ED 249 242 TM 840 556

AUTHOR Hicks, Marilyn M.
TITLE A Comparative Study of Methods of Equating TOEFL Test Scores.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-84-20
PUB DATE Jun 84
NOTE 71p.
AVAILABLE FROM Educational Testing Service, Research Publications, R-116, Princeton, NJ 08541
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *College Entrance Examinations; Comparative Analysis; *English (Second Language); *Equated Scores; Language Tests; *Latent Trait Theory; Measurement Techniques; Sampling; Scaling; Statistical Analysis; Testing Problems; Test Interpretation; *Test Items
IDENTIFIERS Equipercentile Equating; Linear Equating Method; *Test of English as a Foreign Language

ABSTRACT

Six methods of equating Test of English as a Foreign Language (TOEFL) test scores for samples consisting of the usual groups of examinees and groups controlled for native language representation were evaluated in terms of scale stability. The equating methods included three item response theory (IRT) variants (fixed b's scaling, a one-parameter model in which a- and c-parameters were fixed at constant values, and a model in which all three parameters were re-estimated), and three conventional equating methods (Tucker, Levine and Equipercentile). The equating methods were applied to Section II, Structure and Written Expression, and Section III, Reading Comprehension and Vocabulary. For the regular group of examinees, fixed b's IRT equating exhibited the greatest scale stability for both sections with the one-parameter IRT model and Tucker linear equating following in that order. For most equating methods, controlling for native language resulted in increased scale stability relative to the regular group for Section II, but produced more error in Section III. This interaction may be related to the differential performance observed among language groups on Section III in previous studies. Results supported continued use of fixed b's scaling for TOEFL data using a random sample of examinees from the total testing group. (Author)

RR-84-20

RESEARCH REPORT

A COMPARATIVE STUDY OF METHODS OF EQUATING TOEFL TEST SCORES

Marilyn M. Hicks

Educational Testing Service
Princeton, New Jersey
June 1984

A Comparative Study of Methods of Equating TOEFL Test Scores

Marilyn M. Hicks

Educational Testing Service
Princeton, New Jersey

Table of Contents

List of Tables

List of Figures.

Abstract

Six methods of equating TOEFL test scores for samples consisting of the usual groups of examinees tested at each TOEFL administration, and groups of examinees controlled for native language representation were evaluated in terms of scale stability. The equating methods included three IRT variants (fixed b's scaling, a one-parameter model in which a- and c-parameters were fixed at constant values, and a model in which all three parameters were re-estimated), and three conventional equating methods (Tucker, Levine and equipercentile). The equating methods were applied to Section II, Structure and Written Expression, and Section III, Reading Comprehension and Vocabulary. For the regular group of examinees, fixed b's IRT equating exhibited the greatest scale stability for both sections with the one-parameter IRT model and Tucker linear equating following in that order. For most equating methods, controlling for native language resulted in increased scale stability relative to the regular group for Section II, but produced more error in Section III. This interaction between Section III and the controlled group may be related to the differential performance observed among language groups on Section III in previous studies. The results of this study supported continued use of fixed b's scaling for TCEFL data using a random sample of examinees from the total testing group.

A Comparative Study of Methods of Equating TOEFL Test Scores

Introduction

The Test of English as a Foreign Language (TOEFL), which assesses the English proficiency of foreign students desiring to study at colleges and universities in the United States and Canada, is comprised of three sections and seven parts as follows:

I. Listening Comprehension

A. Statements (20 items)

B. Dialogues (15 items)

C. Minitalks (15 items)

II. Stucture and Written Expression

A. Structure (15 items)

B. Written Expression (25 items)

III. Reading Comprehension and Vocabulary

A. Vocabulary (30 items)

B. Reading Comprehension (30 items)

Sections II and III can include 20 and 30 pretest items respectively, interspersed among the operational items. An equated score is reported for each section in addition to a total score. In September 1978, TOEFL adopted item response theory (IRT) methodology in the form of the three-parameter logistic model for the purpose of equating test scores in lieu of conventional linear methods. This implementation of IRT was preceded by a feasibility study which compared equatings based on IRT parameter estimates with those determined from conventional methods for reasonable concurrence (Cowell, 1982). An informal study of scale stability, limited to Sections II and III, was undertaken in November 1979 in which equatings based on original parameter estimates were compared with those derived from a chain of calibrations. Small differences were observed which generated questions that could not be answered by the limited scale of the November 1979 study. In general, it was not possible to determine if the results observed were due to (1) some artifact of the scaling procedures, (2) variability among testing groups, (3) changes in test specifications over the time spanned by the study, or some combination of these factors. Underlying these concerns was the fundamental question of the appropriateness of the IRT model to TOEFL data. Each TOEFL test administration is comprised of over 100 different language groups of varying degrees of affinity to English, the effects of which have been assessed in studies of differential item performance (Angoff & Sharon, 1974; Alderman & Holland, 1980) and factor structure (Swinton & Powers, 1980). Differences observed among language groups in these studies suggested that the basic IRT assumption of the unidimensionality of the latent (ability) space might be violated by TOEFL data. For example, the Swinton and Powers factor analysis of Form YTF4, administered in 1976, indicated that Vocabulary and Reading Comprehension did not constitute a single dimension for non-Indo-European examinees as it did for Indo-European language groups. For the former, performance on Reading Comprehension in Section III was more closely allied to Structure and Written Expression, with Vocabulary defining a separate factor. Furthermore, the factors underlying Section III were less highly differentiated for these examinees than for other language groups. In practice, it appeared that while item parameters might vary among language groups, such parameters could be estimated which characterized the testing group as a whole.

Differential language group performance also has implications for the suitability of some conventional methods of equating TOEFL test scores since most of these procedures are explicitly or implicitly dependent on random sampling from a common population. This study was undertaken to determine the optimal method of equating TOEFL test scores in light of the foregoing considerations.

Background Information

IRT Scaling and Equating of TOEFL Test Scores

The three parameter logistic model for item i,

$$P_i = P_i(\theta) = c_i + (1 - c_i)\{1 + \exp[-1.7a_i(\theta - b_i)]\}^{-1} \qquad (1)$$

specifies the conditional probability of a correct response and requires the estimation of three item parameters, a, b and c, and an ability estimate. A measure of the discriminating power of the item, the a-parameter is related to the slope of the item curve at the point of inflection. The b-parameter is that value on the ability scale midway between the upper and lower asymptotes of the logistic item curve. As a location parameter, it is an index of the item difficulty. The c-parameter is the value of the ordinate at the lower asymptote of the item curve and represents a measure of the tendency to guess on the item.

Using LOGIST (Wood, Wingersky & Lord, 1976), TOEFL parameters are estimated such that thetas are scaled to mean zero and standard deviation one, with b's on the theta scale. If another group of examinees were administered the same item and a similar scaling were applied, any differences in level and spread of ability between the two groups would result in dissimilar values of the b's. The invariance of item parameters across groups and theta estimates across tests will hold only if parameter estimates derived from subsequent groups are placed on some established scale. If a set of items have been scaled on a given group of examinees, estimates based on successive groups can be linearly transformed to the established scale. When old and new forms are linked by a block of common items, the slope and intercept parameters of the line relating the b's can be used to scale all the items in the new form (Marco, 1977). Stocking and Lord (1982) have developed a linear transformation which results from the minimization of the average squared difference between true score estimates and have reported favorable results for this method.

Current TOEFL scaling procedures do not depend on a block of items common to two forms; instead calibrated (scaled) pretested items, selected from many previous test forms, serve as the equating items in each version of the test.

During parameter estimation, the a- and c-parameters for the calibrated items are re-estimated, but the b-parameters are held fixed at the values derived in the initial calibrations. Alluded to as "fixed b's" scaling, the presence of the precalibrated items sets the scale for the noncalibrated items. In common item equating, the equating items are selected to be representative of the total test in content and other specifications; however, the precalibrated items in the fixed b's scaling are chosen to span the range of difficulty and discriminating power of the total test. Implicit in this procedure is the basic IRT assumption that the estimates of difficulties will hold for all testing groups except for scale factors. Plots of item curves, on which were superimposed squares representing the observed proportions of examinees at a given ability level responding correctly to the items (item ability regressions) indicated that, on occasion, some of the precalibrated items (items for which the b's were held fixed) did not adequately reflect the response patterns of the current examinee group. The fit of the newly calibrated items was usually quite satisfactory. (Examples of item ability regressions are given in Figures 13-16).

Once the item parameters are on scale, it is only necessary to calculate the sum of the item curves, the test characteristic function, which specifies true scores as a function of ability. Scores on two tests are then considered equivalent if they depend on the same value of theta. Additional information regarding true score equating (Lord, 1980; pp. 199-205) as applied to TOEFL is outlined in Appendix A.

Some Conventional Equating Methods

Conventional methods of equating include linear and equipercentile equating which are defined as follows:

> equipercentile equating: For a given group of examinees, two scores on separate forms of a test are considered equivalent if their percentile ranks are equal.
>
> linear equating: For a given group of examinees, two scores on separate forms of a test are considered

equivalent ıf they correspond to equal standard score deviates (Angoff, 1982).

In the usual testing situation where separate groups take the two test forms, the strategy utilized in implementing these definitions involves the formation of a synthetic equating population, T, as a weighted composite of the two testing groups, P, the group taking new form X, and Q, the group taking old form Y,

$$T = w_1P + w_2Q. \qquad (2)$$

where w_1 and w_2 are weights assigned to the two groups.

In the case of common item equating, information derived from a set of items, V, common to old form Y and new form X aids in determining the distributions and first two moments of the synthetic equating group. The item set, V, is commonly called the anchor or equating test. Some details of these procedures are given in Appendix B; however, the development of the distributions for population T requires the following assumptions:

a) For equipercentile equating, the conditional distribution of Y given V = v is the same in groups P and Q with a similar assumption for form X,

$$F_P(Y|v) = F_Q(Y|v) \qquad (3)$$

$$G_P(X|v) = G_Q(X|v) \qquad (4)$$

b) For Tucker linear equating, the regressions of Y on V, and X on V is the same for P and Q,

$$E_P(Y|v) = E_Q(Y|v) = av + b \qquad (5)$$

$$E_P(X|v) = E_Q(X|v) = cv + d \qquad (6)$$

and the conditional variances are equal in P and Q,

$$Var_P(Y|v) = Var_Q(Y|v) = \sigma^2 \qquad (7)$$

$$Var_P(X|v) = Var_Q(X|v) = \sigma^2 \qquad (8)$$

Thus, Tucker linear equating, as well as equipercentile equating is based on untestable assumptions since data for the old form in Q or for the new form in P is not available. Braun and Holland (1982) have noted that assumptions (5) through (8) may not be satisfied if the regression system depends on a measurable extraneous variable such as some student background characteristic. Data collected for the period September 1980 through June 1981 reflecting the language group composition for ten administrations of TOEFL indicated, among other things, that language group representation varies across administrations. To the extent that language group membership is related to test performance on TOEFL, these variations may cause the assumption of equality of regressions in Tucker linear equating to be violated. Indeed, Levine (1955) has demonstrated by experiment that the invariance of regression parameters will not hold on parallel tests if the selection of samples is on variables external to the equating design, observing that if the assumptions which are made in deriving the mathematical model for equating are not satisfied, it is probable that its application will result in biased equivalent scores. Levine derived equations to equate tests which have been administered to samples that differ in dispersion and level of ability due to selection on variables extraneous to the equating experiment. His assumptions were presented in terms of the invariance of the true score regression system with the additional constraint that V be parallel to X and Y (Levine,1955; Angoff,1961).

Objectives of the Study

The major objective of this study was the determination of the method of equating TOEFL test scores that will best maintain the stability of the score scale over time, given the variable nature of its testing population. On the assumption that the period of time defined by this study would include test forms for which the test specifications were relatively constant, the following research questions were investigated:

1. What are the effects of population variability? Can they be eliminated by defining an equating group controlled for native language representation?
2. Will alternate methods of scaling IRT parameters produce more stable results than those presently employed? Will a simplified IRT equating model produce better results with TOEFL test scores?
3. How do conventional linear and curvilinear methods compare with IRT equating for the TOEFL population?

Methods

Selecting the Controlled Group

The initial phase of this study sought to determine a set of criteria by which a controlled equating group could be formed in order to compare a variety of equating models with groups as they naturally arise in the TOEFL testing program. In particular, it was required to define a group whose proportionate native language representation could be replicated at each of the experimental administrations. It was hoped that subsets of language groups with similar performance profiles (i.e., similar rank-ordering of item difficulties) could be identified in the expectation of simplifying the sampling process. Preliminary analysis indicated that this approach would not be successful by virtue of variations in item difficulties even among language groups believed to be closely allied. Somewhat similar results were observed in a study of TOEFL item bias by Alderman and Holland (1981), consequently this approach was not pursued further. It also became clear that if it were possible to identify clusters of language groups with similar performance profiles, the composition of these groups would differ for the two sections.

Data was collected on native language groups representing at least one percent of each administration for examinees taking TOEFL at domestic centers for the year previous to this study. These data indicated that large differences existed in monthly representation for Chinese, Arabic, Farsi, Spanish and Japanese speakers. To the extent that item parameter estimates differed among language groups, those estimates might be unduly influenced by over-representation of one or more native languages at any given administration. Likewise, these variations may also violate assumptions basic to some methods of linear equating. The minimum proportions observed in the year preceding the study (to assure availability) for each language group were tallied and a group controlled for native language representation was selected at each administration as given in Table 1.

Table 1.

Native Language Representation for Controlled Equating Group

Language Group	Proportion of Total Controlled Group
Arabic	.207
Bengalese	.006
Chinese	.150
Farsi	.079
French	.018
German	.008
Greek	.020
Hindi	.008
Ibo	.003
Indonesian	.020
Japanese	.110
Korean	.060
Malaysian	.020
Portuguese	.010
Russian	.016
Spanish	.206
Tagolog	.007
Thai	.017
Turkish	.010
Urdu	.009
Vietnamese	.016

Equating Design

Each of the seven experimental administrations of this study was comprised of several subtests which included both operational and pretest items. Section I contained no pretest items; thus, it was not included in this study. The pretest slots of one subtest at each administration contained operational items from the previous form, thus defining an equating chain as shown in Table 2. For some of the equating models this resulted in anchor tests which were internal to the old form and external to the new form. For Sections II and III six types of equating were included in this study as follows:

1. Modified IRT: a- and c-parameters were held fixed at values determined to be representative of current TOEFL data; only the b's were estimated. For Section II, a was fixed at 1.00 and c at .19. For Section III, the fixed value of a was 1.03 and c was .20. Parameters were scaled using the Stocking and Lord characteristic curve transformation (Stocking & Lord, 1982) based on a set of items common to two forms.

2. Fixed b's IRT: This replicated the current TOEFL operational scaling procedures as previously described; b's for the equating or precalibrated items were held fixed at pretested values, only a- and c-parameters were re-estimated on this item set, all three parameters were estimated for the remaining noncalibrated items. The equating items were selected from many previous forms.

2. Three parameters re-estimated: A set of items common to an old and new form facilitated the scaling of all the items. All three parameters were estimated on the new form. As in the case of Modified IRT, using a set of common items, a line relating the parameters of the old (scaled) form and the new form was calculated based on the Stocking and Lord characteristic curve transformation. The parameters of this line was used to place all other items on scale.

Table 2.

Equating Item Links for the TOEFL Experimental Forms

Form	Operational	Pretest Slots
3BTF11	a,a*	
3DTF9	b	a
3DTF10	c	b
3ETF1	d	c
3ETF2	e	d
3ETF3	f	e
3ETF4	g	f
3ETF6(37)		g
3ETF6(38)		a*

4. Tucker linear equating: Tucker parameters were used throughout the chain of equatings.

5. Levine linear equating: Levine parameters were used throughout the equating chain.

6. Equipercentile equating.

TOEFL form 3BTF11, administered in November 1979 was the only relatively recent test edition with linear parameters linked to the TOEFL scale and was therefore chosen as the base equating form in this study. For each IRT and conventional equating condition, a separate 3BTF11 scale was established. For all IRT equatings the experimental form was equated to the appropriate version of the base form. The links served only for the purpose of scaling in the Modified IRT and 3-parameters re-estimated models while in the three conventional equating methods each experimental subtest was equated to the previous form in the chain. The equating group for the fixed b's method was a spaced sample across all subtests of the experimental forms. All other equating groups were necessarily based on the single subtest which served as the link. The input parameters for the regular and controlled groups in fixed b's scaling were operational TOEFL data, i.e., derived from the regular TOEFL testing population. IRT equatings were derived from operational TOEFL computer programs.

A sample size of approximately 1,000 is required for reliable estimation in the three parameter model. As a result, the 3-parameters re-estimated could not be run on the controlled group since, for some administrations, this group represented about one-third (about 300) of the examinees taking any linked subtest form. Consequently, the following conditions were observed in this study:

	Regular Gr.	Controlled Gr.
Modified IRT	x	x
3-parameters re-estimated	x	-
Fixed b's	x	x
Tucker	x	x
Levine	x	x
Equipercentile	x	x

Analysis of Results

The design accounted for an empirical evaluation of the stability of the various equating conditions by utilizing two subtests, 37 and 38, of the final experimental form, 3ETF6. Accordingly, the following items were included in these subtests:

1. A set of items linked to the previous form in the equating chain (subtest 37).
2. A set of items from 3BTF11 as a direct link to the base form (subtest 38).

The equatings derived from the direct link served as the criterion against which each equating chain would be compared using a discrepancy index developed by N. Petersen (Petersen, Marco and Stewart, 1982), and a computer program written by staff in College Board Statistical Analysis. The index is a weighted mean square difference decomposed into the variance of the difference and the squared bias. Thus, if $d_i = (t_i' - t_i)$, where for raw score i, $i = 0, 1, \ldots, n$, t_i' and t_i are converted scores corresponding to the criterion and chain equatings respectively, and f_i is the number of examinees at each score level, then

$$\Sigma f_i d_i^2/n \quad = \quad \Sigma f_i (d_i - \bar{d})^2 /n \quad + \quad \bar{d}^2 \qquad (10)$$

Total Error Squared = Variance of Difference + Squared Bias.

Optimum conditions for the criterion comparisons include equivalent samples, and anchor tests of equal difficulty for the two subtests. All equating comparisons were based on independent samples taking the two 3ETF6 experimental subtest forms. The one exception was fixed b's equating in the controlled group in which case the single subtest samples were not of sufficient size to estimate parameters. Consequently a methodology was adopted which simultaneously estimates a large number of items taken by more than one group of examinees (Lord, 1980, pp. 205-206). Comparisons involving equipercentile equatings were limited to the range of scores actually observed.

Results

Description of Samples Used in the Study

Raw score data. Raw score data for samples in this study are given in Table 3 (R and C refer to regular and controlled groups, respectively). Numbers in this table depend on selection criteria in the computer programs on which these data were based (LOGIST eliminates examinees with perfect scores, the item analysis program produces samples based on a factor of five). Sample sizes reflect an effort to maximize the reliability of the item parameter estimates consistent with the sampling proportions for the controlled group. With the exception of the criterion comparisons for the regular group, a spaced sample was taken across all subtests for fixed b's scaling; all other methods depended on a single subtest. In the 3BTF11 samples, the controlled groups slightly outperformed the regular group and were somewhat more variable. For the experimental forms, the controlled groups were less able overall.

Analysis of regressions of total test on anchor test. Additional information regarding differences between groups can be elicited from the regression of total score on the equating items score. Since the equating items enter into the determination of the converted scores in various ways, analysis of these regressions may illuminate the nature of the differences between some of the equating results for the two groups. These data, including probabilities associated with the null hypotheses usually tested in an analysis of covariance, as calculated from the Wilks-Gulliksen Ancova program (Gulliksen and Wilks, 1950), are presented in Table 4. P(A), P(B) and P(C) represent the degree of confidence in accepting the following hypotheses:

$P(A) = P_{H_o}$ [equal errors of estimate]

$P(B) = P_{H_o}$ [equal slopes]

$P(C) = P_{H_o}$ [equal intercepts].

Table 3.
Raw Score Means, Standard Deviations and Sample Sizes
for All Equating Groups

			II		III	
Form/Group		N	Mean	S.D.	Mean	S.D.
			Base Form - IRT Equating			
3BTF11	R	14068	23.27	7.01	31.43	10.13
	C	7501	23.80	6.80	32.52	10.04
			Base Form - Linear Equating			
3BTF11	R	4580	23.88	7.06	32.03	10.20
	C	2445	24.09	6.93	32.66	9.99
			Experimental Forms - Fixed b's			
3DTF9	R	2283	25.30	6.86	35.77	10.05
	C	1785	25.23	6.86	35.42	9.83
3DTF10	R	1159	25.24	6.68	34.71	10.60
	C	1115	24.69	6.77	32.92	10.97
3ETF1	R	2271	25.45	7.00	36.09	10.15
	C	1849	24.49	7.28	35.12	10.27
3ETF2	R	1774	26.27	6.81	37.10	10.01
	C	1615	25.72	6.84	36.15	9.96
3ETF3	R	2426	25.34	7.09	34.38	9.16
	C	1871	24.76	7.07	33.83	9.15
3ETF4	R	2330	25.38	6.62	37.28	9.65
	C	1709	25.02	6.51	36.74	9.87
3ETF6(37)	R	1011	26.35	6.29	38.20	8.66
	C	1790	24.85	6.90	35.86	9.32
3ETF6(38)	R	988	26.23	6.97	37.26	8.61
	C	1790	24.85	6.90	35.86	9.32

Table 3 (cont'd)

Experimental Forms - Tucker, Levine,
Equipercentile, Modified IRT, 3-parameters Re-estimated*.

			II		III	
Form/Group		N	Mean	S.D.	Mean	S.D.
3DTF9	R	1265	25.39	6.98	35.98	10.20
	C	710	25.20	6.99	35.44	9.95
3DTF10	R	1575	25.98	6.77	35.59	10.48
	C	325	24.88	6.96	33.09	11.03
3ETF1	R	1530	25.96	6.93	36.59	9.88
	C	460	24.92	7.20	35.24	10.27
3ETF2	R	1275	26.62	6.68	37.57	9.90
	C	605	25.94	6.83	36.56	10.00
3ETF3	R	1710	26.14	7.01	35.13	9.20
	C	845	24.76	6.89	33.46	8.70
3ETF4	R	1825	26.24	6.37	38.64	9.21
	C	845	25.40	6.55	37.51	9.11
3ETF6(37)	R	1005	26.44	6.34	38.23	8.66
	C	315	25.38	7.04	36.39	9.51
3ETF6(38)	R	980	25.95	6.49	37.72	8.71
	C	305	24.93	7.04	35.72	9.25

*3-parameter re-estimated based on regular group only.

Table 4

Analysis of Covariance for Regular and Controlled Groups for All Experimental Test Forms

		Section II							Section III						
Form /Group		r	VEE	b	Int.	P(A)*	P(B)*	P(C)*	r	VEE	b	Int.	P(A)*	P(B)*	P(C)*
3BTF11	R	.95**	5.10	1.64	4.59	.71	.93	.61	.95**	1.05	1.74	3.39	.69	.34	--
	C	.95**	5.00	1.64	4.65				.95**	1.03	1.72	3.76			
3DTF9	R	.94	5.54	1.75	3.05	.33	.91	.43	.85	2.89	1.60	8.68	.12	.53	.53
	C	.94	5.91	1.75	2.92				.82	3.20	1.57	9.03			
3DTF10	R	.93	6.03	1.79	2.77	.98	.85	.78	.84	3.22	1.58	8.33	.41	.25	--
	C	.94	6.02	1.80	2.64				.85	3.46	1.65	6.23			
3ETF1	R	.93	6.90	1.92	1.51	.69	.75	.19	.87	2.38	1.61	7.34	.63	.55	.54
	C	.93	7.10	1.93	1.16				.88	2.47	1.58	7.67			
3ETF2	R	.94	5.47	1.74	3.83	.99	.33	.18	.83	3.08	1.60	7.37	.62	.83	.20
	C	.94	5.47	1.78	3.27				.84	2.97	1.61	6.81			
3ETF3	R	.79	18.45	1.44	8.13	.98	.87	.03	.82	2.79	1.41	9.17	.60	.06	--
	C	.78	18.49	1.45	7.45				.78	2.91	1.30	10.95			
3ETF4	R	.78	15.78	1.42	7.67	.09	.91	.25	.80	3.07	1.62	9.26	.49	.42	--
	C	.77	17.45	1.41	7.54				.81	3.20	1.66	8.28			
3ETF6(37)	R	.79	15.01	1.36	8.34	.26	.08	—	.80	2.71	1.40	11.27	.41	.68	.31
	C	.81	16.63	1.48	6.91				.82	2.92	1.42	10.43			
3ETF6(38)	R	.76	17.83	1.32	8.97	.13	.71	.36	.79	2.80	1.43	12.31	.25	.13	—
	C	.77	20.50	1.34	8.38				.84	2.51	1.53	10.02			

* P(A) = P_{H_o} [variance of errors of estimate]

P(B) = P_{H_o} [slopes]

P(C) = P_{H_o} [intercepts]

Not indicated if P(B) < .50

**Spuriously high correlation. Internal anchor test.

No values of P(C) are listed if P(B) < .50. The tests are N dependent, thus small differences may have greater significance with increasing sample size. Samples used in this analysis were those based on the linked subtests, i.e., the samples from the bottom half of Table 3. Focusing on the slopes of the regressions, the two groups were less alike on 3ETF2 and 3EFT6 (37) for Section II. These were the forms, however, on which the slopes demonstrated the greatest concurrence in Section III. On this section, the two groups differed most in 3ETF3 and 3ETF6 (38). In general, the data indicated greater similarity between groups on Section II than on Section III, a significant relationship which impacts on later equating comparisons.

Equivalency of samples. Spiralling of subtest forms (distribution of the subtests in serial order) is intended to assure equivalent samples when more than one form of the test is administered. A rough evaluation of any effects of spiralling on the equating samples for the September (3DTF9) through April (3ETF4) administrations of this study can be achieved through comparisons of mean ability for the fixed b's scaling and for the Modified IRT as given in Figures 1 through 4. Differences in location of the dotted and solid lines relative to the ordinates reflect the fact that each equating method is on a different scale. If mean ability for the fixed b's sample can be considered to be the more reliable estimate for these six administrations (since it is taken across all subtests), then to the extent that the trends in the two sets of data concur, the sample based on the linked subtests can be considered to be representative of its group. The linked subtests graphs (dotted lines) for the regular group appear to be in closer correspondence with the sample taken across all forms than those for the controlled group. For both groups and both sections, an effect due to spiralling can be observed at the April administration where the single subtest sample produced relatively higher means than the fixed b's sample.

The last two points of these plots represent the chain and direct forms, respectively, which were used to determine the stability of the scales.

Figure 1.

MEAN ABILITY FOR ALL EXPERIMENTAL FORMS

SECTION II, REGULAR GROUP

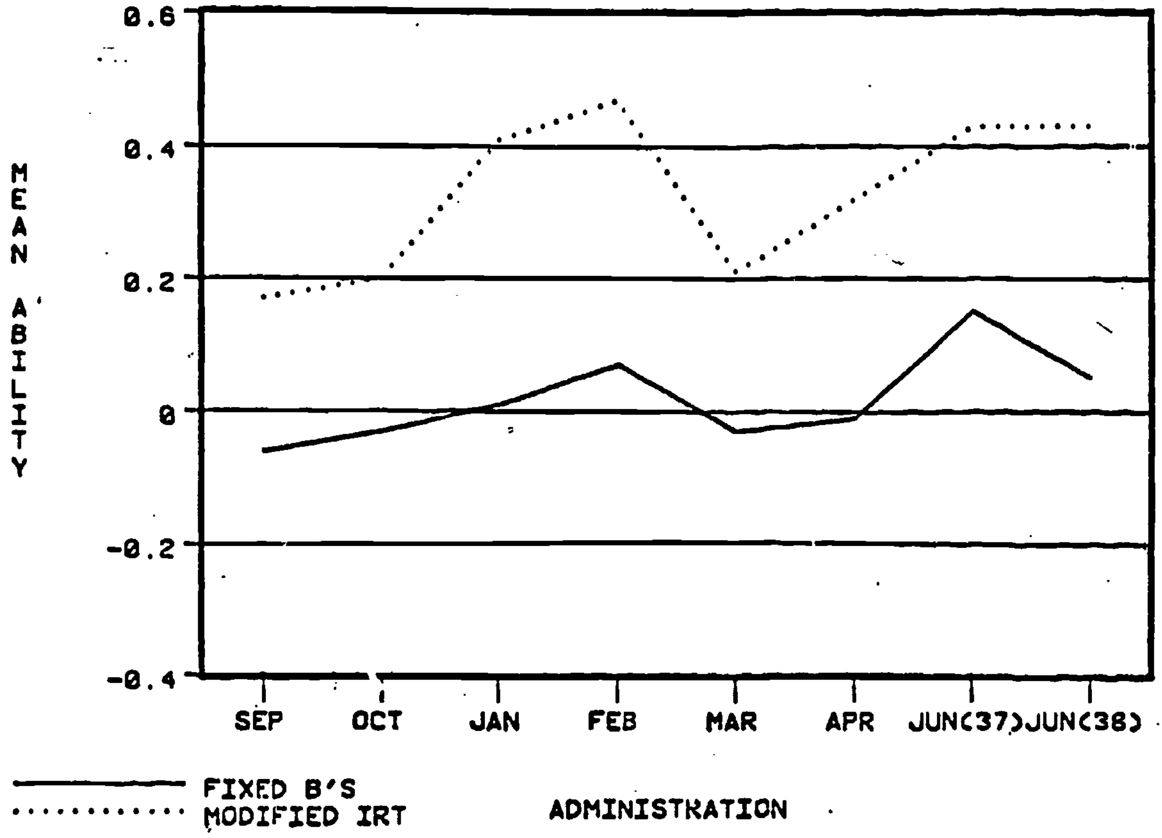


Figure 2.

MEAN ABILITY FOR ALL EXPERIMENTAL FORMS

SECTION III, REGULAR GROUP

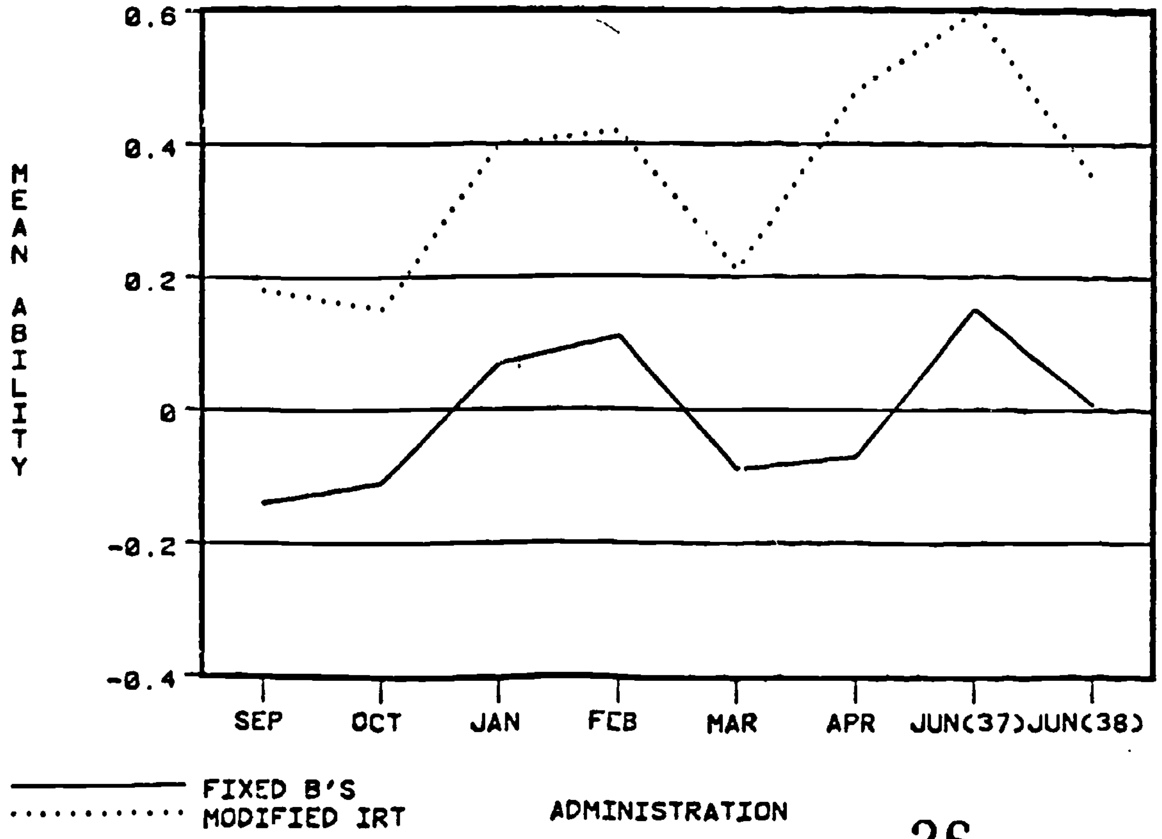


Figure 3.

MEAN ABILITY FOR ALL EXPERIMENTAL FORMS

SECTION II, CONTROLLED GROUP

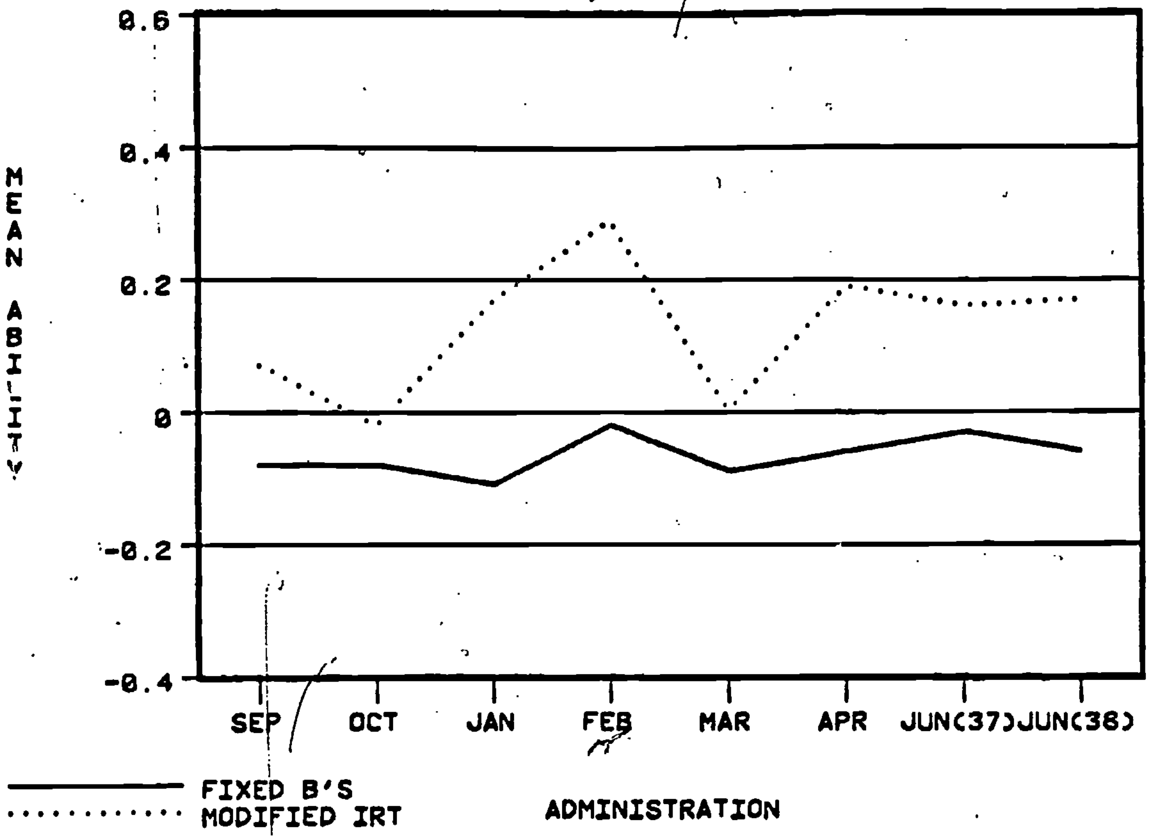
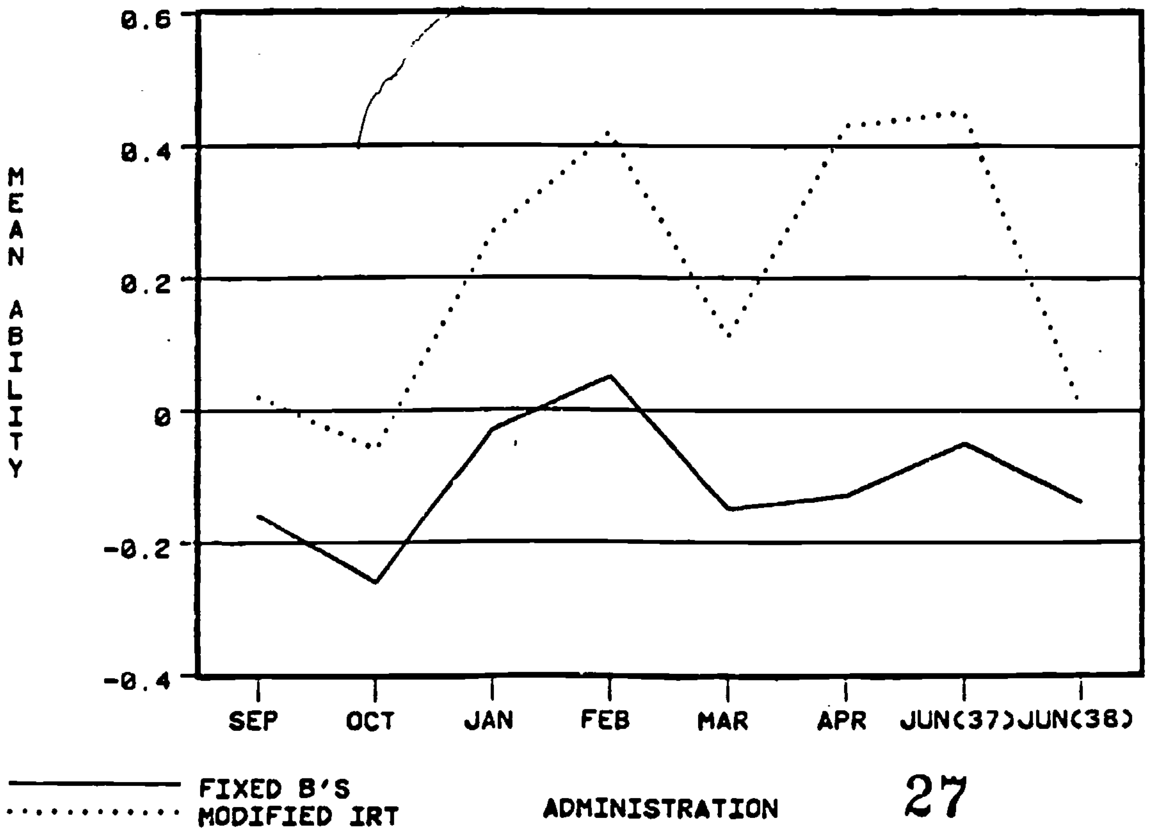


Figure 4.

MEAN ABILITY FOR ALL EXPERIMENTAL FORMS

SECTION III, CONTROLLED GROUP



Sizeable differences in ability exist on these two forms on Section III for both groups, and for the controlled group this difference is very marked.

Characteristics of the Forms Used in the Study

Table 5 lists data relevant to the nature of the forms included in the study in terms of the average difficulty of the operational and equating items. Comparisons of mean deltas (linearly transformed item difficulties, see Angoff & Dyer, 1971) for the operational and equating items indicate that the forms are parallel in terms of mean difficulty with the exception of Section III in 3ETF4, a relatively easy form, and Section III of 3BTF11 which was the most difficult form in the study. Overall, the equating items for the linked forms were slightly more difficult than the operational test, especially in Section III. The characteristics of the forms used in the equating comparisons closely parallel Variation 8 in the Petersen, Marco and Stewart study in that, for some equatings, the base form was slightly more difficult than the test to be equated, and for Section III of subtest 38, the anchor test was more difficult than the operational test. These conditions were found to rather consistently produce greatest error in the evaluation of linear equating (Petersen, Marco & Stewart, 1982, Table 10).

The results described above have obvious implications for the reliability of the equating comparisons in Section III. A common procedure in evaluating the results of an equating experiment has been the use of the identity equating (Levine, 1955; Petersen, Marco and Stewart, 1982). In this case, the base form is re-administered as the final link, and lack of scale stability is evaluated in terms of the departure of the slope of the equating line from unity. Objections to this method involve the possible advantage derived from equating a test to itself in the case of the one-parameter IRT model (Petersen, Marco and Stewart, 1982). An alternative procedure of using two forms, one based on a direct link to the scale and the other the result of the chain, was adopted in this study to circumvent this objection. Equivalent samples for these forms were assumed to be attainable by virtue of spiralling. The second requirement of equivalent difficulty of the anchor tests was difficult to achieve due to current limitations on the availability of items

Table 5

Mean Equated Deltas and Mean r-biserials for Operational and Anchor Tests, Experimental Links

		Section II						Section III					
		$\bar{\Delta}_t$	S.D.	$\bar{r}_t$	$\bar{\Delta}_a$	S.D.	$\bar{r}_a$	$\bar{\Delta}_t$	S.D.	$\bar{r}_t$	$\bar{\Delta}_a$	S.D.	$\bar{r}$
3BTF11	R	11.86	2.22	.52	11.98	2.07	.54	12.60	1.95	.48	12.49	1.94	.47
	C	11.86	2.21	.51	11.98	2.07	.54	12.62	1.97	.47	12.50	1.91	.46
3DTF9	R	11.84	2.22	.54	12.33	1.93	.53	12.20	1.85	.48	12.52	1.81	.49
	C	11.86	2.25	.54	12.23	1.98	.52	12.20	1.82	.47	12.50	1.76	.46
3DTF10	R	11.91	2.09	.54	12.05	1.89	.54	12.26	1.75	.51	12.47	1.50	.48
	C	11.88	2.08	.53	11.93	1.87	.53	12.25	1.76	.51	12.35	1.53	.48
3ETF1	R	11.83	2.33	.54	11.58	2.54	.57	12.28	2.03	.49	12.27	2.22	.52
	C	11.84	2.36	.55	11.62	2.42	.55	12.29	2.00	.50	12.32	2.14	.53
3ETF2	R	11.75	2.25	.52	11.89	2.68	.49	12.41	1.77	.49	12.35	1.92	.48
	C	11.74	2.29	.51	11.87	2.75	.48	12.37	1.75	.48	12.29	1.88	.48
3ETF3	R	11.83	2.24	.55	12.19	1.9	.52	12.37	2.17	.47	12.20	1.70	.48
	C	11.81	2.29	.53	12.09	2.00	.49	12.33	2.19	.44	12.19	1.66	.45
3ETF4	R	11.83	2.44	.52	11.80	2.70	.53	11.99	1.99	.50	12.28	2.24	.47
	C	11.88	2.42	.51	11.84	2.76	.52	11.99	1.99	.50	12.28	2.25	.47
3ETF6(37)	R	11.91	2.12	.50	11.81	2.24	.53	12.13	2.13	.46	12.23	1.66	.46
	C	11.97	2.12	.52	12.08	2.28	.54	12.17	2.15	.49	12.27	1.76	.50
3ETF6(38)	R	11.92	2.08	.51	12.03	1.66	.49	12.10	2.13	.47	12.53	1.83	.38
	C	11.96	2.12	.52	12.07	1.78	.51	12.14	2.13	.46	12.56	1.84	.45

at the extreme ranges of difficulty. The 3BTF11 form was the only relatively recent test edition of TOEFL with linear parameters to the TOEFL scale, and as it turned out, a much more difficult form than those used in the study. For Section III, where a set of reading comprehension items are dependent on a single passage, there is little leverage for manipulation of the level of difficulty. As a result, anchor tests of equivalent difficulty were not attainable for Section III.

The characteristics of the samples and the tests used in the equating comparisons can be summarized as follows:

	Section II	Section III
1.	Base form and operational test are of equivalent difficulty for all equatings.	Base form more difficult than the test to be equated for IRT equatings. For conventional equatings, base form and test to be equated were of equal difficulty.
2.	In the regular group the anchor test on subtest 37 was easier than that on 38. For the controlled group, the anchor tests were of equivalent difficulty.	For both groups, the anchor test on subtest 38 was more difficult than that on subtest 37.
3.	Anchor test roughly equivalent in difficulty to operational test for both groups.	Anchor test relatively more difficult than operational test. Greatest differences in difficulty observed on subtest 38.
4.	Equivalent ability (based on mean theta) on both subtests within groups.	Nonequivalent ability for subtest 37 and 38 within regular and controlled groups.

Given these conditions, comparisons for Section II can be considered to be a valid assessment of the effect of controlling for native language representation and of the stability of the equating methods. The variations observed in the forms for the Section III comparisons exemplify some of the difficulties existing for that section on operational TOEFL. Criterion comparisons for this section will be confounded by substantial departures from optimal conditions and should be interpreted accordingly.

Equating Criterion Comparisons, Regular Group

Discrepancy indices, based on scaled scores, for the regular group are presented in the top half of Table 6. Least error was observed for fixed b's scaling in both sections, with Modified IRT and Tucker equating following in the order of magnitude of error. A positive bias indicates that the criterion, i.e. the direct equating, tended to produce higher scores than the chain, and conversely for negative bias. In Section II, the chain results underestimated the criterion scores, while in Section III, the criterion was overestimated, this latter effect probably due, in part, to the variations in difficulty described above. Indeed, the major effect of the variations observed in Section III was the direction of bias; however, fixed b's equating was the least sensitive to these differences.

The magnitude of the proportion of squared bias for Modified IRT is observed to be quite large for both sections. Although the error for the three-parameters re-estimated model was large compared to other IRT methods, most of this error was due to the variance of the differences. These results are inherent in the models, however. The constant values of the a-parameter in the Modified IRT vary from form to form only by division of the slope of the linear transformation which, in turn, limits the range of the slopes of the test characteristic curves. When compared to the criterion, the major difference is simply a shift in location. As a result, the variability of the differences will be a small portion of the total error. On the other hand,

Table 6

Equating Criterion Comparisons

	Section II			Section III		
Method	Var.	Bias	Error	Var.	Bias	Error
			Regular Group			
Modified IRT	.04	(+)1.04	1.08	.35	(-)1.74	2.09
Fixed b's	.52	(+) .10	.62	.21	(+) .61	.82
3-parameter	3.84	(+)1.64	5.48	3.48	(-)2.36	5.84
Tucker	1.38	(+)2.55	3.93	1.10	(-)1.34	2.44
Levine	3.19	(+)4.02	7.21	2.48	(-)2.20	4.68
Equipercentile	2.00	(+)4.61	6.61	.51	(-)4.41	4.92
			Controlled Group			
Modified IRT	.15	(+)1.04	1.19	1.51	(-)9.23	10.74
Fixed b's	.16	.00	.16	.21	(-) .51	.72
Tucker	.05	(+)1.74	1.79	2.21	(-)2.72	4.93
Levine	.14	(+)2.72	2.86	5.88	(-)4.15	10.03
Equipercentile	.69	(+) .85	1.54	.61	(-)3.86	4.47

the slopes of the test characteristic functions for the 3-parameter model can vary substantially accounting for less systematic difference. These effects can be seen in graphs of the unweighted differences between the criterion and chained results in Figures 5 through 12. For linear equating the graph of the differences is simply a line of negative slope; the greater the absolute value of the slope, the greater the bias.

Standard errors of measurement ranged from 2.92 to 4.03 for Section II and from 2.61 to 3.20 for Section III. Other studies have determined that the standard error of equating is generally less than the standard error of measurement. Equating errors are larger at the tails of the distribution and, among equating methods, largest for equipercentile equating (Lord, 1981(a)). The mean difference (square root of the squared bias) for all criterion comparisons fell within the range of the standard error of measurement. The upper and lower limits of the converted score scale are, in part, determined by the method of equating. For the IRT equatings, these limits are the scaled scores at the upper asymptote of the test characteristic curve of the old form and a lower limit of 20 (see Appendix A). In the equipercentile equatings of this study, the upper and lower limits of the converted scores correspond to the range of observed raw scores. Depending on how the slopes differ in linear equating, greatest differences will generally occur at either or both extremes of the scale. Lists of conversions are given in Appendix C. The frequencies listed there, used to compute the discrepancy indices, are from a representative form of TOEFL. Table C9 in Appendix C presents the scaled scores and standard deviations for each of the experimental samples.

Figure 5.

Unweighted Differences Between Direct and Chain Irt Equatings, Section II, Regular Group

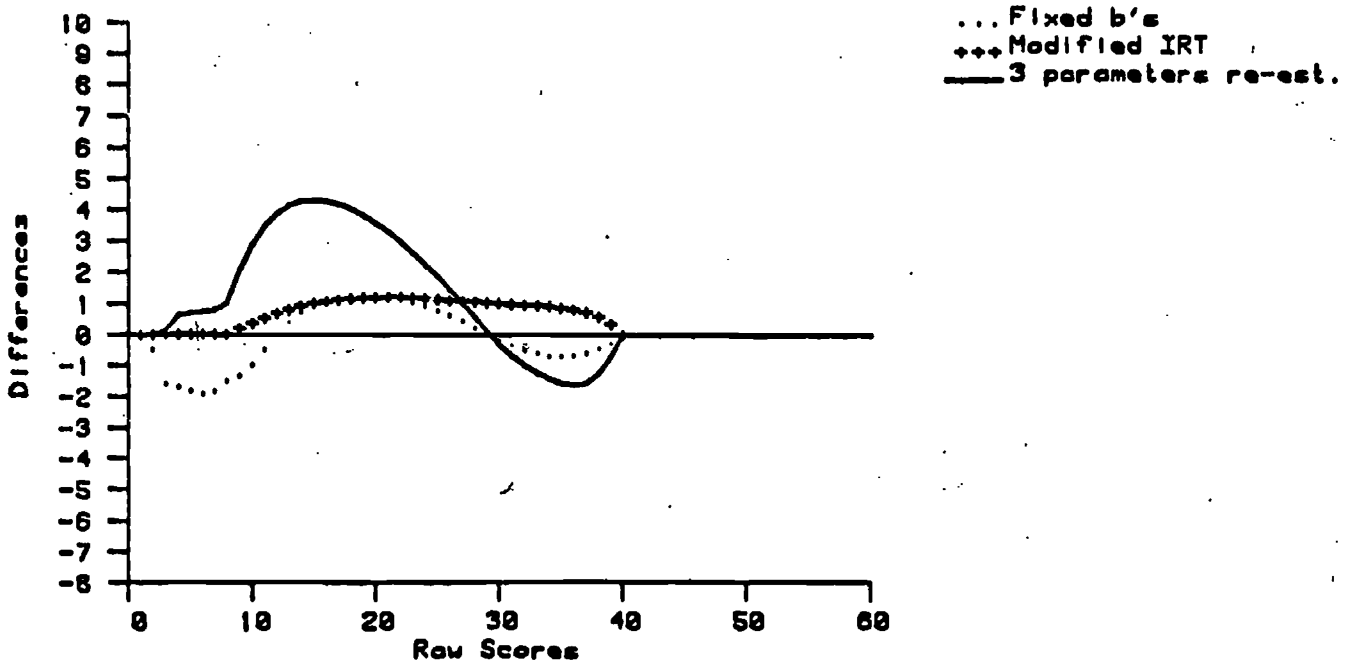


Figure 6.

Unweighted Differences Between Direct and Chain IRT Equatings, Section II, Controlled Group

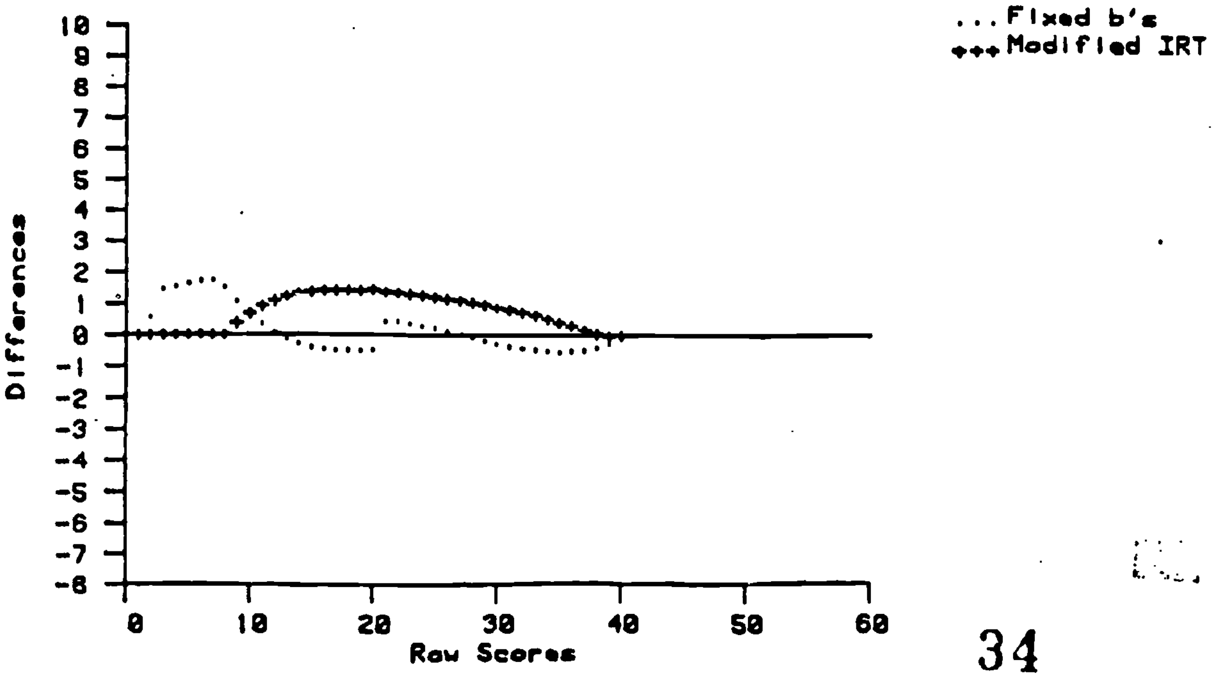


Figure 7.

Unweighted Differences Between Direct and Chain IRT Equatings, Section III, Regular Group

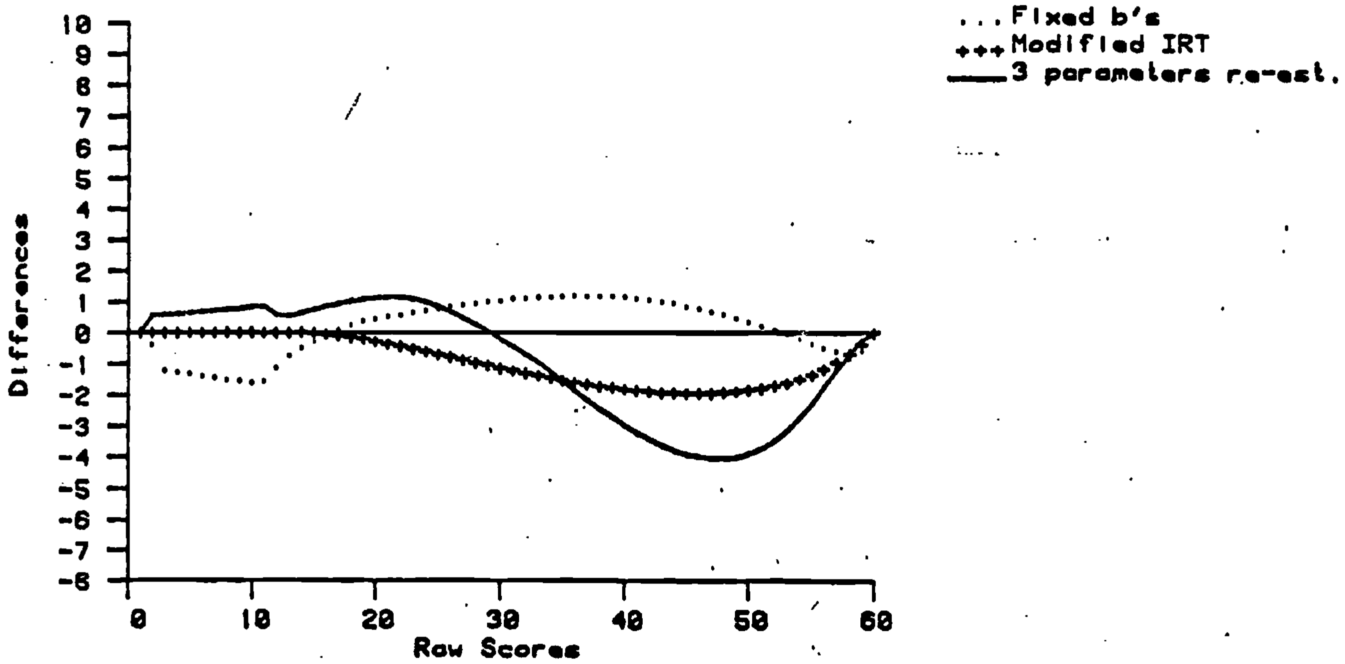


Figure 8.

Unweighted Differences Between Direct and Chain IRT Equatings, Section III, Controlled Group

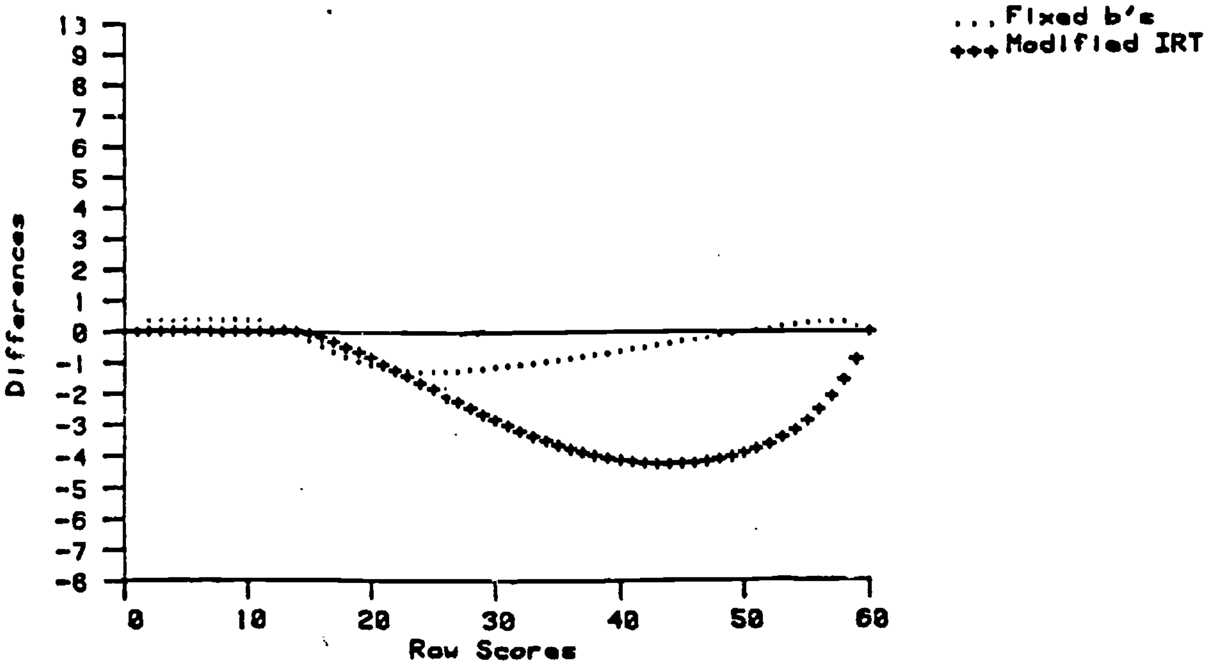


Figure 9.

Unweighted Differences Between Direct and Chain Conventional Equatings, Section II Regular Group

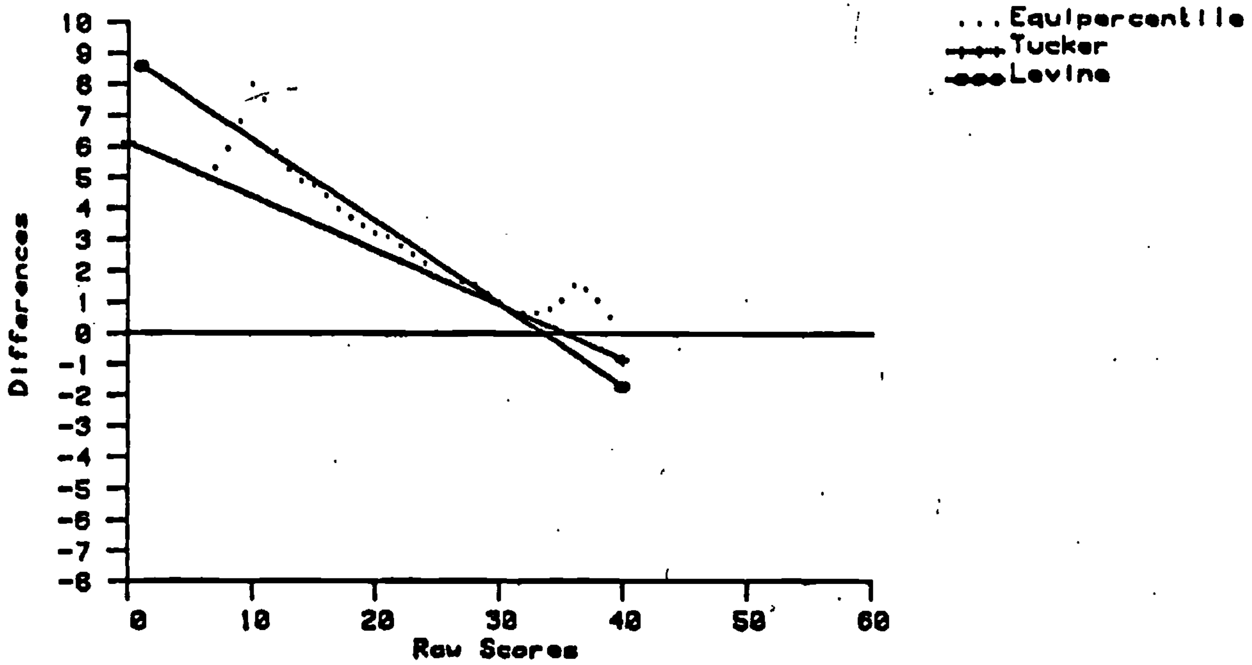


Figure 10.

Unweighted Differences Between Direct and Chain Conventional Equatings, Section II Controlled Group

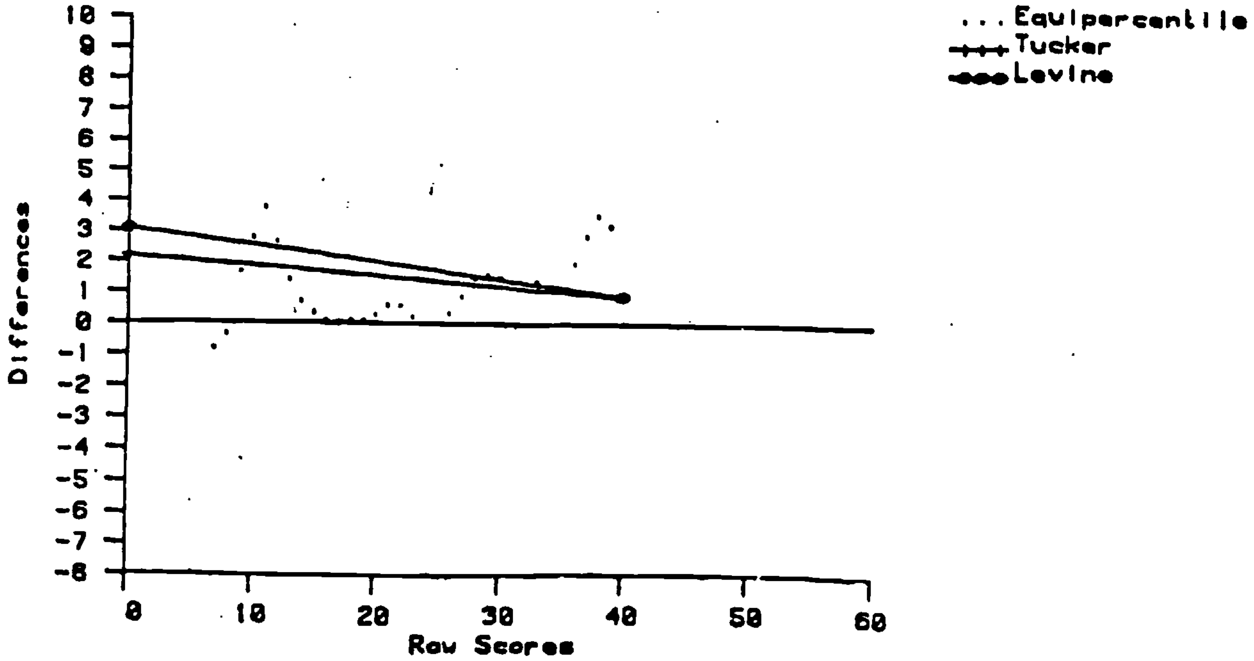


Figure 11.

Unweighted Differences Between Direct and Chain
Conventional Equatings, Section III
Regular Group

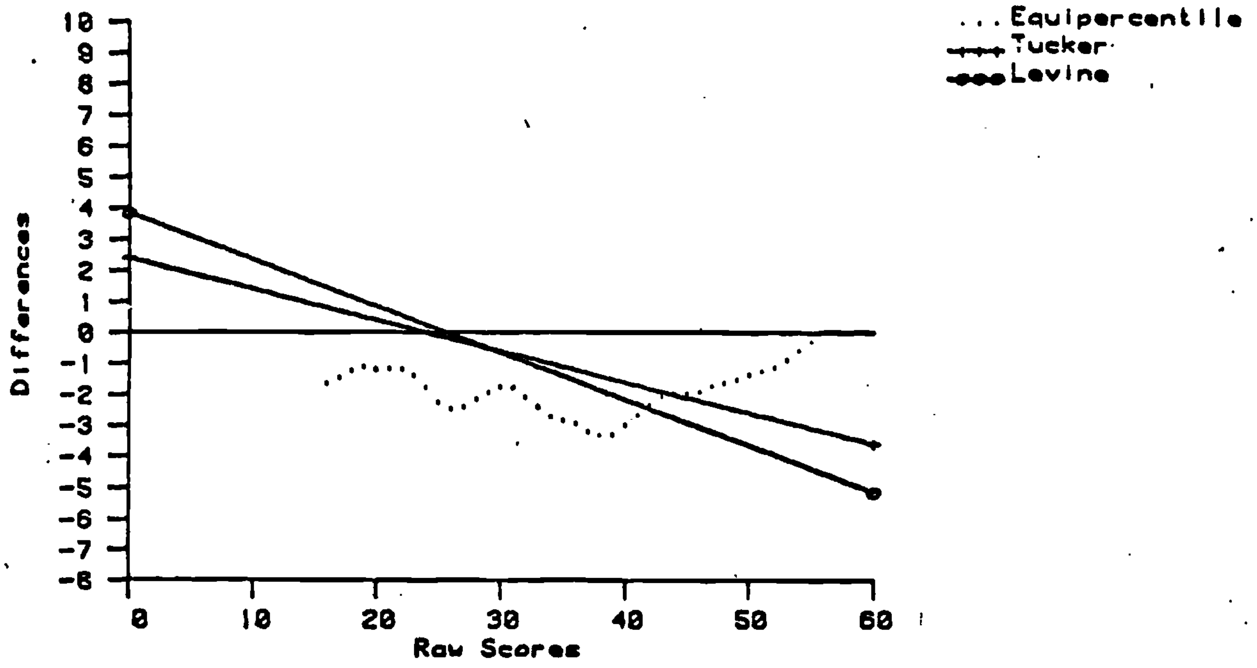
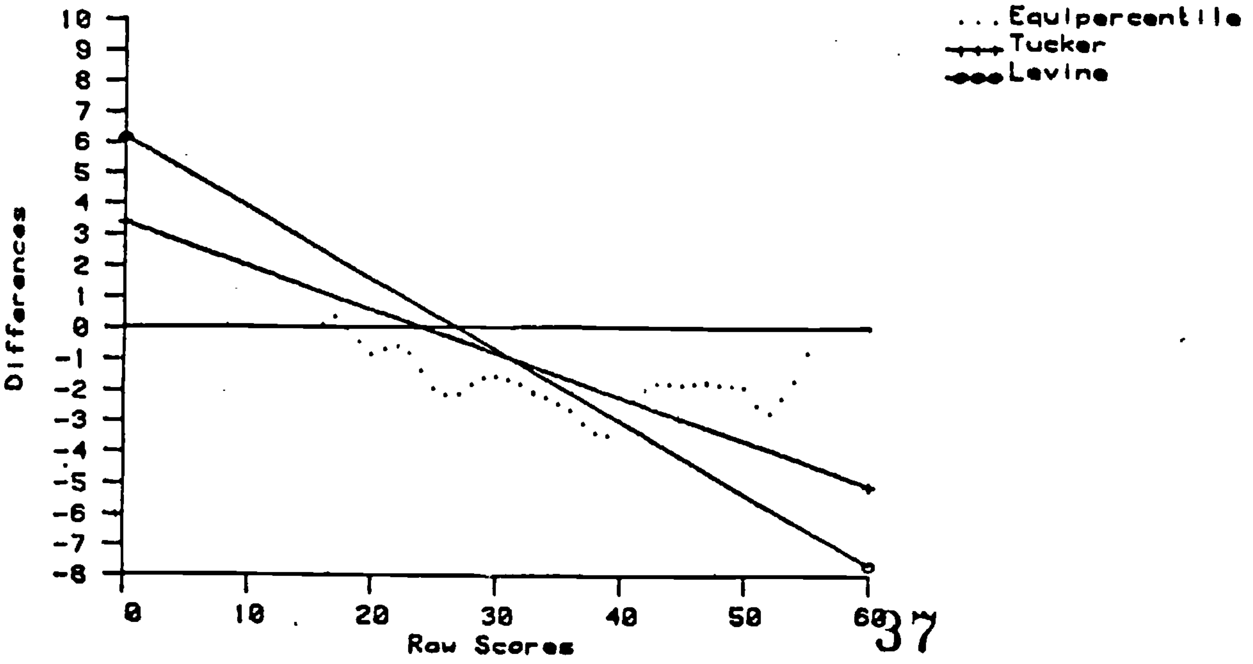


Figure 12.

Unweighted Differences Between Direct and Chain
Conventional Equatings, Section III
Controlled Group



Equating Comparisons, Controlled Group

Results of the equating comparisons for the controlled group are given in the lower half of Table 6 where it can be seen that fixed b's equating again yielded least error for both sections and was less than that observed for the regular group. It can also be observed that controlling for native language produced greater scale stability in Section II for most equating methods, but substantially more error for Modified IRT, Levine and Tucker equating in Section III.

The results for Section III may be related to the effects noted in Table 4, where the regressions of total score on the anchor test indicated greater dissimilarity between the regular and controlled groups on Section III than on Section II. Furthermore, the difficulty of the anchor test on Subtest 38 may have impacted heavily on the criterion comparisons for the controlled group.

Controlling for native language may have affected the dimensionality of Section III in some unexpected way. Since the controlled group (21 native languages out of 154) is more precisely defined in terms of language group composition, the group may have been more sensitive to subtle variations in test content in this section. It is possible that controlling on Section III might have a required a different kind of sampling, perhaps elimination of certain language groups altogether. It is probable that the complexities of the linguistic and factorial relationships of the test, as they impact on native language groups or groupings, militate against any simple method of sampling. Until these relationships are better understood, a random sample of the total group appears to be the most reliable method of sampling TOEFL examinees. However, a more fundamental problem, dealing with the structure of Section III in terms of parallelism across forms and the implications for its construct validity, is suggested by these results (see Swinton & Powers, 1980, pp. 20-21; Alderman & Holland, 1981, p. 18).

Item Ability Regressions

Figures 13 and 14 are item ability regressions for six items from Section III of 3ETF6, subtest 38, based on the Modified IRT with sample sizes of 308 and 988 respectively. Figure 15 displays item curves for the same items derived from the 3-parameters re-estimated model. The vertical lines of these plots denote the standard error of the curve. While it is clear that the larger sample size improves the fit of the item ability regressions, the effect on the equated scores is small indeed; total error amounting to only .003 for comparison of the Modified IRT based on the two sample sizes. Plots of the Modified IRT and 3-parameters re-estimated are quite comparable with the exception of item 24 which was unable to be fit by an average value of the a-parameter. Of the sixty items in this section, six could be identified as requiring a slope parameter other than the average. The discrepancy index for the 3-parameter vs. Modified IRT (based on the smaller sample) is only .09 as listed in Table 8. In the absence of the effects due to methods of scaling or linking, the practical impact on the equated scores of small variations in the a-parameter among a few items seems to be relatively minor.

As noted earlier, fixed b's equating has been observed to occasionally result in poor fit among precalibrated items. Indeed, it was for this reason that the 3-parameters re-estimated model was included in this study. An example of an item better fit by re-estimating the b-parameter is given in Figure 16. This was the most deviant fit of the precalibrated items in these comparisons.

Figure 13. Item ability regressions for six items, Section III, Modified IRT, N = 308.

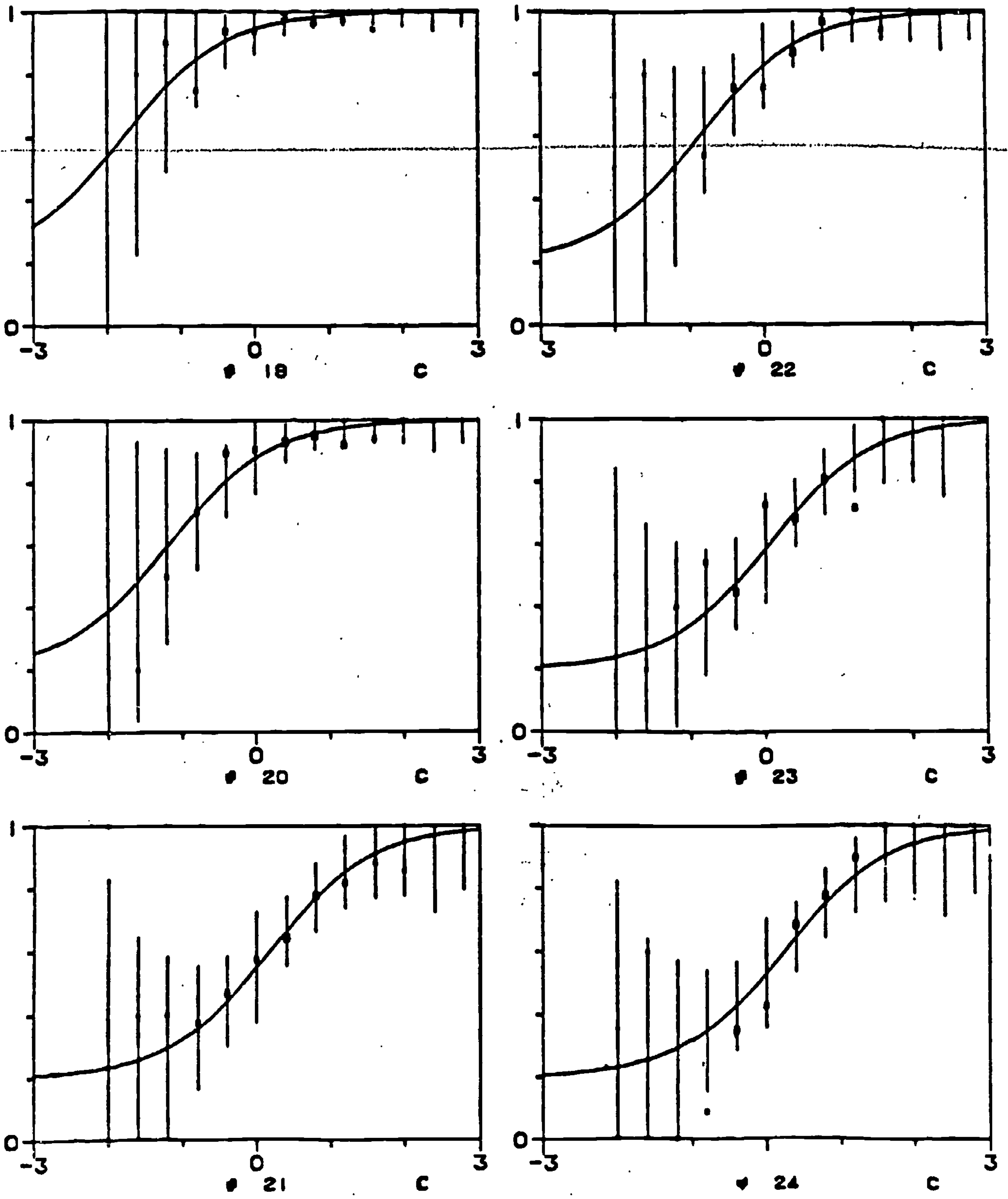


Figure 14. Item ability regressions for six items, Section III, Modified IRT, N = 988.

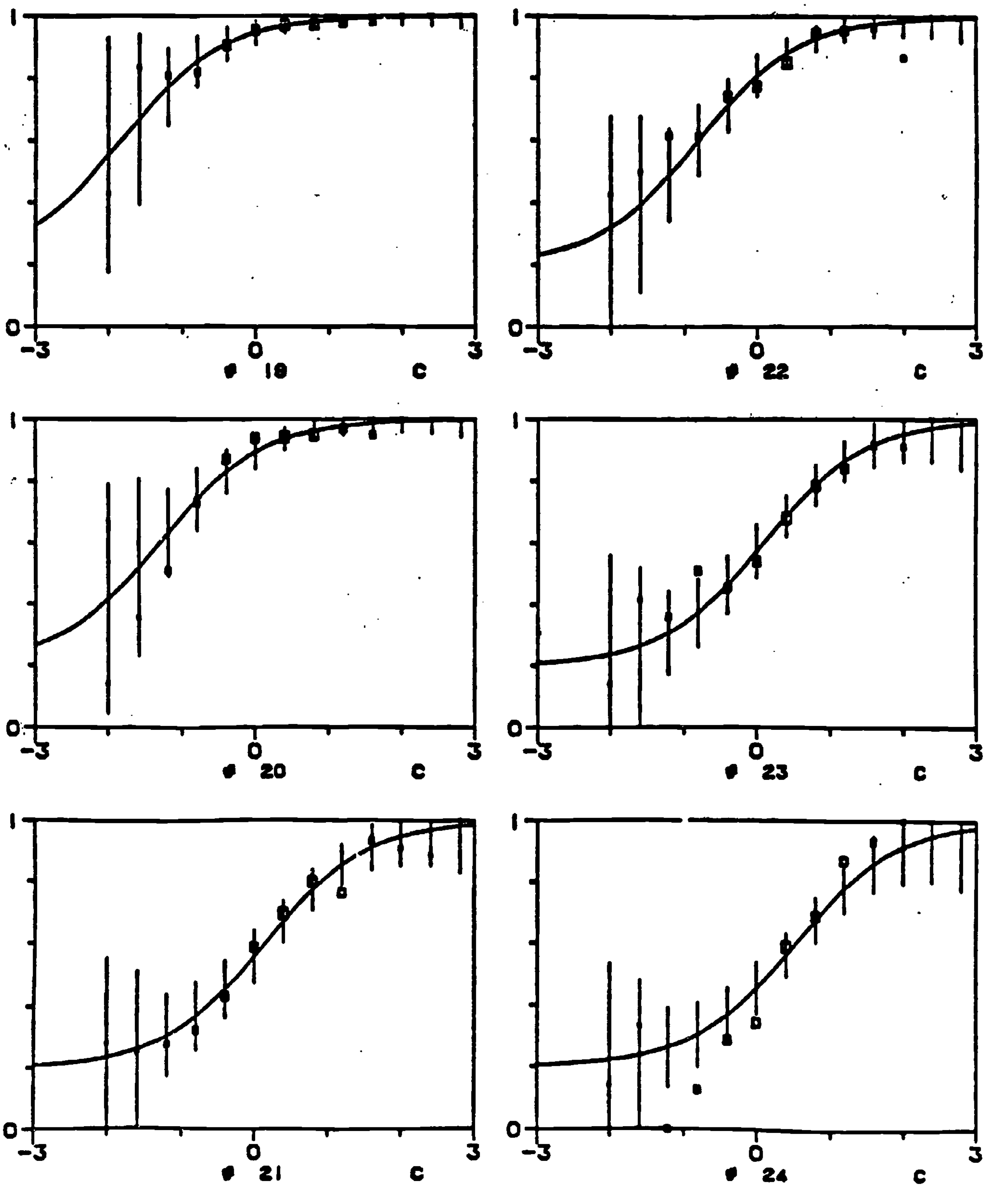


Figure 15. Item ability regressions for six items, Section III, 3-parameters re-estimated, N = 988.

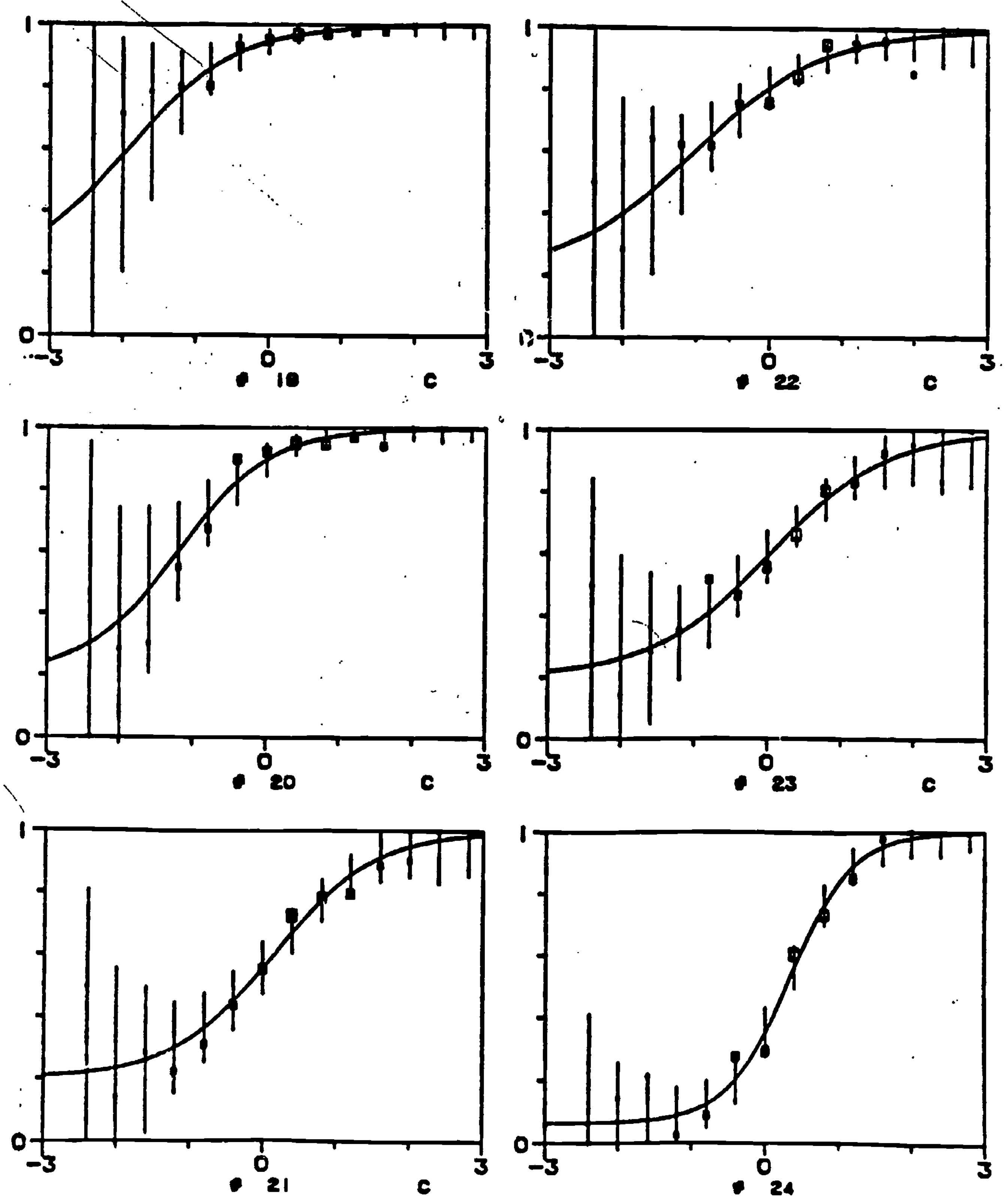
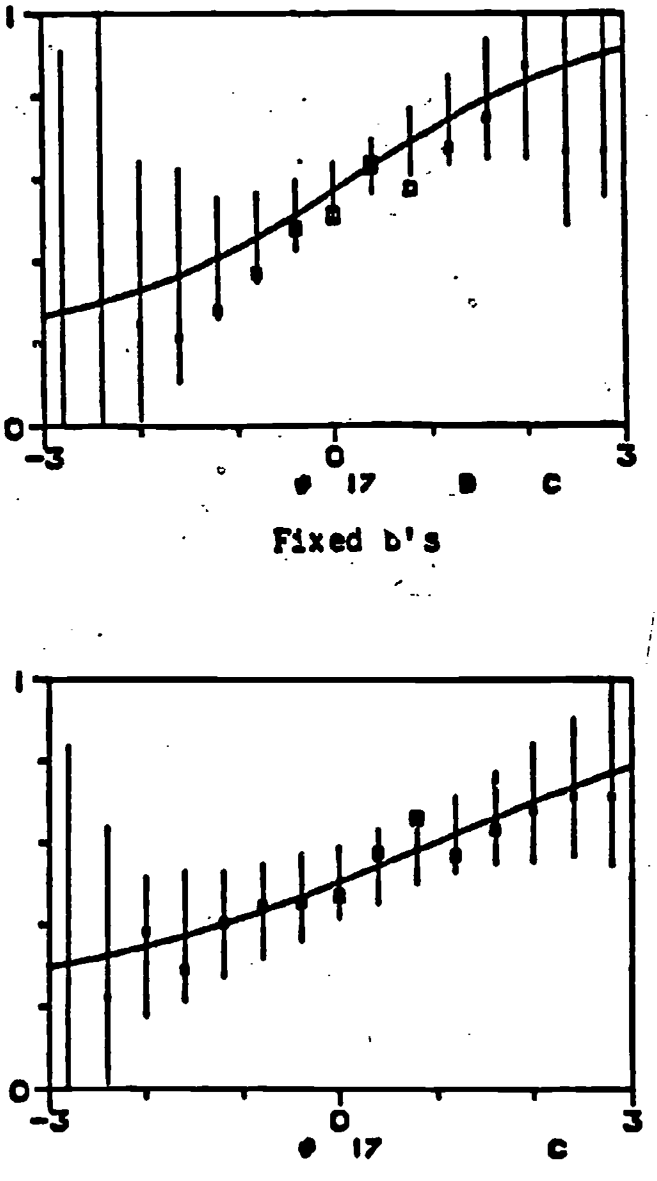


Figure 16. Item ability regressions for an item scaled by fixed b's and 3-parameters re-estimated.



Fixed b's

Three parameters re-estimated

Equatings Based on Separate Language Groups

One of the experimental administrations provided samples of sufficient size to independently equate four language groups. Fixed b's conversions were computed for Arabic, Chinese, Japanese and Spanish examinees taking 3ETF3. These equatings were compared with that based on the total controlled group. Discrepancies as listed in Table 7 were small in general, and the rankings in terms of total error were somewhat different in the two sections.

Methods Comparisons

Discrepancy indices between methods based on the direct and chained results for the regular group are given in Tables 8 and 9. Differences observed in the two tables are illustrative of a major source of error in equating. From Table 8, all else being equal, the various methods produce comparatively similar results, while discrepancies listed in Table 9 reflect, among other things, the variability due to methods of linking the forms. From Table 9, we observe the not too surprising result that Tucker and Levine equatings are the most similar when the effects of linking are taken into account. Among the largest differences observed are the discrepancies between fixed b's and the 3-parameters re-estimated which probably incorporates some of the effects of re-estimating the b-parameters vs. holding them fixed. Modified IRT vs. fixed b's have smallest error among all the IRT comparisons. It can also be observed that the values of total error in Section II tend to be higher than those in Section III. This may be due to the fact that a 41 point observed score scale in Section II, as contrasted with a 61 point raw score scale for Section III is being stretched to one that can theoretically range from 20 to 80.

Table 7

Comparisons of IRT Fixed b's Equating for Four Language Groups with Total Group (Controlled)

	Section II			Section III		
	Var. Diff.	Sq. Bias	Error	Var. Diff.	Sq. Bias	Error
Arabic	.02	.36	.38	.10	.01	.11
Chinese	.10	.01	.11	.04	.04	.08
Japanese	.26	.87	1.13	.05	.14	.19
Spanish	.29	.00	.29	.02	.41	.43

Table 8

Total Error and Squared Bias*, Comparisons of Equating Methods, Direct Results, Regular Group

	Methods					
Methods**	1	3	F	T	L	E
			Section II			
1	—	.27	.44	.26	.07	.29
3	.12	—	.06	.14	.14	.11
F	.12	.00	—	.30	.44	.11
T	.23	.02	.02	—	.08	.28
L	.04	.02	.14	.08	—	.17
E	.03	.04	.03	.11	.00	—
			Section III			
1	—	.09	.07	.52	.31	.27
3	.01	—	.07	.96	.58	.26
F	.05	.01	—	.69	.29	.19
T	.04	.10	.18	—	.12	.51
L	.02	.06	.02	.00	—	.30
E	.02	.00	.00	.08	.05	—

*Total error above diagonal, squared bias below diagonal.

**1-(Mod, IRT), 3-(3-param), F-(Fixed b's), T-(Tucker), L-(Levine), E-(Equipercentile)

Table 9

Total Error and Squared Bias*, Comparisons of Equating Methods, Chain Results, Regular Group

	Methods					
Methods**	1	3	F	T	L	E
			Section II			
1	—	3.95	1.70	2.57	4.60	3.36
3	.37	—	9.68	1.43	1.33	1.47
F	1.01	2.62	—	7.94	11.06	9.35
T	1.12	.20	4.26	—	.36	.17
L	1.42	.34	4.83	.02	—	.35
E	1.64	.44	5.26	.04	.00	—
			Section III			
1	—	1.29	.49	1.26	1.52	1.43
3	.11	—	3.31	1.20	.86	3.66
F	.09	.41	—	2.39	3.09	1.56
T	.13	.48	.00	—	.18	3.79
L	.00	.10	.11	.15	—	3.51
E	.81	.33	1.43	1.53	.73	—

*Total error above diagonal, squared bias below diagonal.

**1-(Mod, IRT), 3-(3-parameter), F-(Fixed b's), T-(Tucker), L-(Levine), E-(Equipercentile)

Discussion and Conclusions

Scale stability using a controlled group. The premise that language group control would improve the stability of the TOEFL scale for various equating methods was generally supported by the outcomes in Section II. The opposite result was observed in Section III, in which case error was increased for Modified IRT and linear equating methods for the group controlled for native language. Only for fixed b's IRT and equipercentile equatings was there a small reduction in error when this type of control was excercised. While marked ability differences on the two subtests may have confounded the comparisons for the controlled group in Section III, the possibility remains that other factors related to the multildimensionality of Section III for some language groups contributed to these results. Based on the findings of this study, controlling for language group representation is not recommended for operational TOEFL at this time.

Fixed b's scaling. The current method of IRT equating by fixing b's produced the greatest scale stability for both groups. It is not surprising that this method of scaling would produce such excellent results in terms of the criterion of this study since the location parameters for half (or more) of the items in each section are fixed with only the a- and c-parameters allowed to vary. Assuming that the b-parameters held for subsequent groups, bias in the a-parameters would be a major source of error. Positive statistical bias does exist for the a's and is greatest for highly discriminating, difficult items (Lord, 1982). In fixed b's scaling, an upper limit of 1.5 is placed on the estimated a-parameter which may reduce the effect of bias for this group of items. Plots of precalibrated vs. re-estimated a's collected over time have exhibited no obvious evidence of bias, differing only in degree of scatter about the line through the origin. A detailed analysis of the precalibrated and re-estimated a's has also failed to detect any evidence of bias. In practical terms, fixed b's equating offers flexibility and item security which cannot be derived from methods of equating based on a block of items common to two forms since compromise of the first form can jeopardize an entire future administration.

Precalibrated items that are identified as seriously aberrant in terms of fit might be treated as noncalibrated and all parameters re-estimated on the current group. Such items could be identified prior to item calibrations by comparing equated deltas based on pretesting with those derived in a preliminary item analysis (equated deltas and b-parameters have been found to correlate very highly, approximately .96). Such a procedure is workable so long as these items remain a small proportion of the precalibrated items, as they are currently.

Modified IRT. Results for the Modified IRT were quite satisfactory for both sections in the regular group. However, as the results for the controlled group indicate, this method may be sensitive to differences in ability. Coupled with the poor performance of the one-parameter IRT model when two tests of unequal difficulty are being equated (Petersen, Marco & Stewart, 1982), Modified IRT is probably not suited to TOEFL data where such variations are likely to occur.

A practical advantage to this method is the smaller sample size required for parameter estimation which would have material impact on the difficulties involved in maintaining a precalibrated item pool. Associated with this is the reduction in computer costs for estimating parameters. Typical costs for running LOGIST (IV) were as follows:

	Modified IRT	3-Parameters
	N = 300	N = 1500
60 items	$10.98	$49.31
90 items	13.02	77.29

It is clear that application of Modified IRT to TOEFL would require acceptance of some inadequately fit items.

3-parameters re-estimated. The relatively large error associated with estimating all three parameters may reflect the "true" effects of the variability associated with TOEFL testing groups. The hypothesis that a less sensitive IRT model might produce better results with TOEFL data was supported by the outcomes for fixed b's as compared to those for the 3-parameters re-estimated model. Fixed b's scaling, as implemented by TOEFL, might be categorized as a less sensitive model by virtue of the constraints imposed on the variation of some of the b-parameters and the limits on the a-parameters. In contrast, a great deal more information about the current group is introduced into the scaling process in estimating all three parameters.

Conventional equating methods. Of the linear methods, Tucker equating produced the best results, outperforming the 3-parameter re-estimated IRT model. It might be concluded that basic assumptions of Levine equating were not met by the data as, for example, the requirement of parallelism of the anchor and operational tests. As noted earlier, the equating conditions of this investigation for Section III were roughly similar to Variation 8 of the Petersen, Marco and Stewart (1982) study which produced large total error. Under optimal conditions, better performance for linear equating might be expected. The results in Section III provide some information regarding the robustness of various equating methods under less than ideal circumstances. In terms of these outcomes, Tucker equating fared rather well, even though it is known to be less than ideal when ability distributions differ.

Limitations of the study. The criterion of this study was the stability of the scale over several links and no attempt was was made to evaluate item fit. Indeed, the Modified IRT version would have been ruled out apriori on the basis of this criterion. The implicit assumption was that all IRT methods would provide reasonable fit to the data. The conclusions from this study are limited by the tenability of this assumption.

Conclusions. The possible dangers and difficulties associated with sampling the extremely complex TOEFL testing population for the purpose of equating were demonstrated in this study with the resulting recommendation for the continuation of random sampling of the total testing group. This recommendation emanates from the results for Section III for the controlled group which were consistent with findings from earlier studies (Swinton & Powers, 1980; Alderman & Holland, 1981), therefore associated with this is the need to evaluate the basic structure of Section III in terms of its parallelism across forms and its construct validity.

Fixed b's, Modified IRT and Tucker linear equating produced satisfactory results for Sections II and III of TOEFL in the regular group of examinees. Owing to its apparent sensitivity to ability differences, Modified IRT should probably not be considered for application to TOEFL. Consequently, fixed b's and Tucker linear equating appear to be the best candidates for equating TOEFL test scores. Each has practical and/or theoretical advantages and disadvantages which can be weighed in terms of program resources and the best interests of the examinee population. The question of item security is of paramount importance to the TOEFL program which administers the tests worldwide, precluding tight control of the security of pretest items. Indeed, problems associated with compromised items in the Far East was the primary reason for adopting IRT in terms of fixed b's scaling. The compromise of a single test form overseas could invalidate the entire form to which it is linked. Fixed b's equating, depending as it does on equating items from many forms avoids this major difficulty. It is clear that tradeoffs between ideal statistical conditions and practical realities cannot be avoided, for there is probably far greater error in compromised equating items than in the occaionally observed poor fit of a few precalibrated items.

In applying IRT to TOEFL data, response patterns of a complex population are being fit by a complex model providing ample opportunities for evidence of the error incurred in analyzing behavioral data via mathematical models. Aside from concerns in terms of meeting the basic assumptions of the model, are questions related to the statistical properties of model parameters (e.g., bias, see Lord, 1981(b); 1982). Associated with this are the effects of some of the artificial constraints on parameter estimation such as the value chosen for the upper limit of the a-parameter, to which can be added unpredictables such as variations in instructional patterns which may be a source of some of the differences observed in item fit of precalibrated items, errors due to test administration, and finally, the social and political factors which can affect the nature of the population. While Tucker linear equating is subject to many of the same sources of error as IRT equating, it depends on far fewer parameters, and as the data in Table 8 demonstrate, is appreciably better, in terms of scale stability, than the 3-parameter IRT model freely applied to TOEFL data, that is, estimating all three parameters for all items. These results imply that TOEFL data probably require some constrained method of IRT equating in order to control for the many sources of variability.

Although ideal equating conditions could not be established in this study for the purpose of evaluating the various equating methods, this lack of optimality may have provided a more accurate reflection of practical implications. This study has demonstrated that a randomly sampled population consisting of 154 or more language groups is viable for a restricted form of IRT methodology. Ostensibly, the assumptions underlying Tucker linear equating are not being seriously violated by TOEFL data. In fact Tucker equating demonstrated a measure of robustness in face of less than ideal circumstances. However global this population, it apparently possesses lawful regularities in its own right amenable to certain statistical operations. The criteria for this conclusion are the empirical results of an equating experiment. While it would have been desirable to establish the suitability of a given equating method to TOEFL via more analytic methods, consistency of results in practical applications is often the only source of methodological validation.

References

Alderman, D. L., & Holland, P. W. Item performance across native language groups on the Test of English as a Foreign Language. (RB-81-16). Princeton, N.J.: Educational Testing Service, 1981.

Angoff, W. H. Basic equations in scaling and equating. In S. S. Wilks (Ed.), Scaling and Equating College Board Tests. Princeton, N.J.: Educational Testing Service, 1961, 120-129.

Angoff, W. H., & Dyer, H. S. The admissions testing program. In W. H. Angoff (Ed.), The College Board Admissions Program: A technical report on research and development activities relating to the Scholastic Aptitude Test and Achievement Tests. New York: College Entrance Examination Board, 1971.

Angoff, W. H., & Sharon, A. T. Patterns of test and item difficulty for six foreign language groups on the Test of English as a Foreign Language. (RB-72-2). Princeton, N.J.: Educational Testing Service, 1972.

Angoff, W. H. Summary and derivations of equating methods used at ETS. In P. W. Holland and D. B. Rubin (Eds.). Test equating. New York: Academic Press, 1982.

Braun, H. I., & Holland, P. W. Observed score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin (Eds.). Test equating. New York: Academic Press, 1982.

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

Gulliksen, H., & Wilks, S. S. Regression tests for several samples. Psychometrika, 1950, 15, 91-114.

Levine, R. S. Equating the score scales of alternate forms administered to samples of different ability. (RB-55-23). Princeton, N.J.: Educational Testing Service, 1955.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980.

Lord, F. M. The standard error of equipercentile equating. (RR-81-48). Princeton, N.J.: Educational Testing Service, 1981(a).

Lord, F. M. Unbiased estimators of ability parameters, of their variance, and of their parallel forms reliability. (RR-81-50). Princeton, N.J.: Educational Testing Service, 1981(b).

Lord, F. M. Statistical bias in maximum likelihood estimators of item parameters. (RR-82-20-ONR). Princeton, N.J.: Educational Testing Service, 1982.

Marco, G. L. Item characteristic curve solutions to 3 intractable testing problems. Journal of Educational Measurement. 1977, 14, 139-160.

Petersen, N. S., Marco, G. L., & Stewart, E. E. A test of the adequacy of linear score equating methods. In P. W. Holland and D. B. Rubin (Eds.). Test equating. New York: Academic Press, 1982.

Stocking, M. L., & Lord, F. M. Developing a common metric in item response theory. (RR-82-25-ONR). Princeton, N.J.: Educational Testing Service, 1982.

Swinton, S. S., & Powers, D. E. Factor analysis of the Test of English as a Foreign Language for several language groups. (RR-80-32). Princeton, N.J.: Educational Testing Service, 1980.

Wood, R. L.,Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (RM-76-6). Princeton, N.J.: Educational Testing Service, 1976.

APPENDIX A

Equating TOEFL Test Scores

Once item parameters have been estimated, test scores x, on a new form of TOEFL are equated to test score, X, on a base form using true score equating (Lord, 1980; pp. 199-205). The equated scores are then placed on the reported score scale through a known linear transformation. If Y is a scaled score, A and B are known constants that linearly transform X, a number right score on the base form, then

$$Y = AX + B \qquad 3(a)$$

$$X = \Sigma P_i(\theta) \qquad 3(b)$$

$$x = \Sigma P_j(\theta) \qquad 3(c)$$

define a transformation Y(x) which equates observed score x, through the elimination of x and θ from the given equations. In practice, this is accomplished by substituting observed number right score on the new form for x in 3(c), then using the known item parameter estimates, solving for θ. Inserting this value of θ in 3(b) and using known item parameters for this form, an equated number right score, X, results. Scaled scores follow from 3(a). Scaled scores are rounded to the nearest integer with those above 80 and below 20 set to 80 and 20, respectively. The total test scaled score is obtained by summing the section scaled scores and multiplying the result by 10/3. The true score distribution is bounded below by Σc_i, thus, the conversions obtained from the equating method above apply only to scores above $x = \Sigma c_i$. For observed scores below this level, where there are relatively few observations, a line relating the c's on the old and new form is calculated.

Appendix B

Some Concepts Underlying Equipercentile and Tucker Linear Equating

In order to implement the definitions of equipercentile and linear equating given on page 4, a synthetic population T is formed as a weighted composite of Group P taking old form Y and group Q taking new form Y. In the case of common item equating, information derived from a set of items, V, common to old form Y and new form X, aids in the determination of the distributions and first two moments of the synthetic equating group. Adopting the development of Braun and Holland (1982), the data necessary to produce this information is given in Table B.

Table B

Distributions Required for Tucker and Equipercentile Common Item Equating

	Old Form Y	New Form X	Common Items
Group P	$F_P(Y\|v)$	$G_P(X\|v)$*	$K_P(v)$
	$E_P(Y\|v)$	$E_P(X\|v)$*	
Group Q	$F_Q(Y\|v)$*	$G_Q(X\|v)$	$K_Q(v)$
	$E_Q(Y\|v)$*	$E_Q(X\|v)$	

* Not observable

In this table, for example, F_P is the conditional distribution of Y given $V = v$ in population P, $K_P(v)$ is the distribution of V in P and E_P is the regression of Y on V in P. The purpose, then, is to derive unconditional distributions of the old and new forms for the synthetic population T, $F_T(y)$ and $G_T(x)$, given the information listed in Table B. $F_T(y)$ can be written

$$F_T(y) = \int F_P(Y|v)dK_P\ w_1 + \int F_Q(Y|v)dK_Q(v)\ w_2.$$

However, F_Q is not observable, but if it is assumed that $F_P = F_Q$, then, F_T is

$$F_T(y) = \int F_P(Y|v)dK_T(v)$$

where $K_T(v) = w_1K_P(v) + w_2K_Q(v)$. Similarly, the distribution function of X is

$$G_T(x) = \int G_Q(X|v)dK_T(v).$$

The equipercentile equating function according to its definition is then,

$$e_Y(x) = F_T^{-1}(G_T(x)).$$

For the case of Tucker linear equating, the function is

$$L_Y(x) = \mu_y + \sigma_x/\sigma_y \, (X - \mu_x).$$

Assuming $E_P = E_Q$ from Table B for X and Y, then

$$\mu_y = \int E_P(Y|v)dK_T(v)$$

$$\mu_x = \int E_Q(X|v)dK_T(v).$$

Formulas for the variances of the synthetic population are similarly derived based on analagous assumptions (i.e., $\mathrm{Var}_P(Y|v) = \mathrm{Var}_Q(Y|v)$, etc.). On the assumptions of linearity of regressions and homoscedasticity of errors, one result of the foregoing is

$$E_P(Y|v) = E_Q(Y|v) = av + b$$

$$\mathrm{Var}_P(Y|v) = \mathrm{Var}_Q(Y|v) = \sigma^2,$$

with analogous formulas for form X. Thus, Tucker linear equating as well as common item equipercentile equating is based on untestable assumptions since data for the old form in Q as well as for the new form in P is not available.

APPENDIX C*

Final Converted Scores

Table C1
IRT Conversions, Regular Group
Section II

RAW	FRQ	MIRT37	MIRT38	3P37	3P38	FB37	FB38
40	58	65.83	65.83	65.83	65.83	65.83	65.83
39	109	64.56	64.90	65.46	64.75	64.60	64.86
38	135	63.27	63.84	64.84	63.59	63.37	63.83
37	167	62.01	62.72	64.02	62.47	62.16	62.76
36	235	60.75	61.56	63.02	61.38	60.96	61.64
35	276	59.5	60.37	61.86	60.28	59.74	60.46
34	326	58.25	59.17	60.58	59.13	58.51	59.22
33	362	57.00	57.94	59.18	57.91	57.26	57.90
32	454	55.74	56.70	57.64	56.62	56.00	56.53
31	438	54.47	55.45	55.98	55.27	54.74	55.11
30	453	53.18	54.19	54.21	53.88	53.48	53.68
29	504	51.89	52.92	52.37	52.46	52.24	52.24
28	513	50.59	51.65	50.50	51.04	51.03	50.82
27	551	49.28	50.36	48.65	49.64	49.85	49.44
26	512	47.97	49.07	46.85	48.26	48.70	48.11
25	508	46.65	47.78	45.12	46.96	47.57	46.82
24	495	45.32	46.48	43.45	45.65	46.47	45.58
23	468	44.00	45.18	41.84	44.41	45.38	44.38
22	422	42.68	43.88	40.29	43.21	44.30	43.21
21	424	41.37	42.57	38.78	42.03	43.23	42.06
20	344	40.07	41.26	37.33	40.87	42.16	40.94
19	302	38.77	39.96	35.93	39.72	41.06	39.82
18	290	37.49	38.65	34.57	38.58	39.95	38.71
17	234	36.22	37.34	33.27	37.43	38.80	37.60
16	211	34.96	36.03	32.02	36.28	37.60	36.50
15	158	33.72	34.72	30.82	35.12	36.34	35.39
14	122	32.51	33.42	29.68	33.95	35.00	34.28
13	115	31.33	32.12	28.62	32.75	33.58	33.18
12	75	30.18	30.85	27.65	31.52	32.07	32.09
11	67	29.09	29.62	26.79	30.26	30.51	31.01

*MIRT=Modified IRT; 3P=3 parameters re-estimated; FB=Fixed b's; E=Equipercentile; L=Levine; 37 and 38 refer to subtests 37 and 38.

Table C1 cont'd

RAW	FRQ	MIRT37	MIRT38	3P37	3P38	FB37	FB38
***	***	******	******	****	****	****	****
10	52	28.08	28.44	26.09	28.98	28.97	29.96
9	35	27.15	27.34	25.60	27.68	27.59	28.95
8	19	26.38	26.38	25.36	26.37	26.46	27.98
7	15	25.15	25.16	24.26	25.06	25.22	27.06
6	8	23.92	23.93	23.09	23.83	23.98	25.92
5	2	22.70	22.71	21.91	22.61	22.75	24.58
4	1	21.47	21.48	20.74	21.38	21.51	23.23
3	2	20.24	20.25	20.00	20.16	20.27	21.88
2	0	20.00	20.00	20.00	20.00	20.00	20.53
1	0	20.00	20.00	20.00	20.00	20.00	20.00
0	0	20.00	20.00	20.00	20.00	20.00	20.00

Table C2

Linear and Equipercentile Conversions
Regular Group
Section II

RAW ###	FRQ ###	T37 #####	T38 #####	L37 #####	L38 #####	E37 #####	E38 #####
40	58	67.35	66.50	68.44	66.72	66.60	66.38
39	109	65.89	65.21	66.89	65.43	65.09	65.60
38	135	64.43	63.92	65.34	64.15	63.64	64.68
37	167	62.97	62.64	63.80	62.87	62.21	63.62
36	235	61.51	61.35	62.25	61.58	60.90	62.43
35	276	60.05	60.06	60.70	60.30	59.70	60.74
34	326	58.59	58.78	59.16	59.02	58.49	59.25
33	362	57.13	57.49	57.61	57.74	57.27	57.92
32	454	55.67	56.20	56.06	56.45	55.95	56.58
31	438	54.21	54.92	54.52	55.17	54.53	55.23
30	453	52.75	53.63	52.97	53.89	52.97	53.95
29	504	51.29	52.34	51.42	52.61	51.38	52.64
28	513	49.83	51.06	49.87	51.32	49.75	51.29
27	551	48.37	49.77	48.33	50.04	48.22	49.84
26	512	46.91	48.48	46.78	48.76	46.74	48.36
25	508	45.44	47.20	45.23	47.47	45.07	46.85
24	495	43.98	45.91	43.69	46.19	43.52	45.73
23	468	42.52	44.62	42.14	44.91	42.12	44.62
22	422	41.06	43.34	40.59	43.63	40.68	43.45
21	424	39.60	42.05	39.05	42.34	39.22	42.27
20	344	38.14	40.76	37.50	41.06	37.70	40.88
19	302	36.68	39.48	35.95	39.78	36.22	39.65
18	290	35.22	38.19	34.40	38.50	34.99	38.65
17	234	33.76	36.91	32.86	37.21	33.65	37.61
16	211	32.30	35.62	31.31	35.93	32.09	36.49
15	158	30.84	34.33	29.76	34.65	30.65	35.38
14	122	29.38	33.05	28.22	33.37	29.42	34.29
13	115	27.92	31.76	26.67	32.08	27.95	33.19
12	75	26.46	30.47	25.12	30.80	26.20	32.03
11	67	25.00	29.19	23.58	29.52	23.23	30.72
10	52	23.54	27.90	22.03	28.23	21.19	29.20
9	35	22.08	26.61	20.48	26.95	20.87	27.68
8	19	20.62	25.33	18.93	25.67	20.54	26.47
7	15	19.16	24.04	17.39	24.39	20.22	25.50
6	8	17.70	22.75	15.84	23.10	-	-
5	2	16.24	21.47	14.29	21.82	-	-
4	1	14.78	20.18	12.75	20.54	-	-
3	2	13.32	18.89	11.20	19.26	-	-
2	0	11.86	17.61	9.65	17.97	-	-
1	0	10.40	16.32	8.11	16.69	-	-
0	0	8.93	15.03	6.56	15.41	-	-

Table C3

IRT Conversions, Controlled Group
Section II

RAW	FRQ	MIRT37	MIRT38	FB37	FB38
***	***	******	******	****	****
40	58	65.83	65.83	65.83	65.83
39	109	64.71	64.66	64.23	64.53
38	135	63.49	63.53	62.95	63.37
37	167	62.24	62.40	61.71	62.21
36	235	60.98	61.27	60.50	61.04
35	276	59.71	60.12	59.27	59.83
34	326	58.42	58.94	58.04	58.58
33	362	57.11	57.75	56.79	57.30
32	454	55.80	56.53	55.53	55.99
31	438	54.48	55.30	54.28	54.67
30	453	53.14	54.04	53.05	53.35
29	504	51.80	52.77	51.84	52.04
28	513	50.45	51.49	50.66	50.75
27	551	49.09	50.20	49.51	49.50
26	512	47.73	48.89	48.39	48.29
25	508	46.37	47.58	47.31	47.11
24	495	45.00	46.27	46.25	45.97
23	468	43.64	44.95	45.21	44.86
22	422	42.28	43.63	44.18	43.77
21	424	40.93	42.31	43.16	42.71
20	344	39.52	40.99	42.15	41.65
19	302	38.25	39.68	41.13	40.60
18	290	36.92	38.36	40.09	39.56
17	234	35.61	37.05	39.03	38.51
16	211	34.31	35.74	37.94	37.46
15	158	33.04	34.43	36.81	36.39
14	122	31.80	33.13	35.62	35.31
13	115	30.61	31.84	34.36	34.20
12	75	29.47	30.56	33.02	33.08
11	67	28.40	29.31	31.59	31.93
10	52	27.43	28.10	30.08	30.76
9	35	26.57	26.95	28.51	29.57
8	19	25.90	25.90	26.87	28.38
7	15	24.69	24.69	25.44	27.20
6	8	23.49	23.49	24.19	25.92
5	2	22.29	22.29	22.94	24.58
4	1	21.08	21.08	21.69	23.25
3	2	20.00	20.00	20.44	21.91
2	0	20.00	20.00	20.00	20.58
1	0	20.00	20.00	20.00	20.00
0	0	20.00	20.00	20.00	20.00

Table C4

Linear and Equipercentile Conversions
Controlled Groups
Section II

RAW	FRQ	T37	T38	L37	L38	E37	E38
***	***	*****	*****	*****	*****	*****	*****
40	58	65.11	65.98	65.86	66.74	64.80	-
39	109	63.85	64.75	64.52	65.46	63.09	66.23
38	135	62.59	63.52	63.18	64.17	61.80	65.31
37	167	61.33	62.30	61.84	62.89	60.77	63.61
36	235	60.07	61.07	60.50	61.60	59.93	61.90
35	276	58.81	59.84	59.16	60.32	59.10	60.18
34	326	57.55	58.61	57.82	59.03	57.81	59.00
33	362	56.29	57.38	56.48	57.74	56.59	57.98
32	454	55.03	56.16	55.14	56.46	55.59	56.91
31	438	53.77	54.93	53.80	55.17	54.46	55.81
30	453	52.51	53.70	52.46	53.89	53.05	54.53
29	504	51.25	52.47	51.12	52.60	51.56	53.14
28	513	49.99	51.24	49.78	51.32	50.14	51.59
27	551	48.73	50.01	48.44	50.03	49.01	49.90
26	512	47.47	48.79	47.10	48.74	47.84	48.19
25	508	46.21	47.56	45.76	47.46	46.57	46.56
24	495	44.95	46.33	44.42	46.17	45.23	45.21
23	468	43.69	45.10	43.08	44.89	43.86	44.06
22	422	42.43	43.87	41.74	43.60	42.51	43.07
21	424	41.17	42.65	40.40	42.32	41.41	42.00
20	344	39.91	41.42	39.06	41.03	40.51	40.79
19	302	38.65	40.19	37.72	39.75	39.61	39.68
18	290	37.39	38.96	36.38	38.46	38.59	38.67
17	234	36.13	37.73	35.04	37.17	37.58	37.61
16	211	34.87	36.51	33.70	35.89	36.40	36.49
15	158	33.61	35.28	32.36	34.60	35.14	35.44
14	122	32.35	34.05	31.02	33.32	33.89	34.56
13	115	31.09	32.82	29.68	32.03	32.30	33.68
12	75	29.83	31.59	28.34	30.75	30.13	32.73
11	67	28.57	30.37	27.00	29.46	28.03	31.77
10	52	27.31	29.14	25.66	28.18	27.76	30.52
9	35	26.05	27.91	24.32	26.89	27.50	29.14
8	19	24.79	26.68	22.98	25.60	27.24	26.89
7	15	23.53	25.45	21.64	24.32	26.51	25.67
6	8	22.26	24.22	20.30	23.03	-	-
5	2	21.00	23.00	18.96	21.75	-	-
4	1	19.74	21.77	17.62	20.46	-	-
3	2	18.48	20.54	16.28	19.18	-	-
2	0	17.22	19.31	14.94	17.89	-	-
1	0	15.96	18.08	13.60	16.60	-	-
0	0	14.70	16.86	12.26	15.32	-	-

Table C5

IRT Conversions, Regular Group
Section III

RAW	FRQ	MIRT37	MIRT38	3P37	3P38	FB37	FB38
***	***	******	******	******	******	******	******
60	8	68.56	68.56	68.56	68.56	68.56	68.56
59	25	67.92	67.54	68.03	67.78	67.30	67.90
58	34	67.19	66.49	67.45	66.76	66.22	66.93
57	53	66.40	65.44	66.91	65.70	65.25	65.89
56	56	65.57	64.39	66.37	64.62	64.33	64.84
55	78	64.71	63.35	65.81	63.53	63.42	63.79
54	102	63.83	62.33	65.20	62.46	62.51	62.73
53	104	62.94	61.32	64.55	61.40	61.61	61.68
52	154	62.05	60.33	63.85	60.37	60.71	60.64
51	158	61.15	59.36	63.09	59.36	59.83	59.62
50	164	60.25	58.40	62.27	58.36	58.95	58.62
49	192	59.34	57.45	61.40	57.39	58.09	57.63
48	224	58.43	56.51	60.49	56.43	57.23	56.66
47	240	57.52	55.58	59.53	55.47	56.38	55.71
46	251	56.61	54.66	58.53	54.53	55.53	54.77
45	259	55.70	53.75	57.51	53.58	54.68	53.84
44	280	54.78	52.84	56.45	52.64	53.84	52.92
43	291	53.87	51.94	55.37	51.72	53.00	52.01
42	287	52.94	51.04	54.27	50.81	52.16	51.11
41	278	52.02	50.15	53.15	49.91	51.31	50.22
40	296	51.09	49.26	52.03	49.02	50.47	49.34
39	307	50.16	48.38	50.89	48.15	49.63	48.47
38	302	49.23	47.50	49.76	47.30	48.79	47.61
37	306	48.29	46.63	48.63	46.47	47.96	46.77
36	312	47.36	45.76	47.52	45.65	47.12	45.93
35	324	46.42	44.89	46.42	44.85	46.29	45.11
34	281	45.49	44.03	45.34	44.06	45.46	44.29
33	311	44.56	43.18	44.27	43.29	44.63	43.48
32	319	43.63	42.33	43.23	42.52	43.80	42.67
31	289	42.71	41.49	42.20	41.75	42.97	41.87
30	262	41.79	40.66	41.19	40.99	42.13	41.07
29	308	40.88	39.84	40.18	40.23	41.28	40.27
28	280	39.98	39.02	39.20	39.47	40.43	39.47
27	266	39.08	38.22	38.22	38.70	39.57	38.66
26	226	38.20	37.43	37.26	37.94	38.70	37.85
25	226	37.33	36.64	36.32	37.16	37.82	37.03
24	203	36.47	35.87	35.41	36.39	36.95	36.22
23	216	35.63	35.12	34.54	35.63	36.07	35.41
22	212	34.80	34.38	33.72	34.87	35.20	34.60
21	178	34.00	33.65	32.96	34.12	34.35	33.81
20	155	33.22	32.95	32.26	33.40	33.51	33.05

Table C5 cont'd

RAW ###	FRQ ###	MIRT37 ######	MIRT38 ######	3P37 ######	3P38 ######	FB37 ######	FB38 ######
19	139	32.46	32.26	31.62	32.71	32.69	32.31
18	123	31.74	31.60	31.04	32.07	31.88	31.63
17	99	31.05	30.96	30.52	31.46	31.10	30.99
16	73	30.40	30.36	30.04	30.90	30.34	30.40
15	67	29.80	29.79	29.62	30.38	29.61	29.87
14	51	29.25	29.27	29.24	29.9	28.91	29.39
13	42	28.77	28.79	28.90	29.45	28.23	28.95
12	17	28.36	28.37	28.41	28.99	27.42	28.53
11	11	27.56	27.58	27.57	28.43	26.60	28.13
10	6	26.72	26.74	26.73	27.56	25.79	27.45
9	7	25.89	25.91	25.89	26.68	24.97	26.57
8	1	25.06	25.08	25.05	25.81	24.15	25.69
7	1	24.23	24.24	24.21	24.94	23.33	24.81
6	1	23.40	23.41	23.37	24.06	22.51	23.93
5	2	22.56	22.58	22.53	23.19	21.69	23.05
4	2	21.73	21.74	21.68	22.31	20.87	22.17
3	1	20.90	20.91	20.84	21.44	20.05	21.29
2	1	20.07	20.08	20,00	20.56	20.00	20.41
1	1	20.00	20.00	20.00	20.00	20.00	20.00
0	0	20.00	20.00	20.00	20.00	20.00	20.00

Table C6

Linear and Equipercentile Conversions
Regular Group
Section III

Raw ###	FRQ ####	T37 ######	T38 ######	L37 ######	L38 ######	E37 ######	E38 ######
60	8	70.97	67.37	71.77	66.62	-	-
59	25	69.97	66.48	70.76	65.76	-	67.27
58	34	68.98	65.58	69.75	64.90	66.21	66.27
57	53	67.99	64.69	68.74	64.04	65.45	66.06
56	56	67.00	63.80	67.74	63.19	64.70	65.00
55	78	66.01	62.91	66.73	62.33	63.91	63.63
54	102	65.01	62.02	65.72	61.47	63.16	62.59
53	104	64.02	61.12	64.71	60.61	62.46	61.59
52	154	63.03	60.23	63.70	59.75	61.73	60.61
51	158	62.04	59.34	62.69	58.89	60.61	59.37
50	164	61.05	58.45	61.68	58.03	59.58	58.23
49	192	60.06	57.55	60.67	57.17	58.81	57.31
48	224	59.06	56.66	59.66	56.31	58.06	56.45
47	240	58.07	55.77	58.66	55.45	57.32	55.60
46	251	57.08	54.88	57.65	54.59	56.54	54.66
45	259	56.09	53.99	56.64	53.73	55.68	53.74
44	280	55.10	53.09	55.63	52.87	54.82	52.86
43	291	54.11	52.20	54.62	52.01	54.00	51.97
42	287	53.11	51.31	53.61	51.16	53.39	51.07
41	278	52.12	50.42	52.60	50.30	52.77	50.13
40	296	51.13	49.53	51.59	49.44	52.13	49.17
39	307	50.14	48.63	50.58	48.58	51.48	48.19
38	302	49.15	47.74	49.58	47.72	50.58	47.27
37	306	48.16	46.85	48.57	46.86	49.60	46.40
36	312	47.16	45.96	47.56	46.00	48.60	45.66
35	324	46.17	45.07	46.55	45.14	47.78	44.95
34	281	45.18	44.17	45.54	44.28	47.04	44.33
33	311	44.19	43.28	44.53	43.42	46.13	43.74
32	319	43.20	42.39	43.52	42.56	45.17	43.10
31	289	42.21	41.50	42.51	41.70	44.16	42.38
30	262	41.21	40.60	41.50	40.84	43.31	41.58
29	308	40.22	39.71	40.50	39.98	42.48	40.59
28	280	39.23	38.82	39.49	39.13	41.73	39.59
27	266	38.24	37.93	38.48	38.27	40.98	38.61
26	226	37.25	37.04	37.47	37.41	40.20	37.73
25	226	36.26	36.14	36.46	36.55	39.15	36.90
24	203	35.26	35.25	35.45	35.69	37.98	36.20
23	216	34.27	34.36	34.44	34.83	36.84	35.46
22	212	33.28	33.47	33.43	33.97	35.87	34.68

Table C6 cont'd

Raw	FRQ	T37	T38	L37	L38	E37	E38
***	****	******	******	******	******	******	******
21	178	32.29	32.58	32.42	33.11	34.94	33.78
20	155	31.30	31.68	31.42	32.25	34.17	32.96
19	139	30.31	30.79	30.41	31.39	33.42	32.32
18	123	29.31	29.90	29.40	30.53	32.87	31.65
17	99	28.32	29.01	28.39	29.67	32.33	30.89
16	73	27.33	28.12	27.38	28.81	31.52	29.88
15	67	26.34	27.22	26.37	27.95	29.82	25.52
14	51	25.35	26.33	25.36	27.09	27.92	-
13	42	24.36	25.44	24.35	26.24	27.04	-
12	17	23.36	24.55	23.34	25.38	26.72	-
11	11	22.37	23.66	22.34	26.52	26.41	-
10	6	21.38	22.76	21.33	23.66	26.10	-
9	7	20.39	21.87	20.32	22.80	-	-
8	1	19.40	20.98	19.31	21.94	-	-
7	1	18.41	20.09	18.30	21.08	-	-
6	1	17.41	19.19	17.29	20.22	-	-
5	2	16.42	18.30	16.28	19.36	-	-
4	2	15.43	17.41	15.27	18.50	-	-
3	1	14.44	16.52	14.26	17.64	-	-
2	1	13.45	15.63	13.26	16.78	-	-
1	1	12.45	14.73	12.25	15.92	-	-
0	0	11.46	13.84	11.24	15.06	-	-

Table C7

Irt Conversions, Controlled Group
Section III

Raw	FRQ	MIRT37	MIRT38	IRT37	IRT38
###	####	######	######	######	######
60	8	68.56	68.56	68.55	68.55
59	25	68.17	67.27	67.36	67.51
58	34	67.68	66.12	66.21	66.50
57	53	67.12	65.03	65.20	65.52
56	56	66.51	63.98	64.24	64.55
55	78	65.86	62.98	63.32	63.59
54	102	65.17	61.99	62.42	62.64
53	104	64.44	61.03	61.53	61.70
52	154	63.70	60.09	60.65	60.76
51	158	62.92	59.16	59.79	59.84
50	164	62.13	58.24	58.95	58.94
49	192	61.33	57.32	58.11	58.04
48	224	60.50	56.42	57.29	57.16
47	240	59.66	55.51	56.47	56.28
46	251	58.81	54.61	55.65	55.40
45	259	57.94	53.72	54.84	54.53
44	280	57.05	52.82	54.03	53.66
43	291	56.16	51.93	53.22	52.78
42	287	55.25	51.04	52.40	51.90
41	278	54.32	50.15	51.58	51.02
40	296	53.38	49.26	50.76	50.14
39	307	52.43	48.37	49.94	49.25
38	302	51.46	47.49	49.11	48.37
37	306	50.48	46.60	48.29	47.49
36	312	49.49	45.72	47.46	46.60
35	324	48.49	44.85	46.63	45.72
34	281	47.48	43.98	45.80	44.84
33	311	46.46	43.11	44.97	43.96
32	319	45.44	42.26	44.14	43.08
31	289	44.42	41.41	43.30	42.20
30	262	43.39	40.57	42.46	41.32
29	308	42.37	39.73	41.62	40.44
28	280	41.35	38.91	40.77	39.55
27	266	40.34	38.11	39.91	38.66
26	226	39.35	37.71	39.05	37.78
25	226	38.36	36.53	38.19	36.90
24	203	37.39	35.77	37.33	36.04
23	216	36.44	35.03	36.47	35.20

Table C7 cont'd

Raw	FRQ	MIRT37	MIRT38	IRT37	IRT38
***	***	******	******	******	******
22	212	35.51	34.30	35.62	34.38
21	178	34.61	33.60	34.78	33.60
20	155	33.73	32.91	33.96	32.87
19	139	32.89	32.25	33.16	32.19
18	123	32.08	31.61	32.38	31.55
17	99	31.31	31.00	31.64	30.98
16	73	30.59	30.42	30.93	30.45
15	67	29.92	29.86	30.27	29.97
14	51	29.31	29.34	29.65	29.53
13	42	28.78	28.84	29.09	29.12
12	17	28.36	28.37	28.58	28.73
11	11	27.53	27.55	27.81	28.20
10	6	26.70	26.72	26.96	27.34
9	7	25.87	25.89	26.11	26.49
8	1	25.04	25.05	25.25	25.63
7	1	24.20	24.22	24.40	24.78
6	1	23.37	23.39	23.55	23.92
5	2	22.54	22.56	22.70	23.07
4	2	21.71	21.73	21.85	22.21
3	1	20.88	20.89	21.00	21.36
2	1	20.05	20.06	20.15	20.50
1	1	20.00	20.00	20.00	20.00
0	0	20.00	20.00	20.00	20.00

Table C8

Linear and Equipercentile Conversions
Controlled Group
Section III

RAW ###	FRQ ###	T37 #####	T38 #####	L37 #####	L38 #####	E37 #####	E38 #####
60	8	72.21	67.10	74.21	66.53	-	-
59	25	71.21	66.24	73.12	65.67	-	-
58	34	70.21	65.38	72.04	64.82	65.27	-
57	53	69.21	64.53	70.96	63.97	65.74	-
56	56	68.21	63.67	69.87	63.12	65.20	-
55	78	67.22	62.81	68.79	62.26	64.21	63.40
54	102	66.22	61.95	67.71	61.41	63.25	61.54
53	104	65.22	61.10	66.62	60.56	62.61	60.31
52	154	64.22	60.24	65.54	59.70	61.97	59.18
51	158	63.22	59.38	64.46	58.85	60.90	58.39
50	164	62.22	58.52	63.37	58.00	59.78	57.82
49	192	61.22	57.67	62.29	57.14	59.13	57.25
48	224	60.22	56.81	61.21	56.29	58.48	56.63
47	240	59.23	55.95	60.12	55.44	57.75	55.99
46	251	58.23	55.09	59.04	54.59	57.00	55.20
45	259	57.23	54.24	57.96	53.73	56.01	54.14
44	280	56.23	53.38	56.87	52.88	55.06	53.23
43	291	55.23	52.52	55.79	52.03	54.19	52.39
42	287	54.23	51.66	54.71	51.17	53.66	51.64
41	278	53.23	50.81	53.62	50.32	53.13	50.79
40	296	52.23	49.95	52.54	49.47	52.47	49.45
39	307	51.24	49.09	51.46	48.61	51.62	48.11
38	302	50.24	48.23	50.37	47.76	50.51	47.10
37	306	49.24	47.38	49.29	46.91	49.37	46.26
36	312	48.24	46.52	48.21	46.06	48.32	45.65
35	324	47.24	45.66	47.12	45.20	47.52	45.04
34	281	46.24	44.80	46.04	44.35	46.76	44.49
33	311	45.24	43.95	44.96	43.50	46.07	43.97
32	319	44.24	43.09	43.87	42.64	45.26	43.43
31	289	43.25	42.23	42.79	41.79	44.34	42.66
30	262	42.25	41.37	41.71	40.94	43.44	41.86
29	308	41.25	40.52	40.62	40.08	42.53	40.91
28	280	40.25	39.66	39.54	39.23	41.78	39.95
27	266	39.25	38.80	38.46	38.38	41.08	38.95
26	226	38.25	37.94	37.38	37.53	40.20	38.00
25	226	37.25	37.09	36.29	36.67	39.05	37.08
24	203	36.26	36.23	35.21	35.82	37.73	36.33
23	216	35.26	35.37	34.13	34.97	36.46	35.63
22	212	34.26	34.51	33.04	34.11	35.63	35.01
21	178	33.26	33.66	31.96	33.26	34.93	34.22
20	155	32.26	32.80	30.88	32.41	34.20	33.31
19	139	31.26	31.94	29.79	31.55	33.41	32.92

Table C8 cont'd

RAW	FRQ	T37	T38	L37	L38	E37	E38
***	***	*****	*****	*****	*****	*****	*****
18	123	30.26	31.08	28.71	30.70	32.60	32.53
17	99	29.26	30.23	27.63	29.85	31.78	32.14
16	73	28.27	29.37	26.54	29.00	30.08	30.17
15	67	27.27	28.51	25.46	28.14	27.52	25.25
14	51	26.27	27.65	24.38	27.29	26.14	-
13	42	25.27	26.80	23.29	26.44	25.83	-
12	17	24.27	25.94	22.21	25.58	25.52	-
11	11	23.27	25.08	21.13	24.73	25.21	-
10	6	22.27	24.22	20.04	23.88	24.90	-
9	7	21.27	23.37	18.96	23.02	-	-
8	1	20.28	22.51	17.88	22.17	-	-
7	1	19.28	21.65	16.79	21.32	-	-
6	1	18.28	20.79	15.71	20.47	-	-
5	2	17.28	19.94	14.63	19.61	-	-
4	2	16.28	19.08	13.54	18.76	-	-
3	1	15.28	18.22	12.46	17.91	-	-
2	1	14.28	17.36	11.38	17.05	-	-
1	1	13.29	16.51	10.29	16.20	-	-
0	0	12.29	15.65	9.21	15.35	-	-

Table C9

Scaled Score Means and Standard Deviations
Modified IRT, Fixed b's, 3-parameter Re-estimated,
Tucker and Levine Equatings

	Regular Group					Controlled Group				
		II		III			II		III	
	N	Mn.	S.D.	Mn.	S.D.	N	Mn.	S.D.	Mn.	S.D.
Modified IRT										
3ETF6(37)	339	49	8.4	50	8.2	314	47	9.3	49	9.1
(38)	329	49	8.3	48	7.4	308	47	9.2	46	8.0
Fixed b's										
3ETF6(37)	1011	49	7.6	49	7.3	1790	47	7.9	47	7.8
(38)	988	48	7.8	47	7.1	1790	47	7.9	46	7.5
3-parameters										
3ETF6(37)	1018	48	9.9	50	8.9	--	--	---	--	---
(38)	988	48	8.3	47	7.4	--	--	---	--	---
Tucker										
3ETF6(37)	1005	48	9.2	49	8.6	315	47	8.9	49	9.5
(38)	980	48	8.4	47	7.8	305	47	8.7	46	8.0
Levine										
3ETF6(37)	1005	47	9.8	50	8.7	315	46	9.4	49	10.3
(38)	980	49	8.3	47	7.5	305	47	9.1	46	7.9