Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 23 Number 16, November 2018 ISSN 1531-7714

Fairness Concerns of Discrete Option Multiple Choice Items

Carol Eckerly, *Educational Testing Service*
Russell Smith, *Alpine Testing Solutions*
John Sowles, *Ericsson*

The Discrete Option Multiple Choice (DOMC) item format was introduced by Foster and Miller (2009) with the intent of improving the security of test content. However, by changing the amount and order of the content presented, the test taking experience varies by test taker, thereby introducing potential fairness issues. In this paper we investigated fairness concerns by evaluating the impact on test takers of the differing testing experiences when items are administered in the DOMC format. Specifically, we described the impact of the presentation order of the key on item difficulty and discrimination as well as the cumulative impact at the test level. We recommend not including DOMC items in exams until the methodology of scoring test takers on these items is revised to address specific fairness concerns identified in this paper.

The Discrete Option Multiple Choice (DOMC) item format was introduced by Foster and Miller (2009) as an alternative to the traditional Multiple Choice (MC) item format for computer administered tests in order to limit test takers' exposure to complete item content. Rather than having access to the stem, key, and all distractors concurrently and then choosing a response, test takers only gain access to response options one at a time as a series of dichotomous true/false responses which are randomly administered to each test taker. Options continue to be administered until a test taker either correctly identifies the key as correct or incorrectly identifies a distractor as correct. After the item has either been scored as correct or incorrect according to this rule, Foster and Miller recommend an additional option be administered with a probability of 0.50 after the item has been scored so test takers are less able to determine the correctness or incorrectness of their responses.

By presenting and scoring items in this manner, test takers will rarely see all of the distractors and the key for each item, and each test taker will have a different testing experience. The rationale behind presenting items in DOMC format is that it may be more difficult for test takers to memorize test content in a way that would seriously compromise the integrity of the test. Foster and Miller (2009) also posit that the DOMC item type may exhibit better measurement properties than traditional MC items by reducing construct irrelevant variance introduced by test-taking skills and cheating.

Limited research has been conducted to determine whether DOMC items are psychometrically comparable to traditional MC items (Foster & Miller 2009; Kingston, Tiemann, Miller, & Foster, 2012). Foster and Miller conducted three experiments using assessment results from introductory psychology students at Brigham Young University. In the first experiment, 39 students responded to items in both traditional MC and DOMC formats; in the second experiment, 150 students responded to items in only the DOMC format; and in the third experiment, 70 students responded to items in both traditional MC and DOMC formats, along with several survey questions. Among the comparisons that could be drawn between traditional MC items and DOMC counterparts, the authors found that most DOMC items were more difficult than traditional MC items, 40% of DOMC items had higher point-biserial correlations than traditional MC items, and test takers on average took 10% less time to respond to DOMC items. Kingston et al. conducted a larger scale experiment with a sample of 802 undergraduate students at Brigham

Young University and the University of Kansas in which traditional MC items were compared to their DOMC counterparts. However, the items presented as DOMC in this experiment were not true DOMC items as described above; answer options were delivered in sequential order such that participants received response options in the same order. The authors reached similar conclusions to those of Foster and Miller regarding point-biserial correlations and item difficulties.

The question of whether response processes to DOMC items fit traditional measurement models has not been addressed in previous literature. In differing contexts, researchers have investigated whether it is possible to learn about the underlying response processes that test takers use to arrive at their final responses in traditional MC items by introducing competing models to those typically used to model response behavior (see, e.g., Deng & Bolt, 2016, and Bolt, Wollack, & Suh, 2012). Presenting items to test takers in a DOMC format facilitates a unique opportunity to learn about the ways in which test takers may arrive at a correct or incorrect response by constraining the steps that a test taker must take to arrive at the response. Whereas test takers can navigate through many different stepwise processes to arrive at the selected option in a traditional MC item, a particular stepwise response process where test takers must respond correctly at each step in order to proceed is imposed for an item in DOMC format. Further, that stepwise response process will have many differing variants which may be administered to test takers.

For example, a four-option item offers 24 possible permutations of response presentation. Thus, we know that a test taker's underlying response process to a DOMC item will necessarily differ depending on the permutation of the item; however, the question remains whether traditional measurement models still reasonably describe these differing response processes. Because a test consisting of DOMC items creates a unique testing experience for each test taker, it is necessary to evaluate the nature of the differing testing experiences and ensure that test takers are not unfairly advantaged or disadvantaged due to the format of administration.

To conceptually compare the DOMC item type to the traditional MC item type, it is helpful to think about the underlying response processes which generate the response data for each item type. For a traditional MC item, test takers have access to all response options at once, and they select the option they believe to be correct. This response process can result from guessing or knowledge about the subject matter addressed in the item, or some combination of both. If an item has four response options, a test taker using true random guessing would answer the item correctly with a probability of 0.25. High ability test takers may be able to recognize the correct response right away, regardless of the attractiveness of the distractors. Test takers of moderate ability may use partial knowledge to eliminate one or more response options and use some combination of partial knowledge and guessing to choose their response.

Similarly, the underlying response process which generates the response data for DOMC items can result from guessing and/or knowledge about the subject matter addressed in the item; however, that process will be different depending on the order in which the test taker receives the response options. For example, if a test taker has no knowledge of the subject matter being assessed and is presented the correct option first, that test taker will answer the item correctly with a probability of 0.5 by using random guessing. However, if the test taker is administered the version of the item with the correct response presented last, the test taker would answer the item correctly with a probability of $0.50^4 = 0.0625$.

Presumably, test takers who sit for exams would generally have a level of ability that would lead them to score higher than they would have using random guessing, but this example highlights the possibility that the same DOMC item could perform differentially depending on the order of response options presented. Given the differing role of guessing for each key position, the following inequality is expected to hold for a 4-option item with one correct answer:

$$P(U=1|\theta, KP=1) > P(U=1|\theta, KP=2) >$$
$$P(U=1|\theta, KP=3) > P(U=1|\theta, KP=4)$$

where U=1 indicates a correct response, θ is test taker ability, and KP indicates key position. Because it is currently recommended that DOMC items should be scored in the same manner regardless of the particular response option presentation of the item, some test takers could be unfairly disadvantaged when items are administered in DOMC format.

Standard 5.16 from The *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014) states that "documentation should be provided to indicate that scores have comparable meaning over alternate sets of test items" (p. 106) when model based psychometric procedures are employed. Because each DOMC item can be presented to test takers in many different ways, it is possible to conceptualize each different order presentation as a different item. Thus, it is necessary to investigate whether scores have comparable meaning across different order presentations.

Methodology

To address whether the DOMC item type introduces fairness concerns, we analyzed data from a test in an IT certification program in which all items are administered in DOMC format. The program had experienced problems with test takers having preknowledge of test content and opted to convert traditional MC items to DOMC format to potentially enhance test security. We investigated how DOMC items compared to their traditional MC counterparts, whether the DOMC item type introduced speededness concerns due to the varying number of response options presented, and whether item-level and test-level statistics varied across different key positions for the same DOMC item. This study differs from previous research regarding DOMC items because it utilizes data from a higher stakes assessment in comparison to earlier studies. Additionally, the sample sizes of items and test takers are relatively large, addressing a limitation in Foster and Miller (2009), and items were presented in the true DOMC format in which response options were randomly administered to test takers, addressing a limitation of Kingston et al. (2012).

Instruments

We analyzed data from two test forms administered in DOMC format and three test forms administered in traditional MC format with items covering the same content. Within each set of forms administered within an item type format (i.e., DOMC and traditional MC), forms were built to be equivalent and there was a great deal of overlap across forms. Each form of the DOMC version of the test consisted of 59 items which either had four total response options with one key, four total response options with two keys, or five total response options with three keys. In order to receive one point for the multiple select items with either two or three keys, test takers had to correctly select all of the keys. On each form, 38 items had one key, 17 items had two keys, and 4 items had three keys. Test takers were not aware of how many keys each item had. Thirty-five items were common to the two forms. Each form of the traditional MC version of the test consisted of 64 items which had between one and three keys and four or five total response options. On each form, 38 items had one key, 22 items had two keys, and 4 items had three keys. Twenty-five items were common to all three forms.

For both the traditional MC administration and the DOMC administration, all items were worth one point and were scored dichotomously. In developing the forms for the DOMC administration, the testing program reviewed all items from the previous traditional MC versions, keeping the ones that were still relevant while discarding the others. In addition, they modified the wording of some of the items to have them better fit the DOMC format. No completely new items were developed for the DOMC administration.

Sample

The sample consisted of test takers who were seeking to become certified in the technical content covered by the test. The certification program was internal to the sponsoring organization, so all test takers were employees of the organization. There were 635 test takers who were randomly assigned to take one of the two forms of the test administered in DOMC format, and there were 2,083 test takers who were randomly assigned to take one of the three forms of the test administered in traditional MC format. The sample of test takers who were administered the traditional MC format may have differed from the sample of test takers who were administered the DOMC format, as the two sets of forms were not administered concurrently. However, the test had been administered for many years and the sponsoring organization reported that the distribution of test taker ability had remained fairly stable over time.

Results

Comparison of DOMC and MC formats

Overall test performance changed substantially when the test forms were administered in the DOMC format compared to the traditional MC format. The average item p-value (proportion of correct response) decreased from 0.54 to 0.38, suggesting that the DOMC test may have been more difficult for test takers. We had access to mappings of 60 traditional MC items to their DOMC counterparts, which allowed for an incomplete analysis of differences in item difficulty at the item level. Figure 1 provides a plot of the p-values for each of these items for the traditional MC administration versus the DOMC administration. Points for all but one of these items lie above the identity line, indicating that they likely were more difficult in DOMC administration than in the traditional MC administration.

Of course, potential differences in the population of test takers who took the MC version versus the DOMC version, including differing amounts of cheating that may have occurred, make it impossible to conclude with certainty that the DOMC items were more difficult. Although test takers had the opportunity to respond to several practice items utilizing the DOMC format prior to taking the test, it is also possible that the decreases in p-value stem partly from test takers' unfamiliarity with the item type. While the results shown here are consistent with previous research suggesting that DOMC items are more difficult than their traditional MC counterparts, we investigated additional questions related to test takers' response processes and whether some groups of test takers were differentially affected by the DOMC item format.

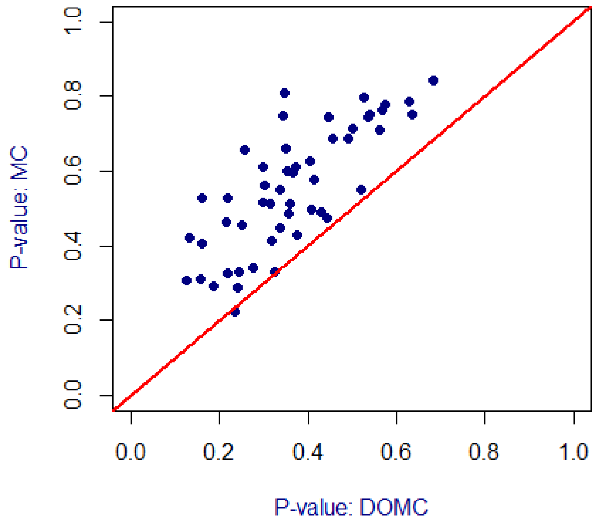

Figure 1. Traditional MC vs. DOMC p-values

Testing Time

Because test takers responding to DOMC items see varying numbers of response options per item, we investigated the length of the test for each test taker based on the number of response options seen. Figure 2 is a plot of the number of response options seen versus the test taker total score on the test. The number of response options seen does not include responses that were shown after an item was scored (which were programmed to occur with a probability of 0.40). Unsurprisingly, there was a positive relationship between test-taker scores and the number of response options seen. Lower ability test takers were more likely to answer items incorrectly earlier in the sequence of response options, terminating the further exposure of remaining response options. Thus it would be more difficult for lower ability test takers to successfully memorize and distribute item content because they would not gain access to a large portion of the response options. Each form of this test consisted of 240 total response options across the 59 items, and the highest number of response options seen by any test taker was 140 (not including any response options shown after an item was scored).

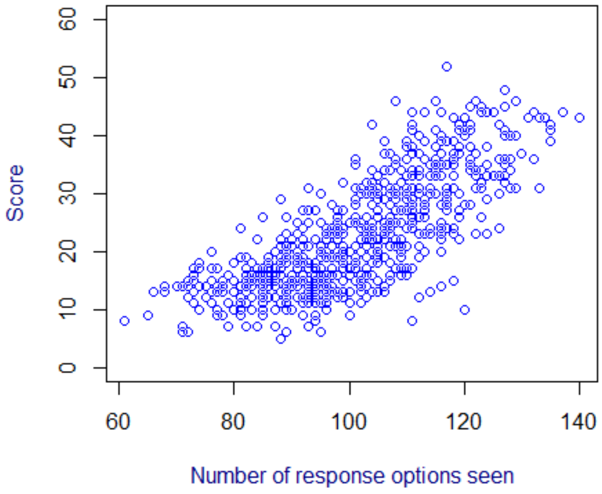

Figure 2. Test taker score vs. number of response options seen

Whereas the varying number of response options based on test-taker ability is an intended consequence of the DOMC item type, Figure 2 also shows that there was quite a bit of variability in number of response options seen for a given score. Thus, we investigated whether test takers who were administered higher numbers of response options were likely to run into time pressure at the end of the test. Figures 3 and 4 address this question, showing test-taker total score versus total time and test-taker total time versus number of response options seen, respectively. Both of these figures indicate that test takers with higher scores and test takers who saw more response options did not seem to run into time pressure at the end of the test. Very few test takers approached the time limit of 95 minutes, and those who did had a high range of scores and number of response options seen, so the time limit seemed appropriate for this particular test. However, the variability in number of response options seen and the positive relationship between total score and the number of response options seen highlight the need for practitioners who employ this item type to carefully consider the time limit to ensure that test takers who effectively have longer tests have sufficient time to complete the test without it being speeded.

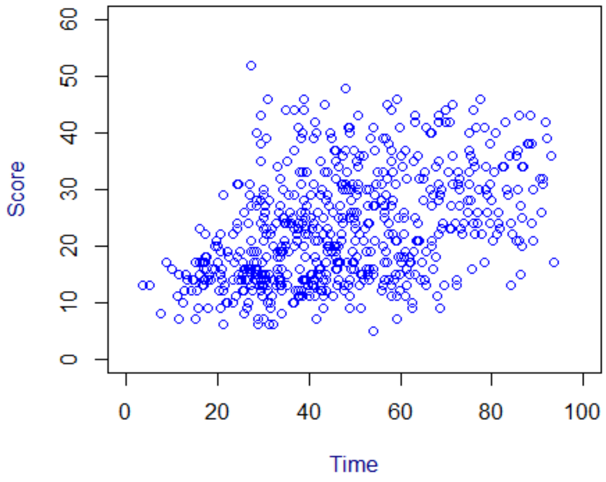

Figure 3. Test taker score vs. time.

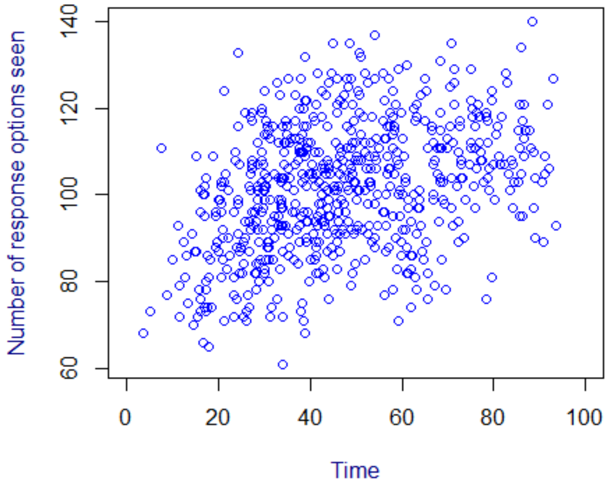

Figure 4. Test taker time vs. number of response options seen

Item-level and test-level statistics

To investigate the effects of differing response option orders on item statistics, we recoded each DOMC item that had only one key into four separate items based on the assigned response order. The assigned response order was one of 24 permutations of the four response options; however, for recoding purposes, we treated permutations which had the same key position as the same item. For example, for an item whose key was “A”, assigned response order permutations ABCD, ACBD, ADBC, ABDC, ACDB, and ADCB were recoded as the same item. For each of these response option presentations, response options B, C, and D would never be seen by the test taker before the item was scored. For response option presentations in which the key is not assigned in the first position, different permutations of distractors may be seen by test takers before the key is presented to them. However, we still based our recoding only on the assigned key position because we wanted to ensure we had a large enough sample size for each recoded item to draw valid conclusions.

Recoding each DOMC item into four separate items based on key position resulted in a 635 test taker by 216 item response matrix for analysis purposes. We performed analyses to evaluate both classical and Rasch item statistics on the recoded items (Rasch, 1960/1980). The Rasch model assumes local independence, meaning that after controlling for test taker ability, no relationship remains between the item responses (Embretson & Reise, 2000). Clearly, recoded items with the same stem have a relationship and can be considered as variant items which are similar but not identical to each other (Woo & Gorham, 2010). In the presence of variant items, test takers should not be administered more than one item from a group of variants to avoid violations of the local independence assumption. These analyses meet this requirement because no test taker responded to more than one recoded item with the same stem; thus the local independence assumption was not violated due to the recoding in this analysis.

Figure 5 shows the frequencies of sample sizes for each of the recoded items used in the analysis. Sample sizes ranged from 65 to 185 responses for each of the 216 recoded items. The cluster of smaller sample sizes shown in the histogram represents recoded items which appeared on one form, and the cluster of larger sample sizes represents recoded items which appeared on both forms. It is worth noting that for each DOMC item, the subsets of test takers responding to each of the four recoded options are randomly equivalent groups because the option presentation order was randomly assigned to each test taker.

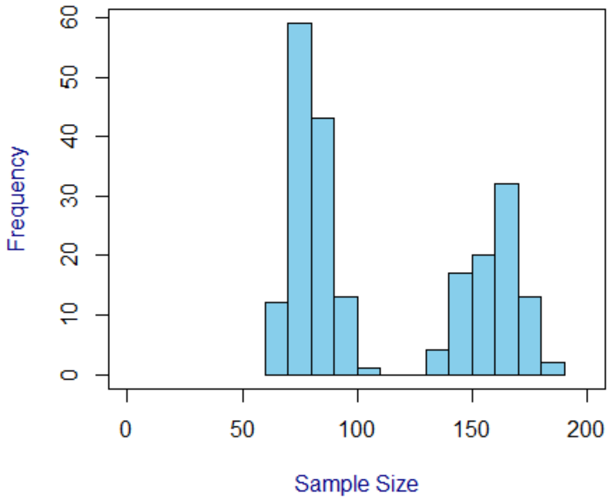

Figure 5. Sample size by item version.

Figure 6 shows p-values for each of the recoded items grouped by key position. As key position increases, p-values generally decrease. Average p-values for recoded items with key position 1, 2, 3, and 4 were 0.64, 0.48, 0.35, and 0.29, respectively (shown in bold red on Figure 6). These results are consistent with our hypothesis that the role of guessing in test-taker responses differs depending on the key position, because guessing plays a larger role in the probability of answering the item correctly with lower key positions. While the assigned option presentation was randomly determined for each item administered to each test taker, the average key position for each test taker ranged from 2.00 to 3.08, suggesting that test takers responded to subsets of items with differing average difficulty. Thus, scoring test takers without taking into account differences in difficulty introduced by key position is likely advantaging test takers with low average key positions and disadvantaging test takers with high average key positions.

The different subsets of recoded items administered to test takers can be conceptualized as representing different forms, and for the purposes of this analysis, we will refer to these differing subsets as different forms. Because test takers effectively received forms of differing difficulty, an item response theory framework is necessary to evaluate results on a common metric and quantify differences in form difficulty.

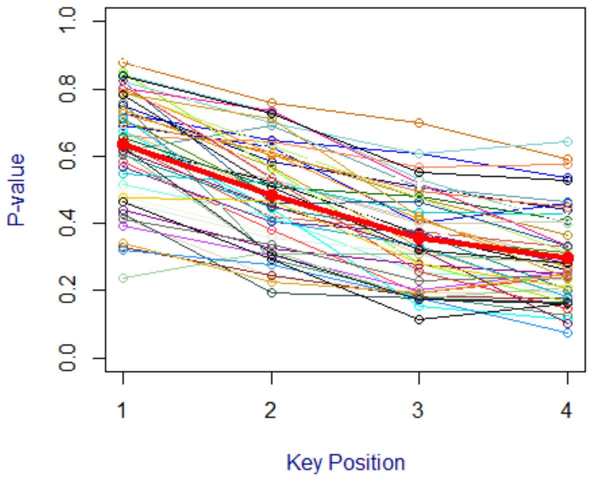

Figure 6. p-value by key position

Figure 7 shows estimated item difficulty parameters from the Rasch model for each of the recoded items grouped by key position. Although different item response theory models may better describe the data resulting from DOMC item administration (e.g., a two-parameter logistic model), we chose to use the Rasch model due to sample size constraints. However, we recognize that there will be systematic model-data misfit based on key position because the Rasch model does not allow for estimation of guessing parameters or variable discrimination parameters. Similarly to the results shown in Figure 6, as key position increases, estimated item

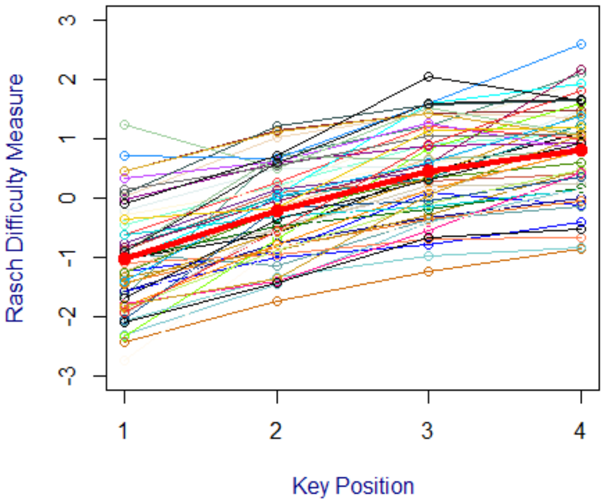

Figure 7. Rasch difficulty measure by key position

difficulty parameters generally increase. Average item difficulty for recoded items with key position 1, 2, 3, and 4 were -1.03, -0.22, 0.44, and 0.81, respectively (shown in bold red on Figure 7).

The Rasch framework allows us to evaluate how these differences in item difficulty by key position manifested themselves at the total test level for individual test takers. Figure 8 shows the distribution of average item difficulty for the complete subset of recoded items each test taker was administered. The distribution of average Rasch item difficulty parameter estimates for the various forms is centered near zero, with a minimum value of -0.40 and a maximum value of 0.32, showing variation in the average difficulty of the items administered to each test taker.

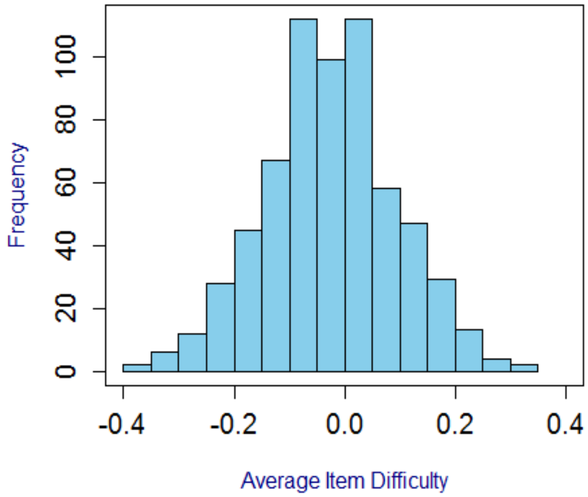

Figure 8. Average item difficulty by recoded form.

To visualize how these differences in average form difficulty affect the test characteristic curves, Figure 9 shows the test characteristic curves for five example forms composed of recoded items. These example forms correspond to the minimum, first quartile, median, third quartile, and maximum average item difficulty. While the first quartile, median, and third quartile example forms show a similar relationship between raw score and Rasch person measure (i.e., theta), the minimum and maximum example forms differ by six points at a Rasch person measure of zero. Thus, if different test takers of equal ability at a Rasch person measure of zero were administered the hardest and easiest form of the test, those who were administered the easiest form would be expected to have raw scores that were six (out of 38) points higher than those who were administered the hardest form. Test takers who randomly received forms with high average item difficulty were clearly disadvantaged compared to test takers who randomly received forms with low average item difficulty because the current scoring approach does not take into account differences in form difficulty introduced by varying key positions of items.

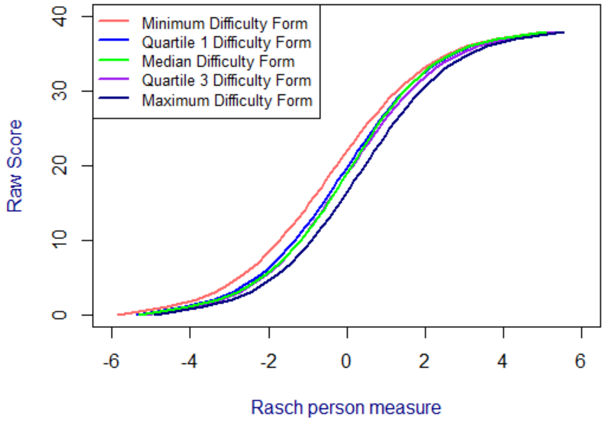

Figure 9. Test characteristic curves: Example forms

In addition to analyzing changes in item difficulty based on key position, we also analyzed changes in item discrimination based on key position by comparing point-biserial correlations. Figure 10 is a plot of point-biserial correlations versus item difficulty for each of the items recoded based on key position. Because we have shown that test takers received forms of differing difficulty, point-biserial correlations were calculated using Rasch person measures (i.e., theta measures) rather than raw scores. Recoded items in which the key was shown in position 1, 2, 3, and 4 are shown in black, red, green, and blue, respectively. Separate linear regression lines are included in the figure to show the relationship between item difficulty and point-biserial correlation for recoded items with the same key position. The negative slopes for each of the regression lines indicate the inverse relationship between item difficulty and point-biserial correlation, and the increasing y-intercepts of the regression lines for increasing key position indicate the direct relationship between key position and item discrimination. Thus, controlling for key position, easier items tended to have higher point-biserial correlations, and controlling for item difficulty, items with higher key positions tended to have higher point-biserial correlations.

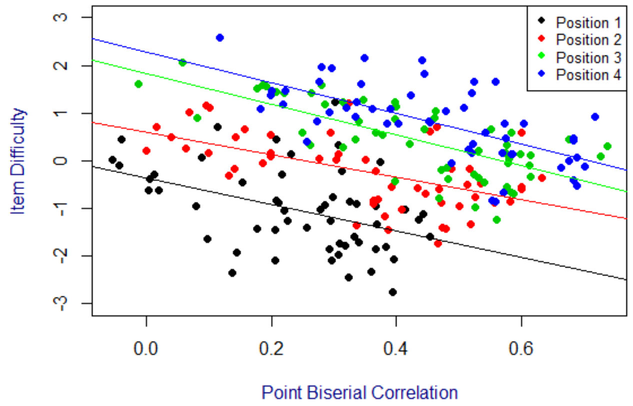

Figure 10. Point-biserial correlation vs. item difficulty by key position

Cronbach's alpha (Cronbach, 1951) is a common measure used to estimate reliability of test scores. The formula for Cronbach's alpha is $\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_C^2}\right)$ where k = number of items, σ_i^2= variance of each item i and σ_C^2 = variance of total test scores. Individual item variances and total test variance can be estimated using item p-values and point-biserial correlations. Because both p-values and point-biserial correlations were shown to be influenced by key position, it stands to reason that estimated reliability for the differing forms shown to test takers would vary. Thus, we estimated Cronbach's alpha for the same example forms shown in Figure 10, corresponding to the minimum, first quartile, median, third quartile, and maximum average item difficulty forms administered to test takers. Cronbach's alpha estimates for these example forms were 0.80, 0.79, 0.83, 0.85, and 0.88, respectively, indicating that harder forms generally had higher estimated reliability.

While we have shown that item-level statistics (i.e. item difficulty and item discrimination) and test-level statistics (i.e., average item difficulty and estimated reliability) vary depending on the key position for the DOMC items, it is also important to analyze how this variability affects individual test-taker measures. Figure 11 displays plots of Rasch person measures using two different scoring methods. Scoring Method 1 estimates item difficulties for the 54 DOMC items without recoding based on key position. It does not take into account potential differences in item difficulty due to key position. Scoring Method 2 utilizes the item recoding described above to estimate item difficulties for the four variations based on key position for each of the 54 DOMC items (for a total of 216 item difficulties), taking into account potential differences in item difficulty due to key position. While the correlation between the Rasch person measures obtained from Scoring Method 1 and Scoring Method 2 is high (i.e., 0.99), it is clear that some test takers are advantaged or disadvantaged due to the particular combination of DOMC item variations they were administered. Test takers who were administered the top 20 easiest or hardest forms are color coded in navy and red, respectively, on the plot to show that Scoring Method 1 overestimates the ability of test takers who were administered easier forms and underestimates the ability of test takers who were administered harder forms.

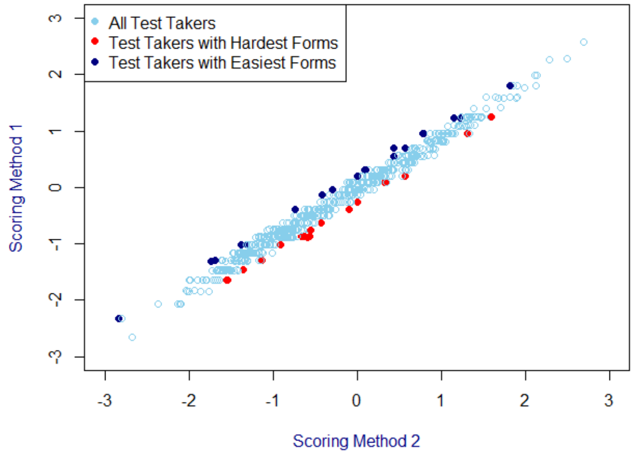

Figure 11. Rasch person measure comparison

Discussion

The DOMC item type was introduced to protect test content from exposure by presenting different subsets of response options to test takers, thereby creating unique testing experiences for each test taker in which not all of the response options are revealed. The validity of test scores depends heavily on test content remaining secure, so efforts to reduce exposure and item harvesting may contribute to the overall health of a testing program. However, we are not aware of any empirical evidence to support the claim that DOMC items improve the integrity of test programs. Further, it remains necessary to ensure that the testing experience is fair to test takers, which is the focus of the current research.

We have shown that item difficulty and discrimination varied substantially for the DOMC items in this dataset, depending on key position, leading test takers to see forms of varying difficulty and reliability. Given that the role of guessing in responding correctly to DOMC items changes depending on the key position, it is reasonable to conclude that these results are not isolated to this dataset. However, the magnitude of variability in difficulty and reliability likely depend on the context of the assessment. For example, assessments which are difficult for the population of test takers will likely see larger effects in item properties due to key position, and some types of test content may be more immune to key position effects than others.

We recommend that testing programs not use DOMC items until a methodology is developed to address the fairness and measurement model fit issues addressed in this paper. As shown in this study, without doing so can introduce significant fairness issues with respect to varying item difficulty and discrimination. One possible strategy to control for difficulty and discrimination could be to include constraints for response presentation order in the DOMC algorithm, ensuring that test takers receive the same number of items with the key in each of the respective positions. This strategy would likely mitigate but not remove the differences in form difficulty for test takers. Further, programs with large enough sample sizes could consider treating each DOMC item as several separate items based on key position, as was done in the analysis here, and either score test takers based on an item response theory model from the recoded item analysis or select items such that the sets of items administered are equivalent. However, we do not recommend using the Rasch model for this purpose because 1) the Rasch model is most likely not appropriate for items in which the key is shown in the first position, as these items are essentially true/false items, and 2) we have shown that item discrimination varies as a function of key position. Lastly, it may be possible to model changes in item performance based on response presentation order and use those models to score test takers probabilistically. Research should be conducted to determine the extent to which any proposed methodology to score test takers on DOMC items may mitigate fairness issues introduced by changing item difficulty based on key and distractor order before these items are used in practice.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Bolt, D. M., Wollack, J. A., & Suh, Y. (2012). Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika*, 77, 263-287.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika.* 16, 297-334.

Deng, S., & Bolt, D. M. (2016). A sequential IRT model for multiple-choice items and a multidimensional extension. *Applied Psychological Measurement*, 40, 243-257.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Foster, D. F., & Miller, H. L., Jr. (2009). A new format for multiple-choice testing: Discrete option multiple-choice. Results from early studies. *Psychology Science Quarterly*, 51, 355-369.

Kingston, N. M., Tiemann, G. C., Miller, H. L., Jr, & Foster, D. (2012). An analysis of the discrete-option multiple-choice item type. *Psychological Test and Assessment Modeling*, 54, 3-19.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: University of Chicago Press. (Original work published 1960)

Woo, A., & Gorham, J. L. (2010). Understanding the impact of enemy items on test validity and measurement precision. *CLEAR Exam Review*, 21, 15-17.

Citation:

Eckerly, Carol, Smith, Russell, & Sowles, John (2018). Fairness Concerns of Discrete Option Multiple Choice Items. *Practical Assessment, Research & Evaluation*, 23(16). Available online: http://pareonline.net/getvn.asp?v=23&n=16.

Authors' Notes:

The authors would like to thank the ETS technical reviewers for comments provided on an earlier version of this manuscript. The original version of this manuscript was written when the first author worked at Alpine Testing Solutions. The opinions expressed are those of the authors and do not necessarily represent the views of the Educational Testing Service, Alpine Testing Solutions, or Ericsson.

A version of this paper was presented at the 2017 Conference on Test Security.

Corresponding Author

Carol Eckerly, Associate Psychometrician
Educational Testing Service
Princeton, NJ

email: ceckerly [at] ets.org