

# Effects of Passage and Item Scrambling on Equating Relationships

Deborah J. Harris  
American College Testing

This study investigated the effects of passage and item scrambling on equipercentile and item response theory equating using a random groups design. For all four tests and for both scramblings used, differences in item and examinee statistics were found to exist between all three forms used (the base form and the two scrambled forms). Up to 50% of the examinees administered a scrambled form would have received a different scale score if the base form equating, rather than the scrambled form equating, had been used to convert their number-correct scores. It is, therefore, suggested that caution be used when scrambled forms are being administered, because in applications such as that studied here, the effects of applying the equating results obtained using a base form to the number-correct scores obtained on a scrambled form can be quite substantial in terms of the numbers of examinees who would receive different scores. *Index terms: context effects, equating, item scrambling.*

In testing situations such as licensure or college admissions tests, it is often very important to have equated scores for examinees that are comparable across test dates and test forms. It is also important, to the extent possible, to prevent examinee copying. Scrambling, the rearrangement of items within a test form, is often used to discourage examinee copying; with scrambling, the same set of items is administered to all examinees. When scrambled forms of a test are used, the question arises as to whether the scrambled form should be equated separately or whether only the base form should be equated, with that conversion then applied to the scrambled form as well as the base form. Although the

measurement literature suggests that item position or context effects exist (Leary & Dorans, 1985; Hambleton & Traub, 1974; Yen, 1980), their effect on equating scrambled forms of a test to a score scale has not been studied.

Philosophical issues also arise as to whether one or two conversions are justified. Because the examinees are all administered the same items, logically there is some appeal to using a single conversion for that particular set of items, regardless of the order in which the items are presented to the examinees. Alternately, given what is known about the existence of context effects (Dorans & Lawrence, 1990; Hambleton & Traub, 1974; Yen, 1980), the fact that identical items administered in alternate orders function differently should not be ignored. Of course, there is no way to know in what order the examinee actually responds to the items; all that is known is the order in which the test publisher intended to present the items.

A question of equity arises when scrambled versions of a test form are administered at the same time the base form is administered, in order to make examinee copying more difficult. If examinees are administered a scrambled version of a test, are they being unfairly advantaged or disadvantaged if the base form conversions are used to obtain scale scores? This possibility of "unfairness" suggests that the use of separate conversions for the base form and the scrambled form may be desirable. However, if two examinees give identical answers to the same items, is it fair for one examinee to receive a higher scale score simply based on the order in which the items were administered? To examine these questions, the present study investigated the magnitude of the

---

APPLIED PSYCHOLOGICAL MEASUREMENT  
Vol. 15, No. 3, September 1991, pp. 247-256  
© Copyright 1991 Applied Psychological Measurement Inc.  
0146-6216/91/030247-10\$1.75

differences between using scrambled or base form conversions.

### Background

The effect of changing the context in which items appear has, in general, been studied in situations other than equating, with section pre-equating being the primary exception. Leary and Dorans (1985), for example, in their review of studies of context effects, discussed issues related to section pre-equating, such as the invariance of the item response theory (IRT) parameters for items administered in two distinct contexts. They concluded that randomly rearranging items or sections of items of the same type did not seem to affect examinee performance under nonspeeded test conditions. They further stated: "The literature has produced evidence of context effects, but has not demonstrated that the effects are so strong as to invalidate test theory or practice that is dependent on an assumption of parameter invariance" (p. 387). However, systematically reordering the items from easy to difficult on one form and difficult to easy on another has shown context effects, so that the practice of scrambling items to create alternate forms has been discouraged (e.g., Hambleton & Traub, 1974). As might be expected, items on speeded tests have also shown greater context effects than items on nonspeeded tests (e.g., Leary & Dorans, 1985).

Using a mathematics and a passage-related reading comprehension test, Yen (1980) examined the effects of altering item context on the item parameter estimates for both the Rasch and the three-parameter IRT model. She consistently found that item parameter estimates obtained from the same context were more similar to each other than item parameter estimates obtained from different contexts. Yen concluded that: "If the scale produced for the trait values is important (for example, if the trait values are supposed to be equated to some other set of trait values), then it is advisable to use large item calibration samples and to maintain the same context for item parameter estimates" (p. 309). This suggests

obtaining separate equating relationships for scrambled forms when using IRT methodology.

Dorans and Lawrence (1990) examined two versions of the Scholastic Aptitude Test (SAT) to ascertain whether statistical equivalence of a base form and a scrambled version held. They employed linear equating to equate one version to the other, and then examined the results to determine whether the identity function (which is what would be expected to result if a test were equated to itself without error) fell within a band of plus and minus two standard errors of equating around the empirical results. Based on their criteria, Dorans and Lawrence concluded that the two SAT test versions did not differ.

The literature thus suggests that context effects appear to be very situation specific. No study in the literature could be found that examined the effects of using a base form versus a scrambled form conversion to obtain scale scores for examinees administered scrambled versions of a test. This study examined this issue, using two scrambled versions of the American College Testing (ACT) Assessment.

### Method

#### Equating Methods

Two equating methods were used in the present study: equipercentile equating with cubic spline smoothing and IRT estimated true score equating. Equipercentile equating is based on Angoff's commonly accepted definition of equating: "Two scores, one on Form X and the other on Form Y (where X and Y measure the same function with the same degree of reliability) may be considered equivalent if their corresponding percentile ranks in any given group are equal" (1971, p. 563). This method uses the total number-correct (NC) score cumulative frequency distributions of both forms and assigns the same scale score to the NC scores on Form X and Form Y if their percentile ranks are the same. Postsmoothing was done using the cubic spline method (Kolen, 1984), because research has shown that this method, when judiciously applied, has the

potential to reduce random error in the equating process.

According to Cook and Douglas, IRT equating methods “characterize equivalent scores on two test forms as those scores which correspond to the same estimated level of the latent trait, ability, or skill underlying both tests” (1982, p. 12). In estimated true score equating under this model, for a given ability level the sum of the estimated item response functions (IRFs) on Form X is considered to be equivalent to the sum of the estimated IRFs on Form Y. Using IRT models requires some stringent assumptions, including unidimensionality and a specified functional form of the IRFs. In this study, the three-parameter logistic IRT model was used, and model fit was not investigated.

#### **Instrument**

The original ACT Assessment (used through June 1989) consisted of four academic tests: English Usage, Mathematics Usage, Social Studies Reading, and Natural Science Reading. The English test consisted of 75 four-option multiple-choice items associated with seven reading passages. The Mathematics test contained 40 five-option discrete multiple-choice items (i.e., items not associated with a passage). The Social Studies and Natural Science tests both consisted of 52 four-option multiple-choice items; 15 items were discrete and 37 were associated with four reading passages. (For more details about the four tests and their content, see the ACT Assessment Program Technical Manual; American College Testing, 1988, p. 93–94.)

#### **Data**

Forms of the ACT Assessment were administered in a random groups design to approximately 25,000 examinees. One of the forms, referred to as the Anchor Form, was used to equate the other forms to the ACT score scale. Form A was considered the base form. Forms B and C contained the same items as Form A, but the items were in a scrambled order. The item responses of the examinees administered these

forms, scored 0/1, were the data used in this study.

#### **Scrambling Schemes**

All four of the tests in the base form (A) were scrambled to create forms B and C. The scrambling orders consisted of the rearrangement of discrete items (e.g., Mathematics); the rearrangement of the passages with the associated items intact (e.g., Form B Social Studies); and the order of the passages left intact with the associated items scrambled (e.g., Form C Social Studies). Extreme scrambling schemes, such as moving the first passage to the end of a test, were avoided.

The two scrambled English versions consisted of interchanging passages and leaving the associated items intact. Form B interchanged passages 1 and 2, 3 and 4, and 5 and 6; passage 7 remained in place. Form C interchanged passages 2 and 3, 4 and 5, and 6 and 7; passage 1 remained in place.

The two Mathematics versions consisted of two different arrangements of random scrambling within blocks of five items. Form B Social Studies interchanged passages 1 and 2, and 3 and 4, with the associated items remaining intact. The discrete item section was randomly scrambled within blocks of five items. The Form C Social Studies kept the passages and their associated items in order, but randomly scrambled the associated items within blocks of five. The Natural Science versions followed the same scheme, by form, as the Social Studies test. These scrambling schemes are depicted in Table 1.

#### **Equating Procedures**

All four tests of Forms A, B, and C were equated separately through the Anchor Form to the ACT Assessment score scale using the equipercentile method. In addition, the Social Studies test of each form was also equated to the ACT Assessment score scale using LOGIST (Wingersky, Barton, & Lord, 1983) and IRT methodology.

The outcome of each equating was a conversion table. For example, the English test on Form B was equated to the English test on the Anchor

**Table 1**  
Scrambling Scheme Used in  
Test Construction

Form A	Form B	Form C
English		
Passage		
1	2	1
2	1	3
3	4	2
4	3	5
5	6	4
6	5	7
7	7	6
Mathematics		
Items in Forms B and C Were Randomly Scrambled Within Blocks of Five		
Social Studies		
Passage		
1	2	1*
2	1	2*
3	4	3*
discrete	discrete*	discrete*
4	3	4*
Natural Science		
Passage		
1	2	1*
2	1	2*
discrete	discrete*	discrete*
3	4	3*
4	3	4*

\*Items were scrambled within blocks of five.

Form (resulting in a conversion of a NC score on Form B to a NC score on the Anchor Form), and through the Anchor Form's NC-to-scale score conversion, to the ACT Assessment score scale (resulting in a conversion of a NC score on Form B to a scale score). The appropriate scrambled form conversion and the base form conversion were applied to the NC scores of examinees who had been administered the scrambled forms.

Composite scores were also of interest in this study. To obtain an examinee's composite score, the four NC test scores were converted using the NC-to-scale score conversion tables that were obtained through the equating process. The scale scores for the four tests were then averaged to obtain the composite score.

## Analysis

NC score statistics were examined across Forms A, B, and C for both total tests and for those sets of items associated with a particular passage. IRT item parameter estimates for the items on Forms A, B, and C were examined for the Social Studies test. NC score equivalents using the scrambled form conversions minus the base form conversions were examined. Scale score moments were compared under both the scrambled and the base form conversions. The percent of examinees who would obtain a different scale score if the base form conversion was used instead of the scrambled form conversion, was computed.

## Results

### NC Score Statistics

NC scores on the total tests and on the items related to a passage were examined for the 4,329 examinees administered Form A, the 4,249 examinees administered Form B, and the 4,195 examinees administered Form C. The groups of examinees administered each of the three forms were assumed, due to the random groups data collection design, to be randomly equivalent. If there were no context effects from rearranging the items, neither NC nor scale score data should differ significantly among the groups for any particular test or passage. Table 2 displays the NC score moments for the English test and for each of the seven English passages. The data displayed in this and subsequent tables have been unscrambled; that is, the moments in the first row of Table 2 refer to the items associated with the first passage in Form A, and the same items in Forms B and C. Table 3 shows NC score moments for the Mathematics test, Table 4 provides data for the Social Studies test, and Table 5 reports results for the Natural Science test.

Tables 2, 4, and 5 show that the passage means differed substantially among the three forms. The *z*s indicate that many of the differences of Form B or C means from Form A means were statistically significant; this was due in part to the

**Table 2**  
Mean and SD of Passage and NC Scores, Standard Error of Mean Differences, and z Ratio for Test of Mean Differences for Forms A, B, and C of the English Test

Passage and Statistic	Form		
	A	B	C
Passage 1, 9 Items			
Mean	6.74	6.85	6.82
SD	1.72	1.70	1.74
SE Difference		.04	.04
z		2.98	2.13
Passage 2, 10 Items			
Mean	7.03	6.99	7.09
SD	1.83	1.81	1.89
SE Difference		.04	.04
z		-1.02	1.49
Passage 3, 15 Items			
Mean	9.86	10.01	9.91
SD	2.89	2.91	2.89
SE Difference		.06	.06
z		2.40	.80
Passage 4, 14 Items			
Mean	9.16	9.22	9.12
SD	2.80	2.74	2.97
SE Difference		.06	.06
z		1.00	-.64
Passage 5, 9 Items			
Mean	5.82	5.56	6.04
SD	2.22	2.46	2.14
SE Difference		.05	.05
z		-5.14	4.66
Passage 6, 9 Items			
Mean	5.16	5.47	4.58
SD	2.37	2.14	2.65
SE Difference		.05	.05
z		6.36	-10.64
Passage 7, 9 Items			
Mean	4.78	4.76	5.20
SD	2.66	2.66	2.41
SE Difference		.06	.05
z		-.35	7.64
Total, 75 Items			
Mean	48.55	48.86	48.75
SD	12.97	12.84	13.26
SE Difference		.28	.28
z		1.11	.70

large sample sizes used in this study. These passage differences tended to cancel out somewhat at the NC score level; that is, test NC score mean differences among the forms were not as great as passage level differences. For comparison,

mean NC scores on the anchor test were 50.34 for English, 18.98 for Mathematics, 27.02 for Social Studies, and 25.07 for Natural Science.

### Equated Scores

Figure 1 displays the differences between the NC score equivalents from equating the base form (A) to the Anchor Form and the scrambled forms (B and C) to the Anchor Form for the 52-item Social Studies test. For example, EQB-A refers to the Form B equivalents obtained using equipercentile equating minus the Form A equivalents obtained using equipercentile equating. If the equating conversions were identical for the base form and a scrambled form, the plotted differences would be a horizontal line at 0. Figure 1 also includes plots of plus and minus two standard errors of the unsmoothed equipercentile equating method for Form A (see Lord, 1982). Because extreme scores provide for poor standard error estimates, due to relatively few examinees scoring at the extremes, only standard errors for the range of 10 to 35 are plotted. Easily obtainable standard errors for the smoothed equipercentile and the IRT methods used here do not exist.

Figure 1 shows that the two equating methods gave more similar results than the two scrambled forms. That is, the equipercentile and the IRT equivalent differences for C-A tend to be more similar to each other than to the equipercentile and IRT equivalent differences, respectively, for B-A. It can also be seen from Figure 1 that the two scrambled forms, B and C, resulted in very different equatings.

The Form C equivalent differences exceeded

**Table 3**  
Mean and SD of NC Scores, Standard Error of Mean Differences, and z Ratio for Test of Mean Differences for Forms A, B, and C of the Mathematics Test (40 Items)

Statistic	Form		
	A	B	C
Mean	19.85	19.81	20.06
SD	8.21	8.31	8.36
SE Difference		.18	.18
z		-.22	1.17

**Table 4**  
 Mean and SD of Passage and NC Scores,  
 Standard Error of Mean Differences, and *z*  
 Ratio for Test of Mean Differences for  
 Forms A, B, and C of the Social Studies Test

Passage and Statistic	Form		
	A	B	C
<b>Passage 1, 9 Items</b>			
Mean	5.44	5.36	5.44
SD	1.84	1.82	1.83
SE Difference		.04	.04
<i>z</i>		-2.02	.00
<b>Passage 2, 10 Items</b>			
Mean	5.80	5.63	5.73
SD	1.92	1.96	1.95
SE Difference		.04	.04
<i>z</i>		-4.06	-1.67
<b>Passage 3, 9 Items</b>			
Mean	4.72	3.67	4.38
SD	2.00	2.21	2.06
SE Difference		.05	.04
<i>z</i>		-23.06	-7.73
<b>Discrete, 15 Items</b>			
Mean	6.00	6.15	6.02
SD	2.68	2.69	2.68
SE Difference		.06	.06
<i>z</i>		2.59	.34
<b>Passage 4, 9 Items</b>			
Mean	4.13	5.54	3.95
SD	2.50	2.08	2.50
SE Difference		.05	.05
<i>z</i>		28.42	-3.32
<b>Total, 52 Items</b>			
Mean	26.09	26.35	25.53
SD	7.89	7.68	7.94
SE Difference		.17	.17
<i>z</i>		1.55	-3.27

the two standard error bands for NC scores of 15 to 30; the Form B equivalent differences were not as large. However, it should not be interpreted that these Form C differences are substantial just because the Form C equivalents exceeded the two standard error bands and that the Form B differences can be ignored. The two standard error bands were presented solely to provide a framework for examining the equivalent differences and were not meant to act as a standard of practical significance.

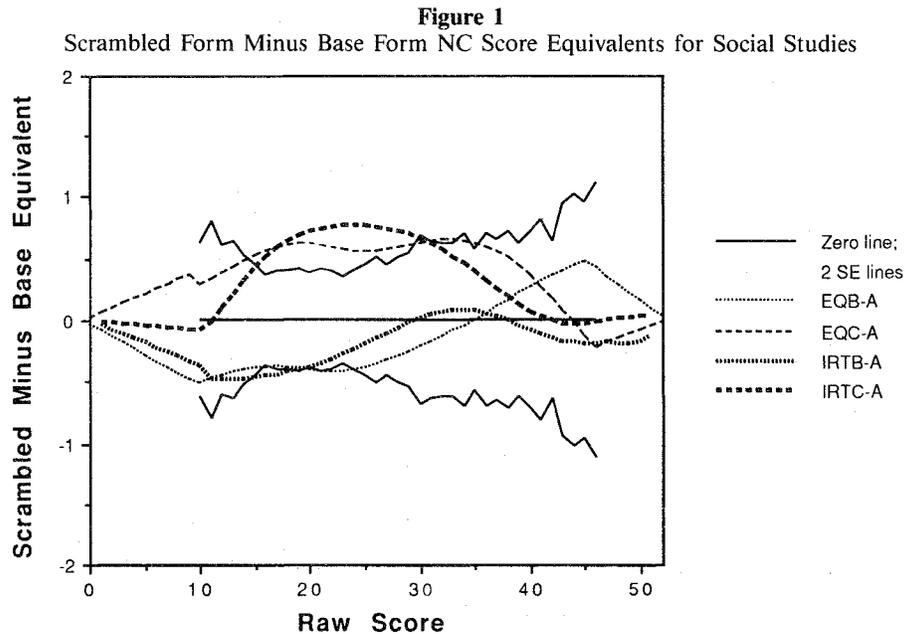
Table 6 gives the scale score moments for Form A, and for Forms B and C, using both the ap-

propriate scrambled form conversions (B or C) and the base form (A) conversions. In all instances, score differences were observed, implying that for no test were the scrambled and the base form conversions identical. Although the impact on group means did not appear unduly large, the fact that mean differences existed using the scrambled and base form conversions implies that some individual examinees would receive different scale scores depending on what conversion was applied.

Table 7 summarizes the percent of examinees affected, which varied widely across tests. The

**Table 5**  
 Mean and SD of Passage and NC Scores,  
 Standard Error of Mean Differences, and *z*  
 Ratio for Test of Mean Differences for  
 Forms A, B, and C of the Natural Science Test

Passage and Statistic	Form		
	A	B	C
<b>Passage 1, 9 Items</b>			
Mean	5.71	5.55	5.69
SD	1.90	2.02	1.93
SE Difference		.04	.04
<i>z</i>		-3.78	-.48
<b>Passage 2, 9 Items</b>			
Mean	4.51	4.70	4.54
SD	2.01	2.02	2.12
SE Difference		.04	.04
<i>z</i>		4.37	.67
<b>Discrete, 15 Items</b>			
Mean	8.05	7.99	8.11
SD	2.69	2.65	2.71
SE Difference		.06	.06
<i>z</i>		-1.04	1.03
<b>Passage 3, 10 Items</b>			
Mean	3.97	3.15	3.90
SD	1.93	1.93	1.93
SE Difference		.04	.04
<i>z</i>		-19.67	-1.67
<b>Passage 4, 9 Items</b>			
Mean	3.15	3.86	3.13
SD	1.94	1.89	1.95
SE Difference		.04	.04
<i>z</i>		17.17	-.47
<b>Total Test, 52 Items</b>			
Mean	25.38	25.24	25.38
SD	7.15	7.04	7.26
SE Difference		.15	.16
<i>z</i>		-.91	0.00



maximum scale score difference was one point. That is, if the scale score conversion established by equating Form A to the Anchor Form were applied to the NC scores from Form B (rather than using the scale score conversion obtained by equating Form B), the most any examinee's score would change would be plus or minus one scale score point. To aid in evaluating these one-point differences, the following approximations of standard errors of measurements can be used: English, 1.52; Mathematics, 2.59; Social Studies, 2.59; Natural Science, 2.53; and Composite, 1.18. (For information on how these estimates were obtained, see American College Testing, 1988, p. 93-94.)

Table 7 shows that up to 50% of the examinees (Form C, Social Studies) would receive different scale scores if the base form conversions were used instead of the scrambled Form C conversions. Although the one-point change is within what would be considered measurement error, it is a source of error that can be determined and controlled. That is, using a base form conversion table for an examinee who was administered a scrambled test form introduces error due to the

difference in item context between the base and scrambled form; using the scrambled form conversion does not introduce this error. Although it is not a recommended practice to place undue emphasis on a single test score, a single point can potentially make a large difference, depending on the purpose for which the examinee's scores is to be used. For example, an incoming college student may be able to participate in collegiate athletics with an ACT composite score of 17, but be ineligible for participation with a score of 16.

### Conclusions

Differences were found between the base form conversions and the scrambled form conversions for all tests. Up to 50% of the examinees who were administered one of the scrambled forms would receive a different scale score if the base form conversion table was used instead of the scrambled form conversion table. Thus, using a base form conversion table with a scrambled form cannot be said to be without impact at the individual examinee level.

Because of this impact, and what the research has shown regarding the existence of context

**Table 6**  
 NC Score Moments for the Tests and Composite

Test, Form, and Conversion	Mean	SD	Skew- ness	Kur- tosis
<b>English</b>				
Form A	18.84	5.33	-.57	.33
Form B: B conversion	18.85	5.30	-.59	.37
A conversion	18.97	5.26	-.56	.34
Form C: C conversion	18.86	5.34	-.57	.34
A conversion	18.89	5.50	-.57	.25
<b>Mathematics</b>				
Form A	17.56	7.48	-.14	-.69
Form B: B conversion	17.59	7.44	-.14	-.65
A conversion	17.50	7.58	-.12	-.72
Form C: C conversion	17.62	7.46	-.17	-.70
A conversion	17.73	7.60	-.12	-.71
<b>Social Studies</b>				
Form A	17.47	7.16	-.05	-.85
IRT	17.54	6.96	-.02	-.78
Form B: B conversion	17.44	7.20	-.10	-.84
A conversion	17.76	6.98	-.12	-.82
IRT B conversion	17.42	7.18	-.11	-.87
IRT A conversion	17.80	6.78	-.08	-.75
Form C: C conversion	17.45	7.21	-.08	-.89
A conversion	16.96	7.24	.01	-.88
IRT C conversion	17.48	7.17	-.12	-.86
IRT A conversion	17.03	7.04	.02	-.80
<b>Natural Science</b>				
Form A	21.45	6.03	-.04	-.57
Form B: B conversion	21.45	6.06	-.05	-.55
A conversion	21.36	6.00	-.06	-.55
Form C: C conversion	21.50	6.04	-.07	-.58
A conversion	21.42	6.11	-.08	-.49
<b>Composite</b>				
Form A	18.96	5.53	-.03	-.60
Form B: B conversion	18.96	5.54	-.07	-.60
A conversion	19.02	5.50	-.07	-.59
Form C: C conversion	18.99	5.59	-.08	-.59
A conversion	18.87	5.67	-.05	-.58

effects, it seems prudent to equate scrambled test forms of tests separately from base test forms and to treat them as distinct forms in such procedures as item analysis and examinee scoring. This recommendation would allow for two examinees, one administered the base form and one administered a scrambled form, to give identical responses to identical items, yet receive different scale scores. This is a problem, but using a single base form conversion with scrambled forms might be more problematic.

This study dealt with only one base form

(Form A) and two scrambled versions of that form (Forms B and C). The two scramblings were selected to minimize the effect of context changes; using more extreme item rearrangements might result in more disparity between the base and the scrambled forms. For instance, if a test is to some extent speeded, the passage or items administered last on the test are most affected. On the English, Social Studies, and Natural Science tests, one of the scrambled forms kept the last passage intact and the other scrambled form transposed the last and next-to-last pas-

**Table 7**  
Percent of Examinees Whose Scale Scores Were Affected by Using Base Form Conversions by Test, and for Composite Score

Form and Test	% Down 1 Point	% Up 1 Point	Total % Affected
<b>Form B</b>			
English	—	12.4	12.4
Mathematics	13.3	4.1	17.4
Social Studies	.7	32.5	33.2
IRT	—	37.9	37.9
Natural Science	9.2	—	9.2
Composite	3.4	9.6	13.0
<b>Form C</b>			
English	7.3	9.9	17.2
Mathematics	—	10.3	10.3
Social Studies	49.3	.5	49.8
IRT	45.3	.4	45.7
Natural Science	7.7	.3	8.0
Composite	14.2	2.9	17.1

sages. The passage means given in Tables 2, 4, and 5 show that in each case the scrambled form with the last passage kept in place had the mean closest to the base form mean for that last passage. If passages were more broadly rearranged than in this study, passage means would likely become even more disparate (see, for example, the differences between the sixth passage means on the English test for Forms B and C, given in Table 2).

Context effects can occur for a variety of reasons in addition to test speededness. Practice effects may cause a passage to appear easier if it is administered second rather than first. Fatigue effects may cause items appearing later in a test to seem more difficult than they would if they appeared earlier in the test. The mathematics items, in which individual items were moved the least, showed less variation across the three forms than did the items on the other three tests.

Due to the nature of context effects, the specific results of this study might not generalize widely. However, given the results found here, it is strongly suggested that those wishing to make use of scrambled test forms to mitigate potential copying while limiting the number of items need-

ed (i.e., it requires twice as many items to use two forms rather than one base and one scrambled form) should at the very least examine the effect that the scrambled form has on equating.

It was found that for the Social Studies tests both the equipercenile equating and the IRT equating methods yielded similar results. This finding implies that, under the conditions in this study (especially the randomly equivalent groups design aspect), the IRT procedure did not appear more context-free than the traditional equipercenile methodology.

### References

- American College Testing. (1988). *ACT Assessment Program Technical Manual*. Iowa City IA:
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington DC: American Council on Education.
- Cook, L. L., & Douglas, J. B. (1982). *Analysis of fit and vertical equating with the three-parameter model*. Paper presented at the annual meeting of the American Educational Research Association, New York.
- Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education*, 3, 245-254.
- Hambleton, R. K., & Traub, R. E. (1974). The effects of item order on test performance and stress. *The Journal of Experimental Education*, 43, 40-46.
- Kolen, M. J. (1984). Effectiveness of analytic smoothing in equipercenile equating. *Journal of Educational Statistics*, 9, 25-44.
- Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research*, 55, 387-413.
- Lord, F. M. (1982). The standard error of equipercenile equating. *Journal of Educational Statistics*, 7, 165-174.
- Wingersky, M. S., Barton, M. A., & Lord, F. M. (1983). *LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. LOGIST 5, Version 1*. Princeton NJ: Educational Testing Service.
- Yen, W. M. (1980). The extent, causes, and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, 17, 297-311.

**Acknowledgments**

*The author thanks Michael J. Kolen, Bradley A. Hanson, and Robert L. Brennan for their comments and assistance.*

**Author's Address**

Send requests for reprints or further information to Deborah J. Harris, American College Testing, P.O. Box 168, Iowa City IA 52243, U.S.A.