Research Report

Does Rearranging Multiple-Choice Item Response Options Affect Item and Test Performance?

ETS RR–19-02

Lin Wang

December 2019

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Senior Research Scientist

Heather Buzick
Senior Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Research Director

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Consultant

Priya Kannan
Managing Research Scientist

Sooyeon Kim
Principal Psychometrician

Anastassia Loukina
Research Scientist

John Mazzeo
Distinguished Presidential Appointee

Donald Powers
Principal Research Scientist

Gautam Puhan
Principal Psychometrician

John Sabatini
Managing Principal Research Scientist

Elizabeth Stone
Research Scientist

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ariela Katz
Proofreader

Ayleen Gontz
Senior Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Does Rearranging Multiple-Choice Item Response Options Affect Item and Test Performance?

Lin Wang

Educational Testing Service, Princeton, NJ

Rearranging response options in different versions of a test of multiple-choice items can be an effective strategy against cheating on the test. This study investigated if rearranging response options would affect item performance and test score comparability. A study test was assembled as the base version from which 3 variant versions were created by rearranging the response options of the items. The 4 versions were administered to randomly equivalent samples of approximately 1,200 test takers in an operational administration. The weighted root mean squared difference and the test characteristic curves were computed from the data to assess the differences between the base and its variant versions. The item-level and test-level results show very small differences between the base and the 3 variant versions.

Keywords Response option rearrangement; item response theory; item performance comparison; test performance comparison

doi:10.1002/ets2.12238

Many large-scale standardized tests administer multiple-choice (MC)-type questions (items) that typically contain four response options, with one option being the correct response (item key) and the other three being incorrect responses (distractors). A test taker provides a response to each test item by either gridding on a paper answer sheet or selecting an option on a computer screen; responses to MC items are often labeled A, B, C, and D (or 1, 2, 3, and 4), and the responses are captured as such for machine scoring. Tests with MC items are prone to various kinds of item key cheating. For example, a test taker may simply copy responses from the answer sheet or computer screen of another test taker sitting next to him or her. A more advanced and prevailing way of item key cheating is often an organized operation in which key creators may obtain test items illegitimately before or during a test administration, work out the item keys, and distribute the keys to test takers who are willing to pay for the item keys. Because the items on a test are usually presented in a fixed sequence, even if a test taker understands little of the content of the test, as long as he or she has the keys recorded in the same sequence as the items on the test and applies the keys in that order (such as A-B-A-C-D-B or 1-2-1-3-4-2 as the keys to six items), he or she is likely to receive a good score (provided the keys are of reasonable quality).

When test scores are used for high-stakes decisions, such as in college admissions or professional license issuances, the popularity of the tests and the importance of the test scores increase, and so does the demand for item keys. Item key cheating has become a highly sophisticated commercial operation that produces and distributes item keys to a large number of paying customers (i.e., fraudulent test takers) in any region of the world, thanks to modern electronic communication technology. For standardized tests that aid in making high-stakes decisions, score validity is of paramount importance to both testing organizations and test score users. Cheating through fraudulent test-taking practices is a serious threat to score validity (Haladyna & Downing, 2004) and test fairness, because test scores obtained through cheating do not represent true proficiency being assessed; test scores from cheating also give cheaters an unfair advantage over honest test takers. Finding effective strategies to reduce the opportunities for item key cheating and to minimize its impact on test results is a significant challenge to testing programs with high stakes.

One effective test security practice against item key cheating is to create more than one version of a test for a test administration. This practice can be accomplished in two ways. One is to place the same items (either individual items or a block of items) at different positions in different versions of the test; the other way is to rearrange the response options to the same test items in different versions (Impara & Foster, 2006). The second strategy in essence generates multiple versions of the same test; therefore, even if the items are presented in the same order on different versions, the order of the response options (including item keys) to an item is changed for different versions. The presentation of the item is partially

Corresponding author: L. Wang, E-mail: lwang@ets.org

changed in that the same response option appears at different positions on different versions. For example, an item key could appear as any of the four possible response options (A, B, C, or D) and so is a distractor. This would make it much harder for a test taker to succeed in cheating simply by following a sequence of item keys, because the response options to an item on his or her version of the test may differ from the response options in the version from which the keys were created by the item key providers. This strategy can be made an integral part of test design, test administration, and data analysis activities of a testing program and is relatively easy to implement. Some institutions also adopt this strategy in their own testing programs. For instance, McGill University (2018) requires that, when making term exams, instructors scramble the order of questions and/or answers to make multiple versions of an exam.

However, there have been concerns about potential context effects on item and test performance induced by varying item positions or rearranging response options. Varying item positions makes a test look different in different versions, and rearranging response options to an item changes the presentation of the item; both practices may affect score comparability among the different versions of a test. Research has been conducted in and outside of the educational measurement field to investigate possible impacts on measurement outcomes due to changing item positions or rearranging response options.

Literature Review

A number of studies reported findings of little impact on item and test performance when item positions were changed (Leary & Dorans, 1985; McLeod, Zhang, & Yu, 2003; Ryan & Chiu, 2001; Vander Schee, 2013). In a comprehensive review of research on item order effect, Leary and Dorans (1985) evaluated prior studies on item order effect, including section order, interaction between item and other factors (e.g., anxiety level), item order on item response theory (IRT) item parameter estimates, and IRT true score equating results. The reviewers concluded that random rearrangement of items or sections of items of the same type under power conditions does not affect test performance; the reviewers suggested that scrambled versions of the same test might be used as a hedge against item key cheating.

Vander Schee (2013) designed an experiment in which three tests on a marketing course were administered at three different times during a semester and each test contained three versions of the same 50 questions, which were arranged in three orders: easy to hard items, hard to easy items, and random; each version (order) was administered to a sample of approximately 150 students of various business-related majors. Measures were taken to minimize the effect of seating arrangement and item order – based test version assignment. The study results suggest that the "test item order and its interaction with gender or major is not a significant factor in student performance" (p. 36) and, therefore, that "faculty who teach introductory marketing courses with students from various business majors should feel comfortable randomizing on multiple-choice exams" (p. 36). In the study by McLeod et al. (2003), the researchers conducted an experiment by randomizing both the order of questions and the order of response options. This study found that the randomization design would reduce cheating with no serious drawbacks. Also, in a study of context effect and differential item function (DIF), Ryan and Chiu (2001) concluded that "the amount of gender DIF and DIF in item parcels tends not be influenced by changes in item position" (p. 73).

There were, however, several studies that reported evidence that item order impacts item performance (Meyers, Miller, & Way, 2009; Way, Carey, & Golub-Smith, 1992; Yen, 1980). Meyers et al. (2009) modeled the item-order effects on IRT-based item difficulty estimates in some K – 12 data and found "measurable effects" related to test equating results due to item position changes. For instance, items on an operational test form would be arranged by difficulty level from easy to hard; the item difficulty parameters were obtained from field trials (pretesting). It was possible that an easy item placed at the end of the pretest could appear at the beginning of the operational test, and vice versa. Interestingly, however, the researchers indicated that the order effects seemed to be mitigated when the hard items from the pretesting forms were placed in the middle of the operational forms. In the study by Way et al. (1992), new reading sets were pretested near the end of the section on a pretest form and were later administered toward the beginning of the section on an operational form; the items were likely to become easier. Conversely, items seemed to become harder if they were pretested near the beginning of the test but appeared toward the end of the operational test.

Golub-Smith (1987) researched the effect of rearranging response options on item performance for an MC test. The author assembled two base versions of the test and, for each base version, created one experimental version by systematically rearranging (scrambling) the response options of the items on the base version. The two base and two experimental versions were spiraled for administration during a pretesting trial; the item response functions and equating functions for all the versions were estimated and compared. The study found that scrambling the options had an effect on item response functions and equating simulations from the items that showed small but real differences. The researcher recommended that this response option scrambling strategy not be implemented for the operational test used for the study.

The item order-related context effect has also been studied in fields outside educational measurement, such as personality assessment (Ortner, 2008; Young, Holtzman, & Bryant, 1954) and marketing survey (Auh, Salisbury, & Johnson, 2003). For example, Young et al. (1954) collected data from eight rating forms constructed from 90 positive and 90 negative 5-point rating scale items; the items were presented in different sequences. The findings indicated that the responses to the items could be "affected by the context within which the items are embedded" (p. 516), and "items descriptive of undesirable personality traits are significantly more susceptible to response shifts with changes in item context than are items referring to desirable traits." (p. 516). Ortner (2008) also found systematic differences in estimated person parameters if item order was systematically changed. Auh et al. (2003) modeled the effects of question order on the output of customer satisfaction in market survey and evaluated the effects in terms of the amount of variation explained by question order; the researchers stated that the question order did not systematically affect the target outcome.

In summary, researchers in educational measurement and other fields have studied the context effects of making multiple versions of the same test on the comparability of the outcomes (e.g., educational test scores, personality scores, survey ratings). Most literature studied impact due to item position change; only limited studies investigated the effect of rearranging response options when item-order change was not feasible. The findings from the studies varied from being affirmative (context effects found) to negative (context effects not found).

The Present Study

Purpose of the Study and Research Questions

The purpose of this study was to evaluate if rearranging response options of MC items would impact the item and test performance. Two research questions guided the study:

1. Does rearranging response options affect item characteristics?
2. Does rearranging response options affect test characteristics and score comparability?

Methods

Design of Rearranging Response Options

The four response options (A, B, C, and D) to each MC item can be rearranged to make different versions of the same item. In the study by Golub-Smith (1987), the base version's item response options, A, B, C, and D, were rearranged to D, C, A, and B for the experimental version; namely, only one experiment version was made from one base version. In the present study, to fully investigate potential effects of changing response option order, each response option of the base version was rearranged in all the other three possible positions to create three experimental (called variant) versions (Variant 1, Variant 2, and Variant 3). Table 1 illustrates this design: the original version is the base and the new versions with the rearranged response options are Variant 1, Variant 2, and Variant 3. For example, response option A appears at the first position in the base, and is then rearranged to the fourth (D) position in Variant 1, the second (B) position in Variant 2, and the third (C) position in Variant 3; Option A can be either the key or a distractor. Response options that follow a logical order (such as chronological events or dates) should not be rearranged.

The test for the study also contained a few multiple-choice-multiple-selection (MCMS) items that were scored dichotomously (1 or 0). Each MCMS item had five response options with three being correct; all three correct response options had to be selected for a score of 1, otherwise a score of 0 was awarded. Table 2 lays out the response option rearrangement plan for the MCMS items; three variants were assembled to be consistent with the MC items.

Test Assembly Design and Test Delivery for the Study

The present study was embedded in a computer-delivered operational test of language proficiency. The test consisted of three sections, with 17 items per section; two sections were operational sections (contributing to test scores), one of which was also used as an internal anchor for linking/equating; the third section was a research section that was not included in computing test takers' scores. The items in the research section were similar to those in the other two sections.

Table 1 The Base and Its Three Variants for Each Multiple-Choice Item With Four Response Options

	Test version			
	Base	Variant 1	Variant 2	Variant 3
Position of response option				
First	A	D	B	C
Second	B	C	A	D
Third	C	B	D	A
Fourth	D	A	C	B
Response option changes		A ↔ D B ↔ C	A ↔ B C ↔ D	A ↔ C B ↔ D

Table 2 The Base and the Three Variants for Each Multiple-Choice-Multiple-Selection Item With Five Response Options

	Test version			
Position of response option	Base	Variant 1	Variant 2	Variant 3
First	A	B	C	D
Second	B	C	D	E
Third	C	D	E	A
Fourth	D	E	A	B
Fifth	E	A	B	C

In fact, the research items had previously been used as operational test items and were eligible for use as external anchor (i.e., linking) items in equating because the parameter estimates of the items were already on the IRT scale of the test and could be used as the reference parameter estimates in scaling and equating. The study was a comparison of the four versions of the research section; all four versions contained the same items, but with the response options of each item ordered differently. The version with the original response option order was the base, from which three variant versions were created by rearranging the response options of each item per Table 1 or Table 2. The four versions were administered to randomly equivalent samples of approximately 1,200 test takers. The research section and its 17 items are referred to as the *study test* and the *study items*, respectively, hereafter.

Data Collection

The data for this study (the study data) were collected from the same operational administration, so that the study data and operational data could be analyzed together; the research section had similar appearance as the two operational sections. The four versions were administered via computer and were assigned to test takers randomly, yielding four randomly equivalent groups (see Figure 1). The test takers' responses to all the operational and study items were electronically recorded and machine scored.

Analysis

Both the operational and study test data were first cleansed by following the same operational data-cleansing procedures to exclude from subsequent data analyses those test takers who attempted too few items, spent too little time on the test, terminated their test sessions abnormally, and so on. Data analyses were performed at both the item and the test levels.

Analysis 1: Item-Level Analysis

The item-level analysis compared the classical and IRT statistics of the study item among the four versions. The classical item statistics include item discrimination, which is indicated by a biserial correlation coefficient (biserial) between item score and total score, and item difficulty, which is indicated by the average item score (AIS) or the percentage of the total test takers selecting the correct responses to dichotomously scored (0 or 1 point) items. Since the four versions contained the same 17 study items, and if the four samples (one sample per version) had equivalent proficiency, then the raw

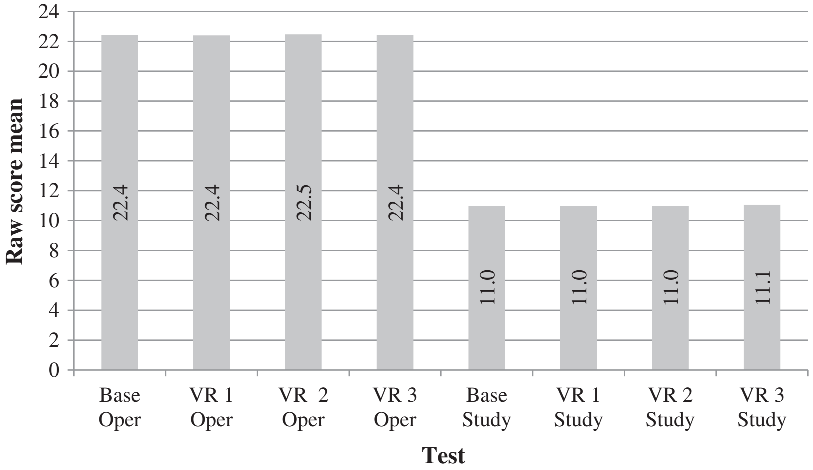

Figure 1 The four study samples' raw score means on the operational (oper.) and study tests for base and variants (VR).

score – based item discrimination (item-total biserial correlation) and item difficulty (AIS) statistics could be compared among the four versions to find out if rearranging the response options impacted item difficulty and discrimination.

The IRT-based analyses included the following two steps:

Step 1: Item calibration. Calibrate the operational items and the study items of the four versions together in the same calibration run with a two-parameter logistic (2PL) IRT model implemented in the ETS proprietary version of PARSCALE (Muraki & Bock, 1999). The parameter estimates of the items from this step needed to be transformed to the IRT scale of test in Step 2.

Step 2: Item scaling. Apply the Stocking and Lord's test characteristic curve (TCC) method (Kolen & Brennan, 1995; Stocking & Lord, 1983; Yen & Fitzpatrick, 2006) to transform (scale) the calibrated IRT item parameter estimates of difficulty (b) and discrimination (a) from Step 1 to the IRT scale of the test; the TCC method required the use of anchor items with reference parameter estimates of a and b from previous administrations. Briefly, the TCC method in this application proceeded in three phases. First, treating the anchor items as a test, the expected raw scores over a range of latent ability (theta) values (typically −4 to 4 in increments of 0.1) from the new parameter estimates (from Step 1) and the reference parameter estimates, respectively, were computed to obtain two TCCs accordingly (see Equation 3). Then, the slope (A) and the intercept (B) of a linear transformation were found that minimized the squared differences between the two TCCs for a given theta value. Finally, the A and B were applied to transform all the items' parameter estimates from Step 1 to the IRT scale of the test.

After Step 2, the study items' parameter estimates were all on the same scale for comparison. The effect of rearranging the response options of an individual item was therefore determined by comparing the ICCs for different versions of the item. The comparison was based on the weighted root mean squared difference (${}_{w}$RMSD) between the two ICCs (one ICC per version) of an item (i) for a population with a standard normal distribution of theta, as follows:

$$w\text{RMSD}_i = \sqrt{\frac{\sum_j \left(P_i^X\left(\theta_j\right) - P_i^Y\left(\theta_j\right)\right)^2 \cdot f\left(\theta_j\right)}{\sum_j f\left(\theta_j\right)}}, \quad (1)$$

where $f(\theta_j)$ is the weight at θ_j from a standard normal distribution. $P_i^X\left(\theta_j\right)$ or $P_i^Y\left(\theta_j\right)$ is a 2PL probability function of item i of version X or Y (such as base and Variant 1) at the ability level θ_j ($-4 \leq j \leq 4$) and is defined as

$$P_i\left(\theta_j\right) = \frac{1}{1 + \exp\left[-Da_i\left(\theta_j - b_i\right)\right]}, \quad (2)$$

where $D = 1.702$ and a_i and b_i are the IRT item parameter estimates of item i. The evaluation criterion is given in the section "Item-Level Results."

Table 3 Descriptive Statistics of the Four Versions of the Research Section Samples

		Operational raw score		Study test raw score	
Version	*N*	Mean	*SD*	Mean	*SD*
Base	1,236	22.41	7.39	10.99	3.86
Variant 1	1,210	22.40	7.27	10.98	3.82
Variant 2	1,283	22.47	7.35	11.00	3.83
Variant 3	1,231	22.42	7.30	11.06	3.84

Analysis 2: Test-Level Analysis

Similarly, the effect of rearranging the response options of the study items was evaluated by comparing the TCCs for different versions of the study test; the comparison was based on expected raw scores computed from the parameter estimates of the study items. For a test of *k* dichotomously scored items, the TCC at the theta level θ_j is computed from

$$\mathrm{TCC}\left(\theta_j\right) = \sum_{i=1}^{k} P_i\left(\theta_j\right), \tag{3}$$

where $P_i(\theta_j)$ is defined in Equation 2. Note that $P_i(\theta_j)$ is also the expected score on item *i* by a test taker of θ_j, $-4 \leq j \leq 4$. In other words, for this study, the TCC for each version at θ_j was simply the sum of the expected scores over the 17 study items. Because the items' parameter estimates for different versions were all transformed to the same IRT scale, the TCCs for the different versions were directly comparable in terms of expected raw score differences across the theta scale.

In IRT true score equating, anchor items are required to scale the items on the new test (see Step 2) before equating the new test scores to the reference test scores. When the response options of the anchor items are rearranged, the equating results could be impacted, and this needs to be investigated. For the study, to assess the possible impact of rearranging response options on the scale scores of the real test, special equatings (not used operationally) were conducted by using the same internal anchor items (scored items; response options were not altered) for the operational equating and different versions of the study test as the external anchor (non-scored items). This was to evaluate if the different versions of the study test would result in comparable raw-to-scale conversion tables.

Results

Proficiency Levels of the Four Study Samples

The random assignment of the four research versions of the study test produced four samples of the test takers. Because the four study samples were all administered the same operational test, the raw score means of the four samples on the operational test were compared to verify if the four samples were equivalent in the proficiency that the operational test measured.

Table 3 and Figure 1 present the descriptive statistics of the four samples that saw the same operational test but different versions of the study test. The raw score means of the operational test (score range 0 – 34) were very similar for the four samples, suggesting that the four samples performed similarly on the same operational test; in other words, the four samples were of similar proficiency levels that the operational test measured. The raw score means of the four versions (score range 0 – 17) of the study test were also very similar, suggesting that the four versions of the study test were of similar difficulty level.

Item-Level Results

Item Statistics Compared

Figures 2 and 3 display the raw score-based item difficulty statistics (AIS, the same as percentage correct for dichotomously scored items) and item discrimination statistics (item-total biserial correlation coefficients), respectively. The four bars for each item represent the item statistics of the base and its three variants. The differences in item difficulty between the base and each variant were all below .04, except for Item 6; the differences in item discrimination between the base

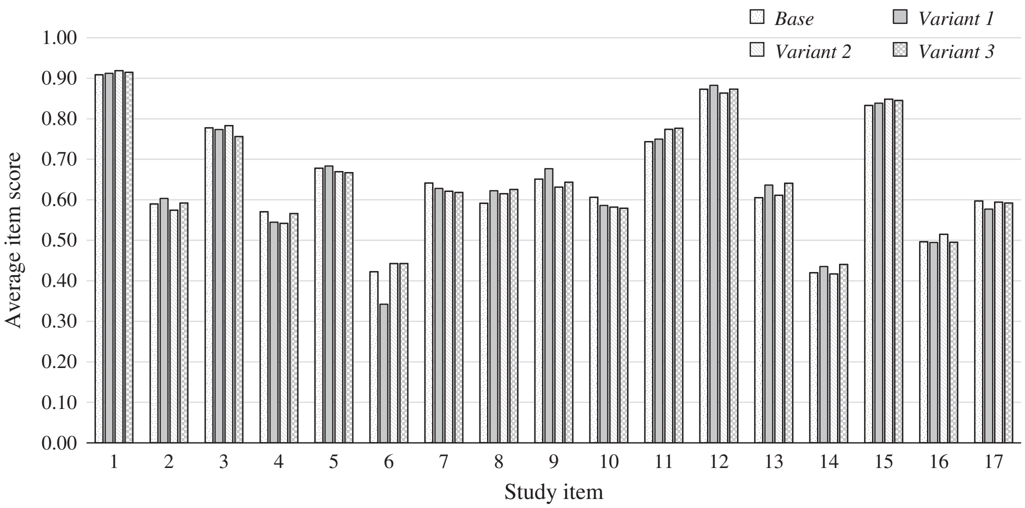

Figure 2 Item difficulty statistics (average item score) of the four versions.

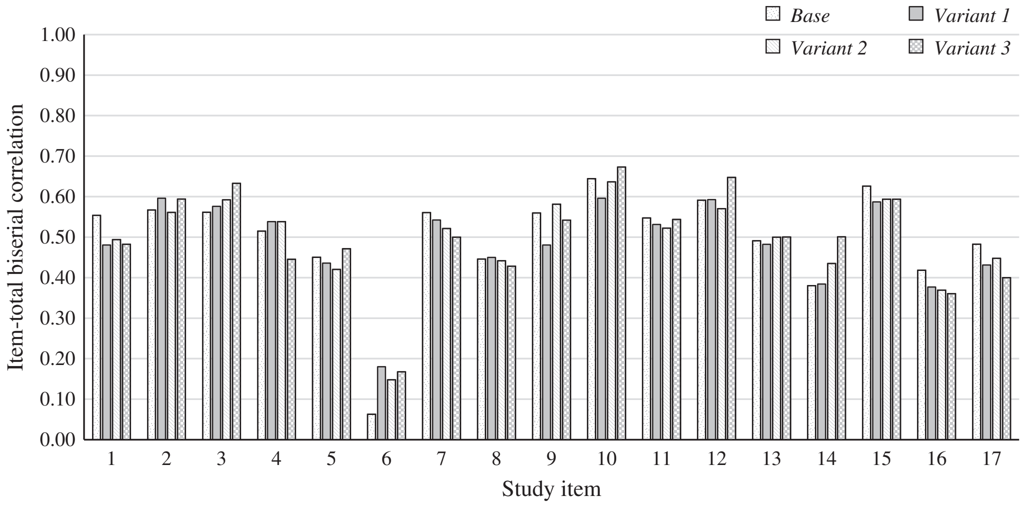

Figure 3 Item discrimination statistics (item-total biserial correlation) of the four versions.

and each variant were below .10 for all items, except Items 6, 9, and 14 (Item 6 had very low discrimination statistics in all four versions: .06 for the base and .18, .15, and .17 for Variant 1, Variant 2, and Variant 3, respectively). Additionally, the differences in item difficulty were below .40 for all except Items 6 and 9, and the differences in item discrimination among the three variances were below .10 for all items except Items 9 and 14. Overall, the differences in difficulty or in discrimination between versions of the same item were much smaller than the differences between items. In fact, the correlations between the base and each variant were all above .97 and .90 for the item difficulty and discrimination statistics, respectively.

Figures 4 and 5 depict the IRT parameter estimates of item discrimination (a) and item difficulty (b) of the base and its three variants. The absolute differences in parameter a ranged from 0 to .26 (Item 12, base vs. Variant 3) between the base and its three variants and from 0 to .35 (also Item 12, Variant 1 vs. Variant 3) among the three variants. The absolute differences in b ranged from 0 to .27 (Item 3, base and Variant 3) between the base and its three variants and from 0 to .34 (Item 9, Variant 1 vs. Variant 2) among the three variants.

Comparisons of the Item Characteristic Curves

The item characteristic curves (ICCs) were compared by computing the weighted ${}_{w}$RMSD (RMSD for short) between each pair of versions of the items (i.e., with different orderings of the response options). The RMSD statistic is on a scale

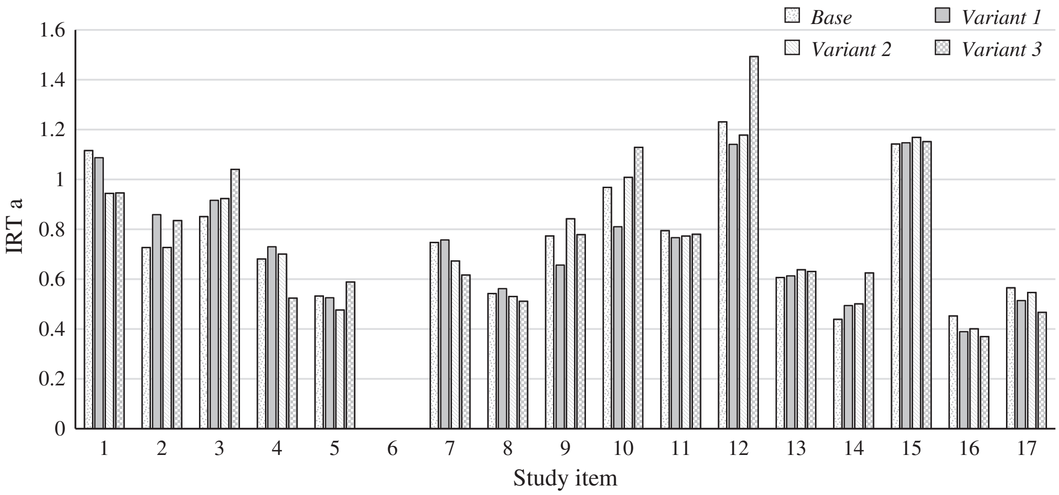

Figure 4 Item response theory item discrimination parameter estimates (IRT *a*) of the four versions.

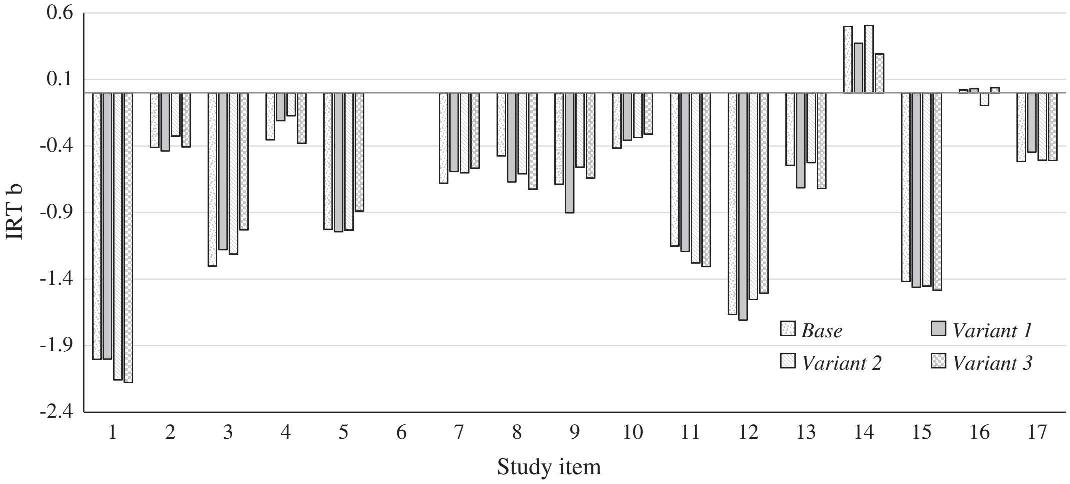

Figure 5 Item response theory item difficulty parameter estimates (IRT *b*) of the four versions.

of 0 to 1 (see Equation 1); the smaller the RMSD statistic is, the more alike a pair of ICCs are. The largest RMSD statistic was .057 (Item 14, base vs. Variant 3) between the base and its three variants and .073 (Item 9, Variant 1 vs. Variant 3) among the three variants; both RMSD statistics were less than .1 (below 10% of the maximum possible value), which was the operationalized criterion.

Distribution of Response Options (Both Keys and Distractors)

The item difficulty analysis in the section Item Statistics Compared was based on the percentage of test takers selecting correct responses (keys). To find out whether rearranging the response options changed the popularity of the distractors, we computed the percentage of test takers selecting each distractor in different versions. For instance, for Item 1, Option D was the correct response, and Options A, B, and C were distractors. Per Tables 1 and 2, Option A on the base version would become Options D, B, and C on Variant 1, Variant 2, and Variant 3, respectively. What percentage of the test takers selected the distractor option A on the base version, D on Variant 1, B on Variant 2, and C on Variant 3? Differences in percentages were calculated between the base version and each variant version for the distractors of the MC items. Most differences were below 4%, with only two being larger (4.8% and 5.1%); the differences were in both directions without particular patterns. Therefore rearranging the response options did not seem to change the popularity of the distractors in general.

Table 4 Summary of the Scale Score Results After Equating

Scale score statistics	Base	Variant 1	Variant 2	Variant 3
Mean	20.27	20.30	20.16	20.26
SD	6.91	6.87	6.94	6.94
Standard error of measurement[a]	2.32	2.30	2.32	2.32
Reliability	0.89	0.89	0.89	0.89
Scaling constant *A*	1.09	1.08	1.09	1.09
Scaling constant *B*	0.05	0.05	0.03	0.05

[a] This is the average conditional standard error of measurement over the range of the scale (0–30).

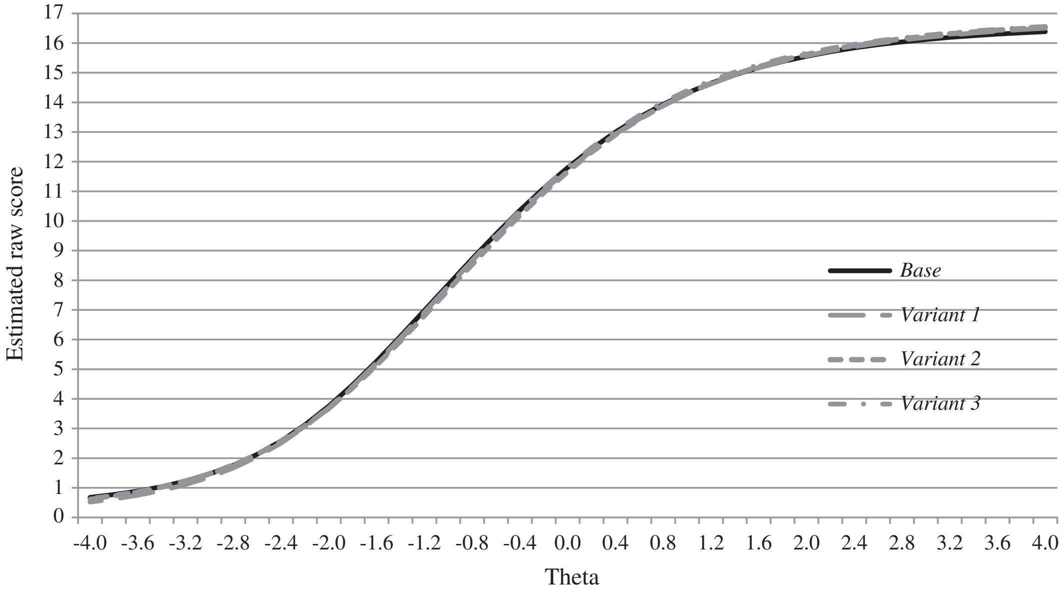

Figure 6 Test characteristic curves of the different versions.

Test-Level Results

Table 4 summarizes the equating results of the four versions from the special equatings where each version was used as an external anchor along with the internal anchor in transforming (scaling) IRT item parameter estimates and IRT true score equating. All the statistics were similar among the four versions.

Figure 6 displays the TCCs of the base and its three variants. The estimated raw scores (0–17) across the ability (theta) scale (−4–4) were computed from the IRT parameter estimates of the 17 study items. The TCCs were very close to one another, indicating that the estimated raw scores of the four versions were similar. Figure 7 plots the differences in TCC between the base and its three variants (base minus a variant); the differences were all very small and within .2 estimated raw score points.

Figure 8 illustrates the raw-to-scale score conversions of the base and its three variants from the special equatings with the same internal anchor and different versions as external anchors. The four conversions were very close to one another. Figure 9 plots the differences in conversions between the base and its variants; all the differences were very small and within .2 scale score points.

Discussion and Conclusion

This study was by design integrated into an operational administration; the different versions of the study test had the same content and statistical properties of the operational test and were administered to highly motivated test takers under the same standardized test administration conditions. The base version and its three variants were randomly distributed to approximately 1,200 test takers, yielding four study samples of similar proficiency level, which was evident in the

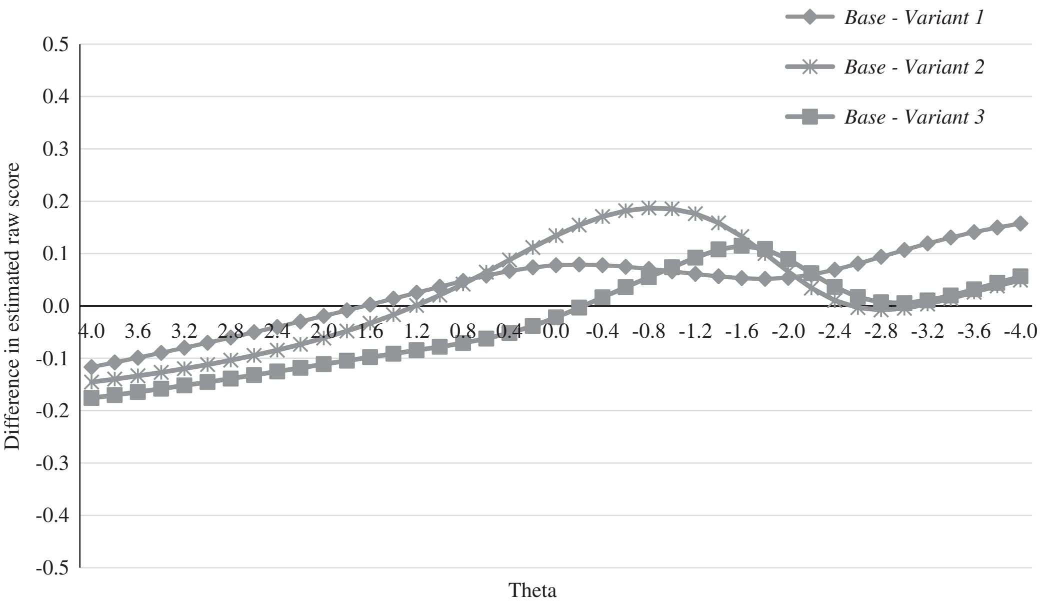

Figure 7 Differences in test characteristic curve between the base and its variants.

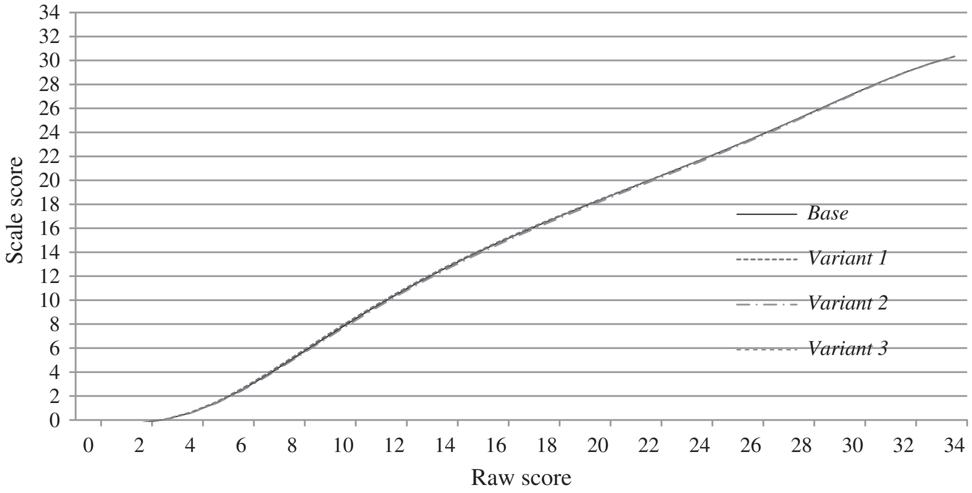

Figure 8 Raw-to-scale conversions of the base and its variants.

operational test mean scores and the mean scores of the four versions (see Figure 1). Each study item's response options appeared in all four possible positions (as A, B, C, or D) for the opportunity to assess if rearranging response options would impact item and test performance.

The fact that the four study samples were equally proficient (per Table 3 and Figure 1) made it possible to compare the item statistics when the items were from different versions. The rearranged response options seemed to have very little effect on the difficulty (AIS) and the discrimination (item-total biserial correlation) of the study items. Furthermore, in the IRT-based analyses, the weighted RMSD statistics were computed to evaluate the differences in each study item's ICC between the base and a variant. The RMSD statistics were less than .10 for all the items. Because RMSD is on a 0–1 metric, differences smaller than .1 were considered very small. Additionally, the small differences in the test takers' selections of rearranged distractors on different versions of the test also supported the finding that the rearranged response options had little effect on the performance of the study items.

Although it is important to be assured that rearranging the response options did not significantly affect individual items' performance, evaluating the performance of the study tests is critical as, ultimately, it is the score on a test, not

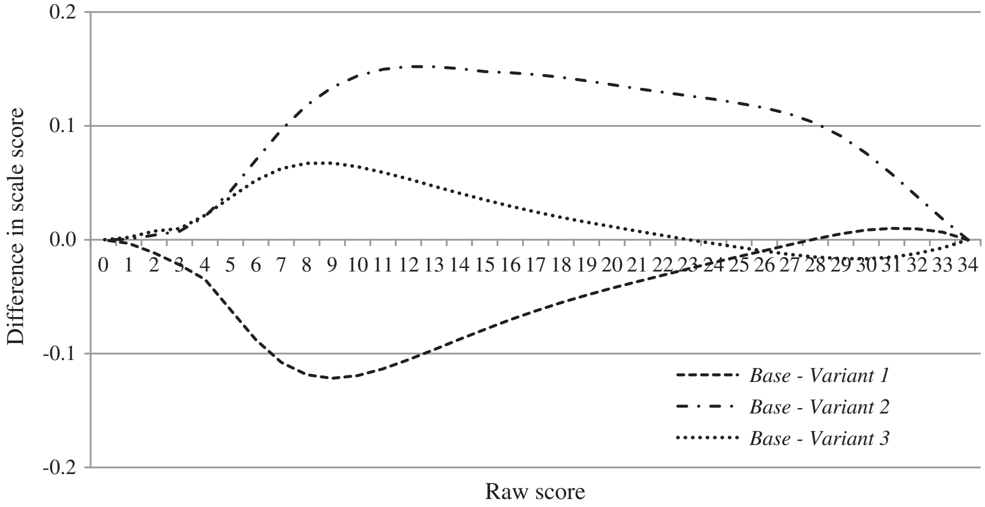

Figure 9 Differences in the raw-to-scale conversions between the base and its variants.

individual items, that is reported for the measured proficiency. Therefore it is critical to have evidence that, when treating each version as a study test, the four versions will show no practically significant score differences. At the test level, the four versions were found to have yielded practically the same estimated (raw) scores corresponding to a theta value on the underlying ability scale (e.g., the TCC differences were all small). Moreover, when the different versions of the study test were used as the external anchors for special equatings (nonoperational), the resulting four raw-to-scale conversions were very similar, with differences in the scale scores all below .2, indicating that the special equatings using the four versions as external anchors yielded comparable scores.

In conclusion, the findings from the study are encouraging in that they did not show practically significant performance differences at either the item or the test level when the response options of the study items were rearranged. In other words, the evidence from the data do not suggest that test takers' scores would have differed significantly, although they took different versions of the same study test. Therefore the strategy of rearranging response options can be considered for operational implementation to enhance test security.

Acknowledgments

The author thanks Samuel Livingston, Hongwen Guo, and Gautam Puhan for their thoughtful comments and suggestions for revision of the original submission. The author also thanks Cindy Nguyen for her contributions to carrying out some of the analyses for the study.

References

Auh, S., Salisbury, L. C., & Johnson, M. D. (2003). Order effects in customer satisfaction modeling. *Journal of Marketing Management, 19*, 379–400. https://doi.org/10.1080/0267257X.2003.9728215

Golub-Smith, M. (1987). *A study of the effects of item option rearrangement on the Listening Comprehension of the Test of English as a Foreign Language* (Research Report No. RR-87-17). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2330-8516.1987.tb00221.x

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*, 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Impara, J. C., & Foster, D. (2006). Item and test development strategies to minimize test fraud. In S. M. Downing & T. M. Halydyna (Eds.), *Handbook of test development* (pp. 91–114). Mahwah, NJ: Lawrence Erlbaum.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating methods and practices*. New York, NY: Springer. https://doi.org/10.1007/978-1-4757-2412-7

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. https://doi.org/10.3102/00346543055003387

McGill University. (2018). *Prevent cheating on exams*. Retrieved from http://www.mcgill.ca/students/srr/honest/staff/exam

McLeod, I., Zhang, Y., & Yu, H. (2003). Multiple-choice randomization. *Journal of Statistics Education, 11*(1), 1–7. https://doi.org/10.1080/10691898.2003.11910695

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*, 38–60. https://doi.org/10.1080/08957340802558342

Muraki, E., & Bock R. (1999). *PARSCALE 3.5: IRT item analysis and test scoring for rating-scale data*. Chicago, IL: Scientific Software.

Ortner, T. M. (2008). Effects of changed item order: A cautionary note to practitioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection and Assessment, 16*, 249–257. https://doi.org/10.1111/j.1468-2389.2008.00431.x

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14*, 73–90. https://doi.org/10.1207/S15324818AME1401_06

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210. https://doi.org/10.1177/014662168300700208

Vander Schee, B. A. (2013). Test item order, level of difficulty, and student performance in marketing education. *Journal of Education for Business, 88*, 36–42. https://doi.org/10.1080/08832323.2011.633581

Way, W. D., Carey, P., & Golub-Smith, M. (1992). *An exploratory study of characteristics related to IRT item parameter invariance with the Test of English as a Foreign Language* (TOEFL Technical Report No. TOEFL-TR-06). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1992.tb01474.x

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. https://doi.org/10.1111/j.1745-3984.1980.tb00833.x

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: Praeger.

Young, H. H., Holtzman, W. H., & Bryant, N. D. (1954). Effects of item context and order on personality ratings. *Educational and Psychological Measurement, 14*, 499–517. https://doi.org/10.1177/001316445401400306

Suggested citation:

Wang, L. (2019). *Does rearranging multiple-choice item response options affect item and test performance?* (Research Report No. RR-19-02). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12238

Action Editor: Gautam Puhan

Reviewers: Hongwen Guoand Samuel Livingston

ETS, the ETS logo, and MEASURING THE POWER OF LEARNING. are registered trademarks of Educational Testing Service (ETS).
All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/