ED 395 944 TM 025 011

AUTHOR	McKinley, Robert L.; Schaeffer Gary A.
TITLE	Reducing Test Form Overla_ of the GRE Subject Test in Mathematics Using IRT Triple-Part Equating. GRE Board Professional Report No. 86-14P.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-89-8
PUB DATE	Apr 89
NOTE	25p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*College Entrance Examinations; Comparative Analysis; *Equated Scores; Graduate Study; *Item Response Theory; *Mathematics Tests; Monte Carlo Methods; *Test Format; Test Items
IDENTIFIERS	Double Part Equating; *Graduate Record Examinations; Linear Equating Method; Test Security; *Triple Part Equating

ABSTRACT

A study was conducted to evaluate the feasibility of using item response theory (IRT) equating to reduce test form overlap of the Graduate Record Examinations (GRE) Subject Test in Mathematics. Monte Carlo methods were employed to compare double-part equating with 20-item common item blocks to triple-part equating with 10-item common item blocks. The two methods were evaluated using a circular design that allowed a form to be equated to itself through a series of other forms. The design was replicated five times. Comparisons between scores on equated forms and scores on the base form indicated that triple-part equating did at least as well as double-part equating. This suggests that it may be reasonable to use IRT equating with the GRE Mathematics Test with smaller common item blocks than are used with the linear equating procedures currently employed, as long as there are more of them. This would result in a substantial reduction in the advantage given to repeat examinees and would significantly decrease the number of items on any one test form affected by compromised security. Items that need to be resolved before the procedure can be implemented are discussed. (Contains 4 figures, 3 tables, and 10 references.) (Author/SLD)

GRE®

GRADUATE RECORD EXAMINATIONS

Reducing Test Form Overlap of the GRE Subject Test in Mathematics Using IRT Triple-Part Equating

Robert L. McKinley
and
Gary A. Schaeffer

GRE Board Professional Report No. 86-14P
ETS Research Report 89-8

April 1989

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

ETS

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

Reducing Test Form Overlap of the GRE Subject Test in Mathematics Using IRT Triple-Part Equating

Robert L. McKinley
and
Gary A. Schaeffer

GRE Board Professional Report No. 86-14P

April 1989

Educational Testing Service, Princeton N.J. 08541

ABSTRACT

A study was conducted to evaluate the feasibility of using item response theory (IRT) equating to reduce test form overlap of the GRE Subject Test in Mathematics. Monte-Carlo methods were employed to compare double-part equating with 20-item common item blocks to triple-part equating with 10-item common item blocks. The two methods were evaluated using a circular design that allowed a form to be equated to itself through a series of other forms. The design was replicated five times.

Comparisons between scores on equated forms and scores on the base form indicated that triple-part equating did at least as well as double-part equating. This suggests that it may be reasonable to use IRT equating with the GRE Mathematics test with smaller common item blocks than are used with the linear equating procedures currently employed, as long as there are more of them. This would result in a substantial reduction in the advantage given to repeat examinees, and would significantly decrease the number of items on any one test form affected by compromised security.

It was concluded that triple-part equating is promising, but it was pointed out that some issues need to be resolved before the procedure can be implemented. Among these are the effect of differing ability distributions across forms to be scaled, and the practicality of constructing six common item sets from each form. It was recommended that an additional study be conducted to examine the effect of differing ability distributions on the equating process, and to investigate the feasibility of combining common item sets into a single scaling.

INTRODUCTION

Many large testing programs use multiple forms of a test. Because these test forms tend to differ slightly in difficulty and other psychometric characteristics, it is necessary to equate the scores obtained from different forms. One way of accomplishing this is through the use of a subset of items, called a common item block, that appears on different test forms. Scores on these common items for groups of examinees receiving different test forms are used to establish the relationship between the different forms.

The use of common items for equating, although a popular method, does have disadvantages. For instance, when an examinee takes the test a second time, even though the examinee receives a different form of the test, if the two forms have common items, the examinee is advantaged the second time because she or he took some of the items earlier. In addition, if the security of a test form were ever compromised, to some extent other forms with some of the same common items also would be compromised.

The GRE Subject Test in Mathematics, like all GRE Subject Tests, uses common items for equating. New forms of the test are equated to one or two old forms. In addition, each new form is subsequently used as the old form in the equating of an even newer form. Forms equated to only one old form are used as the old form in the equatings of two new forms. Forms equated to two old forms are used as the old form in the equating of only one new form. Thus, every form of the test is eventually used in three equatings, once as the new form and twice as the old form, or twice as the new form and once as the old form. Each form therefore has items in common with three other forms. Currently, 20 items are used in common item blocks, and equatings are performed using Tucker and Levine linear equating methods (see Angoff, 1984, for a discussion of these equating procedures). When two old forms are used in the equatings, the two equatings are averaged.

The GRE Subject Test in Mathematics contains only 66 items. As much as 90 percent of a form (60 items) eventually appears on other forms, and any one form may share as much as 30 percent of the items (20 items) with any other form. This is particularly troublesome in view of the fact that, for example, 17 percent of the examinees taking the GRE Subject Test in Mathematics in 1986-87 had taken the test at least once previously. Thus, the problems of test security and repeat test takers having an unfair advantage are magnified with this degree of overlap of items.

The use of item response theory (IRT) based equating might make it possible to reduce the size of the common item blocks required for accurate equating. In theory, with IRT equating it is possible to construct a common item block for every available test form. For example, if there were 10 test forms available, the two 20-item common item blocks currently used could theoretically be replaced with 10 four-item common item blocks, each containing items from a different form. In practice, however, four items are not enough to ensure content representativeness, and are probably not enough to yield precise scalings. Moreover, such a practice would guarantee that a repeat examinee would see at least a few items again.

A more practical procedure might be to use a somewhat smaller number of common item blocks and adjust the size of the blocks accordingly. For example, it might be useful to construct four blocks of 10 items each. Ten items per block is likely to produce better scalings, and increases the extent to which content representativeness can be maintained. In addition, it has the advantage of reducing the number of test forms affected if test security is compromised. Unfortunately, it has one major disadvantage. If there is to be no overlap of common item blocks, a test form equated to four old forms using 10 items per common item block can be used as an old form for at most two new test forms. Thus, at some point it will become impossible to find four old forms to which a new form can be equated.

To avoid this problem, it might be possible to reduce not only the size of the common item blocks, but also the overall number of common items. For instance, if three 10-item common item blocks were used, it would be possible to equate a new form to three old forms, and to later use the form as an old form in the equating of three additional forms. The overall number of common items would be reduced from 40 to 30 at the time of equating, but once the form had been used as an old form three times, the total number of common items would be 60, the same as for the current procedures. But the number of items in common with any one form would be reduced from 20 to 10, thus reducing the advantage for repeaters while not unduly increasing the likelihood of a repeat test taker seeing some items twice.

McKinley and Kingston (1987) demonstrated the feasibility of using IRT true-score equating (Lord, 1980) with the GRE Subject Test in Mathematics. The test was shown to be essentially unidimensional, and item responses were reasonably well fit by the three-parameter logistic (3PL) model. In that study, two forms with 20 items in common were equated using Tucker, equipercentile, and IRT equating. The resulting equipercentile and IRT score conversions were found to be quite similar. Both evidenced a mild degree of curvilinearity, and therefore differed somewhat from the linear equating produced using the Tucker method.

IRT equating might be accurate with smaller common item blocks than are currently used operationally, and the McKinley and Kingston study demonstrated that IRT equating might reasonably be used with the GRE Subject Test in Mathematics. The purpose of this study, therefore, was to determine whether it is possible to reduce test form overlap of the GRE Subject Test in Mathematics through the use of IRT equating using smaller common item blocks.

METHOD

This Monte-Carlo study was based on a circular equating plan, in which a test form was equated to itself through a series of other forms. The extent to which the resulting score conversions matched the base form score conversions served as a criterion for evaluating the quality of the equating. The use of a circular equating scheme to examine scale stability has been employed in other studies of IRT equating methods (e.g., Petersen, Cook, & Stocking, 1983).

Figure 1 shows the relationships among the seven simulated test forms. The number of common items for any two forms is indicated by the number next to the arrow (the arrows indicate the direction of scaling, which is discussed below). Form A is designated as the base form, and Forms B, C, and D are designated as the first tier of forms. Each first-tier form has a different set of 20 items in common with the base form.

Forms E, F, and G are designated as the second tier of forms. Each of these forms has 20 items in common with one of the first-tier forms. In addition, Forms F and G have 20 items in common with the base form, and Form E has 10 items in common with the base form. This design was replicated five times to assess the significance and generalizability of the results.

Item Parameters and Item Response Data

In order to place each form on the same scale and then equate Form A to itself through the other forms, true item parameters for each form were first generated. Estimates of these parameters were then obtained from simulated item response data. These estimates were then used to scale tier 1 forms to Form A and then to scale tier 2 forms to tier 1 forms. This scaling procedure allowed scale drift to occur, and thus provided a need for equating. These procedures are described in greater detail below.

The true item parameters for Form A were obtained in the McKinley and Kingston (1987) study from a calibration of actual test data from an operational test form. The covariance matrix for Form A parameters from that study was then used to generate true parameters for the other six forms. First, parameters for the items that overlapped with Form A were obtained. Then, item difficulty "b" parameters for nonoverlapping items were selected from a standard normal distribution (z-scores) and scaled to the b's for Form A by setting means and standard deviations equal. The item discrimination "a" and lower asymptote "c" parameters for the nonoverlapping items were obtained using a multiple regression model that enabled sampling from the same covariance matrix for each form. Table 1 presents summary statistics for the true item parameters for each form.

True ability parameters ("theta") for examinees taking each form were selected from a standard normal distribution. Thus, the average ability level of examinees taking each form differed only by sampling error. (In actual test situations, the average ability level of examinees taking different forms may differ substantially. Due to budget considerations, however, this additional component was not included in the study. Refer to the Summary and Conclusions section for further elaboration.)

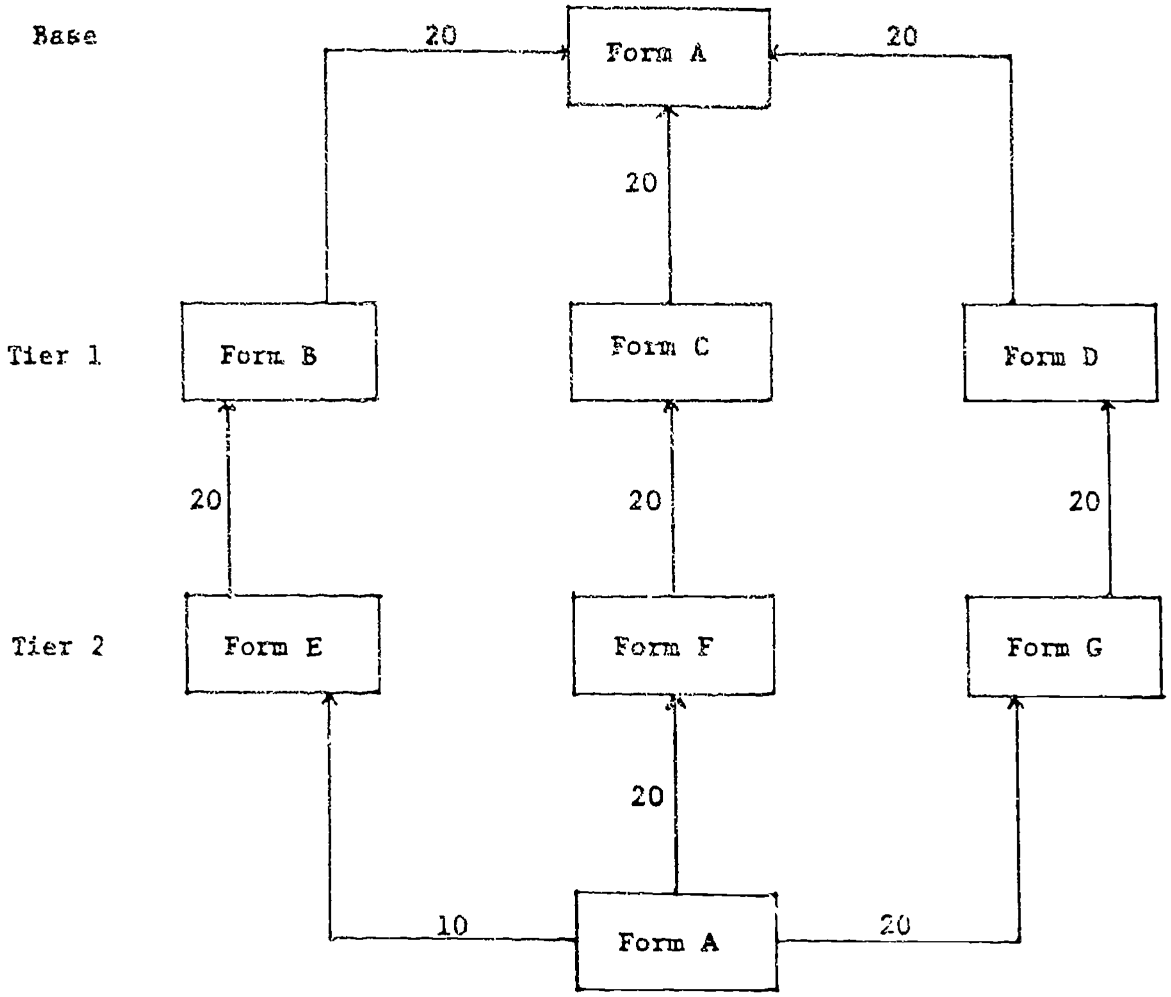

FIGURE 1. Number of common items and direction of scaling for seven simulated test forms

Table 1
Summary Statistics for True Item Parameters

	Mean			Std. Dev.			Correlations		
Form	a	b	c	a	b	c	a-b	a-c	b-c
A	1.07	0.04	0.17	0.32	1.09	0.08	0.28	0.22	0.21
B	1.02	-0.06	0.15	0.33	1.13	0.08	0.15	0.16	0.23
C	1.12	0.36	0.18	0.35	0.98	0.08	0.08	0.21	0.29
D	0.98	-0.07	0.16	0.30	1.15	0.07	0.26	0.24	0.42
E	1.01	-0.10	0.17	0.31	1.00	0.08	0.31	0.18	0.30
F	1.14	-0.05	0.16	0.35	1.18	0.09	0.21	0.17	0.29
G	1.06	0.17	0.18	0.28	1.04	0.08	0.03	0.12	0.18

Item response data were generated to fit the 3PL model for 750 simulated examinees for each test form. This number reflects the approximate number of examinees typically available at equating administrations of the GRE Mathematics test. Right or wrong responses for all items were generated for each examinee to provide a baseline for obtaining estimates of the true item parameters. Based on the true examinee ability parameters and the true item parameters, each examinee had a probability of a correct response to each item under the 3PL model. A random number between 0 and 1 was selected for each examinee-item combination, and if this number was less than the probability of a correct response, the item was scored as correct; otherwise, it was scored as incorrect.

Item analyses were run on each form to evaluate the extent to which the forms were equivalent, and to identify any unrealistic item parameter triplets (e.g., very easy items with high guessing). Only two item parameter triplets were judged to be unrealistic and were modified. Table 2 presents results of item analyses for each form and replication.

The LOGIST (Wingersky, 1983) program was run on item response data separately for each form to obtain estimates of examinee abilities and item parameters. All estimates were scaled using the mean and standard deviation of the ability distributions, which were about the same for all forms. Item-ability regression plots (Kingston & Dorans, 1985) were examined visually for observed proportions correct that fell outside an approximate 95 percent confidence interval around the value predicted by the model. These plots indicated that the 3PL model appeared to fit the data well and confirmed the reasonableness of the data generation procedure.

Scaling and Equating

Because the IRT scale is established independently within individual runs of LOGIST by standardizing the ability estimate distribution, item parameter estimates for different forms tend to be on different scales. In order to equate two forms analyzed in separate LOGIST runs, then, it is first necessary to place the estimates for the two forms on the same scale. In this study, error was intentionally introduced into the IRT scale by using multiple scaling tiers. Averaged multiple IRT equatings (based on separate scalings) were then obtained and evaluated to determine the extent to which the error was corrected.

Pursuant to this aim, scalings were performed in three steps. First, Forms B, C, and D were scaled to Form A. Then, Forms E, F, and G were scaled to the already-scaled Forms B, C, and D, respectively. Finally, Form A (to which all other forms had been scaled, either directly in the case of Forms B, C, and D, or indirectly in the case of Forms E, F, and G) was scaled five times: to Forms F and G, each using 20 common items, and to Forms E, F, and G, each using 10 common items. All scalings were performed using the characteristic curve method (Stocking and Lord, 1983) as implemented in the TBLT program. Note that the scaling of tier 1 to Form A and tier 2 to tier 1 was for the purpose of introducing error. Therefore, only single 20-item scalings were used for each form. Although it would have been more realistic to use 10-item scalings for the triple-part equating procedure, doing so would have made it impossible to determine whether differences in the results for the double- and triple-part equatings were due to the equating methods or differences in the amount of error introduced into tier 2.

Once Form A had been scaled to itself through tiers 1 and 2, it was equated to itself five times. Each equating was based on one of the five scalings describedabove. Equating was performed using the IRT true-score method described by Lord (1980). Because the GRE Subject Test in Mathematics is formula-scored operationally, the equating in this study was done on formula scores. In addition, to facilitate interpretation of results, a linear transformation of estimated true scores to the GRE score scale was also performed, based on the actual original Form A linear conversion parameters.

The equating results based on the two 20-item common block scalings were then averaged, as were the three equatings based on the 10-item common block scalings. For comparison purposes, two of the three equatings based on 10-item common block scalings were also averaged. The same two forms as were averaged in the double-part 20-item scalings were used for the double-part 10-item scalings.

Table 2
Summary of Item Difficulties, Biserial Correlations, and Reliabilities

	Form						
Replication/Statistic	A	B	C	D	E	F	G
Replication 1							
Mean Proportion-Correct	0.58	0.58	0.53	0.59	0.61	0.60	0.56
S.D. Proportion-Correct	0.20	0.20	0.17	0.20	0.19	0.20	0.19
Mean Biserial	0.58	0.59	0.56	0.53	0.57	0.60	0.56
S.D. Biserial	0.11	0.16	0.15	0.13	0.12	0.15	0.15
Reliability	0.93	0.94	0.93	0.92	0.93	0.94	0.93
Replication 2							
Mean Proportion-Correct	0.55	0.59	0.52	0.60	0.61	0.61	0.56
S.D. Proportion-Correct	0.20	0.20	0.18	0.20	0.18	0.20	0.19
Mean Biserial	0.57	0.58	0.56	0.53	0.59	0.60	0.56
S.D. Biserial	0.11	0.17	0.16	0.14	0.13	0.16	0.15
Reliability	0.93	0.93	0.93	0.92	0.93	0.94	0.93
Replication 3							
Mean Proportion-Correct	0.57	0.59	0.53	0.60	0.60	0.59	0.56
S.D. Proportion-Correct	0.20	0.20	0.18	0.20	0.19	0.21	0.19
Mean Biserial	0.57	0.58	0.55	0.55	0.55	0.56	0.55
S.D. Biserial	0.11	0.17	0.15	0.12	0.13	0.16	0.15
Reliability	0.93	0.93	0.93	0.92	0.93	0.93	0.93
Replication 4							
Mean Proportion-Correct	0.56	0.60	0.52	0.59	0.60	0.59	0.56
S.D. Proportion-Correct	0.20	0.19	0.17	0.19	0.19	0.20	0.19
Mean Biserial	0.57	0.58	0.58	0.55	0.55	0.59	0.55
S.D. Biserial	0.11	0.16	0.16	0.12	0.13	0.15	0.14
Reliability	0.93	0.93	0.94	0.92	0.93	0.94	0.93
Replication 5							
Mean Proportion-Correct	0.56	0.58	0.52	0.58	0.60	0.60	0.56
S.D. Proportion-Correct	0.20	0.20	0.17	0.20	0.18	0.20	0.18
Mean Biserial	0.58	0.58	0.56	0.54	0.57	0.62	0.56
S.D. Biserial	0.10	0.17	0.16	0.12	0.15	0.15	0.15
Reliability	0.94	0.93	0.93	0.92	0.93	0.94	0.93

Analysis

If Form A were equated to itself without going through the multiple scaling process described above (i.e., no error in the IRT scale), the resulting score conversion would be given by a 45-degree line through the origin. That is, observed scores and equated scores would be identical. Since in this study Form A was equated to itself, the extent to which the resulting score conversions follow a 45-degree line through the origin is the extent to which the multi-part equating procedures employed were successful in recovering from the error introduced during the scaling process.

Two summary statistics were computed to evaluate the two equating procedures. The first of these was a root mean squared error (RMSE) statistic, which provides an indication of the average amount of deviation of the score conversion from the 45-degree line through the origin. This statistic is given by

$$\mathrm{RMSE} = \left(\frac{1}{750} \sum_{i=1}^{P} N_i (E_i - B_i)^2 \right)^{1/2} , \qquad (1)$$

where P is the number of possible formula scores, E_i is the averaged equated scaled score corresponding to formula score i (obtained either from the double-part equating or the triple-part equating), B_i is the unequated scaled score from the base form corresponding to formula score i, and N_i is the number of examinees in the Form A sample who obtained the formula score i.

The second statistic computed was an indicant of the amount of bias in the averaged score conversions. The amount of bias indicates the extent to which the averaged score conversions tended to be either above or below the 45-degree line through the origin. This statistic was computed as

$$\mathrm{BIAS} = \frac{1}{750} \sum_{i=1}^{P} N_i (E_i - B_i) , \qquad (2)$$

where each term is defined above.

In addition to the RMSE and BIAS statistics, overlay plots were constructed showing the amount of error in the unaveraged score conversions as a function of unequated base scaled score. A separate plot was constructed for each replication for each of the two equating procedures.

RESULTS

Figure 2 displays the differences between 20-item double-part equated scores and base form scores (i.e., equated minus base) for each replication. Figure 3 shows the differences between 10-item triple-part equated scores and base form scores for each replication. Note that in all instances the equated and base scores are the same at the minimum and maximum scores (about 350 and 1050). Equating actually does not occur at these extremes because ability levels cannot be determined.

Figures 2 and 3 illustrate that there was considerable variation in the error patterns for different equatings, regardless of which procedure was used. Interestingly enough, the largest equating errors were observed for one of the 20-item common item block equatings. From this finding it appears that there was not really any advantage to using 20 common items rather than 10.

Figure 4 shows the difference between averaged equated scores (both double-part and triple-part) and base form scores for each replication. The magnitude of the differences generally ranged from zero to about plus or minus 12 points for the double-part equating, except for replication 3, where averaged equated scores were as much as 23 points higher than base form scores at the upper end of the score scale. For the triple-part equated scores, differences generally ranged from zero to about plus or minus eight points, except for replication 5, where averaged equated scores were as much as 20 points higher than base form scores at the upper end of the score scale.

Figure 4 also shows there was no advantage to using 20 common items rather than 10. The errors in the averaged score conversions actually tended to be larger for the double-part equating using 20 common items. It also appears from Figure 4 that the results over replications were more stable for the triple-part equatings, despite the use of fewer common items.

Table 3 shows the RMSE and BIAS statistics obtained for each equating procedure for each replication, as well as the mean and standard deviation over replications of each statistic. From these data it can be seen that, over replications, the triple-part equating procedure yielded less bias and error than did the double-part equating. There was also less variation over replications for triple-part equating, indicating once again that triple-part equating results were more stable than double-part equating results.

For the most part, the RMSE statistics were relatively small, ranging between three and eight. Considering that GRE scores are rounded to the nearest 10 points, these values are for practical purposes almost negligible. In two instances, the double-part equating in replication 3 and the triple-part equating in replication 5, the RMSE values were greater than 10 points.

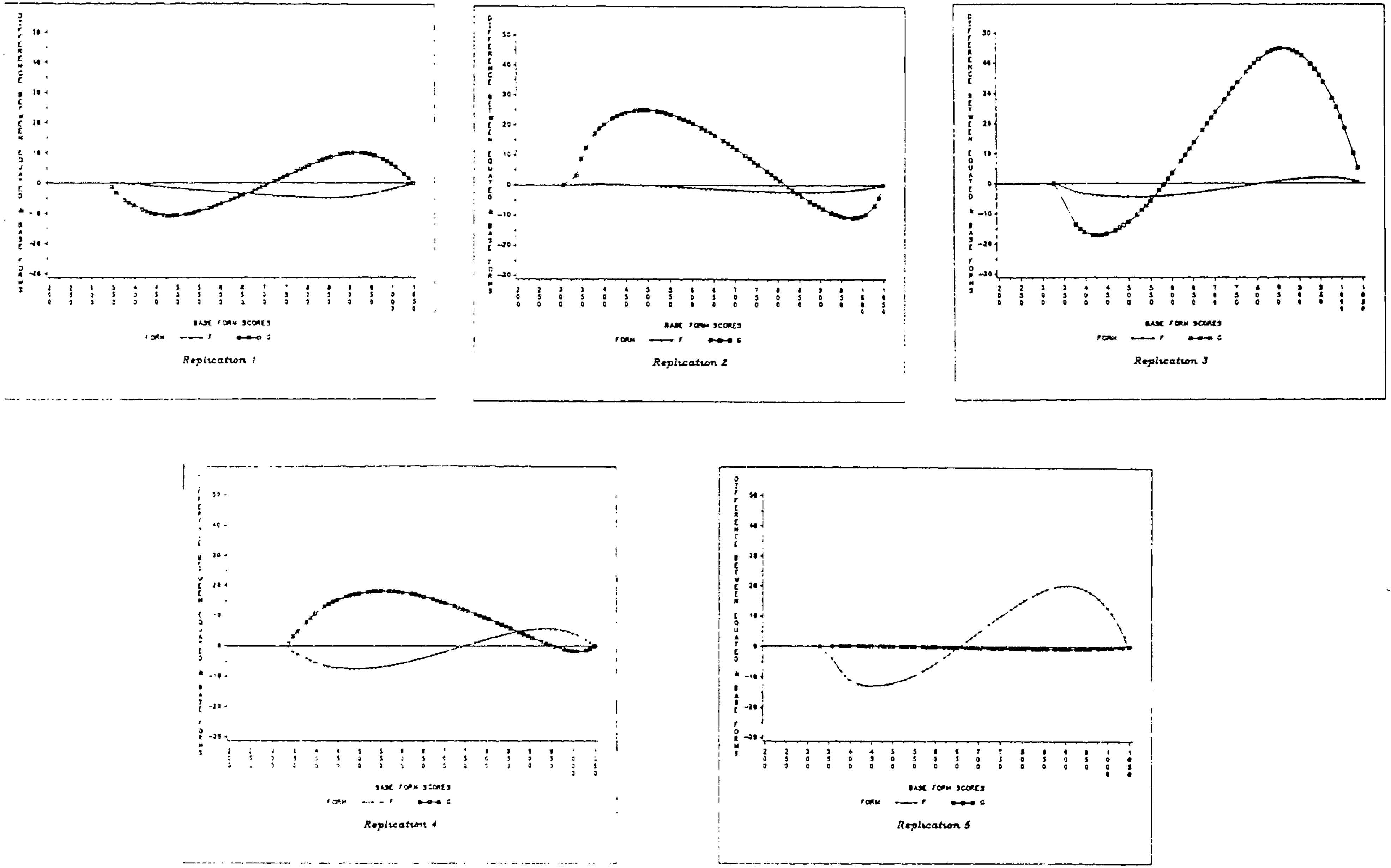

FIGURE 2. Differences between 20-item double-part equated scores and base form scores

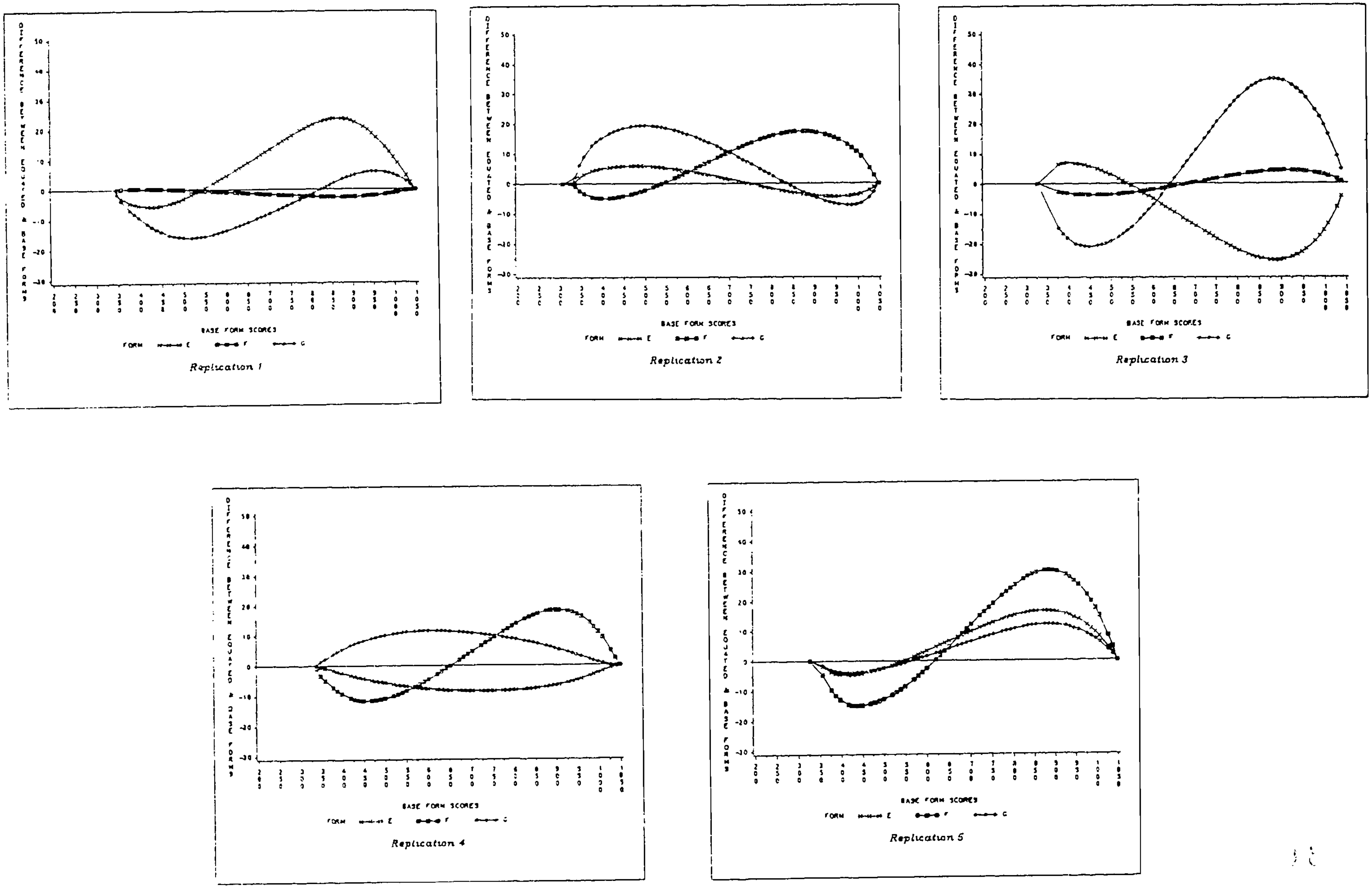

FIGURE 3. Differences between 10-item triple-part equated scores and base form scores

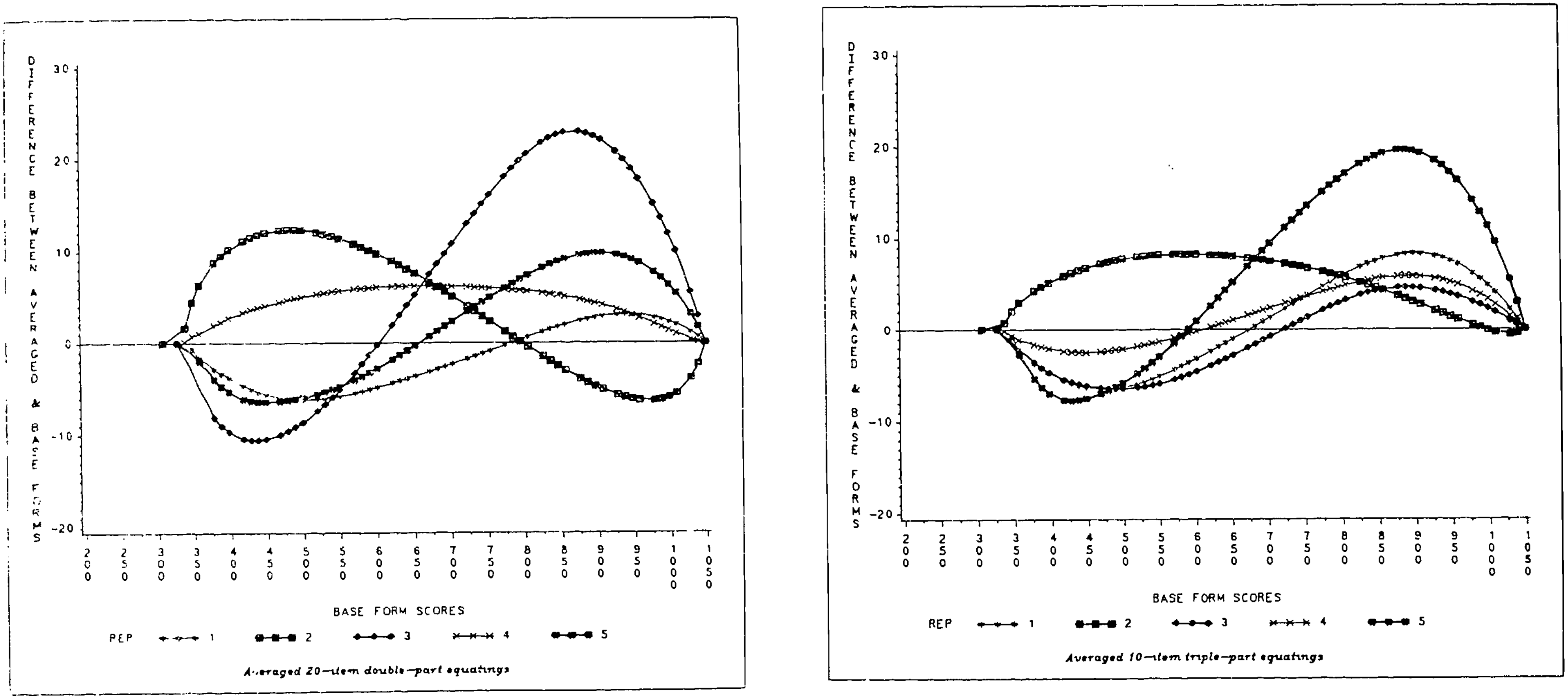

FIGURE 4. Differences between averaged equated scores and base form scores

Figures 2 and 3 illustrate why this occurred. In these two instances, the equating for at least one form exhibited a relatively large difference between averaged equated scores and base form scores, and the two or three difference plots were either above or below the zero difference line at almost all the scores. Thus when the average was taken, one line did not "counteract" another, the averaged line remained relatively far away from the zero difference line, and the RMSE was relatively large. In most other instances, either there were no large errors (e.g., replication 1 double-part equating), or large errors for one equating were offset by opposite-signed errors for another form (e.g., replication 3 for triple-part equating). This finding highlights an advantage of triple-part equating, in that the fewer the number of equatings that are averaged, the more likely it is that large errors will not be offset.

Table 3
Root Mean Square Error (RMSE) and BIAS Statistics

	Replication					Mean	S.D.
Equating/						Over	Over
Statistic	1	2	3	4	5	Reps	Reps
Double-Part (20 items)							
RMSE	3.79	8.26	13.45	5.24	5.85	7.32	3.79
BIAS	-2.05	5.39	6.66	5.05	0.97	3.20	3.63
Triple-Part (10 items)							
RMSE	5.14	6.72	4.29	3.19	11.15	6.10	3.10
BIAS	0.73	6.30	-1.43	1.44	5.91	2.59	3.38
Double-Part (10 items)							
RMSE	5.27	8.85	11.71	7.23	12.02	9.02	2.90
BIAS	-4.01	8.53	3.20	5.55	5.65	3.78	4.75

As was stated previously, for comparison purposes two of the three 10-item equatings were averaged. RMSE and BIAS statistics for this method are shown in Table 3. As can be seen, while the triple-part equating with 10-item common item blocks did at least as well as the double-part 20-item equating procedure, the same cannot be said for 10-item double-part equating. This procedure yielded greater RMSE and BIAS values than either of the other two procedures.

SUMMARY AND CONCLUSIONS

The present study addressed the effects of using IRT equating to reduce test form overlap of the GRE Mathematics test. Monte-Carlo methods were employed to compare double-part equating with 20-item common item blocks to triple-part equating with 10-item common item blocks by equating a form to itself through a series of other forms. Comparisons between scores on equated forms and scores on the base form indicated that triple-part equating yielded less error, less bias, and more stable results than the double-part equating. It was also found that double-part equating with 10-item common item blocks was less satisfactory than either of the other two procedures. This suggests that it may be reasonable to use IRT equating with the GRE Subject Test in Mathematics with smaller common item blocks than current linear procedures now employ, provided there are more of them. This would result in a substantial reduction in the advantage given to repeat examinees, and would significantly decrease the number of items on any one test form affected by compromised security.

However, additional research should be performed before IRT equating procedures are used operationally to reduce test form overlap of the GRE Subject Test in Mathematics. In this study, for example, examinees taking the base and equated forms were randomly sampled from the same ability distribution. However, actual examinees taking different forms of the test are not random samples from the same population; in fact, the ability levels of the two groups may be quite different.

Cook and Eignor (1983) discuss how IRT common item equating may result in less scale drift over time than conventional linear and equipercentile equating methods when new and old form groups differ in ability. Cook and Petersen (1987), however, reviewed several studies of achievement tests and found that IRT equating results were affected when the base and equated form examinees took the test at different administrations and differed in ability level. In one such study (Cook, Eignor, & Taft, 1985), it was hypothesized that examinees taking the test at different administrations differed in the relative recency of their coursework, and this interacted with test content. The test, therefore, may have assessed different constructs for each group of examinees. If similar circumstances apply to the GRE Subject Test in Mathematics, the results found for double- and triple-part equating in the present study might have been different if the base and equated form groups had differed in ability level. Perhaps a follow-up to the present study could be performed to investigate the effects of differing levels of ability for the base and equated form groups.

A related issue that should probably be examined is whether it is really necessary to perform separate scalings for the different common item blocks. The procedure discussed in this study would require the construction of six content-representative common item blocks containing only 10 items, a procedure that might not always be practical, or even possible. However, if all the common item blocks have first been placed on the same scale using a procedure such as TBLT, it may be possible to combine them into a single scaling. If such a procedure is used, for any new form the content-representativeness criterion would only have to be met for the entire set of 30 common items, not for each individual block of 10 items.

The question of whether to perform three scalings or one hinges on the success of each individual scaling. In the design employed in this study, if the IRT parameter estimates for the tier 2 forms are really on the same scale, it seems likely a single, 30-item scaling of Form A would suffice. However, if a factor such as differing ability distributions for different forms introduces error into the scaling process, it may be better to perform multiple scalings. While this issue can be avoided initially by calibrating all current forms in a single run of LOGIST, as more new forms are calibrated and scaled to the original calibration, the issue will become more important. Because of this, it should probably be investigated prior to adoption of triple-part equating.

REFERENCES

Angoff, W. H. (1984). *Scales, norms and equivalent scores*. Princeton, NJ: Educational Testing Service.

Cook, L. L.; & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 175-195). Vancouver, BC: Educational Research Institute of British Columbia.

Cook, L. L., Eignor, D. R., & Taft, H. (1985). *A comparative study of curriculum effects on the stability of IRT and conventional item parameter estimates* (RR-85-38). Princeton, NJ: Educational Testing Service.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225-244.

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement*, *9*, 281-288.

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

McKinley, R. L., & Kingston, N. M. (1987). *Exploring the use of IRT equating for the GRE Subject Test in Mathematics* (ETS Research Report 87-21). Princeton, NJ: Educational Testing Service.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*, 137-156.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201-210.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.