DOCUMENT RESUME

ED 390 890 TM 024 171

AUTHOR	Way, Walter D.; And Others
TITLE	An Exploratory Study of Characteristics Related to IRT Item Parameter Invariance with the Test of English as a Foreign Language. TOEFL Technical Report.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-92-43; ETS-TR-6
PUB DATE	Sep 92
NOTE	45p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Ability; *English (Second Language); *Goodness of Fit; Item Response Theory; Language Tests; Models: Pretests Posttests; *Reading Comprehension; Regression (Statistics); Test Construction; Test Format; Test Items
IDENTIFIERS	Invariance; *Item Parameters; *Test of English as a Foreign Language

ABSTRACT

This study provided an exploratory investigation of item features that might contribute to a lack of invariance of item parameters for the Test of English as a Foreign Language (TOEFL). Data came from seven forms of the TOEFL administered in 1989. Subjective and quantitative measures developed for the study provided consistent information related to the model-data fit of TOEFL test items. For TOEFL Sections 1 and 2, items that were pretested before 1986 exhibited poorer model-data fit than items that were pretested after 1986. For Section 3, reading comprehension, model-data fit appeared to be related to changes in the relative position of items within the sections from the pretest to the final form administrations. Based on the results of the study, it was recommended that the TOEFL program investigate the feasibility of not using pretest item response theory statistics for items pretested before 1986 for Sections 1 and 2 and that guidelines be developed for test developers to use with reading comprehension items to limit change in relative positions of items in the test from pretest to final form administrations. Two appendixes give rules for judging item-ability regression plots and sample item ability regression plots. (Contains 4 figures, 11 tables, and 18 references.) (Author/SLD)

TOEFL®

September 1992

Technical Report

TR-6

An Exploratory Study of Characteristics Related to IRT Item Parameter Invariance with the Test of English as a Foreign Language

Walter D. Way
Patricia A. Carey
Marna L. Golub-Smith

An Exploratory Study of Characteristics Related to IRT Parameter Invariance with the Test of English as a Foreign Language

Walter D. Way
Patricia A. Carey
Marna L. Golub-Smith

Educational Testing Service
Princeton, New Jersey

RR-92-43

Abstract

IRT equating methods have been used successfully with the TOEFL® test for many years, and for the most part the observed properties of items have been consistent with model predictions. However, items that do not appear to hold their IRT pretest estimates do exist. If relationships can be found between features of TOEFL items in pretest calibrations and subsequent lack of model-data fit when these items are used in final forms, steps to eliminate the use of such items in TOEFL final forms can be taken. The purpose of this study was to provide an exploratory investigation of item features that may contribute to a lack of invariance of TOEFL item parameters.

The results of the study indicated the following: (1) subjective and quantitative measures developed for the study provided consistent information related to the model-data fit of TOEFL test items, (2) for Sections 1 and 2, items that were pretested before 1986 exhibited poorer model-data fit than items that were pretested after 1986, and (3) for Section 3 reading comprehension, model-data fit appeared to be related to changes in the relative position of items within the sections from the pretest to the final form administrations.

Based on the results of the study, it was recommended that (1) the TOEFL program investigate the feasibility of not using pretest IRT statistics for items pretested before 1986 for Sections 1 and 2 and (2) that guidelines be developed for test developers to use with reading comprehension items to limit the change in relative positions of items in the test from pretest to final form administrations.

The Test of English as a Foreign Language (TOEFL®) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of more than thirty organizations, public and private, that were concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program, and in 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the GRE Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

✣ ✣ ✣

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. Currently the Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1991-92) members of the TOEFL Research Committee are:

James Dean Brown	University of Hawaii
Patricia Dunkel (Chair)	Pennsylvania State University
William Grabe	Northern Arizona University
Kyle Perkins	Southern Illinois University at Carbondale
Elizabeth C. Traugott	Stanford University
John Upshur	Concordia University

Table of Contents

List of Tables

List of Figures

Introduction

The use of item response theory (IRT) in equating the Test of English as a Foreign Language (TOEFL) has been well documented (Cowell, 1982; Hicks, 1983). The chief advantage of IRT lies in the properties of parameter invariance. If the assumptions of the IRT models are satisfied, item parameters are invariant (subject to a scale transformation). In other words, item parameters are said to be "sample free" (Wright, 1968). It is the advantages of parameter invariance that allow TOEFL final forms to be constr ıcted by combining pretest items from any number of previously administered forms. This is accomplished through the maintenance of a TOEFL item bank. When final forms of the TOEFL test are constructed, pretest item parameter estimates taken from this bank are used to equate the new form to the existing TOEFL scale.

Recently, a new equating design was implemented with the TOEFL test that makes use of a procedure known as IRT preequating (Lord, 1980). As a check on this procedure, fit of the final form data to model predictions based on the pretest item parameter estimates is obtained. This information has uncovered a number of "problem items," that is, items for which the model-data fit was found to be less than acceptable. Depending upon the particular test form and section, anywhere from 8 to 20% of the items have been found to have unacceptable fit. In general, there has been no apparent pattern to the misfit, that is, there has been no consistent tendency for the misfitting items to be overpredicted or underpredicted by the model. For each of the preequated administrations, "parallel" equatings have also been carried out by obtaining new IRT estimates for each of the preequated items and transforming the estimates from the scale defined by the new calibration run to the original IRT scale, that is, the scale defined by the pretest item parameter estimates. The obtained equating conversions from the preequating and the parallel procedures have been consistently close, typically differing by less than one scaled score point per section except in the chance-score range of the scale.

In comparing the results of the preequatings and parallel equatings, it became clear that the items for which the pretest IRT estimates indicated poor fit to the final form data were also the items for which the pretest IRT estimates and the parallel/transformed IRT estimates were most discrepant. In both equating procedures, the problems were related to lack of item parameter invariance between the pretest and final form use. In the preequatings, the problem was manifested in the fit of the data to the model predictions. In the parallel equatings, the problem was manifested in observed violations of the linear relationship that IRT predicts to exist between the item parameter estimates obtained in pretest and parallel calibrations.

Possible Explanations for Lack of Item Invariance

Because successful applications of IRT depend heavily on the accuracy with which individual item parameters are estimated, evidence of lack of item parameter invariance is cause for concern. Theoretically, lack of parameter invariance is explained as a failure to meet the assumptions of the IRT model. However, this theoretical explanation provides little practical guidance in trying to investigate why parameter estimates for some items do not appear to be invariant. Speculations about causes of lack of item parameter invariance have included context effects (Eignor, 1985; Kingston & Dorans, 1984; Yen, 1980), sample characteristics (Cook, Eignor, & Taft, 1988; Golub-Smith, 1986), and interactions between sample characteristics and the IRT estimation procedure (Stocking, 1988). With the TOEFL test, context effects and sample characteristics are particularly pertinent because final form items are drawn from many previously administered forms. Furthermore, each final form of the TOEFL test contains items that differ in terms of what pretest sample they were calibrated in and how long ago the calibrations occurred. It is possible, for example, that the difficulties of some items have truly changed over time because of changes in the way English as a foreign language is taught.

Purpose of the Study

IRT equating methods have been used successfully with the TOEFL test for many years, and for the most part the observed properties of the items have been consistent with model predictions. However, as with any testing program utilizing IRT, it is important for the TOEFL program not only to monitor model-data fit in order to assess item parameter invariance, but also to attempt to understand why items may or may not demonstrate parameter invariance. If relationships can be found between features of TOEFL items in pretest calibrations and subsequent problems with items used in a final form, then steps to eliminate the use of such items in TOEFL final forms can be taken. The purpose of this study was to explore characteristics related to invariance and lack of invariance of TOEFL item parameters. In particular, the study assessed the extent to which parameter invariance appeared to hold for test items based on such factors as the nature of the pretest calibrations, similarities between the pretest and final form calibration samples, and the properties of items themselves, namely, position and content.

Method

Data Source

The data source for this study was seven forms of the TOEFL test administered in 1989. The TOEFL test consists of three separately timed sections: Listening Comprehension (Section 1, 50 items), Structure and Written Expression (Section 2, 38 items), and Vocabulary and Reading Comprehension (Section 3, 58 items). Each test section is further divided into subparts consisting of different item formats. For example, Listening Comprehension contains three item types: statements, dialogues, and minitalks. The Structure and Written Expression section is divided into a part that measures understanding of basic grammar and one that tests knowledge of the grammar of written English. Vocabulary and Reading Comprehension is divided into a part that measures the ability to understand the meaning of words or phrases in sentences and a part that measures the reading comprehension of short passages. Each section of the test is separately scaled and equated using the three-parameter logistic (3PL) IRT model as implemented by the LOGIST computer program (Wingersky, Patrick, & Lord, 1987). For the present study, each of the TOEFL forms examined consisted of items that had been previously pretested and calibrated to the TOEFL IRT scale. In addition, the items in each form used in the study had item parameters reestimated and transformed to the TOEFL IRT scale based on the operational data.

Assessment of Model-Data Fit

To assess fit of the data from each test form to the IRT model, two methods were employed. The first method made use of a graphical model-data fit technique called item-ability regressions (IARs) (Kingston & Dorans, 1985). Although item-ability regressions--and similar techniques such as residual analyses (Hambleton & Murray, 1983)--are typically utilized when both item and examinee parameters are being estimated, in the present study the item-ability regressions were obtained by holding item parameters fixed at original estimates and estimating ability only. Based on visual inspections of these plots, ratings of how well the data fit the original IRT item parameter estimates of each item were made. These ratings were based on a four-point scale that ranged from one (indicating particularly good fit) to four (indicating particularly poor fit). All ratings were carried out by the authors of the study, each of whom has had extensive experience with IRT-based equating methods. To enhance the consistency of the ratings, a set of guidelines were developed, preliminary ratings were made, and a feedback session was held where preliminary ratings were discussed and the guidelines were fine-tuned. Appendix A contains a listing of the rating guidelines. Graphical examples of both a good fitting item and a poorly fitting item, at first (pretest) and second (final form) use, are displayed in Appendix B. Each item was independently rated twice, which resulted in a sum rating for each item that could range from two to eight. The reliability of the rating process was evaluated using two measures of the extent to which ratings agreed: correlations between first and second ratings, and alpha coefficients. The alpha coefficients can be thought of as the proportion of agreement between independent ratings of the same item. These values tend to be higher than correlations between independent ratings because they take into account the variance in the sum of the two ratings for each item. These statistics were calculated by TOEFL section and form. In addition, means and standard deviations of item ratings were compiled by section subpart and form.

The second assessment of model-data fit was obtained using quantitative measures of the difference between item characteristic curves based on the original item parameter estimates for each item, and the reestimated and transformed item parameter estimates. Two statistics were defined, a weighted mean difference statistic (WMD), and a weighted root mean square statistic (WRMS). These statistics are defined for item j as follows:

$$WMD_j = \sum_{i=1}^{31} W_i \, [P_{jold}(\theta_i) - P_{jnew}(\theta_i)] \, , \qquad (1)$$

and

$$WRMS_j = \sqrt{\sum_{i=1}^{31} W_i \, [P_{jold}(\theta_i) - P_{jnew}(\theta_i)]^2} \qquad (2)$$

where $P_{jold}(\theta_i)$ and $P_{jnew}(\theta_i)$ are the probabilities of a correct response for item j using the 3PL model at theta level i and the item parameter estimates for the old and new forms, i represents the midpoint of 31 equal intervals of width 0.2 on the theta scale ranging from -3.1 to +3.1, and W_i represents a weight assigned to interval i defined as the number of estimated theta values in the new form falling in interval i divided by the total number of thetas estimated.

The WMD statistic represents a signed measure of the discrepancies between the predicted probabilities of correct response based on the original (pretest) and reestimated (final form) item parameter estimates. The WRMS statistic represents a more global measure of these discrepancies. The WMD and WRMS statistics are similar to statistics often used in other IRT applications, such as investigations of item bias (Haebara, 1980; Linn, Levine, Hastings, & Waldrop, 1981; Shepard, Camilli, & Williams, 1984).

Because both the subjective item fit ratings and the WRMS statistics are essentially global measures of model-data fit, it was expected that the two indices would be moderately to highly related. However, perfect agreement between the two measures was not expected. The WRMS statistic basically provides a quantitative measure of the consistency of the item characteristic curve (ICC) calculated for an item in two calibrations. If model-data fit is relatively poor, but consistent across two calibrations, the value of the WRMS statistic will be very low, indicating little change in the ICCs. On the other hand, the item fit rating would be high, reflecting the poor model-data fit. Another reason the two measures were not expected to be perfectly related was that even though the item fit ratings were summed across two raters, the resulting scale only ranged across seven possible values. The expected relationships between the WMD statistics and the other two measures were not clear, although it was expected that both low negative and high positive WMD values would be associated with poor model-data fit.

For each of the seven TOEFL forms examined, means and standard deviations of the item fit ratings, the WMD statistics, and the WRMS statistics were compiled for each section of the test and for each of the parts within the test sections. In addition, bivariate relationships between the item fit ratings and the WMD and WRMS statistics were explored.

Relationships Between Item Features and Model-Data Fit

A final set of analyses explored relationships between the measures of model-data fit and features of the test items. The variables explored included (1) the year the test item was pretested, (2) the absolute change in the relative position of the item within the test from the first to the second administration, (3) the sample size used in the original calibration of the item, and (4) the difficulty in estimation of the guessing parameter in the original LOGIST calibrations (defined by the quantity b - 2/a, where b is item difficulty and a is item discrimination).

Results

Assessment of Item Fit

Ratings of Item Fit. The interrater correlations and alpha coefficients for the ratings of item fit are displayed by form and test section in Table 1. These data indicate high agreement between the independent ratings. The interrater correlations ranged from 0.80 to 0.97, and the alpha coefficients ranged from 0.89 to 0.99. With the exception of Section 2 in the February and September administrations, all interrater correlations were at or above 0.90 and all alpha coefficients were at or above 0.95.

Table 2 presents the mean item fit ratings by form and section part. Across all forms, the individual mean ratings range from a low of 3.14 (best) to 6.20 (poorest). Averaging across the seven forms, it appears that the best overall fit was obtained for Section 2 structure, with a mean rating of 4.11, and the poorest fit obtained was for Section 1 statements, with a mean rating of 5.26. Across all forms, the mean item fit ratings were most stable for the vocabulary part, ranging from 3.93 to 4.55. The mean item fit ratings were most variable for English structure, ranging from 3.14 to 5.43.

Item Fit Statistics. The means and standard deviations of the weighted root mean square statistics (WRMS) by form and section part are displayed in Table 3a. The mean values ranged from a low of 0.035 (best) to 0.084 (poorest). As with the item fit ratings, Section 1 statements had the poorest overall fit, with a mean of 0.063. The lowest overall mean WRMS values occurred for Section 2 structure and for Section 3 vocabulary (0.043). As was the case with the item fit ratings, the written expression part had the next best overall fit according to the WRMS statistic. The relative rankings of dialogues, minitalks, and reading comprehension varied slightly according to the two measures, although the overall means of these three parts were relatively similar using either measure.

The means of the weighted mean difference statistics (WMD) are presented in Table 3b. Across forms, the mean WMD values all clustered around zero, ranging from -0.034 to 0.025. There was no apparent pattern in the signs of the mean WMD statistics across the different parts, although WMD standard deviations tended to be larger when the corresponding WRMS values were larger, and smaller in the cases when the corresponding WRMS values were smaller.

Relationships Between Indices. The correlations between the item fit ratings and the two item fit statistics, WRMS and WMD, are presented in Table 4 by form and test section. The correlations between the WRMS and the item fit ratings are moderately high, ranging from 0.68 to 0.86. Combined across months, the correlation between these two measures was 0.74 for Section 1, 0.76 for Section 2, and 0.77 for Section 3. Figure 1 displays the linear regression of the WRMS values on the item fit ratings fr each section. These plots indicate that the scales of the item fit rating are somewhat restricted, particularly at the upper levels (which indicate poorest fit). It can also be seen that for all three sections, few items received total ratings of 3, 5, or 7. These total ratings resulted when the two independent ratings were discrepant. Note that for each section, the items receiving fit ratings of 7 are nearly all below the regression line. One reason for this appears to be that many of the items receiving item fit ratings of 7 were of relatively poor quality from an IRT standpoint. Figure 2 displays the item-ability regression plots for two of these items, both at pretest and at final form. It can be seen that the fit of these items is relatively poor. However, the estimated ICCs for these items were stable from pretest to final form, which was reflected in lower WRMS values than would be predicted from the regression equations.

The correlations between the WMD statistics and the item fit ratings are also provided in Table 4. Since the WMD statistic is a signed measure, its relationship with the fit ratings is somewhat ambiguous. Figure 3 displays bivariate plots of the item fit ratings against the WMD statistics for each section. These plots tend to be V-shaped, with both the highest and lowest WMD values corresponding to item fit ratings of 8, and with relatively few WMD values near zero occurring for items with item fit ratings of 6, 7, and 8.

Item Features and Model-Data Fit

Year of Pretesting. The TOEFL test forms investigated in this study contained items that were pretested between 1981 and 1988. Table 5 displays, for each section, the numbers of items, the mean WRMS values, and the mean item fit ratings (IFR) broken down by the year in which pretesting occurred. For Section 1, both the mean WRMS values and the mean item fit ratings tended to decrease with the time between pretesting and final form use. The mean WRMS value decreased from 0.082 for items pretested in 1981 to 0.051 for items pretested in 1988. The mean item fit rating decreased from 5.95 for items pretested in 1981 to 4.19 for items pretested in 1988. A similar trend occurred for Section 2, although the pattern was less consistent. Table 5 indicates that for Section 2 both the mean WRMS values and the mean item fit ratings were lower for items pretested after 1985 compared to those pretested through 1985. There was no consistent trend between year of pretesting and the model-data fit measures for Section 3.

It is unclear why the relationships observed between the year items were pretested and the model-data fit measures for Sections 1 and 2 were not obtained for Section 3. In the case of Section 1, the relationship may be due in part to the method by which earlier Section 1 items were pretested. Prior to 1986, these items were not pretested within the operational form as they are presently, but were administered separately to examinees in cooperating English language institutes. The samples of examinees used in this earlier pretesting procedure were different from the typical TOEFL testing samples. For Section 2, it is possible that these item types were more easily influenced by changes in instructional methods in English as a second language, and this may have contributed to the relatively poorer fit seen for items pretested between 1981 and 1985.

Change in Item Position. To compare sections with different numbers of items, item position was defined as the item number divided by the total number of items in that section. Item position was defined in this manner because different forms of the TOEFL test may have different numbers of items. Absolute position change was calculated as the absolute value of the difference between these proportions from the pretest to the final form administration. For example, if a listening comprehension item was pretested as item 40 in a 80-item form and administered operationally as item 30 in a 50-item form, the relative change would be the absolute value of 40/80 minus 30/50, or 0.1. Table 6 presents the mean absolute position change for each form and section part. These data indicate that both dialogues and minitalks of Section 1 had the least average change in position (0.11), whereas, Section 2 written expression had the most change (0.19). The correlations among position change, WRMS, and item fit ratings (IFR) by form and section part are presented in Table 7. For Sections 1 and 2, there was no apparent relationship among these measures. However, for the reading comprehension part of Section 3, position change over all months correlated 0.50 with WRMS and 0.41 with IFR. In addition, for four of the forms, the correlations between absolute position change and the two measures of fit were relatively high, ranging from 0.55 to 0.82. Figure 4 displays a plot of the regression of WRMS on position change. In this plot, it can be seen that for most of the items the absolute position change was less than 0.25, although for nine items position change was between 0.35 and 0.40. If these nine items are deleted, the strength of the relationship between absolute position change and WRMS is reduced: the correlation drops from 0.50 to 0.29 and the regression weight changes from 0.23 to 0.15. However, the reduced regression coefficient is still statistically significant at a .0001 level. It should be noted that because the items in this part are in linked sets, item position can change from pretest to final form use both within part and within set. Although the written expression items had the greatest change in location, the correlation across all forms between this variable and each measure of item fit was not significant. For the other parts, there was no consistent relationship among these measures.

Sample Size. Tables 8a and 8b present the mean original calibration sample size by form and section for the best (item fit rating = 2) and the worst (item fit rating = 8) items, respectively. Recall that each form of the TOEFL test consists of items from many different pretest forms, each calibrated in a different LOGIST run. The mean sample size across all forms for the poorest fitting Section 1 items (1,520) was lower than for the best items (1,855). The correlations between calibration sample size and the item fit measures by form and section are displayed in Table 9. For this analysis, only the best and worst items were included. Only the correlation between calibration sample size and IFR for Section 1 (-0.18) was significant. Overall, these data indicate no consistent relationship between sample size and fit. It should be noted, however, that all forms exceeded the minimum recommended sample sizes for parameter estimation, with no sample under 1,000.

Estimation of the Guessing Parameter. For easy items, or moderately easy items with low discrimination, there is often insufficient information at the lower ability levels to estimate the lower asymptote of the item response function (c-parameter). The quantity b - 2/a is the ability level below which stable estimates of c are not obtainable. For items for which this stability criterion falls below a specified level, LOGIST specifies a common c based on the average of all the items below this level. Tables 10a and 10b present the mean quantity b - 2/a by form and section for the best and worst items, respectively. Table 11 displays the correlations between the quantity b - 2/a and the measures of fit. As with calibration sample size, only the best and poorest fitting items were included in this analysis. Overall, there appears to be little relationship among these measures, as only for Section 2 was the correlation (0.20) between b - 2/a and WRMS statistically significant, and in this case the relationship was in the opposite direction from what would have been expected. Furthermore, within test forms the data suggested no consistent relationship between model-data fit and the quantity b - 2/a.

Summary and Discussion

The purpose of this study was to explore features related to item parameter invariance, in particular with regard to IRT equating of the TOEFL test. The assumption of item parameter invariance was assessed by checking the fit of the final form data to model predictions based on the pretest item parameter estimates using both a subjective item rating and a quantitative index. The study assessed the extent to which parameter invariance appeared to hold for TOEFL test items in different sections and different parts within sections. It should be noted that the study was exploratory in nature and focused on investigating only some of many item features that might contribute to lack of parameter invariance. Compiling the necessary data was a difficult task, since much of the pretest information had to be retrieved from archives and large volumes of computer output had to be inspected. Although further, more in-depth investigations of item features could be carried out, such investigations would require greater resources than were available for this study. An extension of the present study might enlist test developers to investigate further item features related to invariance or lack of invariance. For example, items could be classified based on patterns of misfit, and test developers could be provided with the item text, detailed content classifications, the classical statistics, and the IRT-based statistics for the items. Such investigations would be similar to those often carried out in settings related to differential item functioning, and might provide a more solid framework for considering the parameter invariance of TOEFL test items.

One fairly clear finding of the study was that the two methods developed to assess model-data fit produced consistent overall results. The rules developed for subjectively flagging items provided an efficient procedure for identifying those items for ,nich parameters do not hold from pretest to final form use. The WRMS statistics, although based on different criteria, generally confirmed the subjective rating procedure. Interestingly enough, for many items where the item fit ratings and the WRMS statistics were discrepant, the original model-data fit of the item at pretesting was found to be suspect. These results underscore the need for monitoring pretest item calibrations to flag items with questionable model-data fit. IRT-based item flagging procedures have been established with the TOEFL program and are currently part of routine operational procedures.

In comparing the model-data fit of the different TOEFL parts, it was found that Section 1 statements had the poorest model-data fit based on the results of both the subjective and quantitative indices. The other two listening parts, dialogues and minitalks, also tended to exhibit comparatively poor model-data fit. A possible reason for these findings is the method used to pretest Section 1 prior to 1986. As previously mentioned, the examinee samples used to pretest listening comprehension items prior to 1986 consisted of candidates at cooperating English language institutes, most of whom were students in English as a second language programs. Thus, for the Section 1 items, not only length of time but also qualitative differences in the calibration samples may have contributed to poor model-data fit. As Table 5 indicates, model-data fit for items pretested after 1986 appeared to be better, and continued use of more recently pretested items should produce better results for Section 1.

The best model-data fit occurred for the structure, vocabulary, and written expression parts. It is unclear why these parts performed better than the others. One explanation may be the fact that the items in these parts are discrete rather than passage based, and so are less subject to context effects and location effects.

Perhaps the strongest evidence of a relationship between fit and item features occurred for Section 3 reading comprehension, where the relationships between the model-data fit measures and the absolute change in item position were relatively strong. Reading comprehension is the final part of the test; it is the longest part, and the items are in sets, linked together by reading passages. As a result, item position can change both within part and within set. Previous research with the TOEFL test (Bejar, 1985; Secolsky, 1989) has suggested the possibility that Section 3 may be slightly speeded when pretest items are administered. In addition, Kingston and Dorans (1984) found evidence of context effects for reading comprehension items when these items were pretested at the end of the test. What seems to occur at the end of tests involving reading passages when there is no penalty for guessing is that examinees who do not have the time to properly read the final passages will tend to respond randomly to the items referring to those passages. The random responses make these items appear more difficult, which is reflected in the IRT item parameter estimates. If a passage is pretested near the end of the test, and is subsequently administered earlier in an operational form, it is likely that the items relating to that passage will appear easier because now all candidates will have time to properly read the passage. The opposite effect can take place when passages are pretested near the beginning of the section and then administered operationally at the end of the section. In this case, the items will appear more difficult in the operational administration than they did in the pretest administration. Both of these effects seem to have occurred with the TOEFL reading comprehension items.

The results of this study suggest that model-data fit for the reading comprehension part of the TOEFL test may be improved if the position of the items relative to the end of the test is better maintained from the time items are pretested to the time items appear in final forms. In addition, extending the timing limits of Section 3 might reduce the potential effects of speededness on the IRT item parameter estimates obtained for the reading comprehension items administered near the end of Section 3.

With regard to two other features of TOEFL items, calibration sample size and the value of the quantity b - 2/a, there appeared to be no discernable relationships with model-data fit. Perhaps a more useful variable to use in the context of this study would have been the psuedo-standard errors of the item parameter estimates at the time of pretesting data calibrations. However, it was not possible to obtain these data within the scope of this study.

Recommendations

Based on the results of this exploratory study, there are two specific recommendations for operational work related to the TOEFL test. First, the feasibility of not using IRT statistics for Section 1 and Section 2 items pretested before 1986 should be investigated. For example, it is possible to carry out IRT equatings for the TOEFL test when some items do not have pretest IRT statistics. This would allow continued use of items pretested before 1986, although the previously obtained IRT statistics would not be utilized for these items. However, whether or not the use of IRT statistics can be discontinued for items pretested before 1986 will depend partly on the availability of the more recently pretested items in the item pools for each of the three TOEFL sections. Second, attempts should be made to preserve the relative position of reading comprehension items between their pretest and final form appearances. This can be accomplished by setting up rules for test developers to use in selecting IRT-pretested reading comprehension items for TOEFL final forms that will limit the relative movement of items from the pretest appearances to their final form appearances.

TABLE 1 Interrater Correlations (r_{12}) and Alpha Coefficients (α) for Item Fit

Month of Administration	Section 1		Section 2		Section 3	
	(r_{12})	(α)	(r_{12})	(α)	(r_{12})	(α)
February	0.94	0.97	0.80	0.89	0.92	0.96
March	0.96	0.98	0.95	0.97	0.92	0.96
May	0.94	0.97	0.95	0.98	0.96	0.98
July	0.95	0.97	0.97	0.99	0.95	0.98
August	0.94	0.97	0.95	0.98	0.90	0.95
September	0.97	0.98	0.86	0.92	0.95	0.98
October	0.94	0.97	0.94	0.97	0.95	0.97
Overall	0.95	0.97	0.92	0.96	0.94	0.97

TABLE 2 Summary Statistics for Item Fit Ratings

Month of Administration	Section 1						Section 2				Section 3			
	State.		Dialog.		Mini.		Struc.		Wr. Exp.		Voc.		R. Comp.	
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
February	5.05	2.54	5.53	2.47	4.13	2.29	4.64	2.13	4.17	1.79	4.55	2.40	5.28	2.15
March	5.35	2.13	5.33	2.38	4.53	2.64	3.50	2.07	5.12	2.29	4.24	1.98	4.03	2.01
May	5.30	2.45	5.07	2.34	6.07	2.37	3.14	1.41	4.58	2.65	4.17	2.32	4.28	2.15
July	5.70	1.92	4.33	2.47	4.20	2.40	3.64	1.65	5.25	2.36	4.28	2.22	5.97	2.46
August	5.80	2.46	4.93	2.52	5.73	2.12	4.79	2.94	4.46	2.04	3.93	2.05	4.83	2.14
September	4.65	2.58	4.80	2.37	6.20	2.24	5.43	2.47	4.75	2.27	4.10	2.30	5.83	2.62
October	4.95	2.31	4.53	2.53	5.00	2.07	3.64	1.60	4.54	2.47	4.07	2.30	4.62	2.41
Overall	5.26	2.33	4.93	2.40	5.12	2.39	4.11	2.18	4.70	2.27	4.19	2.20	4.98	2.36

TABLE 3a Summary Statistics for Weighted Root Mean Squares

Month of Administration	Section 1						Section 2				Section 3			
	State.		Dialog.		Mini.		Struc.		Wr. Exp.		Voc.		R. Comp.	
	$\overline{x}$	sd	$\overline{x}$	sd	$\overline{x}$	sd	$\overline{x}$	sd	$\overline{x}$	sd	$\overline{x}$	sd	$\overline{x}$	sd
February	0.065	0.055	0.071	0.041	0.038	0.022	0.046	0.029	0.042	0.031	0.047	0.021	0.059	0.035
March	0.064	0.056	0.057	0.031	0.057	0.034	0.036	0.019	0.053	0.028	0.043	0.026	0.043	0.019
May	0.063	0.043	0.050	0.029	0.058	0.031	0.037	0.011	0.040	0.025	0.045	0.023	0.050	0.026
July	0.061	0.034	0.052	0.037	0.048	0.029	0.037	0.021	0.056	0.034	0.044	0.023	0.084	0.054
August	0.062	0.026	0.054	0.032	0.060	0.029	0.057	0.040	0.042	0.019	0.041	0.019	0.050	0.025
September	0.061	0.059	0.063	0.050	0.048	0.025	0.054	0.030	0.046	0.028	0.038	0.020	0.076	0.058
October	0.068	0.039	0.050	0.027	0.050	0.021	0.035	0.020	0.036	0.022	0.042	0.021	0.053	0.029
Overall	0.063	0.045	0.057	0.036	0.051	0.028	0.043	0.026	0.045	0.027	0.043	0.022	0.059	0.040

TABLE 3b Summary Statistics for Weighted Mean Differences

Month of Administration	Section 1						Section 2				Section 3			
	State.		Dialog.		Mini.		Struc.		Wr. Exp.		Voc.		R. Comp.	
	$\overline{x}$	sd	$\overline{x}$	sd	$\overline{x}$	sd	$\overline{x}$	sd	$\overline{x}$	sd	$\overline{x}$	sd	$\overline{x}$	sd
February	-0.000	0.077	-0.015	0.066	0.016	0.030	0.017	0.038	-0.011	0.040	0.008	0.033	-0.007	0.058
March	0.006	0.075	0.006	0.050	-0.014	0.053	0.005	0.025	-0.005	0.046	0.016	0.030	-0.016	0.033
May	0.022	0.056	0.005	0.050	-0.034	0.044	-0.006	0.028	0.003	0.036	0.006	0.037	-0.007	0.044
July	0.016	0.057	-0.012	0.047	-0.008	0.041	0.018	0.017	-0.010	0.051	0.021	0.032	-0.021	0.085
August	0.002	0.051	-0.001	0.053	0.003	0.058	0.015	0.054	-0.009	0.030	0.002	0.033	-0.003	0.044
September	0.005	0.074	-0.000	0.071	-0.006	0.036	0.007	0.041	-0.005	0.043	-0.011	0.032	0.011	0.086
October	-0.020	0.064	0.001	0.043	0.025	0.032	0.013	0.028	-0.009	0.034	-0.004	0.037	0.004	0.045
Overall	0.004	0.065	-0.002	0.054	-0.002	0.046	0.010	0.035	-0.007	0.040	0.005	0.035	-0.005	0.060

TABLE 4 Correlations of Item Fit Ratings with Weighted Mean Differences (WMD) and Weighted Root Mean Squares (WRMS)

Month of Administration	Section 1		Section 2		Section 3	
	WMD	WRMS	WMD	WRMS	WMD	WRMS
February	0.00	0.71[1]	0.21	0.77[1]	0.01	0.75[1]
March	0.09	0.68[1]	-0.12	0.79[1]	0.03	0.86[1]
May	-0.01	0.77[1]	0.19	0.73[1]	0.24	0.82[1]
July	0.12	0.80[1]	-0.13	0.82[1]	-0.07	0.76[1]
August	0.14	0.81[1]	0.31	0.79[1]	0.39[2]	0.80[1]
September	0.07	0.70[1]	0.26	0.75[1]	0.17	0.76[1]
October	0.01	0.83[1]	-0.22	0.74[1]	0.42[2]	0.82[1]
Overall	0.06	0.74[1]	0.08	0.76[1]	0.14[2]	0.77[1]

[1] Correlation is significant at the .0001 level.

[2] Correlation is significant at the .01 level

TABLE 5 Summary Statistics for Weighted Root Mean Square (WRMS) and Item Fit Ratings (IFR) by Year of IRT Calibration

Year of Administration	Section 1					Section 2					Section 3				
	N	WRMS	sd	IFR	sd	N	WRMS	sd	IFR	sd	N	WRMS	sd	IFR	sd
1981	21	0.082	0.063	5.95	2.50	5	0.051	0.034	5.80	2.28	6	0.044	0.028	4.33	2.88
1982	13	0.072	0.050	5.92	2.53	7	0.069	0.042	5.43	2.57	13	0.045	0.025	4.92	1.89
1983	37	0.059	0.037	5.81	2.26	22	0.045	0.023	5.09	2.16	10	0.039	0.019	4.60	2.50
1984	29	0.063	0.045	5.31	2.47	26	0.047	0.032	4.77	2.50	46	0.048	0.030	4.52	2.21
1985	76	0.058	0.032	5.25	2.36	29	0.058	0.035	4.83	2.38	65	0.055	0.040	4.52	2.45
1986	77	0.052	0.035	4.56	2.29	49	0.040	0.024	4.06	2.06	91	0.052	0.024	4.63	2.29
1987	81	0.054	0.033	4.99	2.30	96	0.040	0.023	4.25	2.23	124	0.052	0.039	4.51	2.24
1988	16	0.051	0.030	4.19	2.34	32	0.043	0.022	4.44	2.18	51	0.050	0.030	4.76	2.55

TABLE 6 Summary Statistics for Absolute Position Change

Month of Administration	Section 1						Section 2				Section 3			
	State.		Dialog.		Mini.		Struc.		Wr. Exp.		Voc.		R.Comp.	
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
February	0.18	0.12	0.09	0.07	0.09	0.09	0.13	0.08	0.21	0.14	0.13	0.10	0.13	0.04
March	0.15	0.11	0.14	0.08	0.07	0.02	0.14	0.08	0.18	0.15	0.13	0.10	0.11	0.05
May	0.14	0.12	0.11	0.08	0.11	0.06	0.13	0.09	0.20	0.16	0.14	0.11	0.11	0.05
July	0.15	0.11	0.07	0.05	0.21	0.10	0.11	0.08	0.18	0.13	0.14	0.10	0.11	0.12
August	0.14	0.11	0.11	0.08	0.06	0.05	0.11	0.08	0.19	0.14	0.13	0.11	0.14	0.08
September	0.16	0.10	0.12	0.13	0.07	0.05	0.14	0.10	0.24	0.18	0.14	0.10	0.15	0.12
October	0.17	0.10	0.11	0.08	0.14	0.08	0.12	0.11	0.15	0.13	0.14	0.11	0.11	0.08
Overall	0.15	0.11	0.11	0.08	0.11	0.08	0.13	0.09	0.19	0.15	0.14	0.10	0.12	0.08

TABLE 7 Correlations of Absolute Position Change with Weighted Root Mean Squares (WRMS) and Item Fit Ratings (IFR)

Month of Administration	Section 1						Section 2				Section 3			
	State.		Dialog.		Mini.		Struc.		Wr. Exp.		Voc.		R.Comp.	
	WRMS	IFR	WRMS	IFR	WRMS	IFR	WRMS	IFR	WRMS	IFR	WRMS	IFR	WRMS	IFR
February	0.11	0.36	0.23	0.34	0.29	0.67[2]	0.03	-0.10	-0.16	0.09	0.14	0.05	0.63[2]	0.62[2]
March	-0.33	-0.06	0.65[2]	0.49	0.38	0.22	-0.30	-0.18	0.09	0.01	-0.18	-0.31	0.22	0.25
May	0.16	0.01	0.25	0.07	-0.35	-0.27	0.21	0.69[2]	0.29	0.28	0.20	0.26	0.18	0.28
July	-0.14	-0.09	0.17	-0.14	-0.20	-0.31	-0.02	-0.04	0.35	0.26	0.09	0.03	0.56[2]	0.36
August	-0.13	-0.07	-0.35	-0.14	-0.02	0.51	0.02	0.15	0.33	0.65[2]	0.17	-0.04	-0.02	0.18
September	-0.25	-0.32	0.08	0.08	0.16	0.60[2]	0.37	0.55[2]	-0.01	-0.03	-0.19	-0.22	0.68[1]	0.55[2]
October	-0.16	-0.31	0.17	-0.03	-0.27	0.17	-0.54[2]	-0.45	-0.02	-0.02	0.19	0.19	0.82[1]	0.62[2]
Overall	-0.09	-0.06	0.15	0.11	-0.08	0.02	-0.01	0.10	0.11	0.15	0.06	0.01	0.50[1]	0.41[1]

[1] Correlation is significant at the .0001 level.

[2] Correlation is significant at the .05 level.

TABLE 8a Mean Original Calibration Sample Sizes for Best Items

Month of Administration	Section 1		Section 2		Section 3	
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
February	1847	915	1305	485	1353	505
March	1385	537	1250	226	1263	690
May	1643	887	1274	499	1022	234
July	1546	655	1253	212	1210	304
August	2460	1045	1146	207	1151	192
September	1967	1111	1214	199	1366	569
October	2162	1128	1221	172	1199	412
Overall	1855	950	1237	315	1218	452

TABLE 8b Mean Original Calibration Sample Sizes for Worst Items

Month of Administration	Section 1		Section 2		Section 3	
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
February	1322	697	1041	149	1363	606
March	1368	867	1370	322	1053	149
May	1430	699	1351	341	1258	231
July	1613	765	1516	711	1061	174
August	1806	1021	1302	483	1147	239
September	1046	292	1192	276	1259	474
October	2125	1065	1386	315	1195	296
Overall	1520	846	1325	421	1197	368

TABLE 9 Correlations of Original Calibration Sample Sizes with Weighted Root Mean Squares (WRMS) and Item Fit Ratings (IFR)

Month of Administration	Section 1		Section 2		Section 3	
	WRMS	IFR	WRMS	IFR	WRMS	IFR
February	-0.33	-0.32	-0.32	-0.29	0.10	0.01
March	0.37	-0.01	0.07	0.23	-0.20	-0.14
May	-0.26	-0.13	0.03	0.07	0.34	0.45*
July	-0.00	0.05	0.29	0.27	-0.29	-0.30
August	-0.22	-0.29	0.02	0.23	0.17	-0.01
September	-0.24	-0.52*	0.32	-0.05	-0.12	-0.10
October	0.03	-0.02	0.31	0.34	0.09	-0.00
Overall	-0.11	-0.18*	0.09	0.12	-0.03	-0.02

*Correlation is significant at the .01 level.

TABLE 10a Summary Statistics for Quantity b - 2/a for Best Items

Month of Administration	Section 1		Section 2		Section 3	
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
February	-1.91	0.87	-2.38	0.53	-1.37	0.71
March	-2.68	1.20	-1.51	1.09	-2.07	1.02
May	-2.72	1.64	-2.05	0.99	-1.89	0.88
July	-3.43	1.62	-2.49	1.62	-1.82	0.77
August	-1.86	1.03	-2.83	1.50	-2.28	0.97
September	-2.14	0.91	-2.28	0.83	-2.19	1.05
October	-2.14	0.44	-1.97	1.02	-2.29	1.05
Overall	-2.39	1.21	-2.19	1.17	-2.01	0.96

TABLE 10b Summary Statistics for Quantity b - 2/a for Worst Items

Month of Administration	Section 1		Section 2		Section 3	
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
February	-2.51	1.05	-1.03	0.26	-2.50	1.28
March	-2.20	1.08	-1.94	1.40	-2.41	1.24
May	-2.20	1.15	-2.57	1.13	-2.12	0.61
July	-2.21	1.03	-2.05	0.97	-1.78	0.92
August	-2.61	1.88	-1.78	0.98	-2.27	0.80
September	-2.85	1.70	-2.62	1.32	-2.82	1.72
October	-2.09	0.85	-3.78	3.71	-2.31	1.31
Overall	-2.40	1.32	-2.31	1.71	-2.31	1.25

TABLE 11 Correlations of Quantity b - 2/a with Weighted Root Mean Squares (WRMS) and Item Fit Ratings (IFR)

Month of Administration	Section 1		Section 2		Section 3	
	WRMS	IFR	WRMS	IFR	WRMS	IFR
February	0.02	-0.31	0.75*	0.81*	-0.28	-0.50*
March	0.24	0.21	0.02	-0.18	-0.06	-0.14
May	0.23	0.18	-0.02	-0.22	-0.12	-0.14
July	0.44*	0.43*	0.38	0.17	0.14	0.02
August	-0.19	-0.21	0.41	0.39	-0.01	0.00
September	-0.04	-0.25	0.09	-0.16	-0.10	-0.22
October	0.02	0.03	0.08	-0.38	-0.04	-0.01
Overall	0.09	0.00	0.20*	-0.04	-0.04	-0.14

*Correlation is significant at the .02 level.

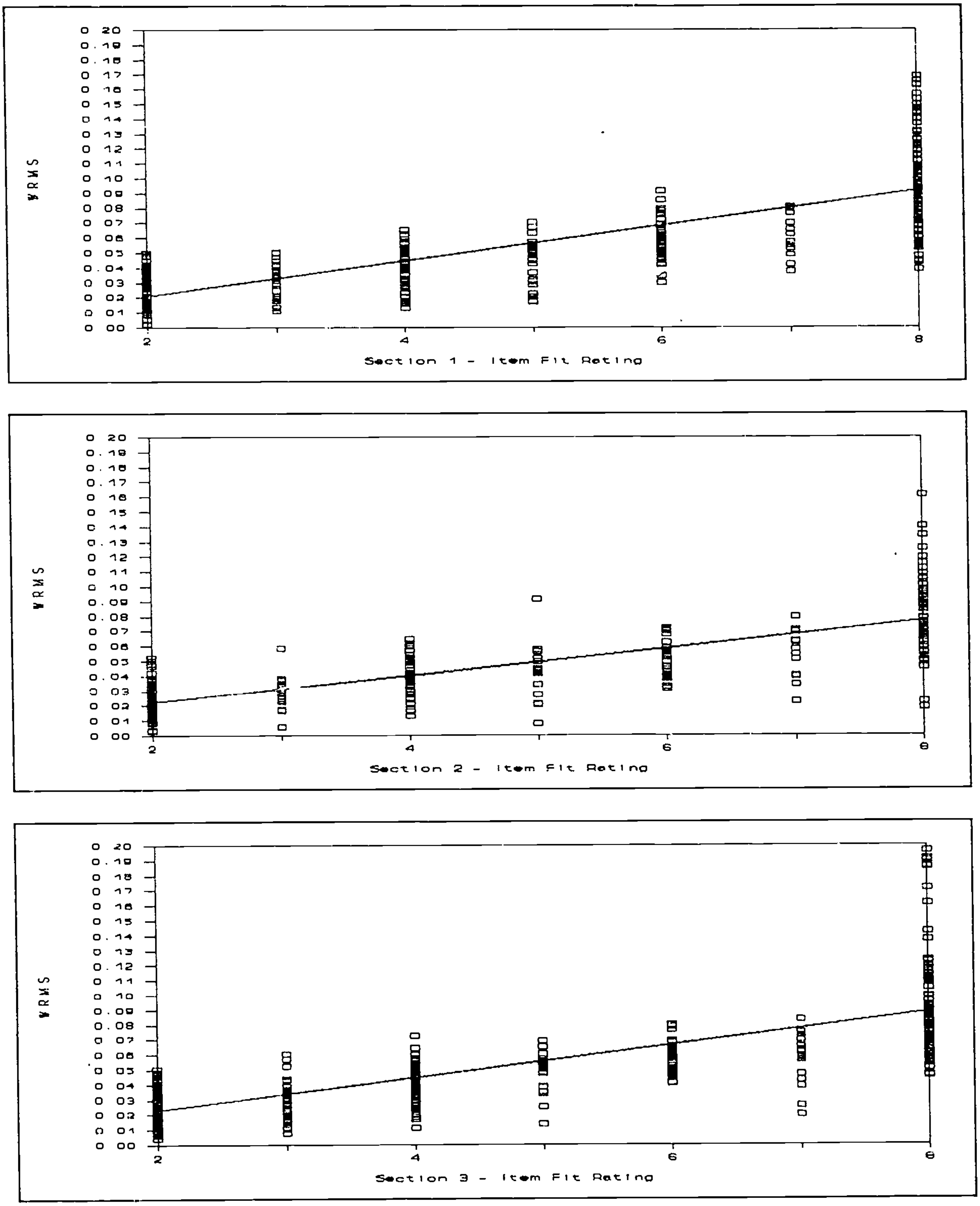
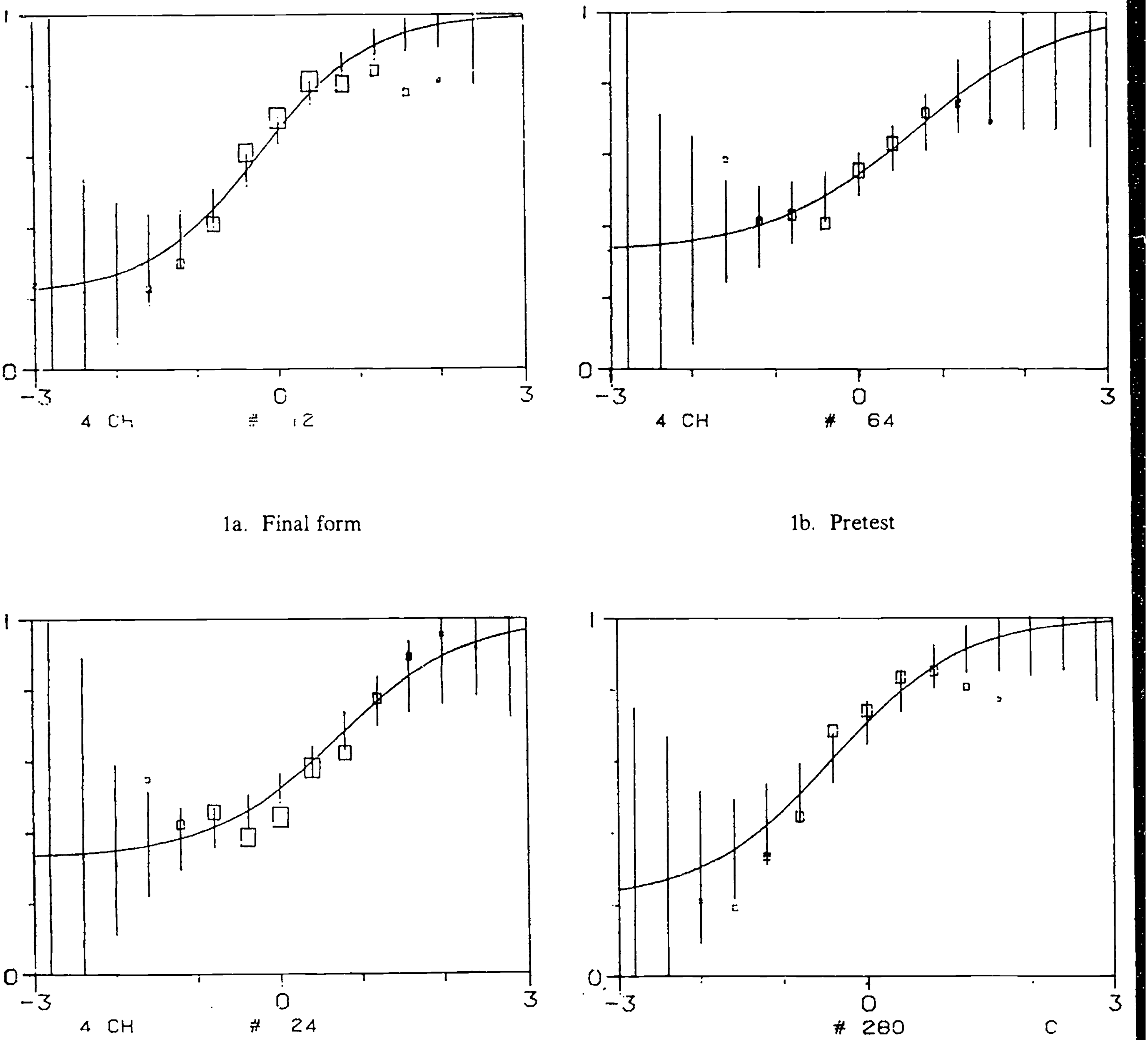


FIGURE 1 Weighted Root Mean Square Plotted Against Item Fit Rating for TOEFL Sections 1, 2, and 3



1a. Final form

1b. Pretest

2a. Final form

2b. Pretest

FIGURE 2 Sample Item Ability Regression Plots for Item Fit Ratings Equal to 7

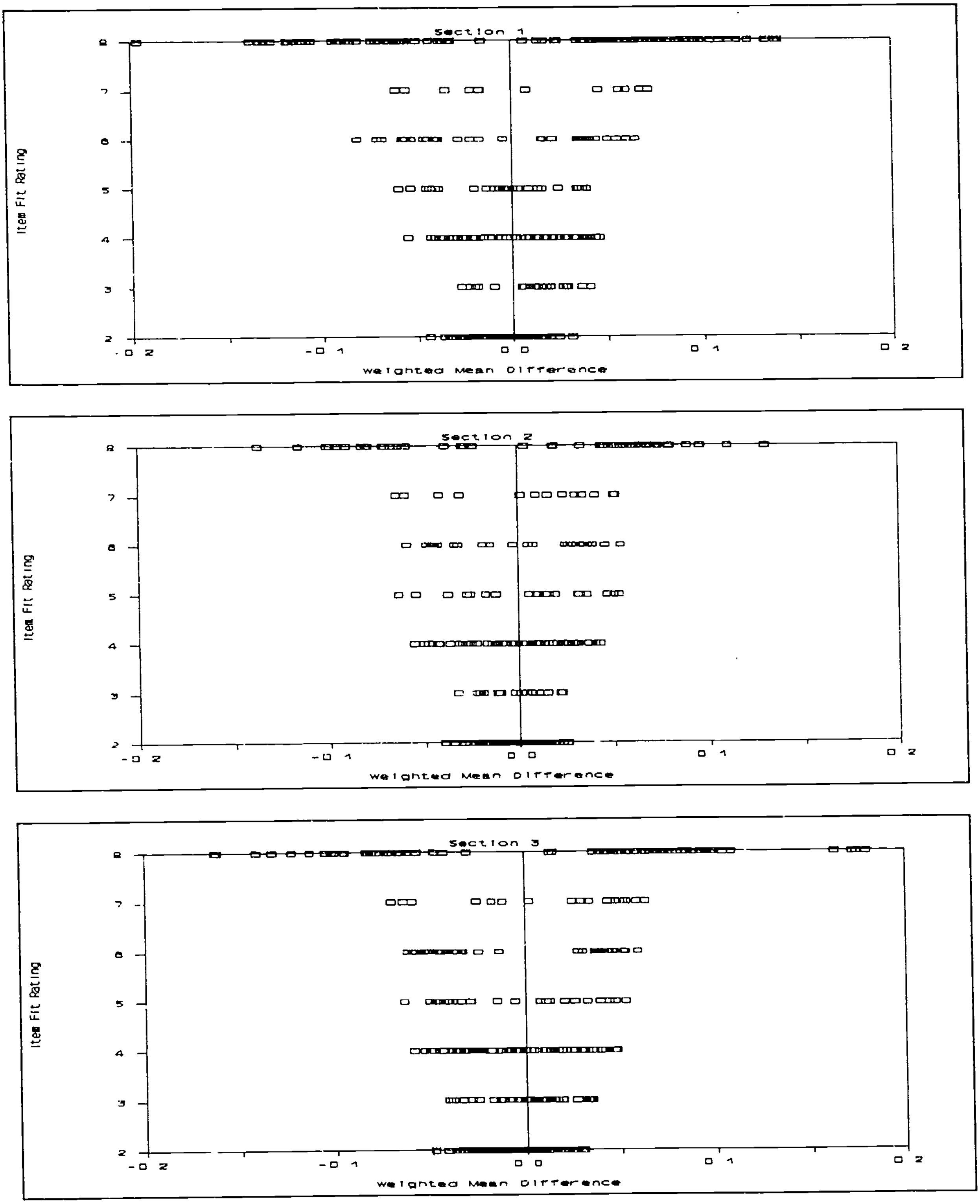


FIGURE 3 Item Fit Ratings Plotted Against Weighted Mean Difference for TOEFL Sections 1, 2, and 3

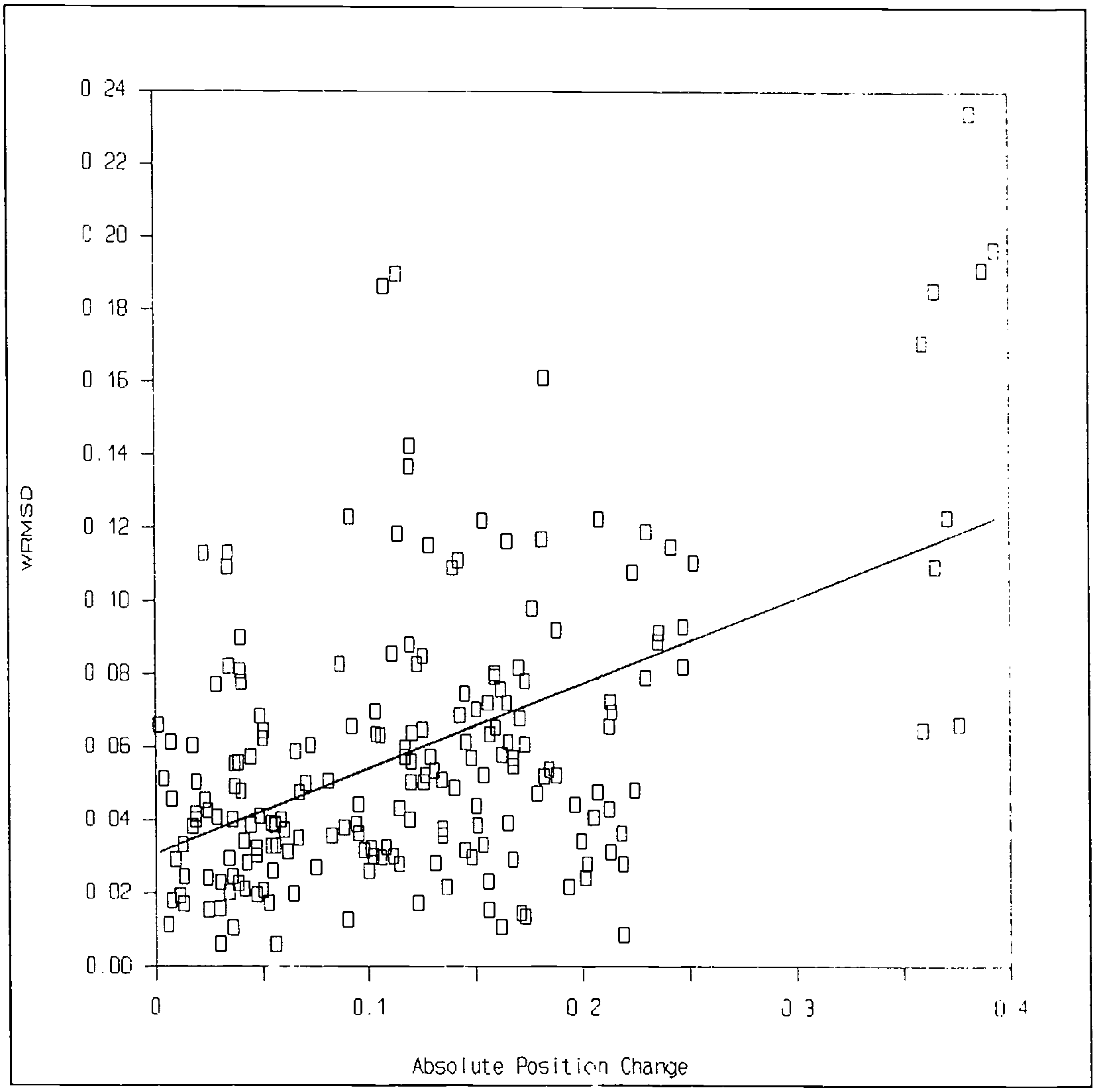


FIGURE 4 Weighted Root Mean Square Plotted Against Absolute Position Change For Section 3 Reading Comprehension Items

Appendix A

Rules for Judging Quality of Item-Ability Regression Plots

For each item, tally points according to the following rules:

1. Add 1 point for each vertical standard error line that does not bisect at least one side of the corresponding box.

2. Add 1 point for each pair of vertical S.E. lines that bisect one side of the corresponding boxes, but that barely do so.

3. Add 1 point for each box that is offset from the corresponding S.E. line with a "noticeable" gap (say, between 1/16" and 1/8").

4. Subtract 1 point if the three boxes around the ability level of 0 indicate good fit to the theoretical line.

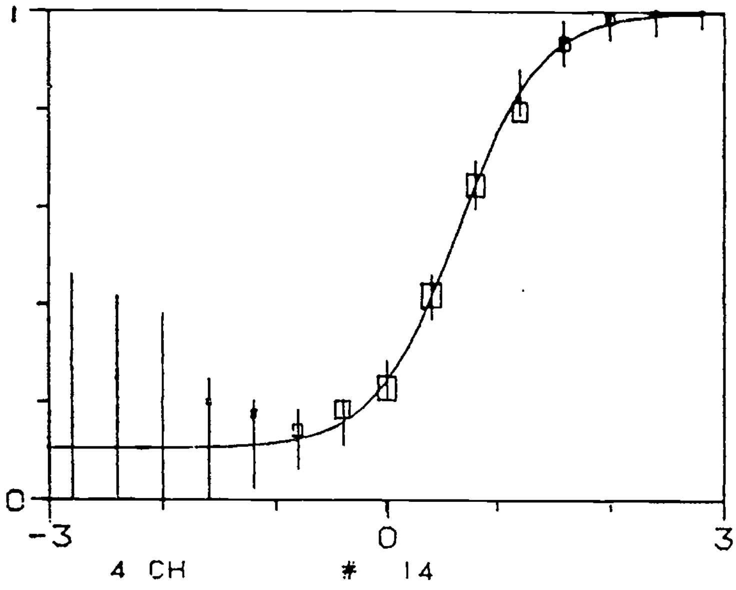
5. Do not count points at extreme ability levels where sample size is "small." (Use your best judgement to define small, and define it consistently for all plots for a given section.)

6. Categorize each item based on the total points tallied:

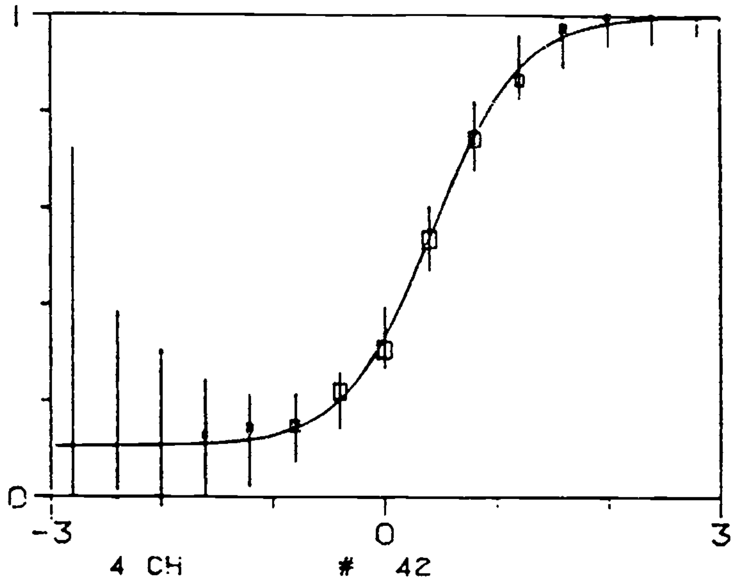
Points	Category	Description
0	1	"Good" Fit
1-2	2	"Fair" Fit
3-4	3	"Marginal" Fit
5+	4	"Inadequate" Fit

Appendix B
Sample Item Ability Regression Plots

Good Item

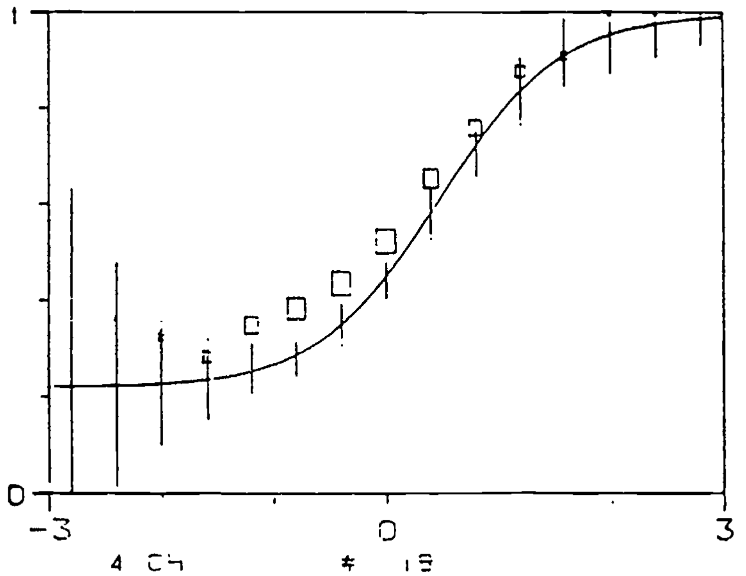


1a. Final form

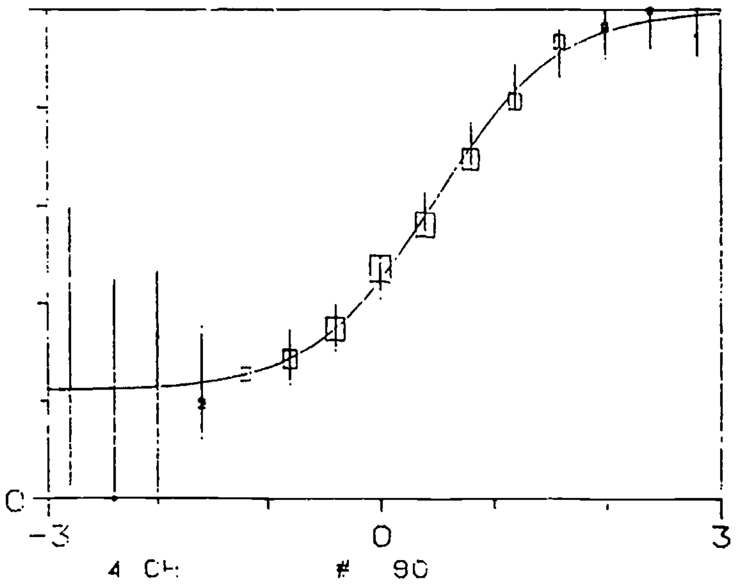


1b. Pretest

Poor Item



2a. Final form



2b. Pretest

References

Bejar, I. I. (1985). *Test speededness under number-right scoring: An analysis of The Test of English as a Foreign Language* (ETS Research Report RR-85-11). Princeton, NJ: Educational Testing Service.

Cook, L. L., Eignor, D. R., & Taft, H. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement* , *25*, 31-45.

Cowell, W. R. (1982). Item response theory pre-equating in the TOEFL testing program. In P. W. Holland & D. B. Rubin (Eds.), *Test Equating* (pp. 149-161). New York: Academic Press.

Eignor, D. R. (1985). *An investigation of the feasibility and practical outcomes of pre-equating the SAT verbal and mathematical sections* (ETS Research Report RR-85-10). Princeton NJ: Educational Testing Service.

Golub-Smith, M. L. (1986). *A study of the effects of examinee native language on TOEFL parameter estimation and equating*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144-149.

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

Hicks, M. (1983). True score equating by fixed b's scaling: A flexible and stable equating alternative. *Applied Psychological Measurement*, *7*, 255-266.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*, 147-154.

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory model fit tool. *Applied Psychological Measurement*, *9*, 281-287.

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

Linn, R. L., Levine, M. V., Hastings, C. N., & Waldrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. *Applied Psychological Measurement*. *5*, 159-173.

Secolsky, C. (1989). *Accounting for random responding at the end of the test in assessing speededness on the Test of English as a Foreign Language* (TOEFL Research Report No. 30). Princeton, NJ: Educational Testing Service.

Shepard, L. A., Camilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 83-138.

Stocking, M. L. (1988). Specifying optimum examinees for item parameter estimation in item response theory (ETS Research Report RR-88-57). Princeton NJ: Educational Testing Service.

Wingersky, M. S., Patrick, R., & Lord, F. M. (1987). LOGIST user's guide (Version 6.0). Princeton, NJ: Educational Testing Service.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. 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.

TOEFL is a program of
Educational Testing Service
Princeton, New Jersey, USA

57906 06443 • Y92M 5 • 275587 • Printed in U S A