-6216/84/020147-08$1.65*

of the magnitude of error attributable to items whose parameters are affected by position.

In computerized adaptive testing (e.g., Weiss, 1982), different examinees are administered different sets of items, and the choice of the item to be administered next depends upon the examinee's responses to items already administered. Generally, a more difficult item follows a correct answer, and an incorrect answer leads to an easier item. Ideally, an adaptive test matches item difficulty with examinee ability level. Computerized adaptive testing holds much appeal because it promises (1) more efficient measurement with fewer items than conventional tests and (2) enhanced opportunities for the introduction of novel item types and the measurement of attributes that conventional paper and pencil testing cannot assess. For computerized adaptive testing to be practical, however, pools of items that are unaffected by the order of their administration are required. The assumption of no location effect needs to be tested.

## Method

### Test Forms

Two operational forms, A and B, of the GRE General Test (previously known as the GRE Aptitude Test) administered in June, 1980 were used in this study. Forms A and B were somewhat different, though both were composed of four separately timed operational sections. The item types, timing, and number of items for these sections are shown in Table 1. Examples of the various GRE General Test item types can be found in Conrad, Trismen, and Miller (1977).

Form A was administered with each examinee receiving one of six different versions of Section V. Each of these six subforms contained approximately one-half of the items from the operational verbal, quantitative, or analytical section of Form B and was administered to approximately 2,500 examinees. Form B was also administered with six different forms of Section V at the same administration at which Form A was administered. Each of these six subforms contained about one-half of the items from the operational verbal, quantitative, or analytical section of Form A and was administered to approximately 1,500 examinees. Thus, each operational item from Form A appeared in one of the six forms of Section V of Form B and each operational item from Form B appeared in one of the six forms of Section V of Form A. This commonality of items was used to study item location effects.

### Item Calibration

These 12 subforms were administered in a spiralled fashion at the June, 1980 administration of the GRE General Test. Spiralling is a term used to describe a test administration practice in which test books are packaged alternating Form A with Form B, such that a selected proportion of the examinees at any testing center take Form A while the rest take Form B (spiralling can also be used with more than two forms). Experience has shown that for the GRE testing program, spiralling results in the assignment of equivalent groups of examinees to each subform.

A total of 10 different item types were administered within each form. Parameter estimates were based on either the set of all verbal items (analogies, antonyms, sentence completions, and reading comprehension), all quantitative items (quantitative comparisons, data interpretation, and regular mathematics), or all analytical items (analysis of explanations, logical diagrams, and analytical reasoning). Each item was calibrated twice, once as an operational item and once when it appeared as a nonoperational item in the fifth section of a form. Hence, a total of six calibrations were made: two verbal, two quantitative, and two analytical.

All item parameter and ability estimates were obtained with the program LOGIST (Wood, Wingersky, & Lord, 1978).

## Results

### Comparisons of Item Difficulty Parameter Estimates

The balance of this paper presents results that demonstrate that item location effects are item-type

Table 1
Description of Test Forms

| Section | Item Type | Timing in Minutes | Number of Items Form A | Number of Items Form B |
|---|---|---|---|---|
| I | Verbal | 50 | 80 | 75 |
| | Analogies | | 18 | 18 |
| | Antonyms | | 20 | 22 |
| | Sentence completions | | 17 | 13 |
| | Reading comprehension | | 25 | 22 |
| II | Quantitative | 50 | 55 | 55 |
| | Quantitative comparisons | | 30 | 30 |
| | Data interpretation | | 10 | 10 |
| | Regular mathematics | | 15 | 15 |
| III | Analytical | 25 | 40 | 36 |
| | Analysis of explanations | | 40 | 36 |
| IV | Analytical | 25 | 30 | 30 |
| | Logical diagrams | | 15 | 15 |
| | Analytical reasoning | | 15 | 15 |
| V | Nonoperational section | 25 | variable | |
| | variable item type | | variable | |

specific. In particular, the focus is on the sensitivity of item difficulty parameters to item location. (Previous research also considered the location effect of items in terms of the *a* and *c* parameters. Those results were ambiguous; see Kingston & Dorans, 1982b.) Later, the impact of item location on IRT precalibration equating results is shown.

*Verbal items.* Table 2 and Figure 1 summarize the effects of item position on IRT difficulty parameter estimates for the verbal items of Forms A and B. The table contains means and standard deviations of item difficulty parameter estimates for the four types constituting the verbal section (analogies, antonyms, sentence completions, and reading comprehension) as well as mean differences (operational mean minus Section V mean), standard deviations of the differences, and their associated dependent-sample *t*-statistics. For all *t*-tests, the degrees of freedom (*df*) are one less than the number of items of each type.

Focusing on mean differences enables the detection of consistent biases that might be induced by item location. Since item parameter estimates are fallible, the standard error of estimate for the *b* parameters provided by LOGIST were used to obtain a standardized difference for each item. This standardized difference was obtained by dividing the observed difference in the *b* parameter estimate by the standard error of the difference. The standard error of the difference was obtained by taking the square root of the sum of the two squared standard errors of estimate for *b* provided by the two LOGIST runs for each item. If the standardized difference is positive, the item is more difficult in its operational location. If the difference is negative, the item is easier in its operational location. Since the preoperational location of an item was always in Section V, the last section of the test, positive differences will be referred to as practice effects, and negative differences will be referred to as fatigue effects.

The mean standardized difference for a given item type provides an index of the direction of that item type's susceptibility to item location effects.

Table 2
IRT Item Difficulty Parameter (b) Estimates for Verbal, Quantitative, and Analytical Items

| Form and Item Type | Operational Mean | Operational S.D. | Section V Mean | Section V S.D. | b(1) - b(2) Mean | b(1) - b(2) S.D. | b(1) - b(2) t |
|---|---|---|---|---|---|---|---|
| Verbal items: Form A | | | | | | | |
| Analogies | .66 | 1.31 | .58 | 1.19 | .07 | .18 | 1.79 |
| Antonyms | .72 | 1.25 | .68 | 1.17 | .05 | .30 | .68 |
| Sentence Completion | -.01 | .88 | .07 | .88 | -.08 | .20 | -1.61 |
| Reading Comprehension | -.04 | .78 | .10 | .73 | -.14 | .26 | -2.72* |
| Verbal items: Form B | | | | | | | |
| Analogies | .45 | 1.30 | .48 | 1.18 | .02 | .24 | .51 |
| Antonyms | .11 | 1.10 | .23 | 1.07 | .12 | .25 | 2.23* |
| Sentence Completion | -.04 | 1.17 | -.00 | 1.13 | .04 | .15 | .90 |
| Reading Comprehension | .27 | 1.05 | .40 | 1.13 | -.13 | .35 | -1.76 |
| Quantitative items: Form A | | | | | | | |
| Regular Math | .69 | 1.39 | .73 | 1.35 | -.04 | .11 | -1.11 |
| Data Interpretation | -1.07 | 1.04 | -.88 | 1.00 | -.20 | .15 | -4.31** |
| Quantative Comparisons | .03 | 1.48 | -.04 | 1.59 | .07 | .38 | 1.02 |
| Quantitative items: Form B | | | | | | | |
| Regular Math | .24 | .92 | .34 | .99 | .09 | .24 | 1.46 |
| Data Interpretation | -.70 | 1.97 | -.31 | 1.92 | .39 | 1.11 | 1.10 |
| Quantative Comparisons | -.36 | 1.38 | -.28 | 1.24 | .08 | .33 | 1.34 |
| Analytical items: Form A | | | | | | | |
| Analysis of Explanations | .15 | 1.06 | .91 | 1.11 | .25 | .24 | 6.58** |
| Logical Diagrams | -.27 | .90 | -.50 | .83 | .22 | .26 | 3.30** |
| Analytical Reasoning | -.35 | 1.20 | -.33 | 1.21 | -.02 | .23 | -.27 |
| Analytical items: Form B | | | | | | | |
| Analysis of Explanations | -.21 | .86 | .09 | .82 | .30 | .18 | 9.88** |
| Logical Diagrams | -.08 | .59 | .07 | .64 | .15 | .13 | 4.33** |
| Analytical Reasoning | -.16 | 1.20 | -.10 | 1.16 | .06 | .19 | 1.14 |

* $p < .05$
** $p < .01$

These mean standardized differences are plotted in Figure 1 for each of the different verbal item types. There are two adjacent rectangles, one for Form A and one for Form B. The height and direction of the bars indicate the degree and direction of the item type's susceptibility to location effects.

Table 2 and Figure 1 indicate a slight practice effect for the antonym items and a slight fatigue effect for the reading comprehension items. These effects cancel out for the verbal test overall. The practical implications of this cancellation will be seen when equating results for the verbal measure are examined.

*Quantitative items.* Table 2 and Figure 2 summarize the effects of item position on IRT difficulty parameter estimates for the quantitative items of Forms A and B. Figure 2 is similar in format to Figure 1. Note in Figure 2 that the amount of practice effect for the quantitative comparison item type in Form B is so small that it is not noticeable.

The most striking feature in this table and figure is the results for the data interpretation items. Table 2 shows that this item type exhibits a profound fatigue effect in Form A and a substantial practice effect in Form B. This former effect, however, disappears in Figure 2 where the data interpretation items appear to have a negligible practice effect in Form A and a slight practice effect in Form B.

Further investigation revealed that the inconsistency in the Form B results was due to a single data interpretation item that had a *b*-value of 1.66 (with a standard error of .16) when administered

**Figure 1**
Mean Standardized Differences Between IRT Difficulty Parameter Estimates Obtained in Operation and Preoperational Locations for Verbal Items

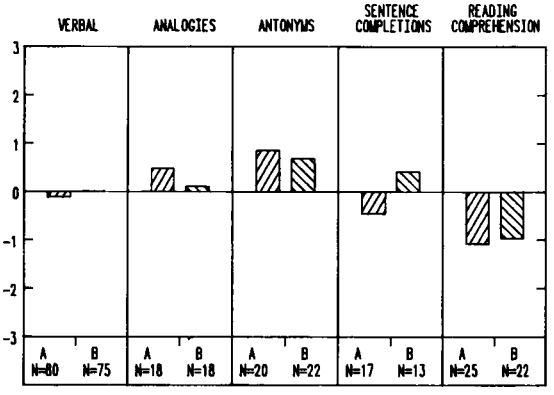
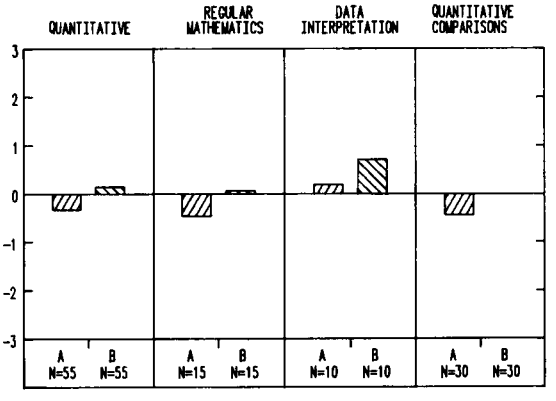



**Figure 2**
Mean Standardized Differences Between IRT Difficulty Parameter Estimates Obtained in Operational and Preoperational Locations for Quantitative Items




operationally and a *b*-value of $-1.75$ (with a standard error of 6.45) when administered nonoperationally, a difference of 3.41. This extreme difference between *b*-values was highly unusual; the median absolute difference for all quantitative items was approximately .15. The discrepancy between *b*-values for the data interpretation items in Forms A and B was much less extreme when this questionable item was not included in the analysis. The mean difference of .39 between *b*-values for the data interpretation items in Table 3 would be a mean difference of only .05.

The standardized difference for the questionable item was only .53. Unfortunately, when this item's parameters are used in equating, it is the unstandardized difference that affects the results.

*Analytical items.* Table 2 and Figure 3 summarize the effects of item position on difficulty parameter estimates for the analytical items of Forms A and B. Note in Figure 3 that the amount of practice effect for the analytical reasoning item type in Form A is so small that it is not noticeable.

In contrast to the results for verbal and quantitative items, where only reading comprehension and data interpretation items appeared to be affected, the analytical results in Table 2 and Figure 3 depict considerable susceptibility to item location. Only the analytical reasoning item type yields acceptable results. Both the analysis of explanations and logical diagrams items exhibit large consistent practice effects. Since these two item types account for about 80% of the analytical measure, the overall effect for all analytical items is quite substantial.

## Equating Comparisons

Equating refers to a statistical process used to transform raw test scores (number correct or formula scores) to a common scale so that scores from two tests, which may differ in their difficulty, can be made more comparable. For this study, Lord's (1980) IRT true-formula-score equating procedure was used.

*IRT true-score equating.* Estimated true formula scores $\hat{T}_A$ and $\hat{T}_B$ on two tests measuring the same ability are related by the equations

$$\hat{T}_A = \sum_{i=1}^{n_A} \hat{P}_i(\theta) - \left[\sum_{i=1}^{n_A} \hat{Q}_i(\theta)/(K_i - 1)\right], \quad (1)$$

and

$$\hat{T}_B = \sum_{j=1}^{n_B} \hat{P}_j(\theta) - \left[\sum_{j=1}^{n_B} \hat{Q}_j(\theta)/(K_j - 1)\right], \quad (2)$$

Table 3
Means and Standard Deviations of Converted Scores Based on Three Different Equating Methods
Form B

| Equating Method | Verbal | | Quantitative | | Analytical | |
|---|---|---|---|---|---|---|
| | Mean | S.D. | Mean | S.D. | Mean | S.D. |
| Spiralling, Linear | 473.29 | 123.35 | 498.63 | 130.31 | 497.36 | 128.91 |
| Spiralling, IRT | 473.27 | 125.14 | 499.75 | 123.38 | 498.12 | 125.44 |
| Precalibration, IRT | 473.81 | 125.47 | 494.81 | 123.65 | 470.29 | 123.25 |

where $K_i$ and $K_j$ are the number of choices per item,

$n_A$ and $n_B$ are the number of items in Forms A and B,

$\hat{P}_i(\theta)$ and $\hat{P}_j(\theta)$ are the item characteristic curves for items $i$ and $j$, and

$\hat{Q}_i(\theta) = 1 - \hat{P}_i(\theta)$.

Using Equations 1 and 2, it is possible to find an estimated true formula score $\hat{T}_B$ corresponding to an estimated true formula score $\hat{T}_A$ for any given ability estimate $\hat{\theta}$. $\hat{T}_B$s and $\hat{T}_A$s that correspond to a given $\theta$ are said to be equated.

IRT true-formula-score equating was used to equate Form B to Form A, which was already on the standard GRE reporting scale of 200 to 900 (since October 1981, the standard GRE reporting scale has been changed to 200 to 800). Form B was equated to Form A twice, once based on Form B parameters obtained when the items appeared in their operational locations, and once when the items appeared in the Section V nonoperational locations. The second type of equating mimics the precalibration equating design that requires resistance to item location effects.

*Verbal measure.* Figure 4 shows the difference between the two Form B equatings for the verbal measure. At each formula score, the converted score based on precalibration equating was subtracted from the converted score based on equating using the operational item parameters. A horizontal line was drawn at zero to provide a no-effect reference. Note that the plot reflects the moderate fatigue effect observed for reading comprehension items in Table 2 and Figure 1. The dip at the upper end of the scale is consistent with the mean difference for

**Figure 3**
Mean Standardized Differences Between IRT Difficulty Parameter Estimates Obtained in Operational and Preoperational Locations for Analytical Items

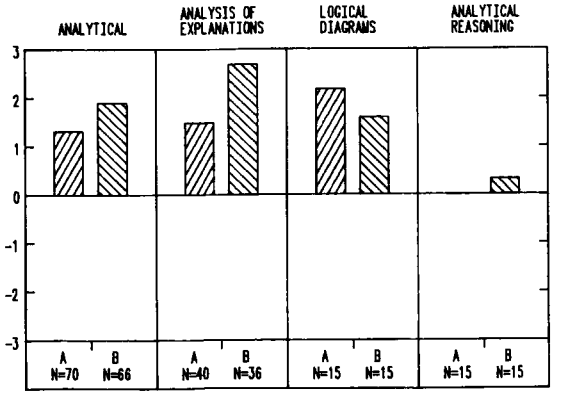
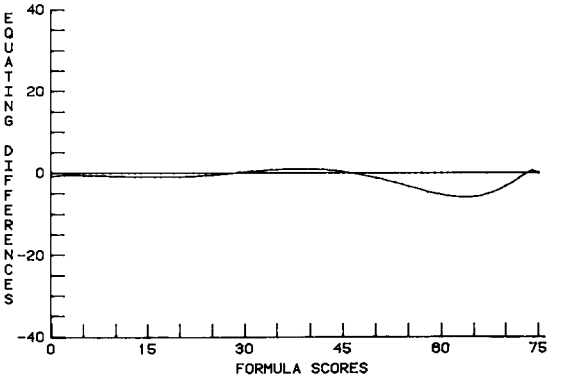



**Figure 4**
Effect of Item Location on Verbal Score Equating



reading comprehension items in Table 2 (.40 in their preoperational location vs. .27 in their operational location). Fortunately, this effect never produces an absolute scaled score difference greater than five, which is one reasonable cutoff for concern given GRE rounding practices. (The GRE General Test reports scaled scores rounded to the nearest 10 points.)

The results in the leftmost section of Table 3 verify that the fatigue effect for reading comprehension items probably would not create any overall scale drift. This table contains the means and standard deviations of converted scores for a population of GRE examinees who took Form B in June, 1980 produced by both types of IRT true-score equating and the operational linear equating. Note that the three equatings of Form B produce comparable summary statistics for this population of examinees. These results do not appear to rule out precalibration for the verbal measure.

*Quantitative measure.* Figure 5 and the middle section of Table 3 contain equating results for the quantitative scale of Form B. In contrast to the verbal results, the effect of the data interpretation items on the quantitative results is noticeable. The mean difference of five points on the quantitative measure in Table 3 reflects the consistent bias evident in Figure 5. The precalibration equating produced a lower mean converted score for this population than would have been obtained from equating based on parameter estimates for items in their operational location.

*Analytical measure.* Figure 6 and the rightmost portion of Table 3 contain the results for the Form B equatings of the analytical measure. The sensitivity of the analytical item types to item position had a profound effect on the equating results. The difference of nearly 30 converted score points in means between the precalibration-based IRT equating and the other two equatings clearly demonstrates that implementation of a precalibration data collection design for these item types would be unwise. Likewise, these item types would be inappropriate for use in an adaptive testing context, using current item response models, because of their susceptibility to item location effects.

## Discussion

The results demonstrate that susceptibility to location effects is item-type specific. Clearly, the analysis of explanations items were susceptible to item location effects. Examination of the format of this item type readily leads to a hypothesis as to why this is so. It is a complex item type with an extensive, complicated set of directions. Most examinees probably have to become very familiar with the instructions before they do well on this item type. Once the directions are understood, it

**Figure 5**
Effect of Item Location
on Quantitative Score Equating

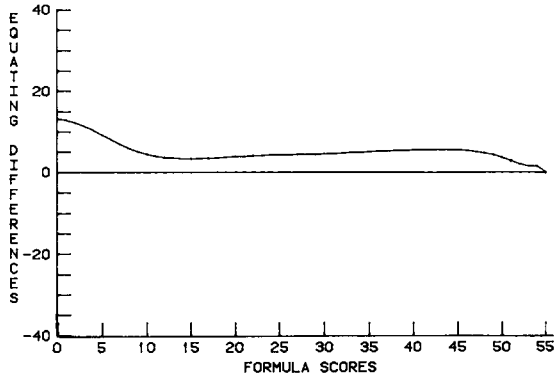
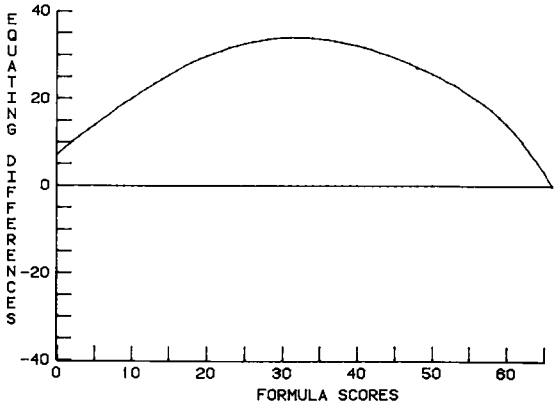



**Figure 6**
Effect of Item Location
on Analytical Score Equating



is relatively easy to deal with the items. Hence, the difficulty of an analysis of explanations item will depend on how many items of that type precede it. This is not the type of item to use in an adaptive testing mode. In fact, on the basis of this research and other studies of GRE General Test item types (Kingston & Dorans, 1982b; Swinton, Wild, & Wallmark, 1983), the GRE analytical section has been revised. Both analyses of explanations and logical diagrams items have been dropped from the test.

Elimination of item types that are not sufficiently resistant to location effects is one solution. An alternative approach might be to give sufficient practice items to all examinees. Clearly, research such as this should be conducted before computerized adaptive testing or precalibration-based equating data collection designs are operationally implemented. Also, psychometricians should attend to the modeling problem. Discarding item types that elicit data that cannot be fit well with existing models is a temporary solution. The development of more general models that incorporate parameters such as item familiarity or item position should proceed.

## References

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement, 17,* 283–296.

Conrad, L., Trismen, D., & Miller, R. (Eds.). (1977). *Graduate Record Examinations Technical Manual.* Princeton NJ: Educational Testing Service.

Kingston, N. M., & Dorans, N. J. (1982a). *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.

Kingston, N. M., & Dorans, N. J. (1982b). *The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory.* (GRE Board Professional Report 79-12bP). Princeton NJ: Educational Testing Service.

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

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

Swinton, S., Wild, C. L., & Wallmark, M. (1983). *Investigation of practice effects on item types in the Graduate Record Examinations Aptitude Test.* (GRE Board Professional Report 80-1cP). Princeton NJ: Educational Testing Service.

Traub, R. E., & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. In D. C. Berliner (Ed.), *1981 Review of Research in Education.* Washington DC: American Educational Research Association.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6,* 473–492.

Whitely, S. E., & Dawis, R. V. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement, 36,* 329–337.

Wood, R. L., Wingersky, M., & Lord, F. (1978). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters.* (ETS Research Memorandum 76-6 [modified 1/78]). Princeton NJ: Educational Testing Service.

Yen, W. M. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17,* 297–311.

## Acknowledgments

*Financial support for this research was provided jointly by the Graduate Record Examinations 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 is gratefully acknowledged, as is the programming assistance of Louann Benton and Christopher Constantini.*

## Authors' Addresses

Send requests for reprints (first author) or further information to Neal M. Kingston, School and Higher Education, Statistical Analysis, 20-P, Educational Testing Service, Princeton NJ 08541, U.S.A., or Neil J. Dorans, College Board Statistical Analysis, Educational Testing Service, Princeton NJ 08541, U.S.A.