DOCUMENT RESUME

ED 390 902 TM 024 215

AUTHOR Hicks, Marilyn M.
TITLE Analyzing the Option Effects of Difficult TOEFL Items with Low Biserials: Methods Developed for Use by Test Assemblers.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-88-15
PUB DATE Mar 88
NOTE 41p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Difficulty Level; Estimation (Mathematics); *Identification; Item Analysis; *Item Response Theory; *Language Proficiency; Language Tests; Test Construction; *Test Items
IDENTIFIERS *Biserial Correlation; *Test of English as a Foreign Language

ABSTRACT

Several exploratory analyses of the fifths data generated by Test of English as a Foreign Language (TOEFL) item analyses were developed in order to evaluate the effects of options on the discriminability of difficult items and to identify difficult items with low, unreliable biserials that had been rejected by test developers, but for which acceptable a-parameters are probably estimable. Intended for use by test assemblers subsequent to an item analysis, the methods were mainly graphical, but included the evaluation of a distance measure and other simple statistics. Localized option effects occur that can impair item discrimination as well as the fit of the item response theory model. The negative impact of these effects on model fit was illustrated, and methods were suggested for analyzing them. An index was also developed to identify very difficult items in which the r-biserial restricts the ability of test developers to construct tests with effective measurement properties at high score levels. Implications of items with nonmonotonic response patterns due to option effects were also discussed. Appendix A presents TOEFL item response functions and Appendix B describes some features of correspondence analysis. (Contains 12 figures and 5 references.) (SLD)

RR-88-15

RESEARCH REPORT

ANALYZING THE OPTION EFFECTS OF DIFFICULT TOEFL ITEMS WITH LOW BISERIALS: METHODS DEVELOPED FOR USE BY TEST ASSEMBLERS

Marilyn M. Hicks

Educational Testing Service
Princeton, New Jersey
March 1988

Analyzing the Option Effects of Difficult TOEFL Items with Low Biserials: Methods Developed for Use by Test Assemblers

Marilyn M. Hicks

Educational Testing Service

Princeton, N.J.

Abstract

Several exploratory analyses of the fifths data generated by TOEFL item analyses were developed in order to evaluate the effects of options on the discriminability of difficult items, and to identify difficult items with low, unreliable biserials which have been rejected by Test Development but for which acceptable a-parameters are probably estimable. Intended for use by test assemblers subsequent to an item analysis, the methods were mainly graphical, but also included the evaluation of a distance measure and other simple statistics.

An effective distracter has the property that examinees are attracted to it in inverse order of ability. To the extent that this ordering is violated for certain ability levels, localized option effects occur which can impair item discrimination as well as the fit of the IRT model. The negative impact of these effects on model fit was illustrated, and methods for analyzing them were suggested. If item writers could account for the factors underlying the interaction between ability level and option responses, it might be possible to modify options accordingly, thereby improving the measurement effectiveness of the item. Departing from the usual reliance on a single index, the approaches in these analyses included, among other things, an evaluation of the biplot generated from a correspondence analysis of the matrix of fifths information, and an analysis of the total option response configuration. Many examples of these analyses were provided.

A significant limitation of the r-biserial for very difficult items which restricts the ability of test assemblers to construct tests with effective measurement properties at high score levels was illustrated. The index developed in this study to identify such items is regarded as an interim strategy until a conventional measure of item discrimination which is optimal over the entire scale of difficulty is developed, a current critical need.

The implications of introducing other dimensions into the test by items with nonmonotonic response patterns due to option effects was briefly discussed. It is possible that application of the procedures developed in the study might provide a method of excercising control over the dimensionality of the measuring instrument at the practical level of item construction.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

OBJECTIVES OF THE STUDY

A current objective of TOEFL (R) Test Development is to increase the production of items at the upper levels of ability. For TOEFL, however, low r-biserials tend to accompany very difficult items. Among other things, the discriminability of a difficult multiple choice item can depend on a complex of option effects. One such effect is the rate at which options attract examinees at each level of ability, which will be shown to impact on the measurement effectiveness of very difficult items. If the associations between ability level and options are such that they impair the item's discriminating power, an obvious expedient is to uncover the nature of those relationships, and then to modify or replace the problematic options accordingly. To the extent that option effects degrade the fit of the data to the IRT model, these approaches might also provide direction for improving item fit.

Based on the foregoing, the main objective of the study was to provide methods of analyzing the relationships between options and ability levels as they affect item discriminability, with the focus on difficult items. The analyses were based on the fifths data (see Figure 1 on page 2) generated from a standard ETS item analysis and are intended for use by test assemblers on a PC subsequent to an item analysis.

In great part, the association of low r-biserials with difficult items stems from the fact that responses are random except for those associated with high ability students, resulting in a low correlation between total score and item responses (Lord and Novick, 1968, p.342). Due to the unreliability of the r-biserial in this instance, an accurate indicator of the discriminability of very difficult items often may not be elicited from standard item analyses. On the other hand, the a-parameter, the IRT discrimination index (see Appendix A, p. 31) can be reliably estimated for such items.

Although TOEFL tests are scaled using IRT parameters, they are assembled based on conventional item statistics. This is so because the tests are only partially calibrated; that is, a subset of the items have item parameters. Since the test assembler's criterion for the inclusion of an item is based on the value of the r-biserial, many usable difficult items are probably being discarded. A subsidiary but related objective of this study was to devise an index that might flag difficult items with acceptable a-parameters in spite of low, unreliable r-biserials. In essence, the study assumed the existence of two sets of difficult items with low r-biserials:

(1) Those for which the r-biserial is a reliable estimate of discriminating power, low values of which might be due to option effects.

(2) Those items for which the low biserial is unreliable, but the item is actually discriminating effectively at very high levels of ability.

Using only the fifths information generated in item analysis, an attempt was made to sort out these two general cases.

Figure 1. Example of fifths information.

RESPONSE CODE	LOW N_1	N_2	N_3	N_4	HIGH N_5
OMIT	0	0	0	0	0
A	101	160	200	217	233
B	21	3	2	1	1
C	78	38	19	11	5
D	49	48	28	20	10
E	0	0	0	0	0
TOTAL	249	249	249	249	249

METHODS OF THE STUDY

The data analyzed in this study consisted of 103 difficult items (delta $\geq$ 13) rejected for inclusion in a test by TOEFL Test Development over the last two years because of low r-biserials (<.20). Deltas are the standard measure of item difficulty used at ETS and represents a transformation of proportion correct to a scale with a mean of 13 and a standard deviation of 4. In addition, 50 items in the same range of difficulty with r-biserials ranging from .20-.39, and 53 items with r-biserials $\geq$ .40 were also analyzed for purposes of comparison. Less than half the total group of items were IRT scaled; all were four-choice items.

Three methods of analyzing fifths information in terms of the objectives of this study are described in this section. They include the analysis of option response profiles, the analysis of option response curves, and biplots from a correspondence analysis of the fifths data. Appendix A, on page 31, briefly describes the item ability regressions, and some relevant terms derived from IRT estimation which will be pertinent in some of the discussion to follow.

Profile Plots. The basic data for all of the methods developed in this study consisted of the fifths information produced by the standard ETS item analysis, an example of which is given in Figure 1 on page 2. The columns represent examinees from five levels of ability (quintiles of the score distribution) and the rows indicate options. Each cell contains the frequency of response to an option, given level of ability. If this matrix is transposed so that the rows are levels of ability, $i = 1, \ldots, 5$ and the columns are options, $j = 1, \ldots, 4$, then this 5 x 4 matrix, N, can be transformed to a matrix P such that a typical element is $p_{ij} = n_{ij}/n_{..}$, the proportion of the total group responding to an option at each ability level. In this matrix representation, each row represents a response profile across options for each level of ability. Omitted responses were not considered in this analysis.

Examples of profile plots for difficult items representing three levels of discrimination are given in the upper left of Figures 2, 3 and 4, on pages 5, 6, and 7, respectively. The numbers, 1-5, label the levels of ability from lowest to highest. It should be noted that the ordinates of these plots are not on the same scale, but this desired comparability was sacrificed for the sake of readability. Some of the features of these plots illustrate their utility in analyzing the effect of options on discriminability.

1. Figure 2, profile plot of a difficult item with a high biserial (.62).

   a. The proportions of examinees responding on the key are strictly ordered with respect to ability, i.e., in the order 1,2,3,4,5.

   b. The differences in the proportions of examinees responding correctly at each ability level are substantial. The ability levels are well separated on the key, tending to assure a high correlation between item performance and total score.

c. A single option, c, serves to draw examinees at a sufficient rate to ensure discrimination on the key, and in reverse order of ability. This might be regarded as a counter-option with content that attracts ability levels inversely relative to the key. Although this type of option is commonly known as a distracter, the term 'counter-option' stresses the optimal property of strict ordering of ability counter to that expected on the key, and serves to distinguish it from non-keyed options that attract examinees in the order expected on the key. While the other options do not have a substantial effect on the distribution of the keyed response, they too are inversely ordered with respect to ability.

2. Figure 3, profile plot of a difficult item with a low-medium biserial (.23).

a. The ability levels are not strictly ordered on the keyed response, but in the order 1 3,4,2 5. In fact, levels 1 and 3, and levels 2 and 4 are virtually indistinguishable on the correct option with obvious implications for the correlations between item response and total score.

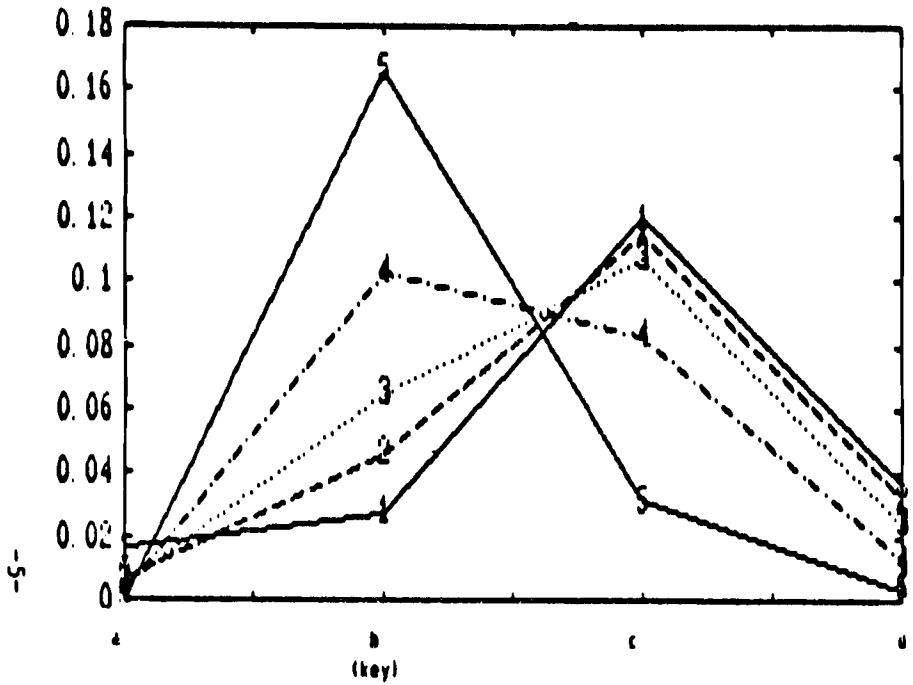
b. No effective counter-option exists. Although option d is the most attractive, it draws examinees other than those at level 5 at about the same rate. Option a is not an effective counter-option, attracting examinees in the order 5,2,1,4,3. Its relatively high attraction for levels 3 and 4 is the primary cause of the observed ordering on the key. The replacement or modification of option a based on information relative to the ability levels it attracts may increase the item's discriminability.

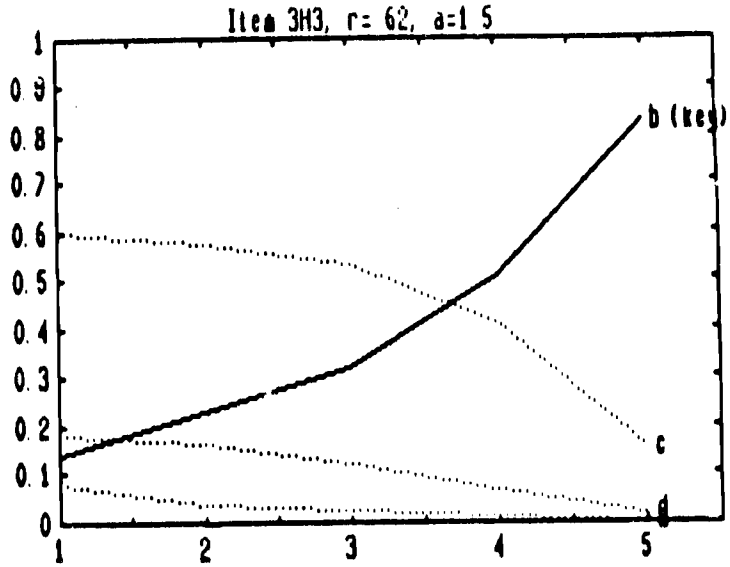
3. Figure 4, profile plot of a difficult item with a low biserial (.14).

a. On the key, all ability levels are responding at the same rate, except for the highest scorers.

b. Option b is a counter-option which attracts examinees in inverse order of ability; but also present is another option, a, which draws examinees in the expected order of ability for a keyed response. This option markedly impacts on the distribution of the correct option. Even though greater numbers of level 5 examinees select this option, they probably represent the lower scorers at this level. While standard item analysis procedures as currently implemented cannot make this important distinction, IRT parameters can (see Appendix A); the a-parameter for this item was calculated to be 1.5, the maximum for TOEFL data. This particular configuration of one relatively effective counter-option, and another option in competition with the key, has been observed to be typical of very difficult items with low biserials but high a-parameters. This item also illustrates the essentially random responses on the key for all levels except the highest scorers, which can only result in a low correlation between correct response and total score.

Figure 2. Profile plot, option response configuration and biplot for a highly discriminating difficult item, (r=.62).





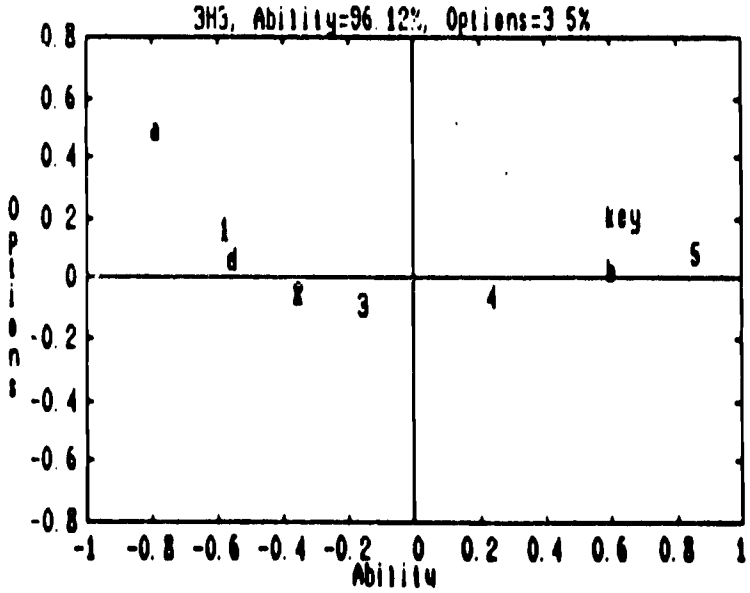
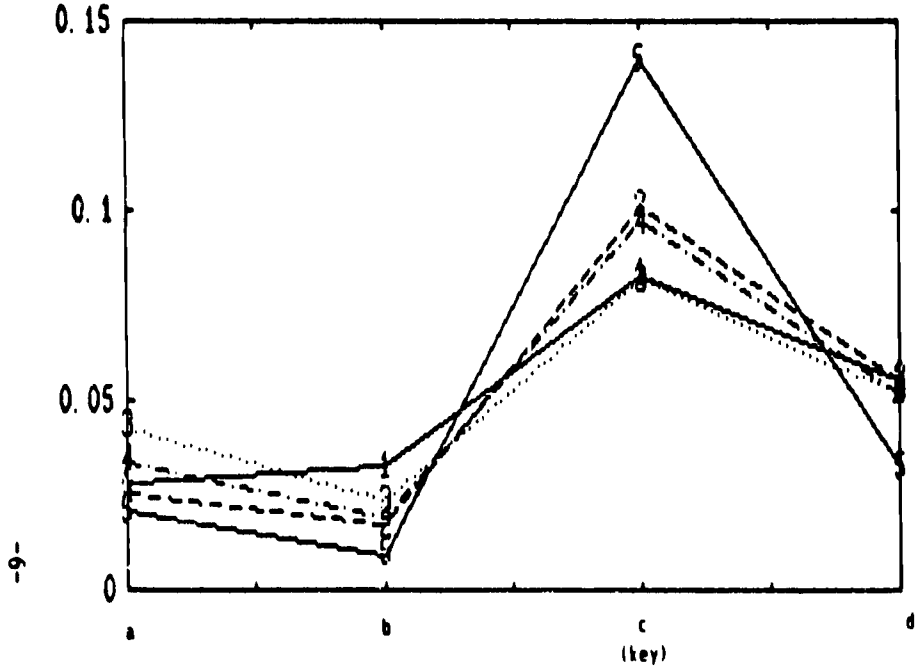
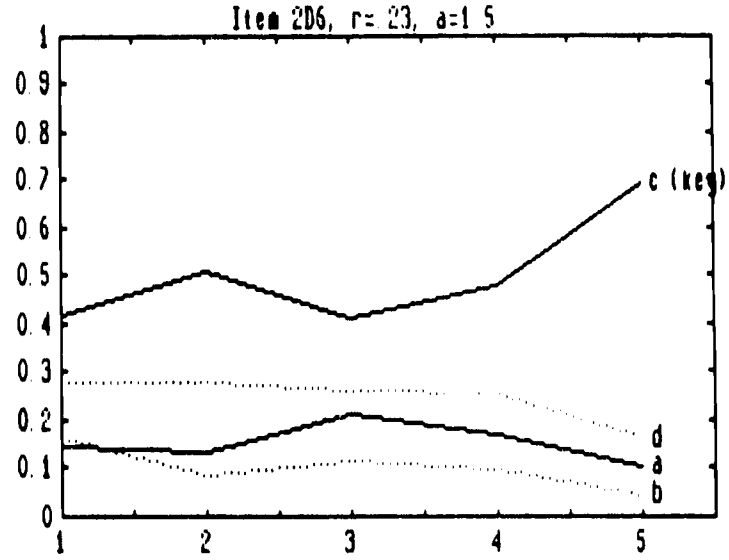
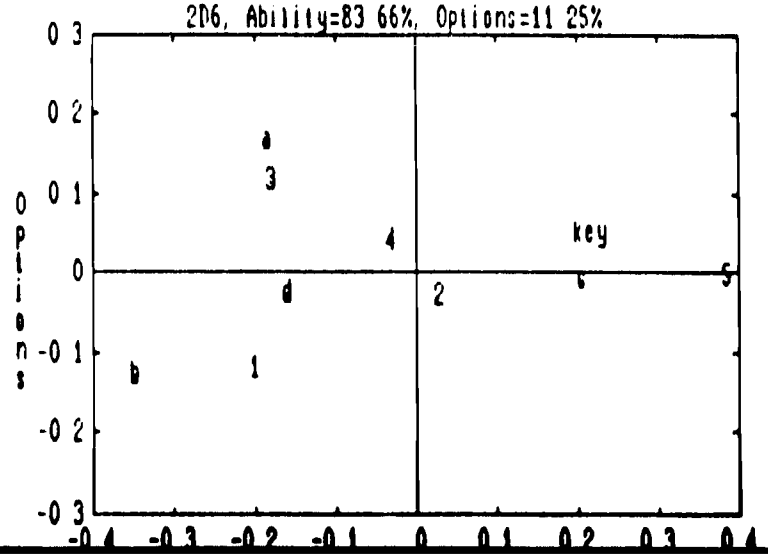


Figure 3. Profile plot, option response configuration, biplot and item ability regression for a difficult item with a moderately low r-biserial (r=.23).







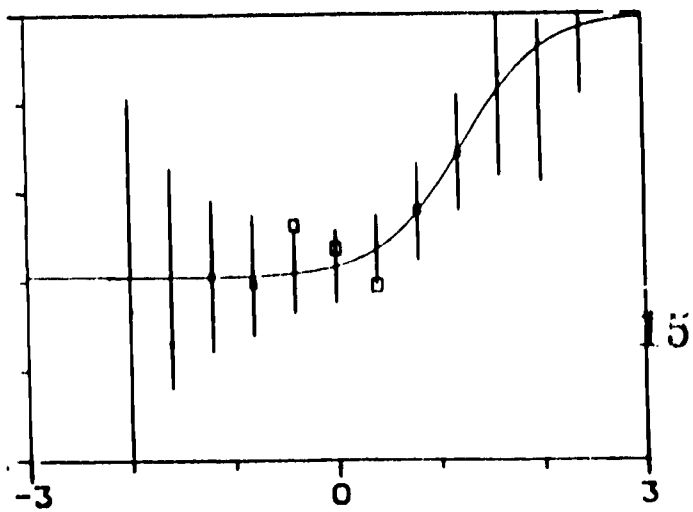
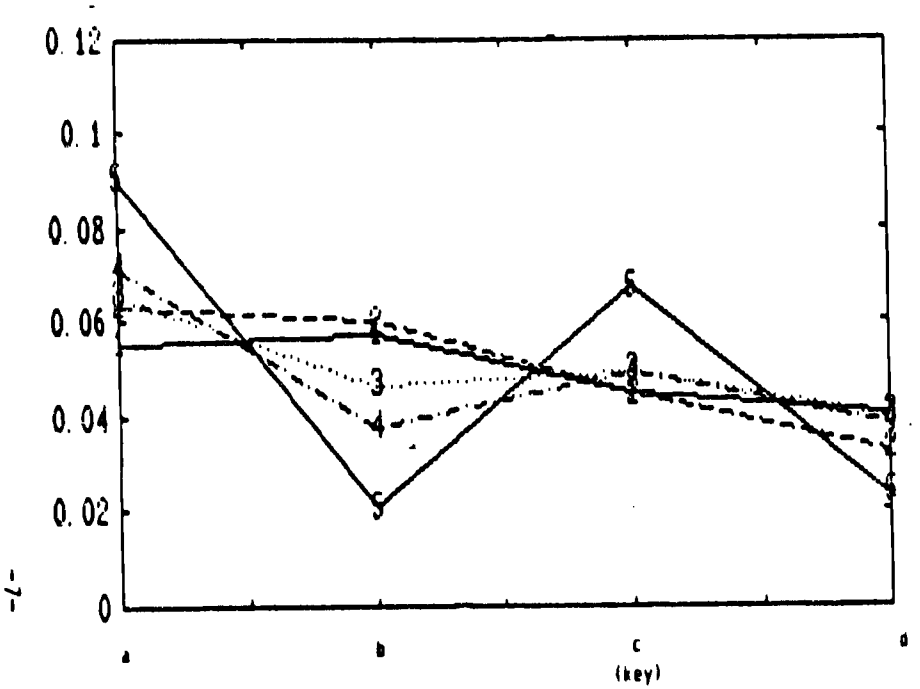
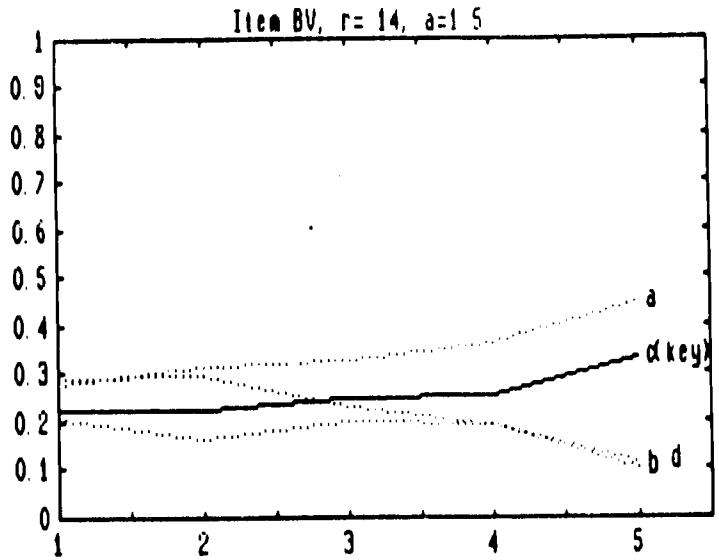
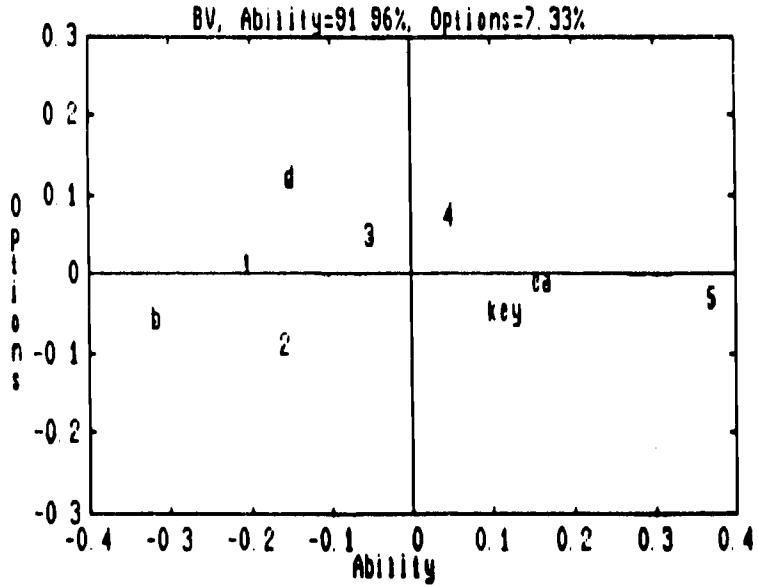


Figure 4. Profile plot, option response configuration and biplot for a difficult item with a low r-biserial (.14).







In spite of the fact that biserials ranged from .14 to .62, in each instance the a-parameter was calculated to be 1.5. The correspondences between r-biserials, a- and b-parameters for these items were:

	r	a	b
Fig. 2	.63	1.5	.58
Fig. 3	.23	1.5	1.38
Fig. 4	.14	1.5	2.16

Aside from the fact that the biserial and the a-parameter are nonlinearly related, these data illustrate one essential difference between them: the biserial is, intuitively, a more global estimate of discrimination while the a-parameter (interpreted in conjunction with the item difficulty) provides information regarding the ability levels at which the item measures most effectively.

The importance of a single effective distracter is well known, but an analysis of the profile plots can generate information about the levels of ability at which these distracters may become ineffective, which can provide direction for remediation of the options, and possibly the item's discriminability. Once an item writer identifies an option that is unduly attractive to a certain ability level, he/she may be able to determine why this is so and change the option accordingly. A new item analysis system, currently under development, is expected to provide values of the slope of the response curve for all options, but detailed information relative to the interaction of options and ability level can be derived from examination of the profile plots (or better, transformations of them). While the profile plots can be analyzed directly, two transformations of the profile matrix, to be described below, can greatly simplify this task.

Item Response Curve. The keyed option response curve (IRC) can be considered the prototype item ability regression obtained from IRT analyses. In the profile matrix, P, each p_{ik}, the proportion of examinees responding to the key, is divided by $p_{i.}$, the proportion of examinees at level i. Even though ability divisions for item analysis are gross compared to the estimates derived from IRT scaling, the resulting curves are very close approximations to the item ability regressions and can often be used to evaluate some cases of poor fit.

Item response curves based on fifths data for the three items in Figures 2-4 are given in the upper right of Figures 2-4, respectively. Each IRC can be identified by the label "key" in these plots. The IRC for the item in Figure 2 (r-biserial=.62) is monotonic increasing in contrast with that for Figure 3 (r-biserial=.23) which clearly reflects the lack of ordering of ability levels on the key as indicated in the analysis of the profile plot. Figure 3 also presents the item response function (IRF) as estimated by LOGIST, the IRT estimation program. Clearly the observed curve (defined by the small squares) does not adequately fit the theoretical curve (the solid line), and the trend

of the observed curve corresponds to the IRC; but from the analyses to be described below, it should be possible to pinpoint the option contributing most to these results.

The effects in Figure 3 demonstrate the extent to which all options are an integral part of measurement on an item. The observed data cannot be fit properly by a logistic curve because of localized option effects, options that do not draw examinees systematically with respect to ability, (not in strict inverse order of ability), with the consequence that the assumption of a monotonic relationship between ability and correct response does not hold for this item. Identification of effective counter-options becomes important in light of these considerations - all options must also work in systematic ways if the assumptions of the IRT model are to be met. A broader approach to the IRT model which recognizes these option effects has been developed by Thissen and Steinberg (1984).

The IRC in Figure 4 is typical of extremely difficult items with low r-biserials, but with satisfactory a-parameters; the curve is flat over levels 1-4 and rises only at level 5, and is nondecreasing, indicating that no levels are being unduly attracted to specific options.

Assessing the Degree of Nonmonotonicity in the IRC. If the intervals on the abscissa associated with the five ability levels of the IRC were to be considered of unit length, then $p_{ik} - p_{(i-1)k}$ is the tangent of the angle formed by the line connecting levels i and i-1, i=2, ..., 5 and the interval on the abscissa. There are four such connecting line segments in these plots; between levels 1 and 2, levels 2 and 3, levels 3 and 4, and levels 4 and 5. An evaluation of these tangents can provide information as to where the item discriminates maximally or minimally, based on the score divisions of the fifths data, but most importantly, a negative tangent can identify ability levels for which there may be an option effect.

The tangent is merely a difference in proportions between 2 adjacent groups, i and i', and the standard error of this difference is:

$$SE = [(p_i(1-p_i))/n_i + (p_{i'}(1-p_{i'}))/n_{i'}]^{1/2} \quad . \qquad (1)$$

For the IRC in Figure 2, the tangents (or equivalently in this case, the difference between adjacent proportions), expressed as a multiple of the standard error of the difference for adjacent levels are:

Levels	(1-2)	(2-3)	(3-4)	(4-5)
Tan:	.09	.09	.19	.32
Tan/SE:	2.23	1.93	3.66	6.66

Maximum discrimination is occurring between levels 4 and 5 (the value of the tangent represents 6.66 standard errors of the difference between the proportions of examinees responding correctly at levels 4 and 5); very effective discrimination is also observed between levels 3 and 4.

For the item in Figure 3:

Tan:	.09	-.10	.07	.21
Tan/SE:	1.61	1.72	1.27	3.88

Again, maximum discrimination occurs between levels 4 and 5 for this difficult item. The option effect described above is flagged by the negative tangent between levels 2 and 3.

For the item in Figure 4:

Tan:	.00	.02	.01	.08
Tan/SE:	.07	.61	.15	2.00.

The tangents reflect the flat curve over most of the ability distribution with a slight rise at level 5.

For this study, localized option effects flagged by a negative tangent were considered significant if the difference in proportions exceeded one standard error of the difference. Consequently, the negative tangent between levels 2 and 3 in Figure 3 would signal an item that should be examined for a non-keyed option that is unduly attractive to certain ability levels, and a determination made as to the factors contributing to this. The choice of one standard error was arbitrary, but the criterion error can vary depending on the degree of accuracy desired. Note that Tan/SE is simply the z-ratio for testing the difference between two proportions, thus, inferences based on normal theory hold if the samples are large, otherwise like the standard error of the IRF described in Appendix A on page 31, these values can be regarded as rough approximations. Typical TOEFL samples for item analyses range from 500 to 1000 or more.

Option Response Configuration. After the IRC has been evaluated for evidence of option effects, the response curves for all other options can be compared with the IRC to determine which options are contributing to any observed nonmonotonicity. Option response curves are presented at the upper right of Figures 2-4 on pages 5-7. For the highly discriminating item in Figure 2, the response curves for options a, d and c are illustrative of effective counter-options, all decreasing while the response curve for the key is strictly increasing. Clearly the most effective option is c, virtually a mirror image or reflection of the IRC.

On the other hand, the option response configuration in Figure 3 reflects the lack of any effective counter-option; response curves for options b and d are relatively flat, with little impact on the key, but option a exhibits a rise at level 3 which accounts for the nonmonotonicity in the IRC at that point; in fact option a is clearly seen to be the most influential option of the set. It too is the mirror image of the IRC and induces the option effect observed for level 3.

The option response configuration in Figure 4 is one that was typically observed for very difficult items with (unreliable) low r-biserials but with high a-parameters. These items usually consist of one option in competition

with the key and two relatively effective counter-options. (Notice how a potential option effect at level 2 is canceled out by options d and b in Figure 4.)

The presence of a quasi-key is almost a necessary condition for very difficult four choice items. As p+ becomes small, and with essentially random responses on two options (which is a common state of affairs), a third non-keyed option must necessarily attract many more examinees than the key. This is the option that usually works as a quasi-key in practical situations.

Apparently, items are rejected in TOEFL test assembly if more high ability students choose a non-keyed option than choose the key, but such a criterion is not viable with extremely difficult items based on the foregoing. As noted above, the ability levels determined by quintiles cannot differentiate among level 5 examinees which is essential with very difficult items. When the data indicate that many of the highest scoring examinees are attracted to an option while few of this group select the key, it is probable that the latter represent the very highest scorers.

Biplots from a Correspondence Analysis. A second analysis generated by a transformation of the profile matrix involved the biplots resulting from by a correspondence analysis of the matrix P. The methods of correspondence analysis are given in detail in Greenacre (1984), and some of its features are outlined in Appendix B on page 32, but it can be characterized as a generalized principal components analysis, the results of which yield a biplot providing a succinct analysis of the relationships between the row and column points of a matrix. Biplots for the three items are given in the lower halves of Figures 2-4.

In a correspondence analysis of these data, the information relating 4 options and 5 ability levels has been reduced to a two-dimensional display. The horizontal axis can be attributed to ability and the vertical axis to option effects. If no option has an unusual attraction for a particular ability level, then the examinee groups will lie on the horizontal axis, ordered with respect to ability. When options exert greater than expected attraction for a given ability level, then scatter along the vertical axis will be observed, and the tendency of an ability level to select a particular option can be evaluated in terms of its proximity to the option point. Unfortunately, in this analysis distance measures between option and ability points are not calculable. A measure of the presence of option effects can also be evaluated in terms of the percentages of the total variance attributed to each axis which is indicated in each plot.

The analysis is profile-sensitive, and the relative placement of the points in the plot can be interpreted in terms of profile similarities; thus for the item in Figure 2, the biplot indicates that the profiles for levels 4 and 5 are comparatively unique, and that these levels tend in the direction of option b, level 5 moreso than level 4. The option response profiles of levels 1,2 and 3 are tending somewhat to options c and d. Option a has no attraction for any level. The differences among profiles for this item account for a substantial amount of variance (as measured by the trace =.27) compared to

values obtained for less discriminating items (items with low biserials tended to result in traces equal to about .04). The trace is a measure of the variance of the profile data generated by correspondence analysis and can be interpreted as a generalized variance, i.e., a weighted variance (see Greenacre, 1984 or Appendix B).

The biplot in Figure 3 reflects the ordering in the profile plot for this item, with levels 1 and 3, and levels 2 and 4 similar in response patterns, and consequently closely located on the plot, relative to the horizontal axis. (The reader should be aware that the biplots are on different scales for the purpose of readability.) The proximity of level 3 to option a clearly reflects the reason for the lack of ordering of ability levels. Both options a and d are unusually attractive to level 4, which also impairs the ordering of ability.

The biplot in Figure 4 illustrates the general case for very difficult items; the response profile for level 5 is markedly different from the others which are essentially random responses with little variability in profile characteristics and is clearly separated from the rest. The plot indicates the preference for options a and c by the top group, in that order. Relative to the horizontal axis, the ability levels are ordered, with no evidence of influential options.

A measure of the presence of option effects can be inferred from the percentages of variance accounted for by each axis; thus, it is clear that the item in Figure 2 is free of option effects since 98% of the variability is accounted for by the ability dimension, while the effect of options accounted for 11.25% of the variance in Figure 3. Based on the data of this study, localized option effects for TOEFL items might be investigated if the ability dimension accounts for less than 90% of the total variance. This value appeared to correspond to results obtained based on the criterion for flagging localized option effects given above.

In a general way, the results of the correspondence analysis of the matrix P provides almost all the information generated by the preceding methods: analysis of the biplot can help to identify localized option effects, and the percentage of the trace accounted for by the second axis can signal option effects. The marked separation of level 5 from the balance of the examinee group observed in Figure 4, typical of very difficult items with low biserials, but with acceptable a-parameters suggested a method, to be described below, for identifying items which could be included in the test.

PDIST, AN INDEX OF THE ESTIMABILITY OF THE A-PARAMETER FOR DIFFICULT ITEMS

It would be helpful if it could be determined from item analysis data whether or not an acceptable a-parameter is estimable for very difficult TOEFL items with low, unreliable biserials. For items that are precalibrated, the TOEFL test assembler need only check the a-parameter to determine whether it can be included in the test. For items that are noncalibrated, an index derived from the profile matrix may prove useful in identifying items for which an acceptable a-parameter can be estimated.

The use of the index will be limited to those items where random responses are observed for groups 1-4, and where only some high level examinees register slightly greater than random responses, which effectively limits its application to items with deltas ≥ 14.0 and r-biserials $<.20$. These are the IRT curves that remain flat over most of the ability range, exhibiting a relatively sharp rise only at the highest ability levels, associated with items very often resulting in an a-parameter of 1.5, the maximum for TOEFL data.

In order to quantify these relationships, the proposed index evaluates the distance between levels 4 and 5 relative to the average distance among levels 1, 2, 3, and 4. Given that levels 1-4 are responding randomly on very difficult items, the average of these distances should be small relative to the separation between levels 4 and 5. If the average of the absolute values of $(p_{1k} - p_{2k})$, $(p_{2k} - p_{3k})$, $(p_{3k} - p_{4k})$ is avd, then:

$$\text{pdist} = \frac{(p_{5k} - p_{4k})}{\text{avd}} \quad . \qquad (3)$$

Pdist is constrained to be positive which assures that level 5 examinees are scoring higher than those at level 4. For items with deltas ≥ 14 and biserials $< .20$, values of pdist were determined that always resulted in estimable parameters for Sections 2 and 3 (see Figures 5 and 6, page 14). These plots suggest that items with pdist values ≥ 4 for Section 2, and ≥ 2 for Section 3 might be considered for inclusion in a test when the biserial is less than .20 and the delta greater than 14. The differences in these cut points reflect the differences in the two IRT scales. Admittedly a small number of items on which to base these determinations, this represented all the items in the study with biserials less than .20 possessing a-parameters. Application of this index may identify difficult items with low biserials for which a's greater than .50 may be estimable.

Figure 5. Distribution of pdist and a-parameters for Section 2 items.

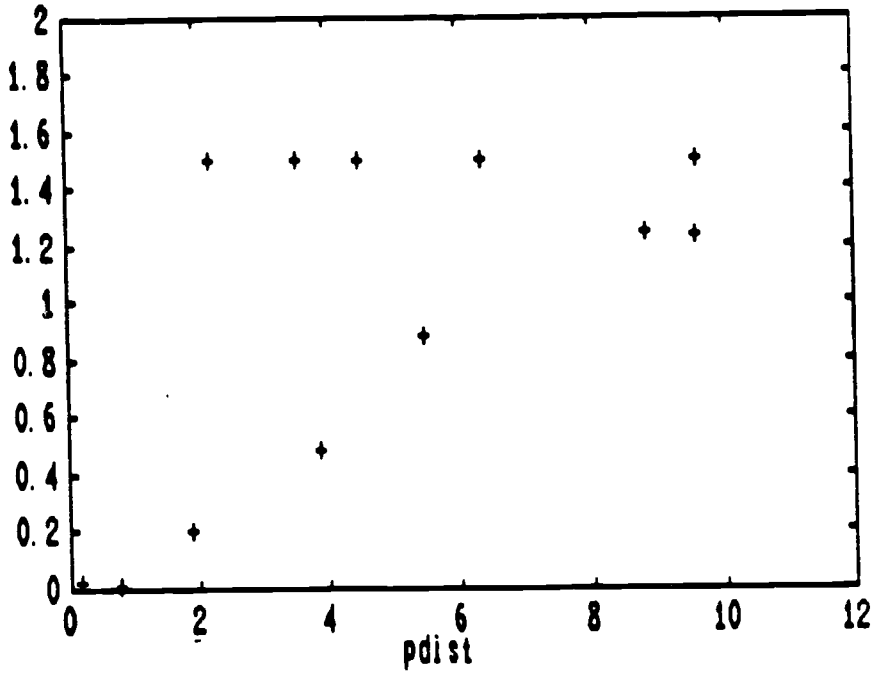
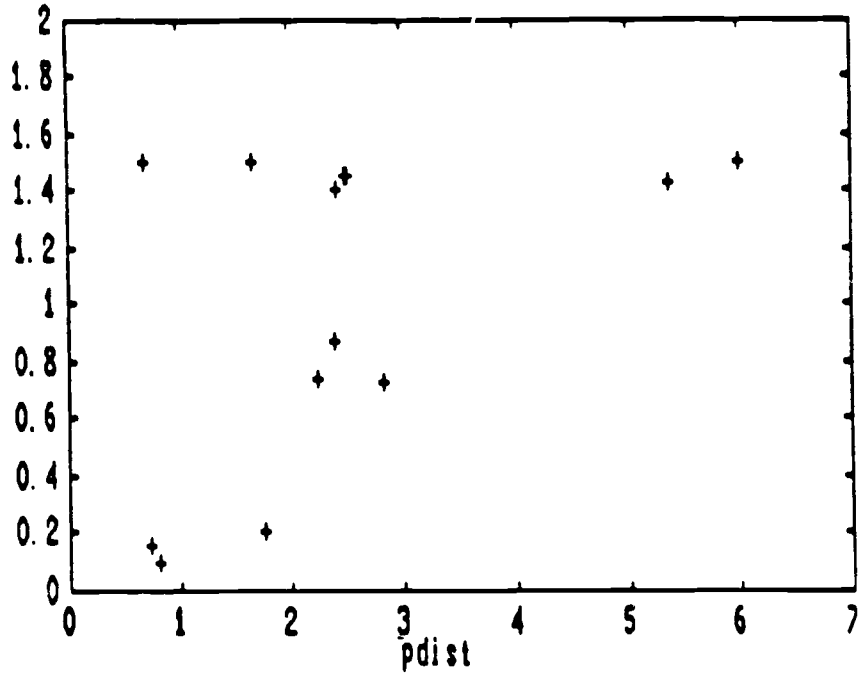


Figure 6. Distribution of pdist and a-parameters for Section 3 items.



SUMMARY DATA AND FURTHER EXAMPLES

Summary Data. Tests of language proficiency tend to yield highly discriminating items over the entire scale of difficulty. The mean biserials for the three sections of TOEFL usually fall in the range .51-.61, though a mean as low as .48 is occasionally observed. Relative to the TOEFL item pool, there are not many items with low r-biserials, and these all tend to be among the most difficult - with b-parameters greater than 2.00. Of the items with low r's, only a small proportion of them have been calibrated since items with r-biserials less than .05 have been automatically eliminated from LOGIST runs in order to avoid problems with convergence. Summary data for difficult items with IRT parameters from Sections 2 and 3 (Structure and Written Expression, and Vocabulary and Reading Comprehension) are given in Table 1 below. The table indicates that the lowest biserials tend to be associated with the most difficult items, but that acceptable a-parameters are estimable for many of them. The values of the trace reflect one of the underlying features of low-r items; the response variability is small.

Table 1.

Summary Data for Difficult TOEFL Items (Precalibrated)*

Mean rbi	a	b	Delta	rbi	trace	n
Section 2						
≥ .40	1.27	.50	14.12	.58	.25	12
.21-.39	.95	1.44	14.33	.31	.07	13
≤ .20	.97	2.22	15.48	.15	.05	12
Section 3						
≥ .40	1.31	.47	13.80	.58	.24	12
.21-.39	.76	1.46	14.63	.33	.13	14
≤ .20	1.04	2.29	14.86	.14	.06	13

*Delta >13.0

Since the group of items with biserials less than .20 and deltas ≥ 13.0 was the focus of this study, sufficient items from Section 1, Listening Comprehension were not available for analysis. Section 1 items result in a very easy scale with a mean delta of 10.7, suggesting that the factors tested in this section have a low threshold of difficulty beyond which effective measurement is not possible. Some of the methods of analysis are also limited to item curves of the type illustrated in Figure 4, usually associated with deltas of 14.5 or greater, few of which are observed in Section 1.

Examples of the Analysis of Six Difficult Items. Six difficult items with biserials less than .20 are analyzed on the following pages as further illustrations of the applications of the methods generated by this study.

Figure 7, Item AA, r=.14, a=1.5, delta=15.0, b=1.9. A visual evaluation of the IRC in Figure 7 on page 17 reveals the presence of option effects at level 3 and possibly level 4. The tangents associated with the difference between adjacent proportions were:

Tan:	.01	-.10	.05	.17
Tan/SE:	.16	2.44	1.25	4.03.

According to the criterion established in this study, the option effect flagged by the negative tangent between levels 2 and 3 is significant. The option response configuration immediately identifies option d as the source of the unsystematic response pattern, representing an almost perfect reflection of the IRC. It is also obvious that the other two options are not effective counter-options.

The biplot also supports option effects for levels 3 and 4. The ability levels are not ordered from 1-5, but in the order 3,4,1,2,5; the lack of ordering clearly determined by the attraction of option d to levels 3 and 4. The percentage of variance accounted for by the ability dimension is only 63% indicating the presence of large option effects - 35% of the variance can be attributed to option effects.

This is a Section 2 item, and pdist was computed to be 3.56. Although this is lower than the cut-point of 4 recommended above for noncalibrated items, an a-parameter of 1.5 was calculated for this item. This is one of the two items in the upper left hand corner of the plot in Figure 5. The item response function from the IRT analysis (at the bottom of Figure 7) indicates that the observed data deviates from the theoretical curve and follows the same trend as the IRC. Analysis of option d in terms of performance by levels 3 and 4 might suggest steps for remediation.

Figure 8, Item DB, r=.08, a=1.5, delta=15.5, b=2.75. Figure 8 on page 18 presents an example of an item effectively discriminating at very high levels of ability in spite of an observed r-biserial of .08. The option response configuration is similar to that given in Figure 4. In this case, two fairly effective counter-options exist as well as the quasi-key (option a). The IRC reveals no localized option effects.

Pdist for this section 3 item was 4.33 and an a-parameter is estimable. It was calculated to be 1.5 with a b-parameter of 2.75. The item response function produced from IRT estimation (not shown) demonstrated a good model fit. The biplot reflects the lack of option effects by the amount of variance (98%) attributed to the ability dimension alone, as well as the strict ordering along this axis.

Figure 7. Item AA, option response configuration, biplot and item ability regression (r=.14, a=1.5, delta=15.0. b=1.9).

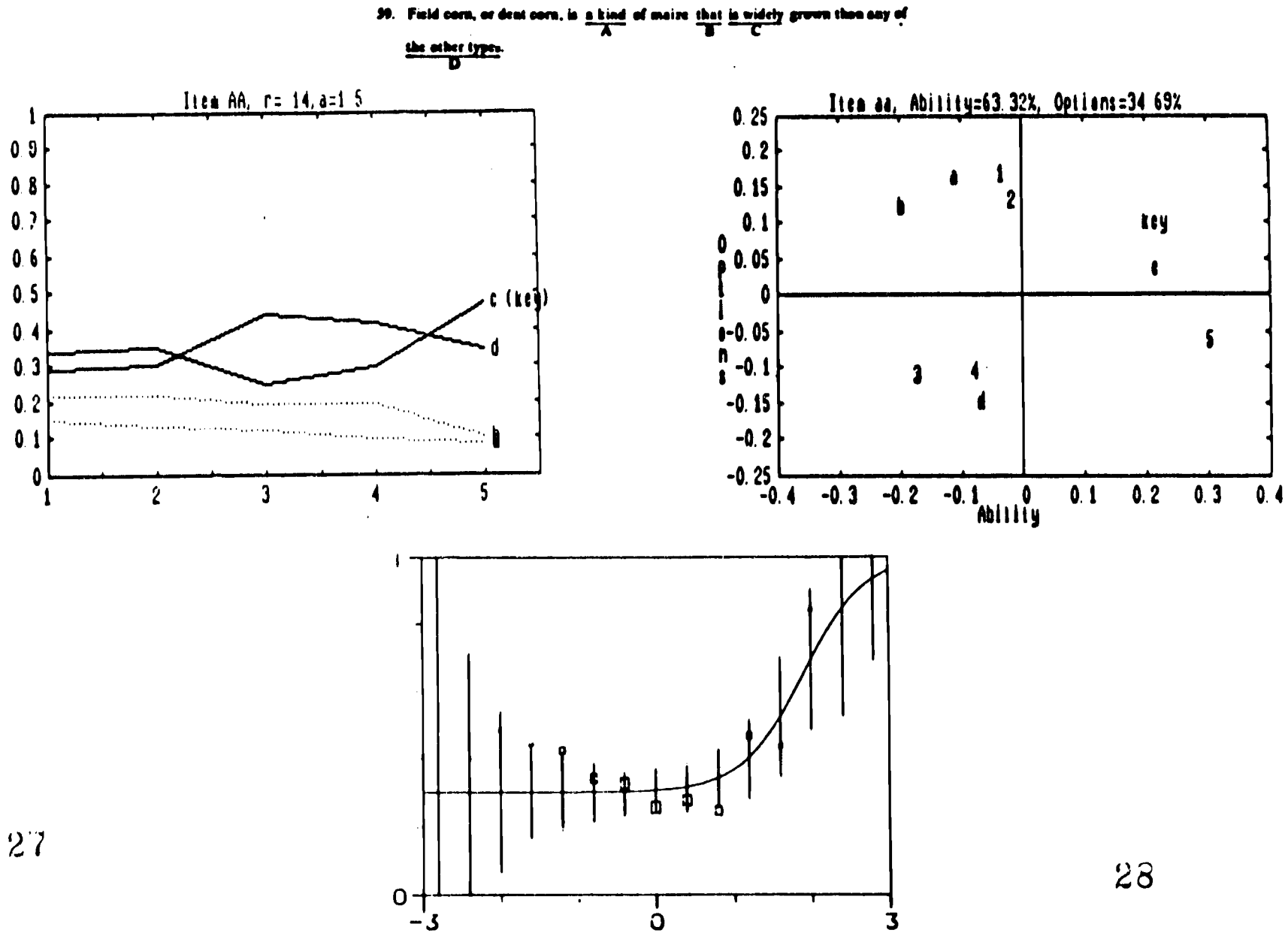
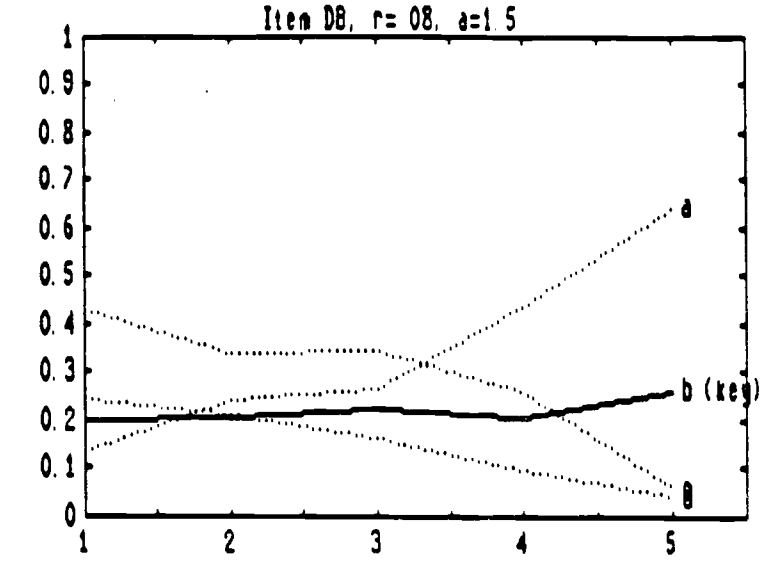


Figure 8. Item DB, option response configuration and biplot, r=.08, a=1.5, delta=15.5, b=2.75.



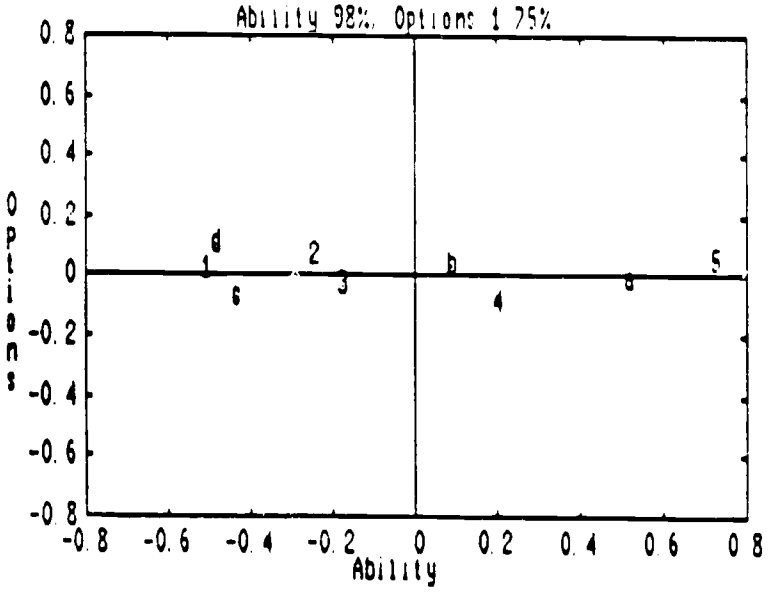


Figure 9, Item BR, r=.10, delta=16.6, noncalibrated. In Figure 9 on page 20, the negative tangent between levels 4 and 5 in the IRC flags a localized option effect at level 5. Examination of the option response configuration confirms that option a functions as a quasi-key, virtually paralleling the IRC, but option d is not an effective counter-option at level 5; it is unduly attractive to this group. Again, the curve defined by the reflection of the IRC quickly identifies the problematic option. If a downturn in the option d curve at level 5 could be effectuated, then the same configuration of quasi-key and two relatively effective counter-options would result as in previous examples of difficult items with high a-parameters.

The biplot for this item reflects the lack of optimal ordering of ability levels - level 5 precedes level 4 and its proximity to option d indicates its preference for that option. The percentages of variance accounted for by the ability dimension (87.74%) and options (9.69%) also point to option effects which upon remediation might improve the item. Pdist for this section 3 item was 1.07; thus, an acceptable a-parameter is probably not estimable.

Figure 10, Item AT, r=.05, delta=14.8, noncalibrated. In Figure 10 on page 21, option d exerts the greatest negative impact on the key at levels 2 and 4, and is clearly seen to impair the ordering of ability levels in the biplot. Pdist was calculated to be 1.9 for this Section 2 item; consequently, an acceptable a-parameter is probably not estimable.

Figure 11, Item AS, r=.18, a=.21, delta =14.8, b=2.85. The option response configuration as well as the biplot for the item in Figure 11 on page 22 indicates no option effects, simply flat profiles for all options. Option d is the least effective counter-option and might be a candidate for replacement. This example illustrates the possible utility of these plots in the absence of localized option effects; it may identify a single option that is the best candidate for replacement or remediation with the possibility of improving the r-biserial.

Figure 12, Item BK, r=.10, delta=14, noncalibrated. Option effects are observed for levels 3 and 4 in the option response configuration in Figure 12 on page 23, with option d the obvious offender. The biplot confirms these relationships in terms of the percentage of variance attributed to the option dimension. Pdist was calculated to be 7.25 for this Section 3 item; thus, while an a-parameter is estimable, it is likely that the fit to the model will not be optimal.

Figure 9. Item BR, option response configuration and biplot, r=.10, delta=16.6, noncalibrated.

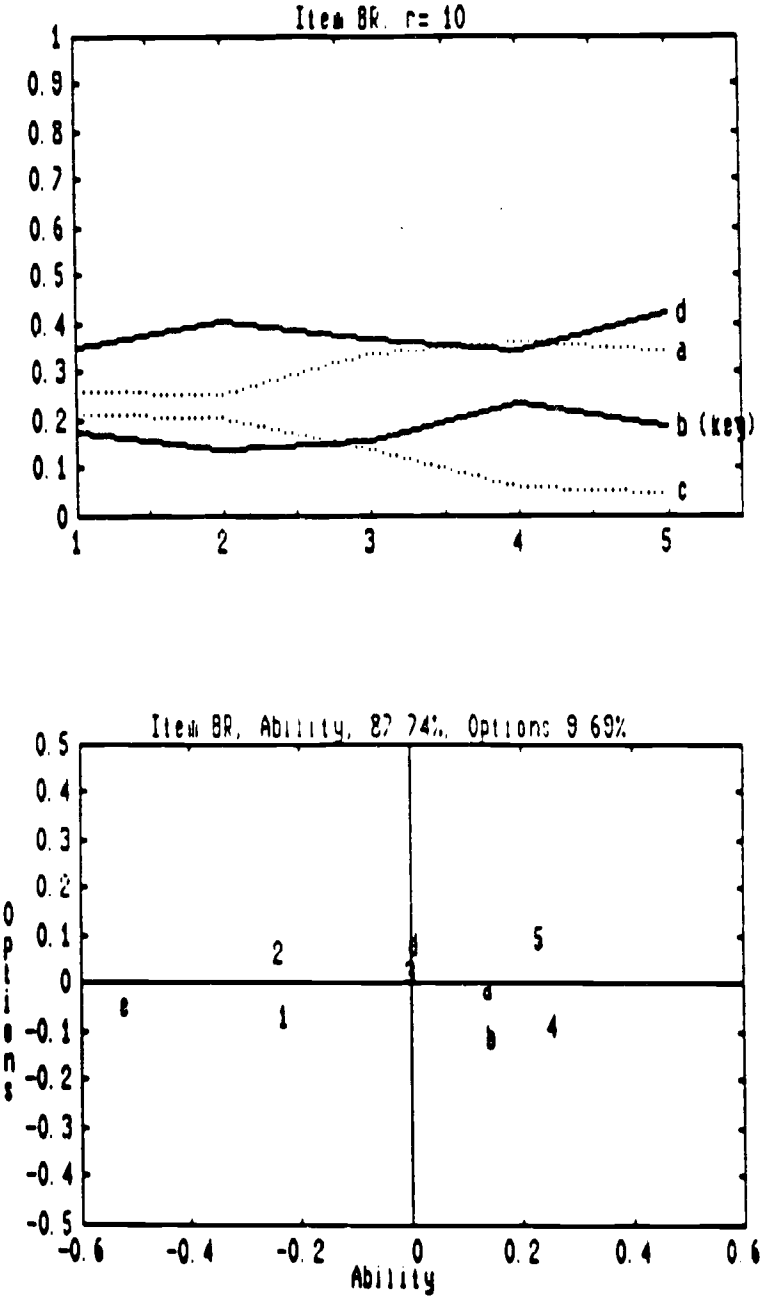


Figure 10. Item AT, option response configuration and biplot, r=.05, delta=14.8, noncalibrated.

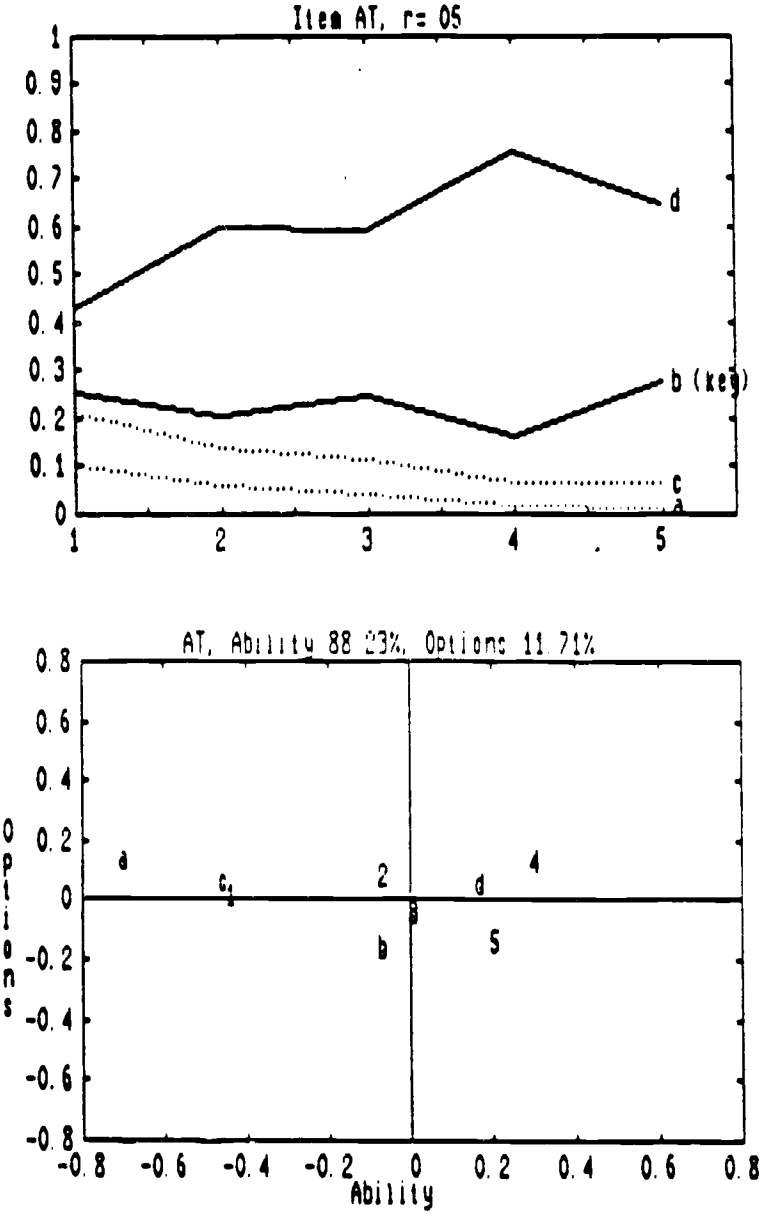
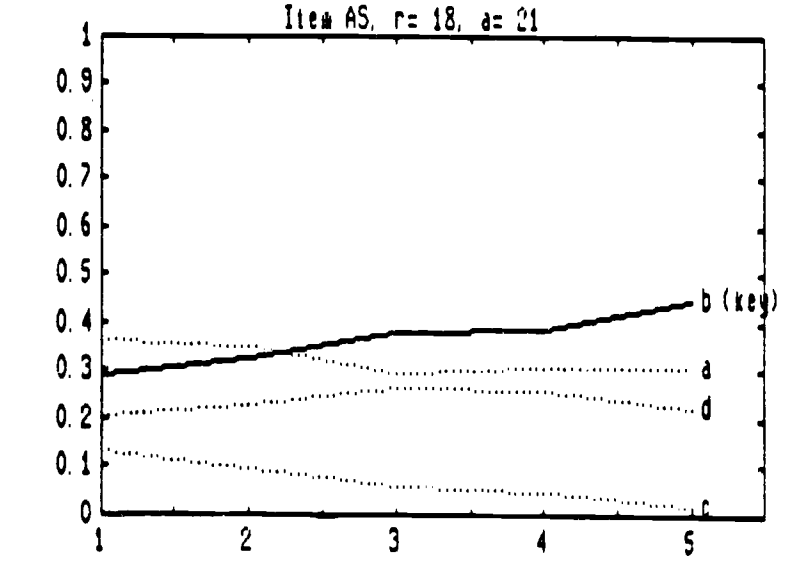


Figure 11. Item AS, option response configuration and biplot, r=.18, delta=14.8, b=2.85.



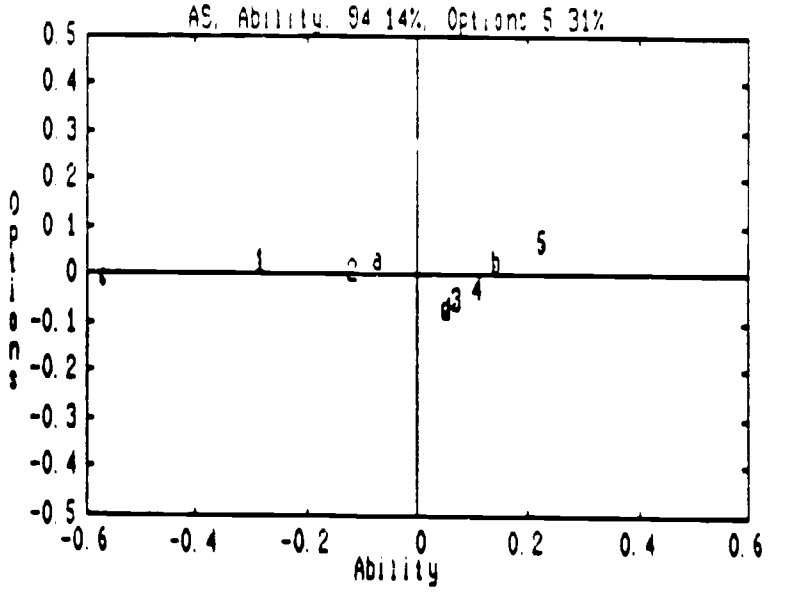
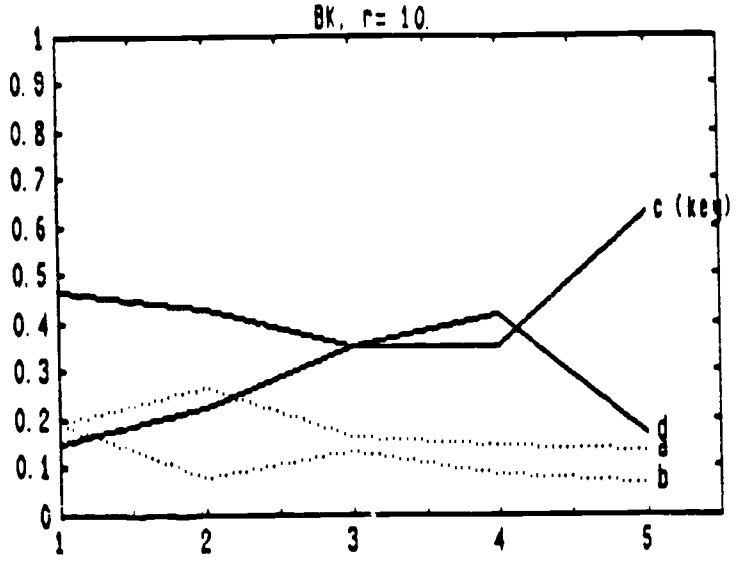
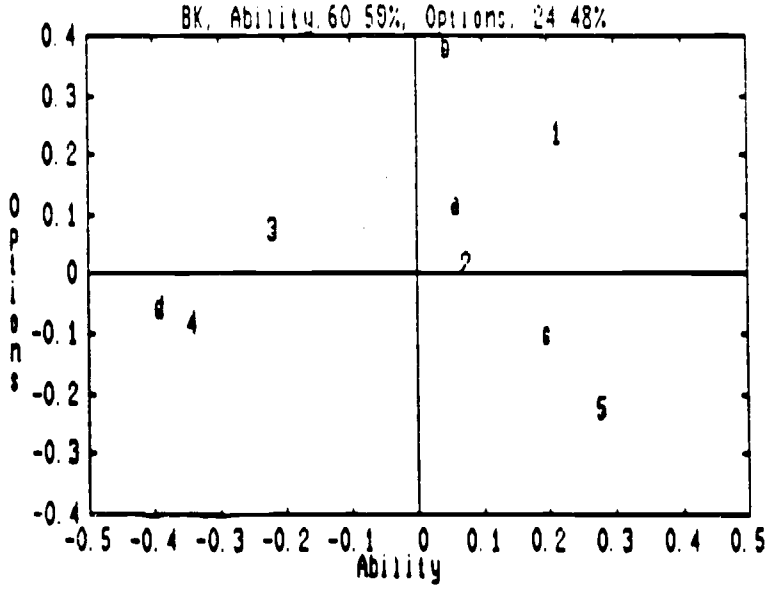


Figure 12. Item BK, option response configuration and biplot, r=.10, delta=14.0, noncalibrated.





DISCUSSION

Summary. Two methods of analyzing fifths data generated by TOEFL item analyses were developed in order to evaluate the effects of options on the discriminability of difficult items, and to identify difficult items with low, unreliable biserials for which acceptable a-parameters are probably estimable. Intended for use by test assemblers subsequent to an item analysis, the methods were graphical, but also included the evaluation of a distance measure.

The analysis identified certain option response configurations for difficult items which are probably discriminating effectively in spite of (unreliable) low r-biserials. A recurring configuration of options for acceptable four-choice items of very high difficulty was comprised of relatively effective counter-options and a "quasi-key", an option that draws examinees in the order of ability expected on the key and at a higher rate. The criteria for this judgment were characteristics of item analysis data for items with acceptable a-parameters. An index was suggested for use with TOEFL data which might identify such items. The index is scale dependent and limited to items with deltas greater than 14.0 and biserials less than .20.

The negative impact of localized option effects on IRT item fit was illustrated, as well as the importance of the quality of the entire response configuration - the key and all options. Evaluation of the option response configuration may also provide explanation for unusual values of the c-parameter. Hypotheses have been generated regarding irregularities in the observed curve which often occur at the lower levels of ability as in Figures 7, 10 and 12, however these may be due to localized option effects such as those described for those items.

While the original intent of this investigation was to focus on the application of correspondence analysis to these problems, many other ways of evaluating the fifths matrix surfaced during the course of this study, but the most effective appeared to be the analysis of the option response configuration described above. It has the advantages of dealing with untransformed data, and in most cases, ease of interpretation. All of the methods developed, except pdist, are applicable to items at any level of difficulty. The methods are also limited to those cases where one only option impacts negatively on the key, which is often the case. It may not be practical to attempt to disentangle interactions among several options.

Analyses leading to the identification of ineffective options may be considerably simpler than identifying the correct option revision, but it is hoped that these detailed analyses might make that task somewhat easier. The approach in these methods departs from reliance on a single index, however the often complex relations among options probably require an exploratory approach in the evaluation of their effects.

Several investigators have recognized the inadequacy of the logistic model in the presence of what are termed in this study "localized option effects" (Sympson, 1986; Thissen and Steinberg, 1984); however, the methods that have been generated to deal with them are fairly complicated. If such items are not

overly abundant within a given test, then it would be far simpler to improve the fit at the level of item construction as suggested by procedures given above. Indeed, the analysis at this level might provide further insight into factors that have a differential impact on measurement along the ability continuum.

Further Implications of Nonmonotonic Keyed Responses. It has been shown that the nonmonotonicity of option response curves can be detrimental to effective measurement, and that critical to good item discrimination is the requirement that the keyed option increase with ability. These two conditions are clearly interdependent. Underlying these relations is the fundamental and most heuristic assumption of item response theory which constrains the probability of a correct response to increase with ability, in this case a single ability or latent trait. When the correct response is not monotone and a dip occurs in the observed curve, the intrusion of another dimension or latent trait is implied (i.e., by violation of the assumption); thus the quality of item discrimination and the unidimensionality of the test are directly related.

One plausible hypothesis for the option effects described in this study might be based in inhibitory learning effects such as proactive inhibition, in which case there is interference from previous learning with the result that many lower level examinees, unimpeded by this difficulty, score higher on such items. Distracters are present that capitalize on this temporary confusion, clouding measurement with the artifacts of the learning process. It is also possible that certain inhibitory learning effects may be idiosyncratic to particular language groups. If a test contains a sufficient number of items of this type, then many lower level examinees will receive higher than expected total scores, reflecting the contamination of the measurement of English language proficiency with another factor or dimension. If this is a reasonable explanation of some of these effects, and if such items can be categorized, then they might be consigned to a diagnostic instrument where individuals at certain ability levels having this difficulty could be identified, but perhaps the major implication of applying these methods is the possibility of exercising some control over the dimensionality of the measuring instrument at the very practical level of item construction.

Implementation of the Methods of this Study. Subsequent to an item analysis, the following steps might be taken by TOEFL Test Development:

1. Apply pdist to any Section 2 and 3 item with delta $\geq$ 14 and r-biserial $<$ .20. Items with values of pdist $\geq$ 4 for Section 2, and $\geq$ 2 for Section 3 should be considered as acceptable for inclusion in a final form.

2. For items not meeting the criterion in (1), evaluate the Option Response Configuration, as described in this study, in order to determine which options might be remediated, or whether the item should be completely reworked or scrapped.

3. For items of any difficulty with low biserials, the evaluation of the Option Response Configuration should help to identify localized option effects which might guide in the remediation of the problematic option.

4. It might be desirable to categorize these option effects in terms of the frequencies at each level, and most importantly, in terms of factors contributing to them.

Further Research. This study has illustrated a significant limitation of the r-biserial, and points out the need for a 'conventional' measure of discrimination that can adequately assess this characteristic at any level of difficulty. This is a critical need for TOEFL test developers who, because of this difficulty, are unable to identify many acceptable difficult items for test assembly. The distance measure suggested above for identifying discriminating difficult items is necessarily gross, since it is based only on relationships among quintiles. Furthermore, it has no generality since it is dependent on the IRT scale of the particular test; thus it is regarded as an interim procedure designed to meet a pressing and immediate need. A method involving finer divisions of the score scale, and relating discrimination and item difficulty, should be considered for development, and might include some adaptation or modification of the evaluation of tangents as given above. In any case, a more effective assessment of conventional item discrimination is clearly needed.

It would be informative to determine how useful these methods may be in practical applications; consequently, a follow-on study of the effectiveness of item revisions made on the basis of these analysis might be considered.

REFERENCES

Greenacre, M. J. (1984) Theory and applications of correspondence analysis. N.Y.: Academic Press.

Lord, F.M. & Novick, M. R. (1974) Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

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

Thissen, D. & Steinberg, L. (1984) A response model for model choice items. Psychometrika, (49), 501-519.

Sympson, J.B. (1986) Extracting information from wrong answers in computerized adaptive testing. Paper presented at annual meeting of American Psychological Assoc., Washington, DC.

APPENDICES

Appendix A

TOEFL Item Response Functions or Item Ability Regressions

Item response functions (IRFs) for TOEFL are computed based on the three parameter logistic model (Lord, 1980, eq. 2-1, p.12). The equation specifies the probability of a correct response as a function of ability, and each parameter (a, b and c) indexes a characteristic of the IRF. The a-parameter, related to the slope of the IRF, is a measure of the discriminating power of the item. For TOEFL items, the range is 0-1.5 with the value 1.5 indicating maximum discrimination. The b-parameter locates the curve on the horizontal or ability axis, thereby defining the difficulty of the item. The range of the b-parameters for TOEFL is approximatelv -2.5 to +2.5, but higher absolute values are often observed. The c-parameter, the height of the lower asymptote of the curve, reflects the tendency to guess. The means of the c-parameters range from .15 to .21 across the three sections of TOEFL. An example of an IRF is given in the lower right of Figure 3.

In the plots of the IRFs generated for TOEFL, the theoretical curve given by the equation cited above is denoted by a solid line. On these plots, the ability axis ranges from -3 to +3, with a mean of zero, thus items near -3 are very easy and very difficult items are those near +3. An observed curve (small squares) consisting of the actual proportion of examinees at each ability level responding correctly to an item is superimposed on the IRF, and the adequacy of model fit is assessed by the correspondence of the two curves. The plots also include vertical lines representing a rough estimate of the 95% confidence interval around the IRF at selected ability levels which aid in the evaluation of model fit. The IRF in Figure 3 indicates that fit is most seriously affected by the group of examinees at ability level near -.5 since the small square representing those examinees is located beyond the limits of this interval.

Appendix B

Some Features of Corrrespondence Analysis

The basic mathematical tool of correspondence analysis and its variants is the singular value decomposition (SVD) of a nonsymmetric matrix. The following brief descriptions of some of the elements of correspondence analysis are taken from Greenacre (1984). The ordinary SVD is given by

$$\underset{IxJ}{A} = \underset{IxK}{U}\ \underset{KxK}{D_s}\ \underset{KxJ}{V'} \qquad U'U = V'V = I \qquad (1)$$

where U and V are the right and left singular vectors respectively of A, and K is the rank of A. U contains the eigenvectors of AA' and V contains the eigenvectors of A'A, D_s is a diagonal matrix of singular values, the square roots of the eigenvalues of A. The ordinary SVD can be considered a special case of the generalized SVD:

$$B = N D_s M' \qquad N'D_r^{-1}N = M'D_c^{-1}M = I \qquad (2)$$

where D_c^{-1} and D_r^{-1} are diagonal matrices, expressing the right and left singular vectors N and M in the metrics D_r^{-1} and D_c^{-1} respectively. D_s has the same meaning as above. An important feature of the SVD is that the right and left singular vectors define bases for the coordinates of the columns and rows of the relevant matrix.

The simplest form of data utilized in correspondence analysis can be represented in a two way contingency table, N (IxJ, i = 1, ..., I; j = 1, ..., J), with the columns defining categories of a variable and the rows representing objects or individuals for whom a set of frequencies, n_{ij}, have been observed. The matrix is transformed to P by dividing each element by n.., the total number of frequencies. A vector r, of row totals, containing elements $\Sigma_j p_{ij}$, and a vector c of column totals consisting of elements $\Sigma_i p_{ij}$ define row and column centroids.

In correspondence analysis of fifths data as given in the study, each row of P represents a profile across choices of options for a given ability level. It is expected that the profiles of adjacent ability levels would exhibit greater similarity than widely separated ability levels. Likewise, the columns represent profiles of responses on a given option across ability levels.

These row and column profiles define two clouds of points in J and I (weighted) Euclidiean dimensional space. The total inertia (a weighted variance) is given by

$$in(I) = in(J) = Tr[D_r^{-1}(P - rc')D_c^{-1}(P - rc')'] . \qquad (3)$$

The total inertia is also given by the sum of the singular values $\Sigma_k s_k^2$ where the sum is from k = 1, ..., K, the rank of P - rc'. The purpose of the

analysis is to determine the K*, ($K^* \leq K$) dimensional subspaces of the row and column clouds which are closest to the given points in terms of weighted sum of squared distances. The lowest rank approximation (i.e., K*) in this least squares sense can be shown to be the singular vectors, in the metrics D_c^{-1} and D_r^{-1}, corresponding to the largest singular values of $P - rc'$. In these subspaces, the K* right and left generalized singular vectors of $P - rc'$ are the principal axes of the row and column clouds, respectively. The correspondence analysis of $P - rc'$ involves those steps in the solution of equation (2), where $B = P - rc'$. The actual solution involves the ordinary SVD of

$$D_r^{-1/2}(P - rc')D_c^{-1/2} = U D_s V' \quad U'U = V'V = I \qquad (4)$$

and (2) results from the transformation

$$N = D_r^{1/2}U; \quad M = D_c^{1/2}V. \qquad (5)$$

Significant results of the analyses are the biplots of the coordinates of the row and column points. In this context, the coordinates of the row points with respect to the basis M is

$$F = D_r^{-1}ND_s. \qquad (6)$$

Likewise, the coordinates of the column points with respect to the basis N is

$$G = D_c^{-1}MD_s. \qquad (7)$$

In general, the interest lies in the relative position of these coordinates and not in M and N which define the dual problem. Presentations of both coordinate matrices in a single plot (biplot) are feasible due to the geometric correspondence of the row and column points, in terms of position and in terms of inertia. The overall purpose of correspondence analysis is to explicate the geometry of a group of high dimensional points through an approximate low-dimensional display (Greenacre, 1984).