Research Report

Examining an Alternative to Score Equating: A Randomly Equivalent Forms Approach

Chi-Wen Liao

Samuel A. Livingston

April 2008

ETS RR-08-14

Examining an Alternative to Score Equating: A Randomly Equivalent Forms Approach

Chi-Wen Liao and Samuel A. Livingston

ETS, Princeton, NJ

April 2008

Abstract

Randomly equivalent forms (REF) of tests in listening and reading for nonnative speakers of English were created by stratified random assignment of items to forms, stratifying on item content and predicted difficulty. The study included 50 replications of the procedure for each test. Each replication generated 2 REFs. The equivalence of those 2 forms was evaluated by comparing the raw-score distributions focusing on the greatest difference in the cumulative distributions. For listening, 10 replications produced cumulative distributions that differed at some point by more than 0.10, and 4 replications produced differences greater than 0.15. For reading, only 3 replications produced differences greater than 0.10. The difference between the results for listening and reading reflects the greater variation, within strata, in the difficulty of the listening items. The REF procedure may become more effective if item difficulty can be predicted more accurately.

Key words: Equating, item selection, randomization, test construction, parallel forms, item difficulty

Acknowledgments

The authors would like to thank the group of content experts (including Elizabeth Ashmore, Trina Duke, Melissa Locke, Guinn Roberts, Mary Schedle, and Barbara Suomi) who worked diligently for almost two years to identify item difficulty drivers, develop rating schemes, code items, and develop the test specification blueprint. Without their work, the randomly equivalent forms procedure could not have been carried out and studied empirically. The authors would also like to thank Ying Feng, who wrote computer programs to perform stratified random selection of items and produced all the statistics for this project. The project would not be possible without her dedication and excellent programming skills.

The equating of scores on alternate forms of a test requires empirical data collected in a pattern that links the two forms. Whether the data are used in an item response theory calibration or in a score-level equating calculation, the data must link performance on the two forms of the test. Such a link can be provided in several ways. One way is to administer both forms to the same examinees. Another way is to administer the two forms to groups selected to have the same distribution of ability. Another way is to include a common set of anchor items in both forms. Yet another way is to administer the items on the two forms to overlapping samples of examinees, allowing all the items to be calibrated on a common scale. However, in some practical testing situations, all of these data collection designs may be impossible. In such a situation, how can scores on alternate forms be made comparable?

One solution may be to find a way to assemble test forms of equal difficulty without requiring precalibration of the items. The purpose of this study is to investigate the extent to which this goal can be accomplished by stratified random assignment of items to test forms from a pool of items for which no empirical difficulty information is available. Our term for this procedure and the test forms it produces is randomly equivalent forms (REF). Can this procedure produce test forms of equal or nearly equal difficulty? If so, it will be valuable for use in situations where items are not precalibrated and score equating is not possible.

Methodology

The REF approach is a procedure for assembling two or more test forms at the same time by stratified random sampling of items from an item pool. (For an early application of this technique, see Yoakum & Yerkes, 1920.) The strata, which the test developers call *bins*, are specified in terms of item content, format, and estimated difficulty (predicted from several features of the item content). Randomly equivalent test forms are assembled by a computerized procedure that randomly assigns the items in each bin to different forms.

The Item Pools for the Study

The items for this study come from tests that assess the listening and reading proficiency of nonnative speakers of English. Each test consists of 100 multiple-choice questions. There are four item types for listening and five item types for reading. The content specifications for these tests were being revised, and two prototype forms of each test, based on new content specifications (see Tables B1 and B2 in Appendix B), were being administered in a field tryout.

These prototype forms, identified as Forms C and D, were not assembled using the full REF technique, but the stratification of items on the basis of content and difficulty was essentially the same as in the REF procedure.

The new content specifications for the tests called for items to be classified into bins based on predicted item difficulty levels. For each item type, these predicted difficulty levels were obtained by applying a prediction formula over the item difficulty features of items. The item difficulty features and item prediction formulae were developed in a separate analysis called the *difficulty drivers* analysis. (See a summary of the analysis process in Appendix A.)

To assemble Forms C and D, test developers first wrote items and coded each item with item features using the rating system developed from the *difficulty drivers* analysis. Then, they applied difficulty prediction formulae to item features and determined items' levels of difficulty, such as very easy, easy, medium, difficult, or very difficult. Items were then manually assembled by test developers to become Forms C and D based on the new content specifications in Tables B1 and B2. These items were the item pool for the REF study.

Data Collection

The two prototype forms, Forms C and D, were administered in a field tryout in November 2004 to a sample of 1,958 examinees in two different countries where these tests are heavily used. The participants in one country had all taken one operational form of each test, Form A9, six months previously. In that country, the participants were randomly assigned to two groups, one group taking each of the two prototype forms. The participants in the other country took Form A9 the same day that they took Form C or Form D. They were randomly assigned to four groups: one group took Form A9, then Form C; a second group took Form A9, then Form D; a third group took Form C, then Form A9; and the fourth group took Form D, then Form A9. Participants' scores on the prototype forms were not reported.

Table B3 in Appendix B shows a comparison of the participants' scaled scores on Form A9 with the scaled scores of all examinees taking any operational form of these tests within the previous two years. In comparison to the full two-year population, the examinees participating in the study were somewhat stronger, on the average, in listening (by about 0.3 to 0.4 *SD*) but only slightly stronger in reading (less than 0.1 *SD*).

Screening the Data

The data analysis was preceded by a screening step intended to remove from the data set any participant whose score on either test in the study—listening or reading—was so discrepant from the corresponding score on the operational test form that both scores could not plausibly represent a valid assessment of the participant's English language skills. The screening was based on the statistic

$$\frac{|\, z_X - z_Y \,|}{\sqrt{2 - 2r}}$$

where z_x and z_y are the examinee's standardized scores on the forms taken operationally and in the study, and *r* is the correlation between them. This statistic, computed from a bivariate normal distribution, has a normal distribution with *M* = 0 and *SD* = 1. Any participant for whom the absolute value of the statistic was greater than 2.8 in listening or reading was excluded from the data. This value would screen out approximately 0.5% of a population in which there were no true outliers.

The REF Procedure

The items from the prototype forms were collected into two item pools (one for listening and one for reading) from which to assemble REFs. A computer program with a stratified randomization algorithm then assembled pairs of REFs. The strata were the bins described above, which classified the items according to their predicted difficulty. The computer program assigned half of the items in each bin to REF 1 and the other half to REF 2. The assignment was random, but with two constraints. The first constraint was that each of the two REFs was to include, as nearly as possible, the same number of items from each of the two prototype forms (i.e., a difference of no more than one item). This constraint was applied separately to each bin. The second constraint required that item sets based on a common stimulus were to be assigned intact to REF 1 or REF 2. An item set could not contribute some of its items to REF 1 and the others to REF 2.

The procedure for assigning the items *in each bin* to REF 1 and REF 2 is described below:

- Randomly select an item or item set from the bin; assign it to REF 1.

- Randomly select another item or item set; assign it to REF 1.
- Continue until REF 1 has been assigned half of the items that came from one of the two prototype forms.
- From then on, select randomly among items from the other prototype form, assigning the selected items to REF 1.
- Continue until the number of items assigned to REF 1 equals the specified number of items for that bin.
- Assign the items remaining in the bin to REF 2.

This procedure was applied separately to the items in each bin. When the items from all the bins had been assigned, the REF procedure was complete.

Analysis Methods

The REF procedure was replicated 50 times to produce 50 pairs of REFs. The objective of the analysis was to compare the difficulty of REF 1 and REF 2 in each pair of REFs and to summarize these results over the 50 pairs of forms. To determine whether REF 1 and REF 2 were equivalent in difficulty, it was necessary to compare the score distributions on REF 1 and REF 2, either in the same group of examinees or in equivalent groups of examinees. If we had been able to administer the full item pool—both prototype forms—to each examinee, we would have been able to compute each examinee’s scores on both REF 1 and REF 2. However, each examinee took only one prototype form (i.e., half the item pool) during the field tryout.

Our solution to this problem was to create examinee teams of equal ability. The procedure for creating the teams was as follows:

- In the group of examinees taking each prototype form, determine the equipercentile correspondence between raw scores on the prototype form and scaled scores on the operational form, A9. Use this correspondence to give each examinee a scaled score on the prototype form.
- Use the scaled scores on the prototype forms to match examinees taking the two different prototype forms. High-scoring examinees who took one prototype form were paired with high-scoring examinees who took the other prototype form, etc.

The examinee teams were formed before we began constructing REFs. In forming the teams, we paired examinees of similar ability, so that each team's performance would be like that of a single examinee taking both REF 1 and REF 2. The matching procedure was done separately for the listening and reading tests. Because of the requirement that the two examinees on each team were to be of similar ability, not all examinees taking the prototype forms could be included in the matching procedure. About 91% of the examinees in the smaller group were included in the matching procedure for the listening test, and about 96% for the reading test.

Table B4 in Appendix B shows the means and *standard deviations* of the examinees' scores on Forms C and D, before and after the matching procedure. The scaled-score means after the matching procedure were almost exactly equal, as intended. The raw-score means, however, were not, indicating that there was a substantial difference in difficulty between the two prototype forms, especially for the listening test. (Note that the REF procedure was not used to assemble the two prototype forms.)

The score distributions that we used to compare the difficulty of REF 1 and REF 2 were computed from the raw scores of these two-person teams. Each team consisted of two examinees of very similar ability who had taken different prototype forms, one taking Form C and the other taking Form D. The two examinees contributed equally to the team's score on each REF, because each REF took half its items from prototype Form C and the other half from Form D. Therefore, any differences in the distributions of the raw number-correct scores on REF 1 and REF 2 could be attributed to differences in difficulty between REF 1 and REF 2.

Measures of Similarity

The 50 replications of the REF procedure yielded 50 pairs of score distributions of REF 1 and REF 2 for comparison. We compared the two score distributions in each pair by computing statistics that indicated the similarity of the two distributions. We chose statistics expressed in units that are meaningful to readers who are not familiar with the score scale for these particular tests.

One such statistic is the Kolmogorov D-statistic, the largest difference between the cumulative distributions. This statistic can range from .00 to 1.00; it represents a proportion of the examinee population. It answers the question, over the whole score range, of what the largest difference is between REF 1 and REF 2 in the proportion of examinees attaining a given score. On a test used with a pass/fail cut score, the statistic can be interpreted as the largest difference between the pass rates, over all possible choices of a pass/fail cut score. Typically, this largest difference occurs in the middle of the score range, where most of the scores are located.

In addition to the Kolmogorov D-statistic, we computed several statistics expressed in terms of the standard deviation of the scores. One such statistic was the standardized mean difference (sometimes referred to as the effect size)—the absolute value of the difference between the means, divided by the average value of the two standard deviations. We also computed the size of the difference between REF 1 and REF 2 in the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles of the raw-score distributions, expressing these differences in standard-deviation units.

Results

Table 1 summarizes the differences between REF 1 and REF 2 in terms of the Kolmogorov D-statistic: the largest difference over the score range in the proportion of examinees achieving a given score. The table summarizes the values of this statistic over 50 replications of the REF procedure for each test (listening or reading). Figures 1 and 2 present this information graphically.

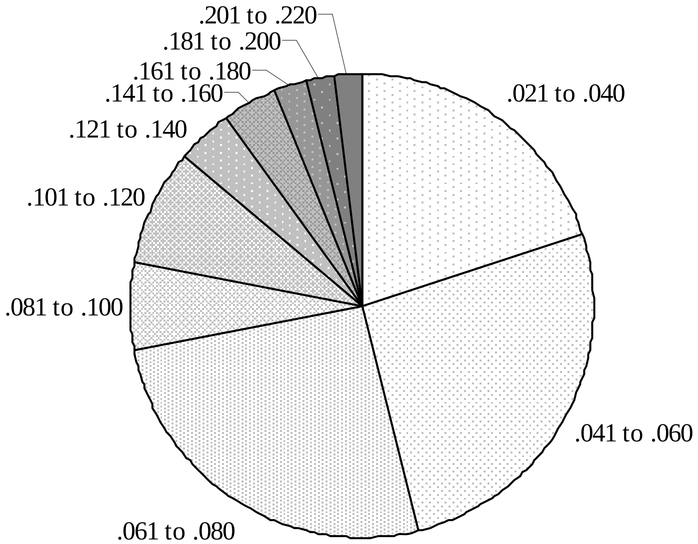

***Figure 1*. Distribution of D-statistic between REF 1 and REF 2: listening test.**

Table 1

Distribution of D-Statistic (Largest Difference in Cumulative Score Distributions) Over 50 Replications of Randomly Equivalent Forms (REF) Procedure

Largest difference between cumulative distributions for REF 1 and REF 2	Number of times observed in 50 replications	
	Listening test	Reading test
.000 to .020 (0 to 2% of the examinees)	0	2
.021 to .040 (2% to 4%)	10	12
.041 to .060 (4% to 6%)	13	16
.061 to .080 (6% to 8%)	13	12
.081 to .100 (8% to 10%)	3	5
.101 to .120 (10% to 12%)	4	2
.121 to .140 (12% to 14%)	2	1
.141 to .160 (14% to 16%)	2	0
.161 to .180 (16% to 18%)	1	0
.181 to .200 (18% to 20%)	1	0
.201 to .220 (20% to 22%)	1	0

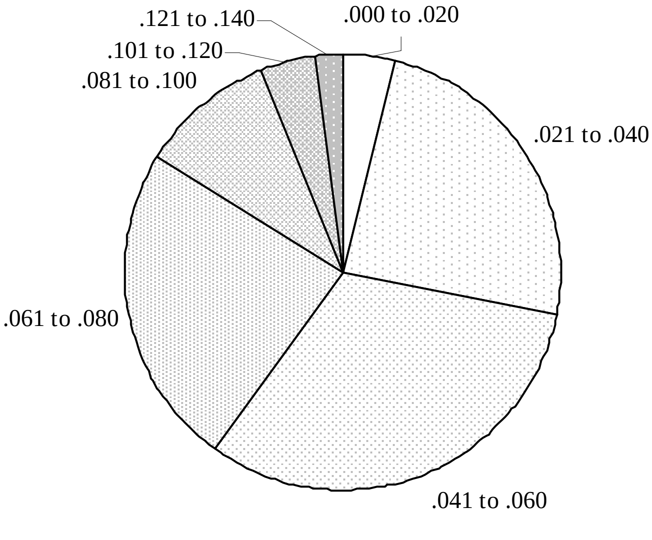

Figure 2. **Distribution of D-statistic between REF 1 and REF 2: reading test.**

Over the 50 replications of the REF procedure for the listening test, the value of the Kolmogorov D-statistic ranged from .025 to .206, with a mean of .070. Ninety percent of the replications (45 of the 50 replications) produced values no larger than .125. For the reading test, the values of the Kolmogorov D-statistic were smaller, ranging from .018 to .128, with a mean of .057. Ninety percent of the replications produced values no larger than .088.

The largest difference between the REF 1 and REF 2 distributions usually occurs in the middle of the score range, where most of the scores are located. Appendix C shows some examples of paired listening score distributions selected from the 50 replications of the REF procedure. Figures C1, C2, and C3 represent small (D-statistic = .04), medium (D-statistic = .07), and large (D-statistic = .17) differences, respectively.

Table 2 summarizes the distribution, over the 50 replications, of the standardized difference between the mean scores on REF 1 and REF 2. For the listening test, the standardized difference between the mean scores ranged from 0.01 to 0.45, averaging 0.12 over the 50 replications of the REF procedure. Ninety percent of the replications (45 of the 50 replications) produced values no larger than 0.26. For the reading test, the standardized difference between the means ranged from 0.00 to 0.25, averaging 0.09 over the 50 replications. Ninety percent of the replications produced values no larger than 0.18.

Table 2

Standardized Difference Between Mean Scores on REF 1 and REF 2 Over 50 Replications of Randomly Equivalent Forms (REF) Procedure

Standardized difference between mean scores on	Range of values over 50 replications	Mean over 50 replications	45 of 50 replications within
Listening test	0.01 to 0.45	0.12	0.26
Reading test	0.00 to 0.25	0.09	0.18

Tables 3 and 4 show the standardized differences between REF 1 and REF 2 with regard to selected percentiles of the score distributions. Table 3 shows this information for the listening test; Table 4 shows the same information for the reading test. These results are similar to those for the standardized difference between the mean scores.

Table 3

Distribution of Standardized Difference Between Selected Percentiles of Scores on REF 1 and REF 2 Over 50 Replications of Randomly Equivalent Forms (REF) Procedure for Listening Test

Percentile of score distribution	Standardized difference between REF 1 and REF 2		
	Range of values over 50 replications	Mean over 50 replications	45 of 50 replications within
5th	0.00 to 0.27	0.11	0.20
10th	0.00 to 0.41	0.10	0.20
25th	0.01 to 0.39	0.12	0.24
50th	0.01 to 0.52	0.15	0.29
75th	0.00 to 0.59	0.15	0.36
90th	0.01 to 0.53	0.16	0.35
95th	0.01 to 0.48	0.15	0.26

Table 4

Distribution of Standardized Difference Between Selected Percentiles of Scores on REF 1 and REF 2 Over 50 Replications of Randomly Equivalent Forms (REF) Procedure for Reading Test

Percentile of score distribution	Standardized difference between REF 1 and REF 2		
	Range of values over 50 replications	Mean over 50 replications	45 of 50 replications within
5th	0.00 to 0.34	0.10	0.20
10th	0.00 to 0.34	0.10	0.20
25th	0.00 to 0.41	0.11	0.20
50th	0.00 to 0.32	0.10	0.20
75th	0.00 to 0.26	0.12	0.20
90th	0.01 to 0.31	0.11	0.19
95th	0.00 to 0.24	0.11	0.20

To understand why the REF procedure produced more forms of nearly equal difficulty for the reading test than for the listening test, we examined the actual difficulty of the items in each bin, as indicated by the percentage of correct answers. The REF procedure can be expected to work best when the items within each bin are highly similar in their actual difficulty.

Figures 3, 4, 5, and 6 and Tables 5, 6, 7, and 8 show how similar, in actual difficulty, the items or item sets in each bin were. The bin numbers indicate the predicted difficulty of the items in the bin; the higher the number, the greater the predicted difficulty. The means and standard deviations in the tables are those of the actual difficulty values. Figures 3 and 4 each contain a separate data point for each item (expressed by the symbol of an open diamond). The height of the data point indicates the percentage of correct answers to the item; a data point appearing high on the graph indicates an easy item. The solid diamond in each column represents the mean proportion-correct for all the items in the bin. Figures 5 and 6 contain a single data point for each item set. The height of the data point indicates the overall percentage of correct answers to the items in the set. The items in these bins were not assigned individually to REF 1 or REF 2; item sets were kept intact during the REF procedure. The data points in all four figures are placed into columns to represent the grouping of items into bins.

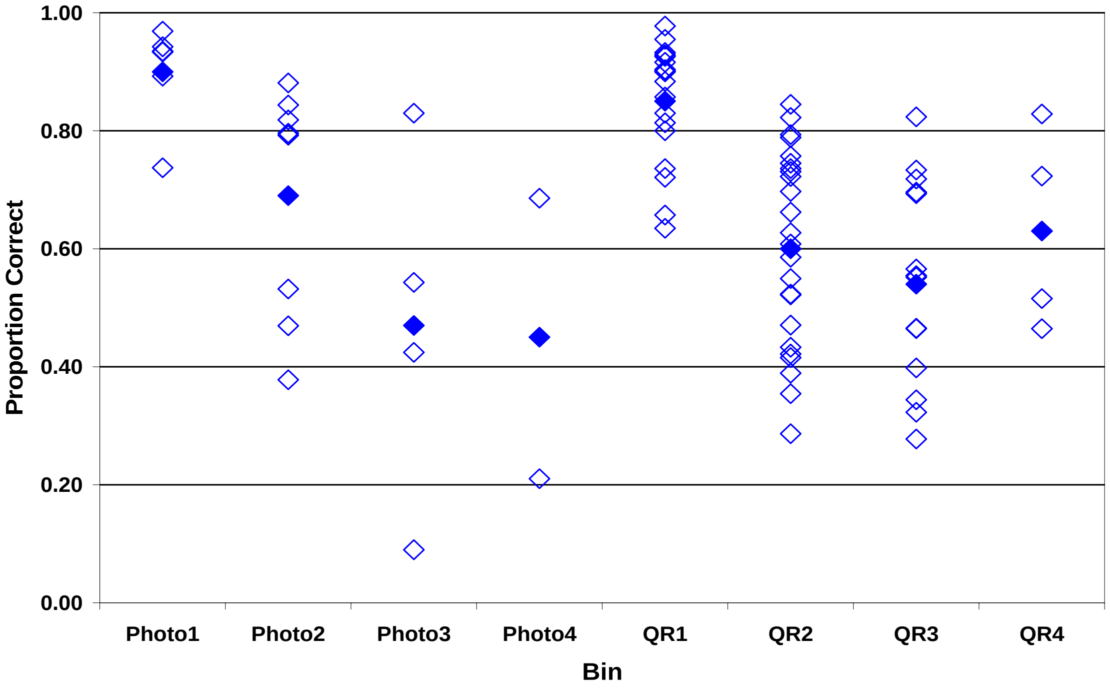

Figure 3. **Actual difficulty of items in each bin: listening discrete items.**

Within each item type, the items or item sets in the higher-numbered bins were predicted to be more difficult (i.e., to have lower percentages of correct answers). The observed percentages of correct answers generally upheld this prediction, but with some exceptions. In particular, the listening item types, conversation and talk, included in Figure 5 did not follow the predicted pattern.

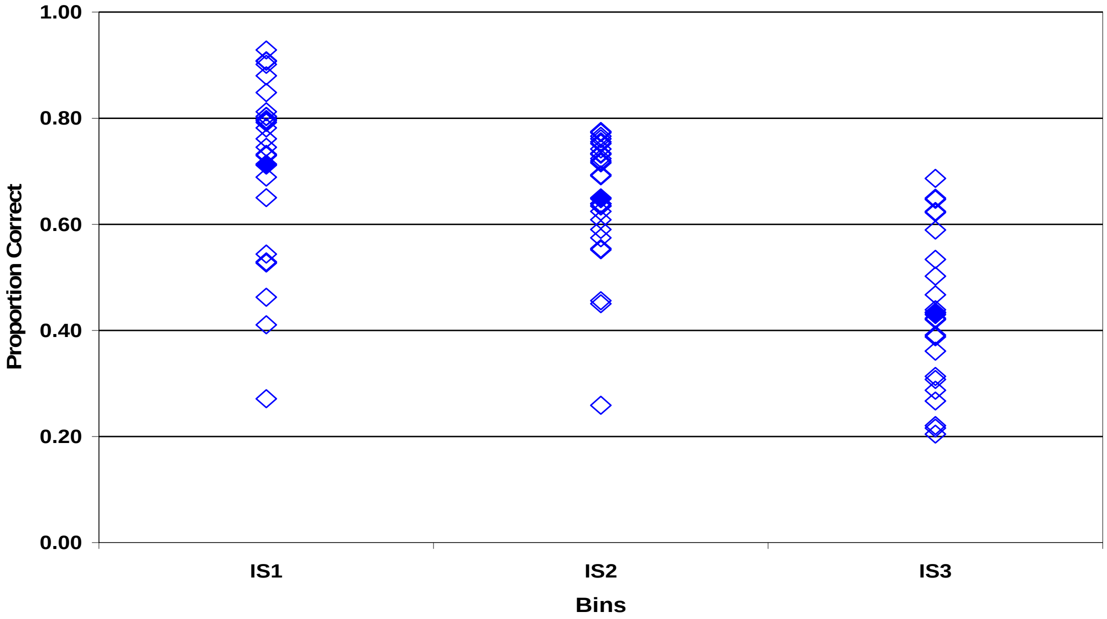

***Figure 4.* Actual difficulty of items in each bin: reading discrete items.**

Tables 5, 6, 7, and 8 display the means and standard deviations of the proportion-correct statistic for the items in each bin. Tables 5 and 6 show the results for the discrete items, which were assigned individually to bins; Tables 7 and 8 show the results for the item sets. The standard deviations in Tables 7 and 8 are of the proportion-correct values of the item sets, not of the individual items. The two bins for the more difficult photo items (see Table 5) show large standard deviations, indicating that the difficulty of these items was not accurately predicted. Note, however, that there were only four items in bin Photo 3 and only two items in bin Photo 4. In general, item difficulty was predicted better for the items in the reading item sets than for the items in the listening item sets, as indicated by the means and standard deviations for each bin.

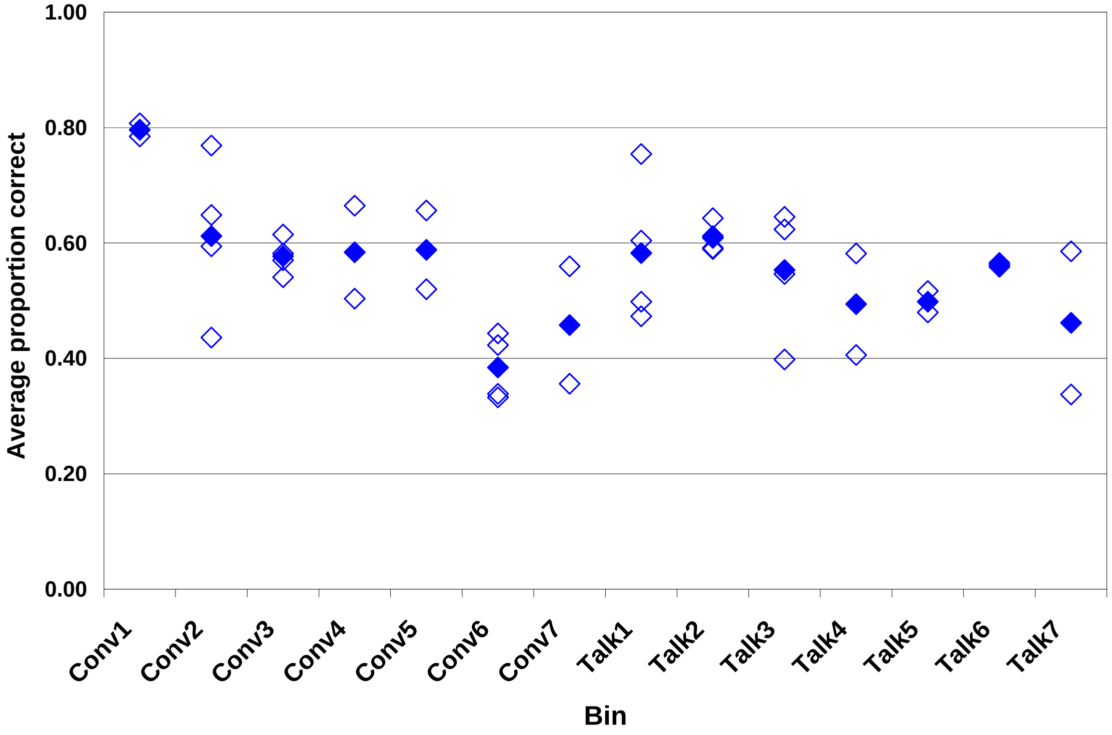

Figure 5. **Average difficulty of item sets in each bin: listening items sets.**

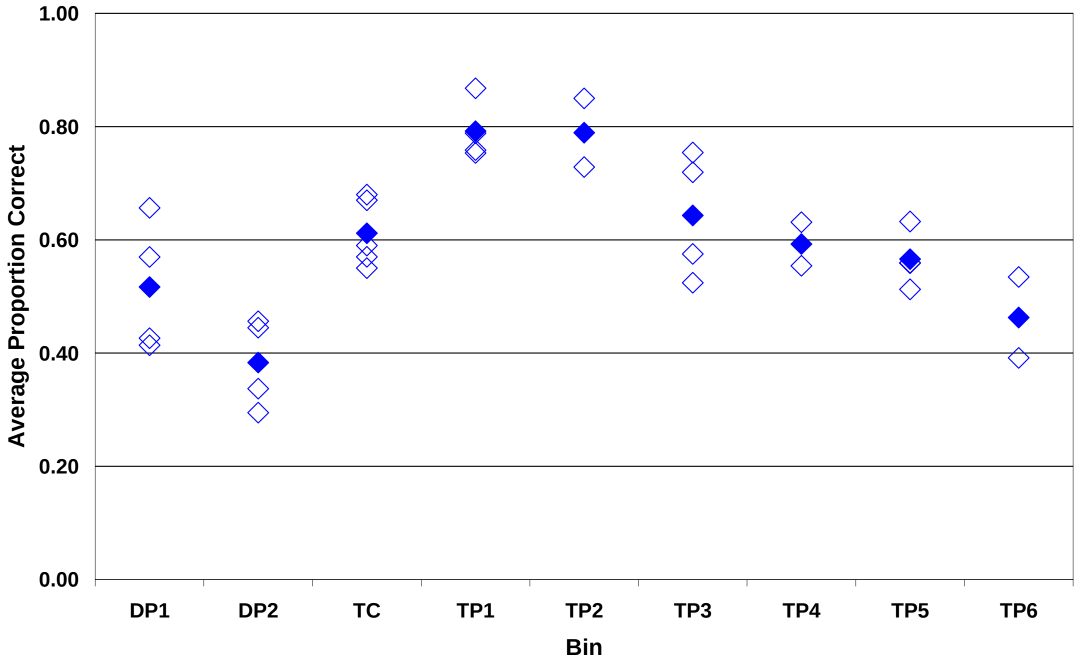

Figure 6. **Average difficulty of item sets: reading item sets.**

Table 5

Predicted and Actual Difficulty of the Items in Each Bin: Listening Test, Discrete Items

Test and type of item	Bin name	Number of items	Predicted difficulty	Mean	*SD*
Photo	Photo1	6	e	0.90	0.08
	Photo2	8	m	0.69	0.20
	Photo3	4	d	0.47	0.31
	Photo4	2	vd	0.45	0.34
Question and response	QR1	18	e	0.85	0.10
	QR2	24	m	0.60	0.16
	QR3	14	d	0.54	0.17
	QR4	4	vd	0.63	0.17

Note. e = easy, m = medium, d = difficult, vd = very difficult.

Table 6

Predicted and Actual Difficulty of the Items in Each Bin: Reading Test, Discrete Items

Test and type of item	Bin name	Number of items	Predicted difficulty	Mean	*SD*
Incomplete sentence	IS1	28	e	0.71	0.17
	IS2	28	m	0.65	0.12
	IS3	24	d	0.43	0.15

Note. e = easy, m = medium, d = difficult.

Discussion and Conclusion

For the listening test, the REF technique did not consistently produce forms of equal or nearly equal difficulty. The largest difference between the cumulative percentages (the Kolmogorov D-statistic) was greater than .10 in 11 of the 50 replications, greater than .12 in 7 of those replications, and greater than .14 in 5 of them.

For the reading test, the results were better. The largest difference between the cumulative percentages was greater than .10 in only 3 of the 50 replications, greater than .12 in only 1 replication, and was never greater than .14.

Table 7

Predicted and Actual Difficulty of the Items in Each Bin: Listening Test, Item Sets

Type of item	Bin name	Number of item sets	Number of items in each set	Predicted difficulty	Mean	*SD*
Conversation	Conv1	2	3	ve, e, m	0.80	0.02
	Conv2	4	3	e, e, m	0.61	0.14
	Conv3	4	3	e, m, m	0.58	0.03
	Conv4	2	3	m, m, d	0.58	0.11
	Conv5	2	3	m, d, d	0.59	0.10
	Conv6	4	3	m, d, vd	0.38	0.06
	Conv7	2	3	d, d, d	0.46	0.14
Talk	Talk1	4	3	ve, e, m	0.58	0.13
	Talk2	4	3	e, e, m	0.61	0.02
	Talk3	4	3	e, m, m	0.55	0.11
	Talk4	2	3	m, m, d	0.49	0.12
	Talk5	2	3	m, d, d	0.50	0.03
	Talk6	2	3	m, d, vd	0.56	0.00
	Talk7	2	3	m, d, d	0.46	0.18

Note. ve = very easy, e = easy, m = medium, d = difficult, vd = very difficult.

Even the results for the reading test do not establish the REF technique as a satisfactory alternative for score equating—where score equating is possible. However, where score equating is not possible, the REF technique may be preferable to the other alternatives, such as the assumption that examinee groups are of equal ability or the assumption that forms created by other procedures are of equal difficulty.

Examining the actual difficulty values (not the predicted difficulty values) of the items within each bin revealed that items in the same bin sometimes differed greatly in difficulty. This variation appears to have been the main reason for the less-than-optimal performance of the REF procedure. In general, the predictions of item difficulty for reading items were better than those for listening, and that was why REF differences in reading were smaller than in listening.

Table 8

Predicted and Actual Difficulty of the Items in Each Bin: Reading Test, Item Sets

Type of item	Bin name	Number of item sets	Number of items in each set	Predicted difficulty:	Mean	*SD*
Double passage	DP1	4	5	e, e, m, m, d	0.52	0.12
	DP2	4	5	e, m, m, d, d	0.38	0.08
Text completion	TC	6	4	e, m, m, d	0.61	0.05
Traditional passage	TP1	4	2	e, m	0.79	0.05
	TP2	2	3	e, m, m	0.79	0.09
	TP3	4	3	e, m, d	0.64	0.11
	TP4	2	3	e, m, vd	0.59	0.05
	TP5	4	4	e, m, d, d	0.57	0.05
	TP6	2	4	m, m, d, vd	0.46	0.10

Note. ve = very easy, e = easy, m = medium, d = difficult, vd = very difficult.

To make the REF procedure more effective, it will be necessary to predict item difficulty more accurately, so as to substantially reduce the within-bin variation. The rating procedure used in this study was not effective enough to make the REF approach work. It was also extremely time consuming. A possible alternative might be to adapt an essay-scoring computer program, such as ETS's *e-rater*® to predict the difficulty of items. Designed to predict ratings of the quality of the writing in student-written essays, *e-rater* structural features of the language as predictors. It is calibrated on a sample of essays, using experts' ratings of those essays as a criterion. After calibration, *e-rater* is then used to predict the ratings that would be given to other essays. The proposed application of *e-rater* to predict item difficulty would use empirical measures of the difficulty of sample items as a criterion. After calibration, *e-rater* would be used to predict the difficulty of other items. If this approach results in accurate prediction of item difficulty, the REF technique might then consistently produce forms of nearly equal difficulty.

References

Yoakum, C. S., & Yerkes, R.M. (Eds.). (1920). *Mental tests in the American army*. New York: Holt.

Appendix A
A Summary of Item Difficulty Drivers Analysis

A group of content experts studied items from the existing operational forms of each test. There were four item types for listening and five item types for reading. For each item type, the experts identified features of the items that, they believed, were associated with item difficulty and developed a system for rating each item on each feature. Those features varied between item types. The rating was either a simple 1 or 0 (feature present or absent) or a classification on a scale of 0 to (at most) 3 for each feature. These features were referred to as *difficulty drivers*.

The statisticians used the response data from operational forms of these items to determine formulae for predicting the difficulty of an item from its difficulty drivers. They derived a separate formula for each item type by performing a logistic regression analysis in which each observation was a single examinee's response to one of the items of that type. The dependent variable was the examinee's item score (1 or 0); the independent variables were the examinee's scaled score and the content experts' ratings of the item—the difficulty drivers for that item. The regression coefficients from this analysis were used to create the formula for predicting item difficulty.

For each item type, the statisticians applied the difficulty prediction formula to all possible combinations of ratings of the difficulty drivers. Statisticians and content experts then divided the predicted-difficulty scale into intervals, which they used to classify the items into difficulty levels: very easy, easy, medium, difficult, and very difficult.

The content experts decided to use the difficulty levels to establish bins for assembling test forms. For discrete items (items that were not part of item sets), each bin corresponded to a single difficulty level. For item sets, each bin corresponded to a combination of item difficulty levels (e.g., two easy items and one medium item).

The content experts used the bins to create the new test specifications, specifying the number of items from each bin to be included in the test. These specifications are shown in Tables B1 and B2 in Appendix B.

Appendix B

Test Specifications

Table B1

Test Specification for a Randomly Equivalent Form (REF) of the Listening Test

Discrete items			
Item type	Bin name	Number of items	Difficulty
Photo	Photo1	3	e
	Photo2	4	m
	Photo3	2	d
	Photo4	1	d
Question & response	QR1	9	e
	QR2	12	m
	QR3	7	d
	QR4	2	vd

Item sets				
Item type	Bin name	Number of item sets	Number of items in each set	Difficulty
Conversation	Conv1	1	3	ve, e, m
	Conv2	2	3	e, e, m
	Conv3	2	3	e, m, m
	Conv4	1	3	m, m, d
	Conv5	1	3	m, d, d
	Conv6	2	3	m, d, vd
	Conv7	1	3	d, d, d
Talk	Talk1	2	3	ve, e, e
	Talk2	2	3	e, e, m
	Talk3	2	3	e, m, m
	Talk4	1	3	m, m, d
	Talk5	1	3	m, d, d
	Talk6	1	3	m, d, vd
	Talk7	1	3	d, d, d

Note. ve = very easy, e = easy, m = medium, d = difficult, vd = very difficult.

Table B2

Test Specifications for a Randomly Equivalent Form (REF) of the Reading Test

Discrete items			
Item type	Bin name	Number of items	Difficulty
Incomplete sentence	IS1	14	e
	IS2	14	m
	IS3	12	d

Item sets				
Item type	Bin name	Number of item sets	Number of items in each set	Difficulty
Double passage	DP1	2	5	e, e, m, m, d
	DP2	2	5	e, m, m, d, d
Text completion	TC	3	4	e, m, m, d
Traditional passage	TP1	2	2	e, m
	TP2	1	3	e, m, m
	TP3	2	3	e, m, d
	TP4	1	3	e, m, vd
	TP5	2	4	e, m, d, d
	TP6	1	4	m, m, d, vd

Note. ve = very easy, e = easy, m = medium, d = difficult, vd = very difficult.

Table B3

Scaled Score Means (SD) of the Test Population and Forms C and D Samples

	Population (*N* = 6,121,672)	Form C (*N* = 1,008)	Form D (*N* = 950)
Listening	314 (84)	338 (81)	345 (80)
Reading	267 (92)	268 (90)	273 (90)

Table B4

Means and Standard Deviations of Scores of Examinees Taking Prototype Forms

	All examinees		Examinees included in evaluation of REF procedure	
Listening test	Form C	Form D	Form C	Form D
Number of examinees	1,008	950	865	865
Scaled score mean *(SD)*	338 (81)	345 (80)	342 (84)	343 (81)
Standardized mean difference (scaled scores)	0.09		0.01	
Raw score mean *(SD)*	56.5 (14.0)	62.9 (14.1)	57.2 (14.5)	62.5 (14.3)
Standardized mean difference (raw scores)	0.45		0.37	
Reading test	Form C	Form D	Form C	Form D
Number of examinees	1008	950	911	911
Scaled score mean *(SD)*	267 (89)	272 (90)	269 (89)	270 (90)
Standardized mean difference (scaled scores)	0.06		0.01	
Raw score mean *(SD)*	56.6 (15.5)	59.0 (14.5)	56.9 (15.5)	58.5 (14.5)
Standardized mean difference (raw scores)	0.16		0.11	

Appendix C

Examples of Score Distributions of Randomly Equivalent Forms

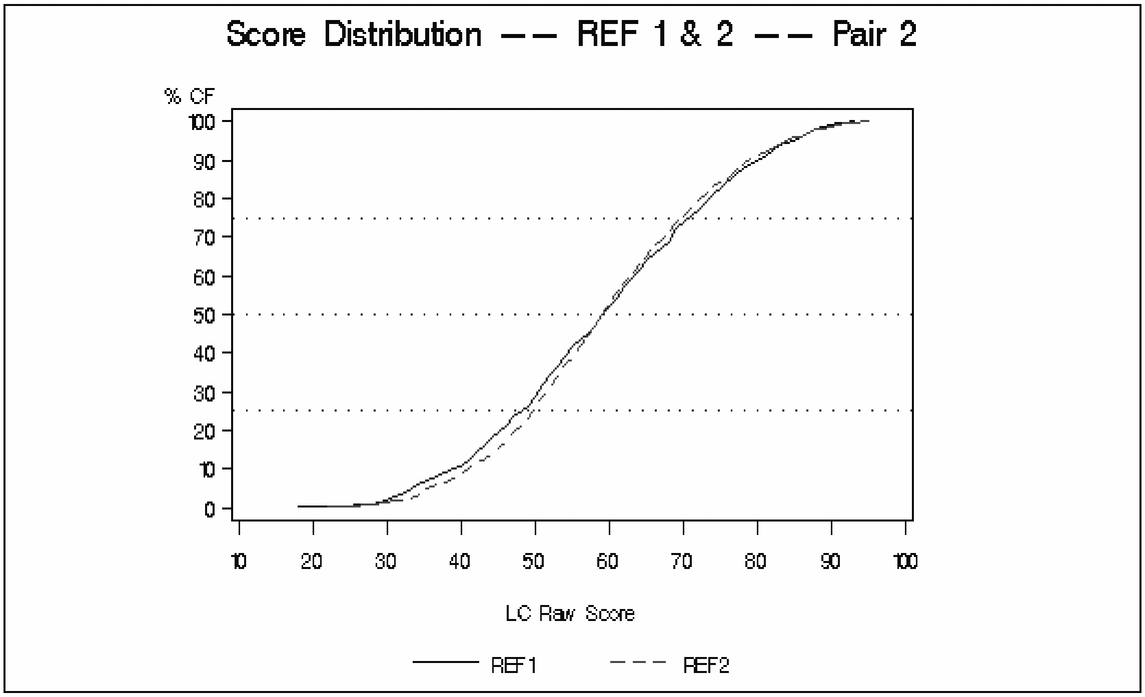

Figure C1. **An example of small D-statistic (= .04).**

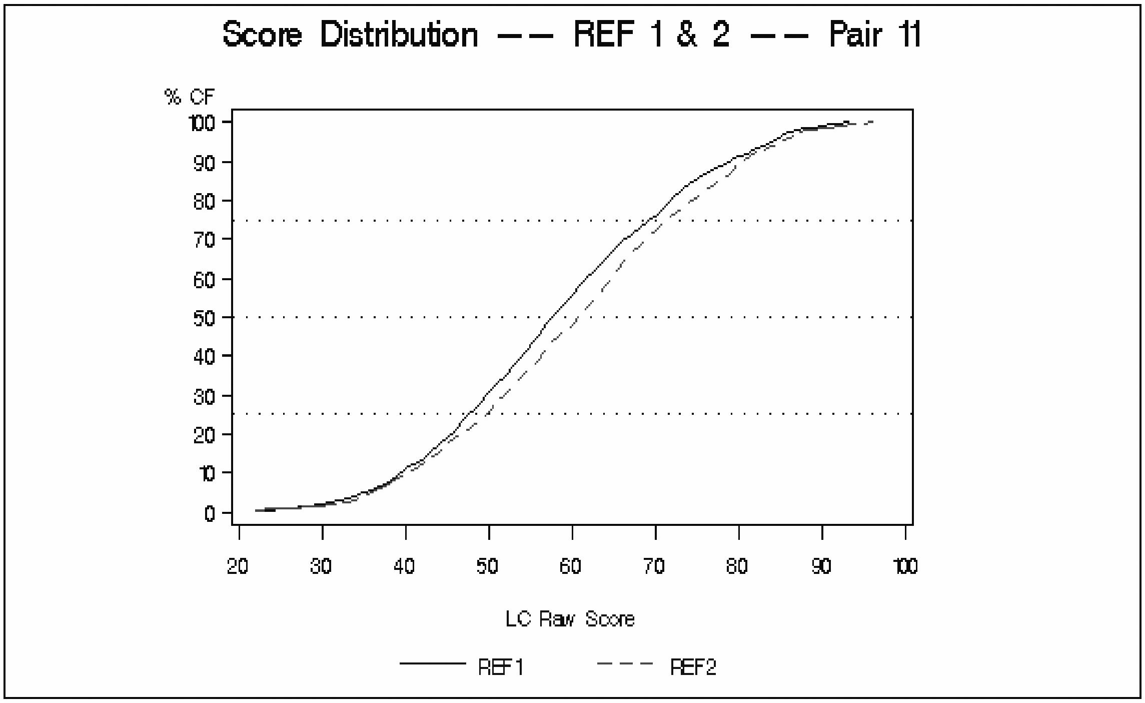

Figure C2. **An example of medium D-statistic (= .07).**

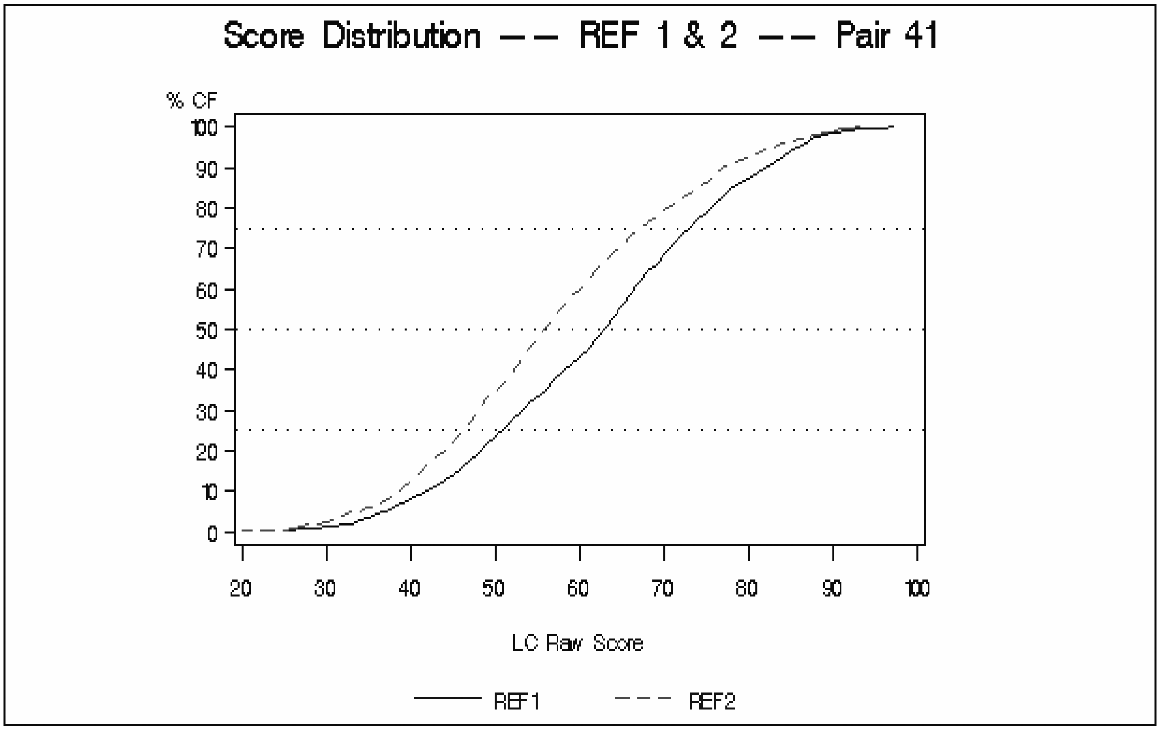

Figure C3. **An example of large D-statistic (= .17).**