DOCUMENT RESUME

ED 382 644 TM 023 078

AUTHOR Boldt, R. F.
TITLE Simulated Equating Using Several Item Response Curves.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-57; TOEFL-TR-8
PUB DATE Jan 94
NOTE 36p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Cutting Scores; English (Second Language); *Equated Scores; *Item Response Theory; Research Methodology; *Sample Size; *Simulation; Statistical Analysis; Test Items
IDENTIFIERS Anchor Tests; Rasch Model; *Test of English as a Foreign Language; Three Parameter Model

ABSTRACT

The comparison of item response theory models for the Test of English as a Foreign Language (TOEFL) was extended to an equating context as simulation trials were used to "equate the test to itself." Equating sample data were generated from administration of identical item sets. Equatings that used procedures based on each model (simple item response, three-parameter logistic, and Rasch models) were accomplished under several conditions and the results were compared. Conditions varied by sample size, anchor test difficulty, and the TOEFL section equated. The largest discrepancies between scores identified as comparable occurred for the three-parameter logistic (3PL) and the modified Rasch models at the lower extreme scores, and for the simple models at the upper extreme score. If it is true that the cut scores for most institutions occur in the mid-score ranges, the findings suggest that the 3PL should not be used if equating samples are substantially reduced from the present size. The other models are promising for small sample equating, with the one-parameter logistic models being most promising. Six tables and six figures illustrate the analysis. (Contains 10 references.) (SLD)

TOEFL®

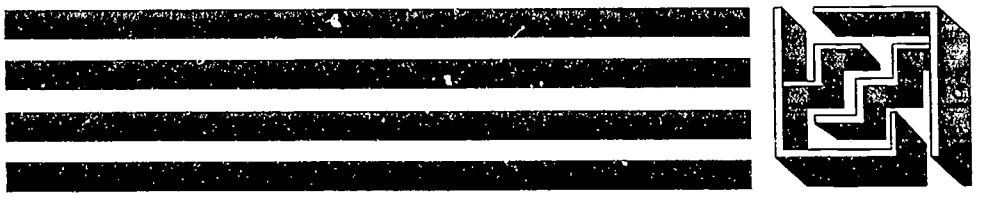
January 1994

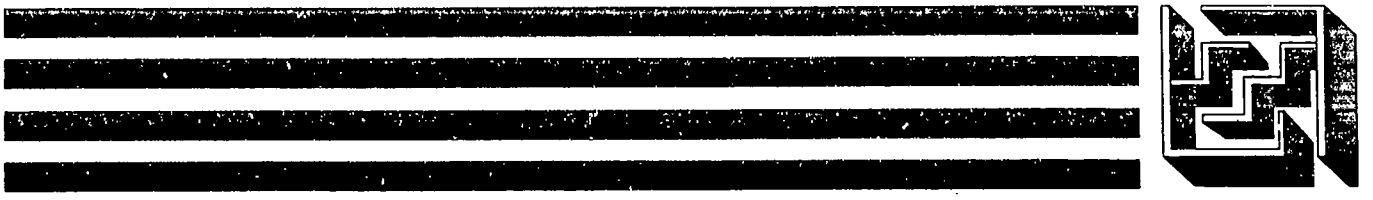
Technical Report

TR-8

Simulated Equating Using Several Item Response Curves

R. F. Boldt





SIMULATED EQUATING USING SEVERAL ITEM RESPONSE CURVES

R. F. Boldt

Educational Testing Service
Princeton, New Jersey

RR-93-57

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

ABSTRACT

Previous research has shown that a single factor, factor being used in the sense of factor analysis, gave a very good account of item covariances within TOEFL® sections. This result is consistent with the assumption that the product of a person parameter and an item parameter models the probability that the person would pass the item. The assumption forms a simple item response model. A subsequent cross-validation study using this model supported the efficacy of that assumption by predicting item success as accurately as did the 3-Parameter Logistic model, and a modified Rasch model. The purpose of the current study was to extend the comparison of models to an equating context. "Equating" is a statistical process that identifies comparable scores from parallel tests administered to different populations. In an operational context, equating serves to facilitate comparison of scores generated on different forms of a test.

The present study consisted of simulation trials designed to "equate the test to itself." That is, equating sample data were generated from administration of identical item sets. It is useful to do this as a test of model validity, because if the same item sets are used to equate, an accurate equating would identify equal scores as comparable. Discrepancies between comparable scores signify error model misfit or random error.

Equatings that used procedures based on each model were accomplished under several conditions and the results were compared. The conditions varied by sample size, anchor test difficulty, and the TOEFL section equated. In order to compound the difficulty of the equating task, results were based on equating samples that were mismatched in performance on a correlated measure.

Most discrepancies between comparable scores were largest at the extremes. The largest discrepancies between scores identified as comparable occurred for the 3PL and modified Rasch models at the lower extreme scores, and for the simple models at the upper extreme score. For the 1,000-case sample, most were in fractions of score points. As expected, 3PL equatings exhibited the largest discrepancies for the 100-case sample. The simple item response model yielded the most discrepancies that were in excess of the standard error of measurement, in part because with that model the maximum discrepancies occurred at the top of the score range, where the standard errors of measurement approach zero. Imposing an upper bound on the probability of correct response in the simple model markedly reduced its errors.

TOEFL scores are used for educational decisions. If it is true that most institutions' cut scores occur in the mid-score ranges, the present study suggests that 3PL should not be used if equating samples are substantially reduced from the present size. The other models are promising for small-sample equating, with the one-parameter logistic models being most promising.

The Test of English as a Foreign Language (TOEFL®) was developed in 1963 by the 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 data confidentiality will be protected.

Current (1992-93) members of the TOEFL Research Committee are:

James Dean Brown	University of Hawaii
Patricia Dunkel	Pennsylvania State University
William Grabe	Northern Arizona University
Kyle Perkins (Chair)	Southern Illinois University at Carbondale
Linda Schinke-Llano	Millikin University
John Upshur	Concordia University

Table of Contents

List of Tables

List of Figures

Introduction

The purpose of this report is to explore self-equatings of TOEFL sections using several item response curves. Some of the conditions were unfavorable to successful equating. Hence, the study provides a comparison of how well the curves work for equating under difficult conditions.

Each equating used groups that were selectively mismatched in score on a correlated variable. Other conditions under which the equatings took place are as follows: (a) the representativeness of the equating subtest (difficult items, easy items, and a representative set of items) was varied, (b) different sizes of equating samples (1,000 and 100 examinees) were used, and (c) different TOEFL sections were equated. For each TOEFL section, four subsets of items were used as equating tests: (i) the odd-numbered items (producing a representative half-length test), (ii) the easiest half of the items, (iii) the most difficult half of the items, and (iv) all of the items.

The six item response curves used were of two general types. Three of the curves used a logistic function of ability to express the probability of item success, and three curves assumed that whatever functions of ability drive the item response curves are proportional across items. This second type of response curve uses relatively few parameters whose estimation requires only simple numerical processes. The purpose of this study was to compare the efficacy of the two types of functions.

The study consisted of simulations that involved "equating the test to itself." For example, where test A and test B were equated, the equating process calculates, for each test, the expected score, given a particular level of ability (as defined by the equating model). Scores for tests A and B "correspond" if they are the expected scores for the same level of ability. If A and B are exactly parallel, then equal scores should correspond. In this study, the equated tests were indeed exactly parallel, as they were the same test. That is, the equating sample data used identical item sets and were randomly using data from one TOEFL administration. Thus the tests equated are exactly parallel, and discrepancies between comparable scores signify model misfit or random error. Different equating methods based on different item response curves could lead to different sets of score correspondences, and the research question was to learn how well the different item response curves produced correct correspondences under the various conditions.

Background

A prior study, which stimulated the present research, attempted to find and identify clusters of examinees with similar patterns of success on TOEFL items (Boldt, 1989). The intent of that study was to (a) identify common item-success patterns, (b) sort examinees into groups whose performance best matched the common patterns, and (c) compare the groups' background data statistics. For example, one hypothesis to be examined was that cluster membership would be related to previous language experience. The study entailed factor analyzing item covariances with the expectation that the occurrence of multiple groups would be reflected in multiple factors.

A surprising outcome of the 1989 study was that no major group effects were evident within the sections of the TOEFL test; each set of item covariances was very well approximated by a single factor. This result suggested a very simple mathematical assumption to use in an item response model (Boldt, 1989), an assumption that was used with the second type of model mentioned above. This assumption was that the probability of success of a person with an item equals the product of a parameter that characterizes the person and a parameter that characterizes the item. Combining a single item parameter and the person parameter through simple multiplication is the feature that makes the assumption so simple. Some other item response models use several types of item parameters and relate such parameters to the probability of item success using more complicated functions (e.g., logistic).

The model suggested by the 1989 study was simple, and because it fit TOEFL data, it was potentially useful. But the phi-coefficients used in that study are not used in equating; even if phi-coefficients are consistent with a single factor model, other statistics might not be equally consistent with that model. Therefore a further validation study of the model was undertaken by Boldt (1992). That study evaluated the efficacy of the model for predicting item responses, test scores, and test score means and standard deviations. Only very simple computations were needed--a few sums and ratios--for estimating parameters and forecasting item results.

The "predictions" were made for three item response models and their accuracy was compared. The models comprised the simple one described above, a modification of the Rasch model, and the 3-parameter logistic model (3PL), which is currently used for TOEFL section equating. The first two models are simpler than the 3PL model in that they feature approximately one parameter per item, rather than three as in the 3PL model. These models were described in detail in the Boldt (1992) paper. The important outcome of that project was that predictions from the three models were about equally accurate. That is, the simple

models performed as well as the more complex 3-parameter logistic model. As a result, this study explored their consistency with types of statistics more directly related to maintaining score comparability across forms.

One way to conduct such exploration is to administer a series of forms to different populations, starting and ending with the same form. Where this design is used, the correct transformation for changing scores on the starting form to scores on the ending form is known--scores on the starting form should be equivalent to scores on the ending form because they are the same form. Several equating methods can be compared and the model that leads to best agreement with that known transformation can be identified. In practice, however, such a study requires data from operational administrations. Unfortunately, the necessary readministration needed for self-equating is counter to TOEFL policy, which firmly precludes repetitive operational administration of test items. For this reason, the present study simulated TOEFL administrations of parallel forms by drawing item data samples from the same source population. That is, each sample of data represented the product of an administration of TOEFL.

Method

The database from which the samples were selected comprised 5,000 cases selected at random from the cases used for the May 1992 equating. Two sample sizes were used: 1,000 and 100.

Samples. Two samples were needed for each equating, and the models were to be tested by using mismatched samples. The amount of mismatching was determined by reference to raw score means and standard deviations from a dozen previous administrations. Samples for the present study were chosen so that the means approximated the highest and lowest means from the dozen previous administrations, and the standard deviations were similar to those of the samples with the extreme means. To do this, the cases for a sample being used to equate a given section were selected based on scores derived from other TOEFL section scores. Using a separate variable in this way avoids curtailing the errors of measurement in the equating data. The selector for the Section 1 samples was the Section 3 score with 20 being the lowest score for the high group and 47 being an upper bound for the low group.[1]

[1] The TOEFL consists of three sections with raw score ranges from 0 to 50 for Section 1, 0 to 38 for Section 2, and 0 to 58 for Section 3. Thus, the source for the Section 1 high sample comprised those examinees in the 5,000-case database whose Section 3 scores ranged from 20 to 58; the source for the Section 1 low sample comprised those examinees in the 5,000 case database whose Section 3 scores ranged from 0 to 47. The selectors for the other two sections and the bounds were as follows: the Section 2 selector was the Section 1 plus the Section 3 score, with lower and upper bounds of 26 and 80; the Section 3 selector was the Section 1 score, with lower and upper bounds of 20 and 46.

The highest and lowest mean raw scores and standard deviations from the dozen operational administrations follow: (a) the Section 1 low mean was 30.7(30.4) with a standard deviation of 9.2(9.6), and the high mean was 34.0(34.2) with a standard deviation of 9.6(9.7); (b) the Section 2 low was 24.9(24.5) with 7.2(6.1), and the high was 27.2(27.1) with 7.1(7.0); and (c) the Section 3 low was 37.0(37.0) with 11.1(11.1), and the high was 39.3(39.6) with 10.3(10.4). The numbers in parentheses are the means and standard deviations of the portions of the 5,000-case database that satisfy the cut scores. That is, the highest Section 1 mean from the dozen operational administrations whose data were examined was 30.7, with the standard deviation from that administration being 9.6; for the 5,000 cases in the database whose Section 2 scores were 20 or above, the mean score for Section 1 was 30.4 and the standard deviation was 9.7. The selectors and cut scores were chosen, using trial and error, to provide the close match that can be noted between the extreme means from previous administrations and the corresponding means from the truncated 5,000-case database, which are provided in adjacent parentheses. The standard deviations that resulted from the truncations, which are recorded in parentheses, closely approximated those from the operational administrations, with Section 2 having the greatest discrepancy. The restriction in variation noted for Section 2 is undoubtedly due to the truncation. However, using a high enough cut score to obtain a good match between standard deviations would have yielded a poor match between means.

Equating Items

Four sets of equating items were used. One set, which consisted of all the items in the test, was used only with the 1,000-case sample and provided a minimal error condition. This equating set represents the tests exactly.

A second set of equating items consisted of the odd-numbered items. This set of items was used with 100-case samples to provide a set of equating items that are representative of the test. These odd-numbered items comprise a half-length test that is broadly representative of the total test. This set of equating items is referred to as the "Broad Equater." It is more representative of equating sets than either the "Hard Equater" or the "Easy Equater," which are described in the following paragraph.

The third and fourth sets of equating items were half-length tests containing the most and least difficult items. The item difficulty mismatch between these half-tests and the total test surely exceeds the amount of mismatch that would occur in practice--probably even from major blunders in selecting equating items. These sets of equating items are referred to as the "Easy Equater" and the "Hard Equater."

Equating Item Response Models

Six procedures for computing the correspondence between equated test scores were used, each based on a different assumption about the item response curve. As previously noted, three of the procedures used some form of the logistic function and three used the assumption that item response curves are proportional through at least some part of the ability range.

The logistic-based curves were all implemented using the PC version of LOGIST® (Wingersky et al, 1987). They differ by the constraints imposed on relations among item parameters. LOGIST is a computer program that infers an ability score, which is often referred to as "theta," and three item parameters, which are often referred to as "a," "b," and "c." "Theta," which relates to examinees, is considered the parameter that indexes ability. The "a" parameter indexes the maximum rate of change of the probability of item success with change in theta; "b" is a location parameter indexing item difficulty and is the theta value at which the rate of change of the item response curve with respect to change in theta obtains its maximum value; "c" is the value the item response curve approaches as theta approaches minus infinity.

For the item response curve used by the TOEFL program, referred to as "3PL," LOGIST estimates a, b, and c for each item. Thus the number of parameters to be estimated is approximately 3 times the number of items, plus the number of examinees.

In contrast with the 3PL model, the Rasch model sets all "c" values to zero and all "a" values to an arbitrary constant. With the Rasch model the only parameter free to vary across items is the "b" parameter. Thus the number of parameters to be estimated is equal to the number of items plus the number of examinees. This model is referred to as "RSCH."

A model that is intermediate to 3PL and Rasch treats the "a" and "b" parameters the same as with the Rasch model, requires the "c" parameter to be the same for all items, but allows the "c" parameter to be other than zero. The value of the "c" parameter is estimated. In this study, as in previous studies, that value proved to be approximately .2. This modification of the Rasch model is referred to as "MODR."

Though the earlier project (Boldt, 1992) used only the 3PL, modified Rasch and unbounded PIRC models (see immediately below), the Rasch model was added because of its popularity and simplicity. It is a one-parameter model, and has the advantage that a good deal of related statistical theory has been developed. The addition of the Rasch model would not have occurred had it not been convenient to use, because it had

previously been eliminated as a candidate for TOEFL equating. There are a large number of comparison studies of the Rasch and 3PL models. An entree to this literature can be found in the bibliographies by Divgi (1986, 1989), who favors 3PL, and Henning (1989), who favors Rasch.

The fourth model used was the new, simple model referred to earlier in this report. In its original formulation, Boldt (1992) assumed that the probability of item success was proportional to the value of the item parameter value multiplied by some monotonic increasing function of ability throughout the ability range. The item response curves are proportional, hence the acronym "PIRC" for "proportional item response curve." Under this model, item parameters are proportional to the item difficulties, and person parameters are proportional to the number right scores. This model will be referred to as "PIRC," and in text the acronym may be preceded by "unrestricted."[2]

A fifth model, the second of the PIRC-related models, was used because it was noted in the 1992 study that the probability of response success estimated using parameters from the previous (fourth) model could exceed one in a few cases. It was also noted in that study that the item success rate far exceeded 90% in instances where the estimated item success probability exceeded one. Thus, a computed item success "probability" in excess of one forecast a very high likelihood of item success, even though a probability in excess of one is really out-of-range. Hence for this model, computed "probabilities" of item success that were in excess of a value of one were reduced to a value of one when used in equating computations. To be consistent with this practice, the estimation procedure treated those instances where the estimated probability of success exceeded one as missing data. This model will be referred to as "PIRC-1," where the 1 refers to the value of the upper bound of the estimated probability of item response.[3]

The sixth model, and the third of the PIRC-related models, was introduced to avoid the use of item success probabilities of

---

[2] A constant of proportionality was chosen, norming the vector of item parameters to an arbitrary length. Item parameters consisted of item difficulties prorated to have an arbitrary sum of squares; person parameters that were consistent with this norming were chosen.

[3] For this version of the PIRC model the estimation procedure began by using item response parameters that were proportional to item totals. As with the previous version of PIRC, the item and person parameters can be multiplicatively adjusted in a compensating fashion, and the sums of squares of item parameters were standardized at each iteration, thus particularizing the scale.

The iteration by which estimates were derived is as follows. Using initial estimates of parameters, row and column sums of item scores were cumulated, except where estimated item scores exceeded one, then estimated item scores were substituted for the actual item score. Iteration was necessary because the change in parameters could affect those instances in which the estimated probability of item success exceeded one. The iterations converged quickly, in the sense that successive iterations returned proportional item parameters in a very short time. It was much faster than any of the LOGIST-based procedures.

one. That the occurrence of some event has probability one implies that the event is certain to occur. But certainty is impossible to demonstrate for a number of reasons, hence it is more appropriate that probabilities in our model not reach one. For this reason the sixth model, like the fifth model, assumes that the probability of item success is estimated by the product of item and person parameters, as long as that estimate does not exceed a bound. In contrast with the fifth model, the bound in the sixth model is less than one, and it is estimated, rather than arbitrarily chosen. The principle on which the estimation was based was: When the product of person and item parameters exceed the bound, the success rate should equal the bound. Thus, if .95 were taken as the upper bound to item success, then .95 was assigned as the probability of item success to all item-person combinations for which the product of item and person parameters exceeded .95, and, in addition, .95 of the responses to which that upper bound was assigned would be correct. This model will be referred to as "PIRC-e," where the E refers to the fact that the value of the upper bound of the probability of item response was estimated.[4]

Equating Procedures

Once the samples had been drawn, equatings using each model were carried out in a three-step procedure. First, parameters were estimated for the high and low samples using procedures appropriate to the particular equating model.

Second, item parameters were converted so that examinee parameters would be on a common scale. This step was necessary because the scale of resulting parameters is, to some extent, arbitrary. For logistic parameters this conversion is accomplished using the procedure described by Stocking and Lord (1983), which can be invoked using LOGIST software. For PIRC the conversion is accomplished by prorating the set of item parameters determined on one sample so that they have the same sum of squares as the set of parameters determined on the other sample (keeping in mind that the items are common to the two samples).

Third, a broad range of values of examinee parameters was chosen, and expected test scores were computed for each value

---

[4] This version of the PIRC model used two sets of iterations for estimating parameters, one occurring within the other, as follows. Beginning with an arbitrary choice of upper bound, (.95 was a good guess for any TOEFL section in these data), item and person parameters that were calculated using a procedure like that of the first PIRC model, except that estimated item scores were substituted for item scores when the estimates exceeded the upper bound (.95 if that were the guess). This procedure returned a set of item and person parameters that could be used to calculate the proportion of successes for those cells where the product of person and item parameters was in excess of the upper bound. If that proportion exceeded the upper bound, then the upper bound was increased slightly and the procedure was repeated. For each estimate, only a few repetitions of this procedure were needed to identify upper bounds that led to matching proportion pass for those responses to which the upper bound probability was assigned.

which, if graphed, yielded a test-characteristic curve. Scores from the two tests are "equated" when they arise from the same value of the examinee parameter. This procedure is called "theta equating" in the case of LOGIST (Cook & Eignor, 1991; Lord, 1980).

In the present study, equal scores should be equated because the item data are from the same tests. Hence, equal examinee parameters, which imply equated scores, should also imply equal scores. If graphed, equated scores should yield a linear plot that, extended as necessary, passes through the origin. Such a line will be referred to as the "ideal line."

Several characteristics of each set of equated scores and related statistics were developed as follows:

(1) Discrepancies at the medians of the 5,000-case databases were computed. "Discrepancy" refers to the absolute value of the difference between equated scores, consistent with the notion that the larger the value, the greater the departure of the result from that expected, if the equating were perfect.

(2) A maximum discrepancy was found for every equating, as was the average of the raw scores used to compute that maximum.

(3) Scale score discrepancies implied by raw score discrepancies were computed for selected equatings, as were standard errors of measurement (SEM). SEM indexes score variation due to unreliability, and was included as an aid to interpreting the magnitude of the departures of the equating line from the ideal. Score levels were taken into account during all scale score conversions and SEM references. Raw score to scale score conversions were taken from the test analysis report for the form used in the study (ETS, 1992).

Results

The results of the equatings are presented in Tables 1, 3, and 5. Interpretations of the entries in these tables differ only as to the TCEFL sections of interest. Tables 2, 4, and 6 from the Test Analysis are included, to facilitate interpretation of the results. These tables, Tables 5 through 7 in Appendix B of the test analysis (ETS, 1992), give converted scores and conditional SEMs (C.S.E.M in the test analysis) for each raw score value for the section to which the table applies.

The first data line in Table 1 indicates that, using Section 1 data from the 1,000-case sample and all of the items as the equating set, the 3PL model yields a discrepancy of .05 at the

median of the 5,000-case database, a maximum discrepancy of .46, and an average of 11.6 for the equated scores that yielded the maximum discrepancy. The results in this and all tables are given in raw score points.

Also in the upper section of Table 1, for "Easy items in the equater," the first data line indicates that, using Section 1 data from the 1,000-case sample and the easy set of equating items, the 3PL model yields a discrepancy of .14 at the median of the 5,000-case database, a maximum discrepancy of .37, and an average of 11.7 for the equated scores that yielded the maximum discrepancy.

The bottom section of Table 1 differs from the upper section in that the 100-case sample was used. Also, all of the items served as the representative set for the 1,000-case sample, but only the odd-numbered items served as the representative set (broad equater) for the 100-case sample.

Table 1 reveals that at the median of the 5,000-case database, which occurred between raw scores 33 and 34, the largest discrepancies were associated with the PIRC model, with the greatest being .88 (PIRC-e, broad equater, 100-case sample), i.e., less than a raw score point. That this difference is small relative to SEM on the TOEFL Section 1 scale can be seen as follows. At the raw score value of 33, Table 2 reveals that one raw score point--which is greater than .88 of a raw score point--translates to about a half a TOEFL scale score point, which is in turn considerably less than the SEM obtained at that level (approximately two and a third scale points).

Figure 1 displays the Section 1 equating in which the .88 discrepancy occurred. In this figure, and in all that follow, the scores on both axes are expected raw scores. The abscissa is the raw score for the low group; the ordinate is the raw score for the high group. The points are the two raw scores expected at various levels of ability. A perfect equating would lie along a straight line from the origin (0,0) to a point where the x and y coordinates were both 50 (50,50). Note in the figure that the equating line does not pass directly through the point (30,30), even though it passes through the origin. Thus it is departing from the ideal line as one moves up and to the right along the line. Between 33 and 34, the median of the 5,000-case database, the discrepancy was .88 of a raw score point, and that is the largest discrepancy noted in all the equating for Section 1 at the database median.

Figure 1 displays several aspects of equating using the PIRC-e model. Note first that the line does not extend to a score of 50 on either axis. This is a consequence of an upper probability bound less than one: The item probabilities whose sums are plotted in Figure 1 are constrained by the model to be

less than one. Therefore the sum of probabilities, i.e. the expected test score given ability level, is, regardless of the ability level involved, less than the number of items--50 in Section 1. A similar effect will be seen in the plots for 3PL and MODR (Figures 2 and 3), but at the lower end of the equating. In those models, because the expected value for any item is greater than zero, the sum of the item probabilities of success can never be zero. Hence the 3PL and MODR cannot reach the origin. However, in both of these models the item probability of success approaches one at the upper level of ability, so they both approach the maximum score.

With the PIRC-e model, as with all the PIRC models, the equating line passes through the origin and is relatively straight along much of its path. When the upper bounds for the probability of correct response are not equal, the path of the equating line does not return to the ideal line, but would do so if those bounds were equal. It will be seen that for the unrestricted PIRC model the path of the equating line is simply a straight line through the origin, not necessarily coincident with the ideal line. For PIRC-1, the equal upper bounds for the probability of correct response always bring the path back to the ideal line at its upper right-hand limit.

A number of maximum discrepancies in Table 1 exceed the SEM when converted to the section scale. These are as follows: (a) all equatings using the PIRC-e model, (b) all equatings using 3PL in the 100-case sample, (c) the equatings using the unrestricted PIRC model and all the items as the equating set for the 1,000-case sample, and the broad equater set for the 100-case sample. The Rasch and MODR models performed well most consistently.

The largest maximum discrepancy in Table 1 was for the 3PL model using the easy item equater with the 100-case sample. Figure 2 displays this equating. As expected, the path of the equating line never reaches the origin, but it does approach the (50,50) point at the upper end. However, the lower asymptotes for the forms are different, therefore its maximum error for the plot is at the lower end of the path.

In Table 3, the largest discrepancy at the median of the 5,000-case database was 1.28 for the broad equater using a sample size of 100 and equating model PIRC-e. The median of the 5,000-case database was between 26 and 27. As can be seen in Table 4, a discrepancy of 1.28 on the raw score scale at the raw score level of 26 implies approximately .6 of a point on the section scale--less than a quarter of an SEM. In contrast with Section 1, conversion of the maximum discrepancies to the standard score scale yields discrepancies that are smaller than the SEM, but other results are similar. PIRC-e yields the largest discrepancies for all samples, and 3PL yields large discrepancies for the 100-case sample. PIRC-1 yields relatively large

discrepancies for the broad and easy equating sets for the 100-case sample. As in Table 1, the Rasch and MODR models performed well most consistently.

In Table 5 the 3PL model yielded the largest discrepancy at the median of the 5,000-case database for Section 3, occurring between raw scores 39 and 40, which was .89. This discrepancy occurred using the easy equater with the 100-case sample. Table 6 reveals that one raw score point difference between 39 and 40 yields approximately half a point difference on the Section 3 scale, a value considerably less than the SEM obtaining at that level, which is approximately two and a half.

As with the other tables, in Table 5 large maximum discrepancies occur when using PIRC-e. None of these discrepancies is large enough, when converted, to exceed the SEM when the 1,000-case sample is used, but converted discrepancies obtained using the 100-case sample do exceed the SEM. No other converted discrepancy exceeds the SEM. The 3PL model again yielded relatively large maximum discrepancies for the 100-case sample. The Rasch and MODR models also yielded large discrepancies but in a range where the change in scale score with a point change in raw score is small, and the SEM is large. That is, when discrepancies for Rasch and MODR were converted to the section scale, they were smaller than the SEMs.

Figures 3 through 6 display equatings for which the maximum discrepancies did not exceed the SEM, but they are included to display the characteristics of models not previously displayed. The particular equatings displayed were those in which the models' largest maximum discrepancies occurred.

Figure 3 presents an equating for Section 3 using the MODR model. Like 3PL it fails to extend the equating line to the origin, but displays the full upper test score range to the (58,58) point, there being 58 items in Section 3. The maximum discrepancy occurs in the lower score range.

Figure 4 displays an equating of Section 1 data using the PIRC-1 model. This plot reaches both the origin and the (50,50) point as expected. Note that the path of the equating line is essentially straight after departing from the origin. It increasingly departs from the ideal line as the score level increases, and one can note that it definitely misses the point at (40,40). However, the upper bound of one for the probability of correct item response brings the curve back to the ideal line, which it meets at (50,50). PIRC-1 was the most successful of the PIRC models in this study.

Figure 5 displays an unrestricted PIRC equating for Section 1. Note that it is straight, passing through the origin and reaching its maximum departure from the ideal line at the upper right-hand corner of the plot.

Figure 6 is a plot of RSCH equating for Section 3. Like PIRC-1 the equating line begins at the origin and moves up to the (50,50) point, departing slightly from the ideal line in the middle score range. This departure can best be noted at the (20,20) and (30,30) points. RSCH was the most successful model in this study.

Summary and Discussion

This report is the third in a series that explores the possibility of using the PIRC model for simplified equating purposes. The first study, (Boldt, 1989) was an attempt to detect groups with different patterns of item difficulty, and to relate group membership to demographic variables using latent structure analysis. The weakness of the group structure detected in that study stimulated the surmise that proportional item response curves (PIRC) might be useful with TOEFL data. Use of the PIRC model offered simple numerical procedures and, because it requires only one parameter per item, offered reduced clerical demands and smaller pretest sample sizes. The smaller pretest sample sizes could, in turn, allow increased yields of pretested items for the same pretest examinee volume.

Further evaluation of the PIRC assumption was provided in a succeeding report (Boldt, 1992), in which the efficacy of several models, including PIRC, for predicting examinees' item success and raw score statistics was cross-validated. Besides PIRC, the study used the 3PL and modified Rasch item response curves. The models displayed approximately equal cross-validity. Hence research continued, using tasks more closely related to equating.

The third and present project was originally conceived with a different self-equating design than the one used here. That design, one that is more commonly used to examine the stability of equating, would have required administering a test form and then equating that form to itself through several other test forms. Operational data had to be used, if the study were to be feasible, therefore reusing the original form operationally was required. But this reuse would have violated TOEFL policy. The current simulation was therefore performed instead.

The present study differs from the usual self-equating study, which does not intentionally use extreme sample variations. But the present study took special pains to use samples that were mismatched in ability to a degree that represented extremes of variation normally encountered across administrations. This was done because equating under ideal conditions of equivalent samples seldom occurs; a study done under such ideal conditions would have very limited implications for the models being compared.

Aside from variation in samples' score levels, factors that can affect equating errors are sample size and the extent to which the equating set represents the total tests. These factors were minimized when all the items were used as the equating set with the 1,000-case sample. Under these conditions the Rasch and MODR models were superior. In defense of 3PL it should be noted that 1,000 is not a large number of cases for that model--indeed it has been regarded as a minimum. As for the PIRC models, it seems unlikely that they would be more effective with even larger sample sizes. Even if they were, TOEFL already has facilities in place to use 3PL, Rash, or MODR for equating with large samples and, without more favorable PIRC results from the present study, there is no incentive to change them.

It has been noted that the maximum discrepancies are expected through the high test score range when the unrestricted PIRC model is used; the PIRC-1 and PIRC-e models were added in the hope of reducing equating errors at the upper score range. It is also true that many fewer examinees achieve high scores, and that TOEFL educational decisions occur more frequently in the mid-score range. Hence, discrepancies at the level of the medians of the 5,000-case section score distributions were examined. The tables reveal that indeed the PIRC discrepancies are smaller in the mid-score ranges than in the upper-score ranges, but are not consistently smaller than those of the logistic models.

The easy and hard equater sets were used to reveal the models' possible sensitivity to these adverse equating conditions. Indeed, for the 1,000-case samples the discrepancies were slightly larger for these equater sets than when all items were used. This slight disadvantage of the biased equater sets might be due to the fact that they contained only half of the items, in contrast to the broad range equater sets which, in the 1,000-case samples, contain all of the items. No overall special disadvantage for the biased equater sets was noted when the biased equater sets were used with the 100-case samples, where the broad equater set has the same number of items as the easy and hard sets.

The effect of reducing the sample size to 100 was expected to be, and was, most noticeable with results from the 3PL model, which consistently produced the larger discrepancies. This result is to be expected because, with this model, many parameters are determined for each equating sample, and the equated scores are identified only after that estimation is completed. Thus, the estimation can entail extensive capitalization on chance, which can emphasize sample differences, thus precluding appropriate association of scores by the equating program. There are, however, well-known contraindications for, and strictures against, the use of 3PL with the small sample used here. It has been pointed out that the 1,000-case sample is in the lower range of what is usually recommended for that model.

The important research question for this study was whether the PIRC model should be considered as a basis for reduced sample-size equating. Based on the present results the answer is that it should not. But, the study does suggest that using Rasch or MODR methods be seriously considered to gain the advantages of small sample size. Use of the existing software, LOGIST, and the current scales could continue. It would be necessary, however, to conduct some resampling studies to determine the combined effects of model and equating sample-size change.

Though the PIRC model is probably not a promising basis for TOEFL equating, it might prove useful elsewhere. One possible context for PIRC is that of licensing or certification. Test programs in licensing and certification may entail very small samples, with the important discriminations occurring in the low score range, where smaller discrepancies were noted when PIRC was used. Resampling studies could explore the efficacy of the PIRC models and compare them with the Rasch model for this type of application.

One interesting aspect of the PIRC model is that it has a rather simple multidimensional extension. Indeed, a multidimensional extension of the unrestricted PIRC model, when applied to summed items, produces a factor model that has often been used for raw test scores. Thus the model could provide a bridge between item and test statistics in a multidimensional context. For example, even given the result of the Boldt (1989) study, a much later informal, confirmatory model-sampling factor analysis of TOEFL detected a factor related to item difficulty. This effect was so weak that the potential value of a one-dimensional PIRC model for equating was not considered overly impaired.

Finally, the policy of not readministering items poses no undue disadvantage when reconsidering equating models. This policy is dictated by very real and severe test security problems. Leakage of test items and forms occurs and the availability of cheap, rapid communication procedures affords their wide dissemination. Indeed, when such leakage occurs it seems likely that the item parameters would be affected. If the item parameters were affected, the possibility of self-equating is destroyed because self-parallelism no longer exists. Thus, a self-equating experiment using experimental data is not feasible when significant test leakage occurs between administrations. Clearly, one must seek a data source other than operational tests. Perhaps the pretest system might be used for this purpose. Though it would not be feasible to administer operational TOEFL forms within that system, a system of miniature tests that incorporate properties under investigation could be used to answer specific research questions.

Table 1

For Each Model, Discrepancies at the Value of the Source Sample's Median and the Maximum Discrepancies(Score Level) for Section 1

Model	Disc. at Median	Max Disc.(Score)
Equating Sample Size of 1,000		
All Items in the Equater		
3PL	.05	.46(11.6)
MODR	.00	.27(9.6)
Rasch	.00	.01(5.2)
PIRC-1	.60	.85(47.5)
PIRC-e	.75	1.52(45.2)
PIRC	.83	1.21(48.9)
Easy Items in the Equater		
3PL	.14	.37(11.7)
MODR	.17	.27(10.0)
Rasch	.15	.15(33.2)
PIRC-1	.04	.57(48.0)
PIRC-e	.13	1.34(46.1)
PIRC	.15	.23(49.6)
Hard Items in the Equater		
3PL	.06	.54(11.6)
MODR	.16	.27(10.0)
Rasch	.15	.15(34.9)
PIRC-1	.23	.67(47.6)
PIRC-e	.34	1.38(46.1)
PIRC	.15	.22(49.4)
Equating Sample Size of 100		
Broad Equater (odd-numbered items)		
3PL	.16	2.09(12.2)
MODR	.33	.59(10.4)
Rasch	.24	.24(32.3)
PIRC-1	.67	1.03(40.9)
PIRC-e	.88	1.64(46.8)
PIRC	.67	.98(49.0)
Easy Items in the Equater		
3PL	.13	2.10(12.2)
MODR	.48	.61(10.4)
Rasch	.02	.07(44.6)
PIRC-1	.11	.64(46.4)
PIRC-e	.43	1.58(49.9)
PIRC	.14	.21(49.4)
Hard Items in the Equater		
3PL	.12	1.90(12.2)
MODR	.20	.89(17.8)
Rasch	.27	.28(36.8)
PIRC-1	.02	.40(41.8)
PIRC-e	.04	1.54(46.9)
PIRC	.14	.21(49.6)

Table 2
Scale Scores and SEMs for TOEFL
Section 1 Raw Scores

Raw Score	Scale	SEM	Raw Score	Scale	SEM
0	24.09	2.73	26	47.62	2.42
1	24.99	2.73	27	48.14	2.41
2	25.9	2.73	28	48.65	2.39
3	26.8	2.73	29	49.16	2.38
4	27.7	2.73	30	49.67	2.37
5	28.61	2.73	31	50.18	2.35
6	29.51	2.73	32	50.7	2.34
7	30.42	2.73	33	51.22	2.32
8	31.32	2.73	34	51.76	2.31
9	32.21	2.73	35	52.31	2.29
10	32.97	2.73	36	52.88	2.27
11	34.29	2.65	37	53.47	2.24
12	36.08	2.64	38	54.09	2.21
13	37.61	2.63	39	54.73	2.18
14	38.93	2.62	40	55.42	2.14
15	40.07	2.6	41	56.15	2.1
16	41.07	2.59	42	56.93	2.05
17	41.96	2.57	43	57.78	1.99
18	42.76	2.55	44	58.72	1.91
19	43.5	2.53	45	59.76	1.82
20	44.18	2.51	46	60.94	1.71
21	44.82	2.5	47	62.28	1.56
22	45.43	2.48	48	63.85	1.35
23	46	2.46	49	65.69	1.02
24	46.26	2.45	50	67.75	0
25	47.1	2.43			

Table 3

For Each Model, Discrepancies at the Value of the Source Sample's Median and the Maximum Discrepancies(Score Level) for Section 2

Model	Disc. at Median	Max Disc.(Score)
Equating Sample Size of 1,000		
All Items in the Equater		
3PL	.04	.73(10.0)
MODR	.02	.04(9.6)
Rasch	.00	.02(2.2)
PIRC-1	.33	.76(34.0)
PIRC-e	.71	1.35(32.5)
PIRC	.56	.78(32.2)
Easy Items in the Equater		
3PL	.42	.84(10.1)
MODR	.33	.36(9.4)
Rasch	.29	.35(18.8)
PIRC-1	.36	.49(36.1)
PIRC-e	.79	1.40(32.4)
PIRC	.62	.86(37.2)
Hard Items in the Equater		
3PL	.37	.61(19.7)
MODR	.33	.48(19.5)
Rasch	.25	.41(15.7)
PIRC-1	.56	.64(25.4)
PIRC-e	.31	.99(34.9)
PIRC	.71	.98(37.0)
Equating Sample Size of 100		
Broad Equater (odd-numbered items)		
3PL	.46	1.40(34.3)
MODR	.67	.67(26.6)
Rasch	.61	.64(24.3)
PIRC-1	.85	1.27(31.9)
PIRC-e	1.28	2.05(35.3)
PIRC	.48	.67(32.3)
Easy Items in the Equater		
3PL	.35	1.43(34.3)
MODR	.20	.30(34.9)
Rasch	.07	.10(20.1)
PIRC-1	.74	1.18(31.9)
PIRC-e	.79	1.22(31.3)
PIRC	.46	.63(37.3)
Hard Items in the Equater		
3PL	.58	1.87(33.8)
MODR	.21	.21(27.1)
Rasch	.06	.19(6.6)
PIRC-1	.11	.62(32.5)
PIRC-e	.52	1.85(35.5)
PIRC	.52	.73(37.4)

Table 4
Scale Scores and SEMs for TOEFL
Section 2 Raw Scores

Raw Score	Scale	SEM	Raw Score	Scale	SEM
0	17.7	3.37	20	44.68	2.84
1	18.84	3.37	21	45.51	2.86
2	19.98	3.37	22	46.31	2.89
3	21.13	3.37	23	47.1	2.92
4	22.27	3.37	24	47.86	2.95
5	23.42	3.37	25	48.63	2.97
6	24.56	3.37	26	49.4	2.98
7	25.71	3.37	27	50.19	2.98
8	26.85	3.37	28	51.01	2.98
9	28.38	3.22	29	51.86	2.96
10	31.51	3.13	30	52.75	2.93
11	33.8	3.06	31	53.72	2.89
12	35.57	2.99	32	54.77	2.82
13	37.08	2.93	33	55.93	2.72
14	38.43	2.88	34	57.25	2.59
15	39.66	2.84	35	58.79	2.4
16	40.81	2.81	36	60.69	2.11
17	41.88	2.8	37	63.27	1.62
18	42.87	2.8	38	67.93	0
19	43.8	2.82			

Table 5

For Each Model, Discrepancies at the Value of the Source Sample's Median and the Maximum Discrepancies(Score Level) for Section 3

Model	Disc. at Median	Max Disc.(Score)
Equating Sample Size of 1,000		
All Items in the Equater		
3PL	.12	.28(15.8)
MODR	.05	.41(13.1)
Rasch	.01	.03(12.5)
PIRC-1	.10	.48(56.1)
PIRC-e	.03	.92(56.6)
PIRC	.06	.09(57.5)
Easy Items in the Equater		
3PL	.01	.61(21.0)
MODR	.21	.40(13.1)
Rasch	.11	.15(48.7)
PIRC-1	.12	.76(50.8)
PIRC-e	.54	1.08(53.5)
PIRC	.42	.61(57.1)
Hard Items in the Equater		
3PL	.35	.38(33.1)
MODR	.28	.41(13.1)
Rasch	.12	.12(41.2)
PIRC-1	.35	.40(36.7)
PIRC-e	.24	.89(54.7)
PIRC	.47	.69(56.9)
Equating Sample Size of 100		
Broad Equater (odd-numbered items)		
3PL	.29	1.31(53.8)
MODR	.04	.91(18.4)
Rasch	.23	.42(27.5)
PIRC-1	12	.90(55.7)
PIRC-e	.10	1.76(53.7)
PIRC	.17	.26(57.6)
Easy Items in the Equater		
3PL	.89	1.44(21.9)
MODR	.87	.88(40.4)
Rasch	.94	1.25(29.1)
PIRC-1	.37	.83(55.8)
PIRC-e	.45	1.21(54.0)
PIRC	.09	.14(57.5)
Hard Items in the Equater		
3PL	.52	1.33(53.8)
MODR	.41	2.31(20.7)
Rasch	.74	1.44(16.8)
PIRC-1	.51	1.09(55.6)
PIRC-e	.68	1.52(53.7)
PIRC	.11	.16(57.3)

Table 6
Scale Scores and SEMs for TOEFL
Section 3 Raw Scores

Raw Score	Scale	SEM	Raw Score	Scale	SEM
0	19.43	2.64	30	45.01	2.52
1	20.17	2.64	31	45.57	2.52
2	20.92	2.64	32	46.11	2.51
3	21.67	2.64	33	46.65	2.49
4	22.41	2.64	34	47.18	2.48
5	23.16	2.64	35	47.71	2.47
6	23.9	2.64	36	48.23	2.45
7	24.65	2.64	37	48.75	2.43
8	25.4	2.64	38	49.26	2.41
9	26.14	2.64	39	49.78	2.39
10	26.89	2.64	40	50.3	2.37
11	27.64	2.64	41	50.83	2.34
12	28.38	2.64	42	51.36	2.32
13	29.13	2.64	43	51.9	2.29
14	31.03	2.61	44	52.45	2.26
15	32.58	2.61	45	53.01	2.22
16	33.92	2.6	46	53.59	2.18
17	35.11	2.59	47	54.19	2.14
18	36.2	2.59	48	54.82	2.09
19	37.2	2.58	49	55.48	2.03
20	38.12	2.58	50	56.18	1.97
21	38.99	2.57	51	56.94	1.9
22	39.8	2.57	52	57.76	1.81
23	40.56	2.56	53	58.68	1.71
24	41.28	2.56	54	59.72	1.58
25	41.96	2.56	55	60.96	1.42
26	42.62	2.55	56	62.51	1.21
27	43.25	2.55	57	64.57	0.89
28	43.85	2.54	58	67.1	0
29	44.44	2.53			

Figure 1

TOEFL Section 1 Equating: 100 Cases

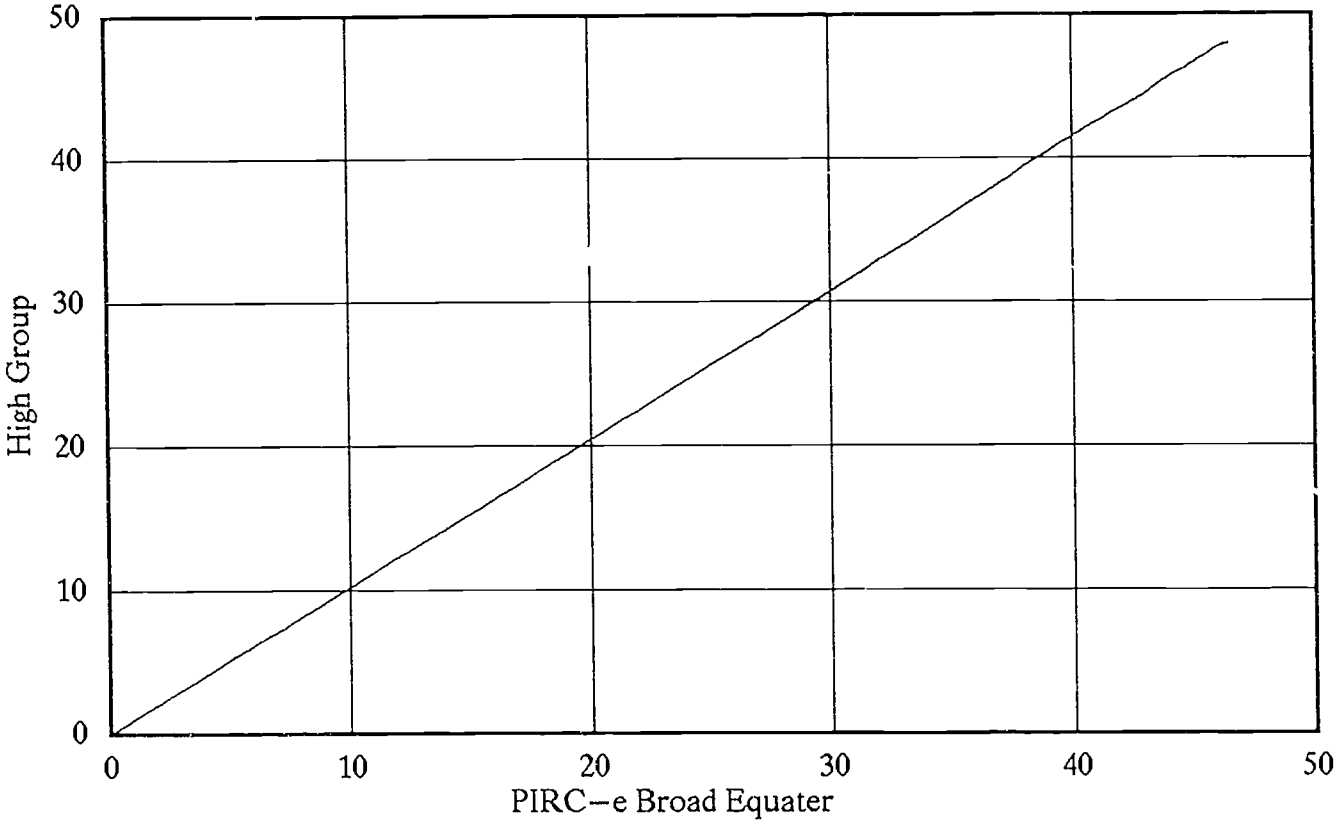


Figure 2

TOEFL Section 1 Equating: 100 Cases

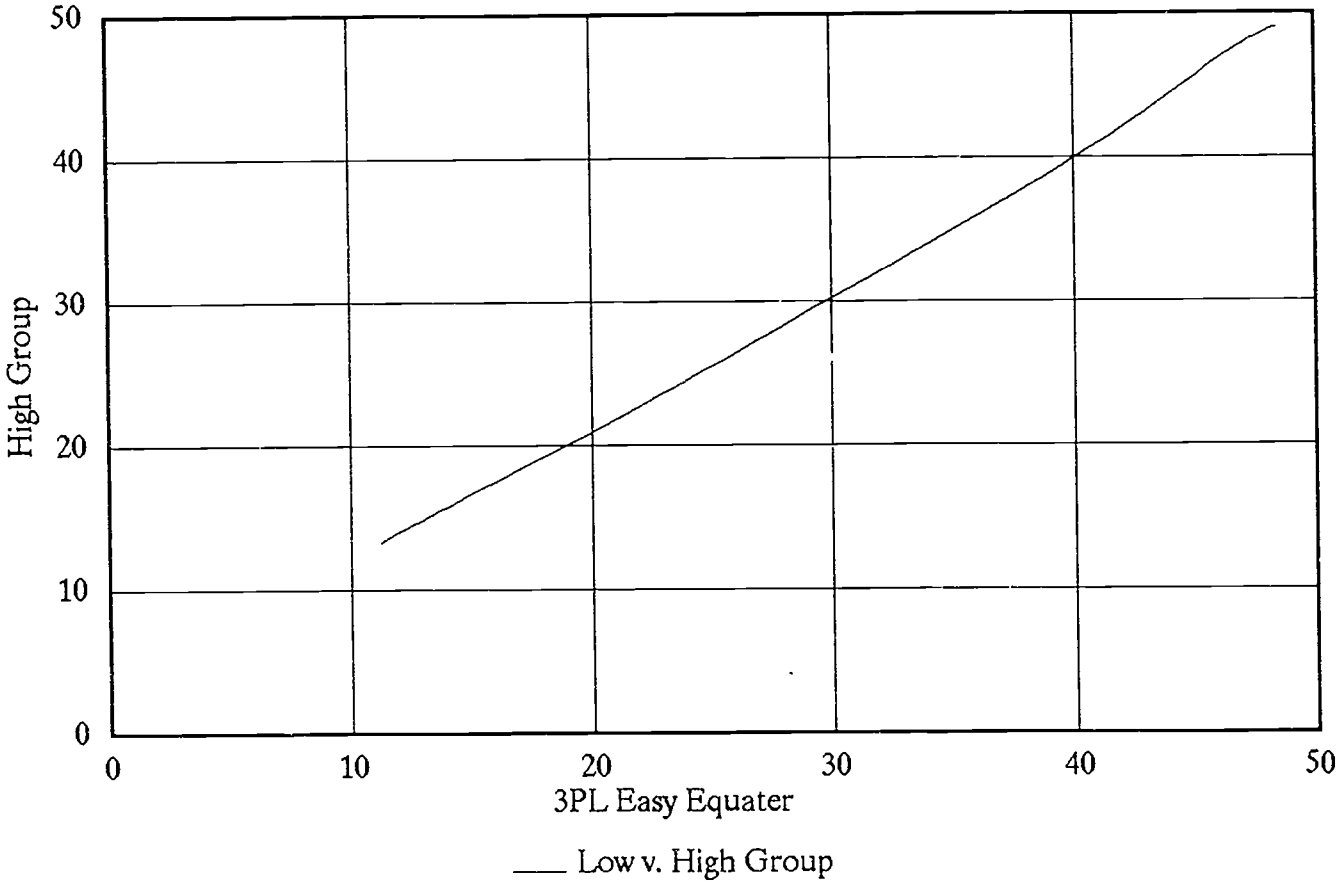


Figure 3

TOEFL Section 3 Equating: 100 Cases

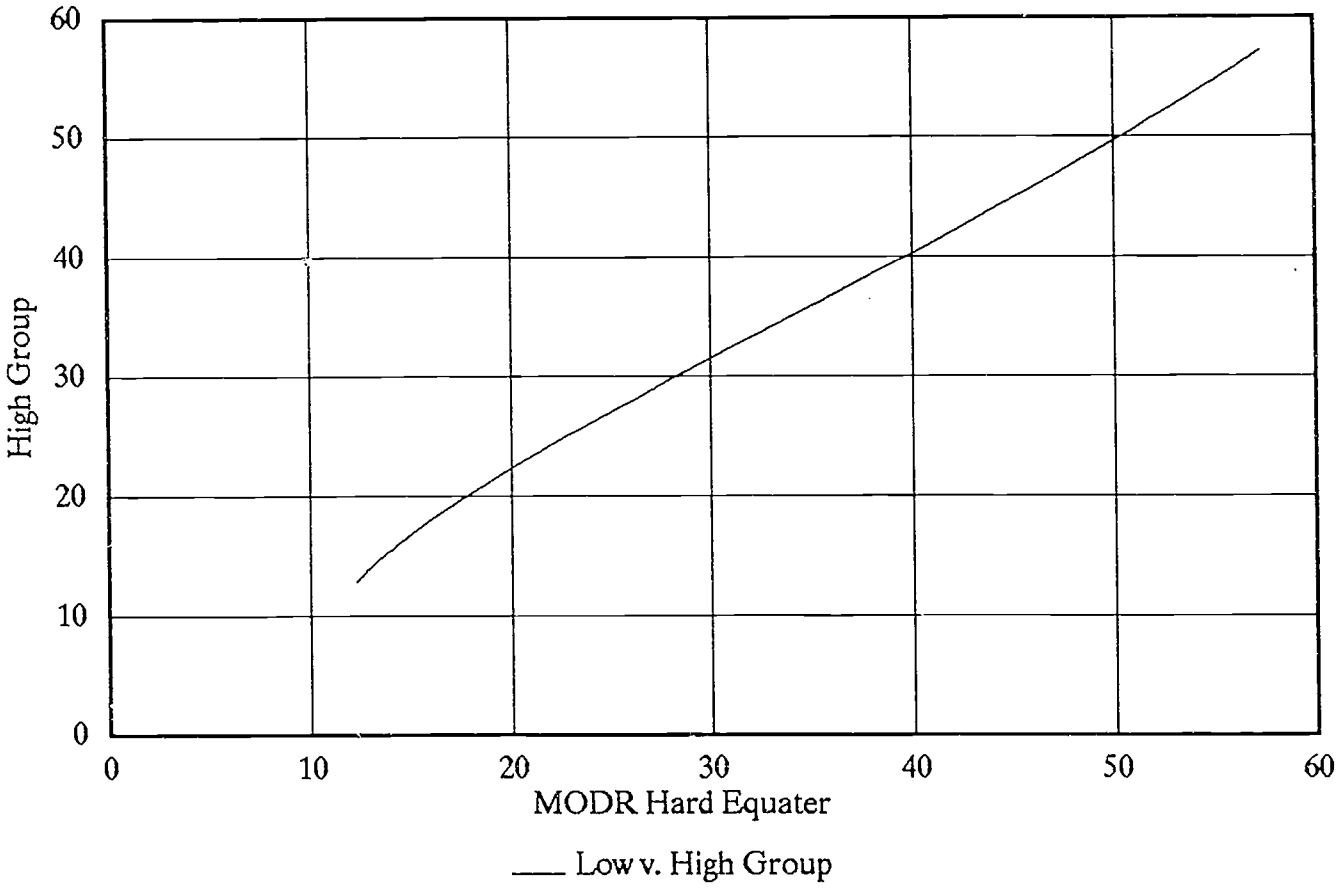


Figure 4

TOEFL Section 1 Equating: 100 Cases

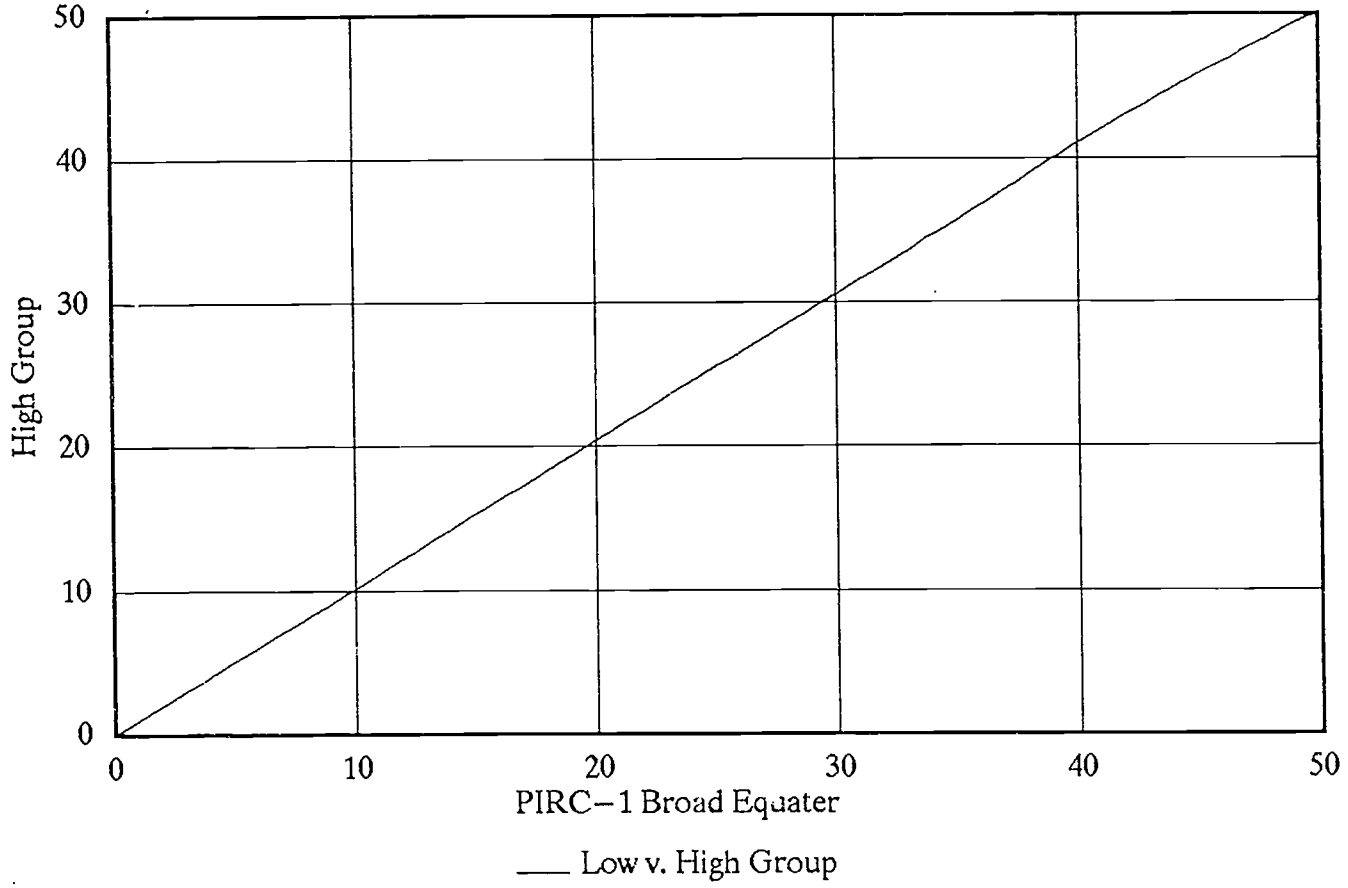


Figure 5

TOEFL Section 1 Equating: 1000 Cases

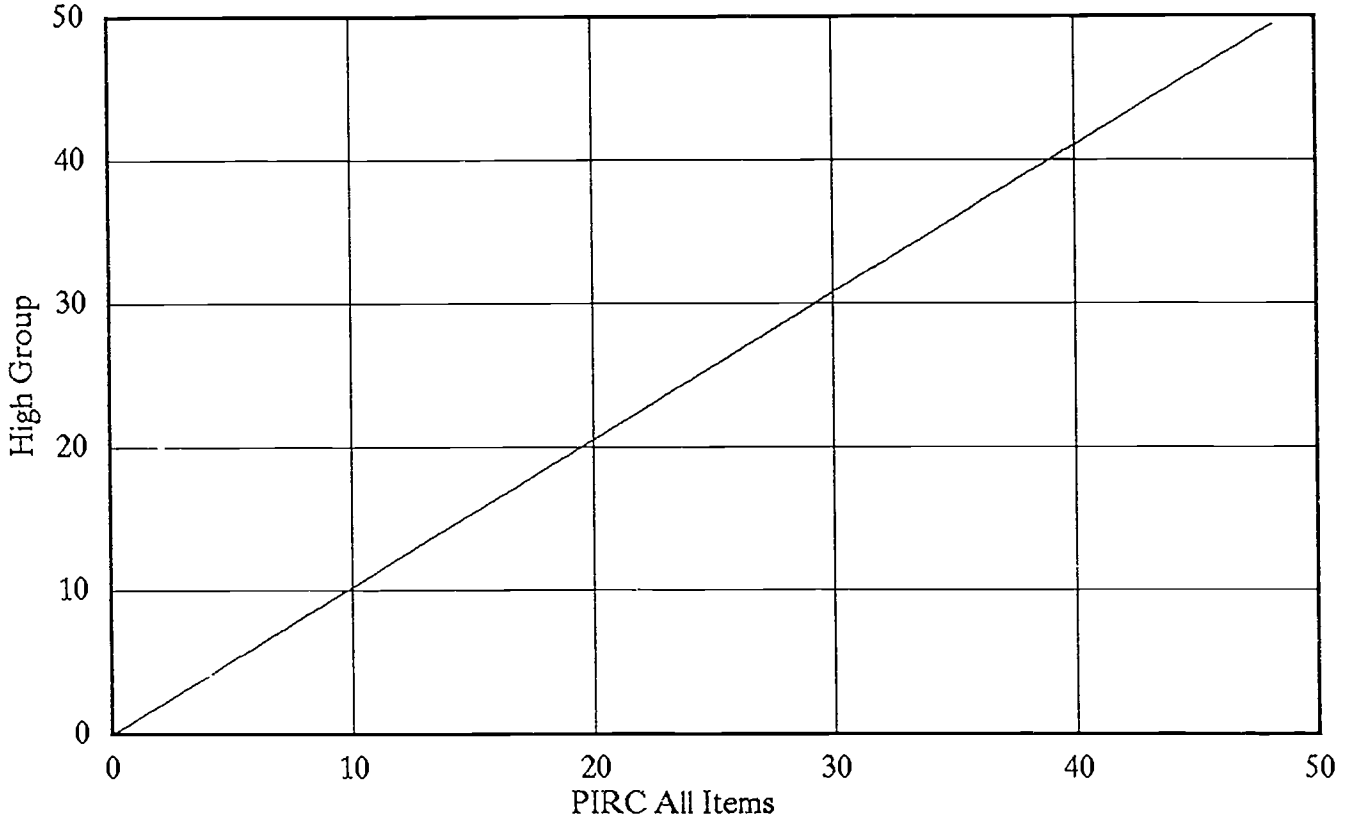
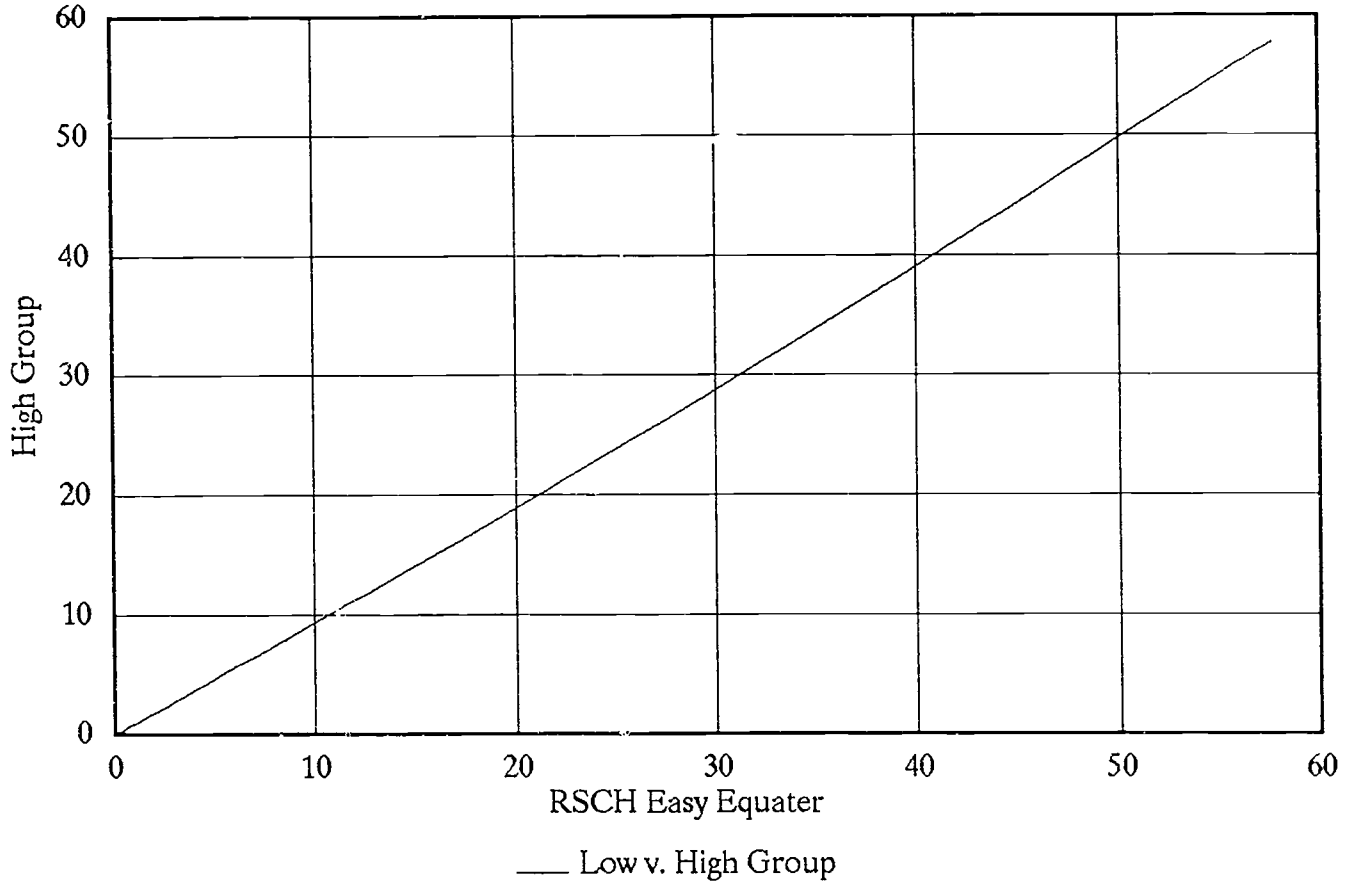


Figure 6

TOEFL Section 3 Equating: 100 Cases



REFERENCES

Boldt, R. F. (1992). Cross validation of item response curve models using TOEFL data. Language Testing, 9, 79-95. (Also, TOEFL Technical Report 4 and Educational Testing Service Research Report 91-33. Princeton, NJ: Educational Testing Service.)

Boldt, R. F. (1989). Latent structure analysis of the Test of English as a Foreign Language. Language Testing, 6, 125-142. (Also, TOEFL Research Report 28 and Educational Testing Service Research Report 88-27. Princeton, NJ: Educational Testing Service.)

Cook, L. L., & Eignor, D. R. (1991). IRT Equating Methods. Educational Measurement: Issues and Practice, 10(3), 37-44.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. Journal of Educational Measurement, 23, 283-298.

---------- (1989). Reply to Andrich and Henning. Journal of Educational Measurement, 26, 295-299.

Educational Testing Service. (1992). Test of English as a foreign language test analysis: Form 3OTF05. (Statistical Report SR-92-128. Princeton, NJ: Author.)

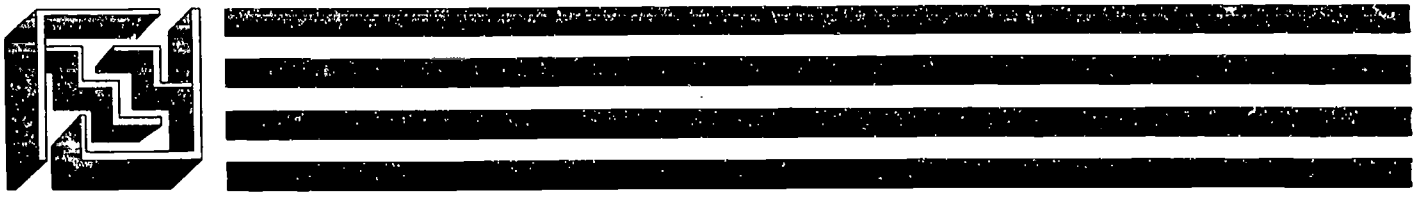
Henning G. (1989). Does the Rasch model really work for multiple choice items? Take another look. A response to Divgi. Journal of Educational Measurement, 26, 91-97.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale NJ: Erlbaum.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-120. (Also, Educational Testing Service Research Bulletin 82-25. Princeton, NJ: Educational Testing Service.)

Wingersky, M. S., Patrick R., and Lord, F. M. (1987). Logist user's guide (Version 6). Princeton, NJ: Educational Testing Service.

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



57906-06445 • Y14M 5 • 275603 • Printed in U.S.A.