Analysis of the Discrete Option Multiple Choice Item: Examples from IT Certification

Carol Eckerly
Alpine Testing Solutions

Russell W Smith
Alpine Testing Solutions

John Sowles
Ericsson

September 7, 2017
Paper presented at the Conference on Test Security,
Madison, WI

Introduction

The Discrete Option Multiple Choice (DOMC) item format was developed by Foster and Miller (2009) as an alternative to the traditional Multiple Choice (MC) item format in order to limit examinees' exposure to complete item content. Rather than having access to the stem, key, and all distractors concurrently and then choosing a response, examinees only gain access to response options one at a time as a series of dichotomous true/false responses which are randomly administered to each examinee. Options continue to be administered until an examinee 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 (2009) 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, examinees will rarely see all of the distractors and the key for each item, and each examinee will have a slightly different testing experience. The idea is that it would be harder for examinees to memorize exam content in a way that would seriously compromise the integrity of the exam. Foster and Miller 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, and whether response processes to DOMC items fit traditional measurement models (Foster and 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 format; 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 format, 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 examinees took 10% less time to respond to DOMC items on average. Kingston, Tiemann, Miller, and Foster conducted a larger scale experiment with a sample of 802 undergraduate students at Brigham Young University and the University of Kansas where 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 had similar conclusions to those of Foster and Miller regarding point-biserial correlations and item difficulties.

Because an exam consisting of DOMC items creates a unique testing experience for each examinee, it is necessary to evaluate the nature of the differing testing experiences and ensure that examinees are not unfairly advantaged or disadvantaged due to the format of administration. Standard 5.16 from The Standards for Educational and Psychological Testing (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 examinees in many different ways (e.g., an item with four response options item offers 24 possible permutations of response presentation), it is possible to conceptualize each different order presentation as a different item. Because it is currently recommended to score items in the same manner regardless of the order presentation, it is necessary to investigate whether scores have comparable meaning across different order presentations. To that end, this paper investigates questions relating to the differing testing experiences of examinees. We analyzed data from an exam in an IT certification program that administered all items in DOMC format. Prior to administering the items in DOMC format, the program had problems with examinees having preknowledge of exam content and opted to convert traditional MC items to DOMC format in order to enhance exam security. This study differs from previous research regarding DOMC items due to its high stakes nature. Additionally, the sample size of items and examinees is large, addressing a limitation in Foster and Miller (2009), and items were presented in the true DOMC format where response options were randomly administered to examinees, addressing a limitation of Kingston et al. (2012).

Sample

The sample of examinees who were administered the DOMC version of the exam consisted of 635 examinees who took one of two 59-item forms of the exam. All items were administered in DOMC format, where items either had 4 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, examinees had to correctly select all of the keys. On each form, 38 items had one key, 17 items had two keys, and four items had three keys. Examinees were not aware of how many keys each item had. Thirty-five items were common to the two forms. The sample of examinees who were administered the traditional MC version of the exam consisted of 2,083 examinees who took one of three 64-item forms of the exam. The exam consisted of both traditional MC (with 1 key and 4 total response options) and multiple select MC (with 2 or 3 keys and 4 or 5 total response options), where 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.

Results

Overall exam performance changed substantially when the exam forms were administered in the DOMC format compared to the traditional MC format. Figures 1 and 2 show the distribution of item *p*-values (proportion of correct response) for both the traditional MC and DOMC formats. We did not have access to the mapping of the traditional MC items to their DOMC counterparts, so we were unable to analyze differences in difficulty at the item level. As indicated above, previous research has shown that DOMC items tend to be more difficult than their traditional MC counterparts as evidenced by lower proportions of examinees answering correctly. While the results shown here are consistent with previous research showing that DOMC items are more difficult than their traditional MC counterparts, we investigated several additional questions related to the increased difficulty to shed some light on what may have led the items to be more difficult and whether some groups of examinees were differentially affected by the DOMC item format.

Figure 1

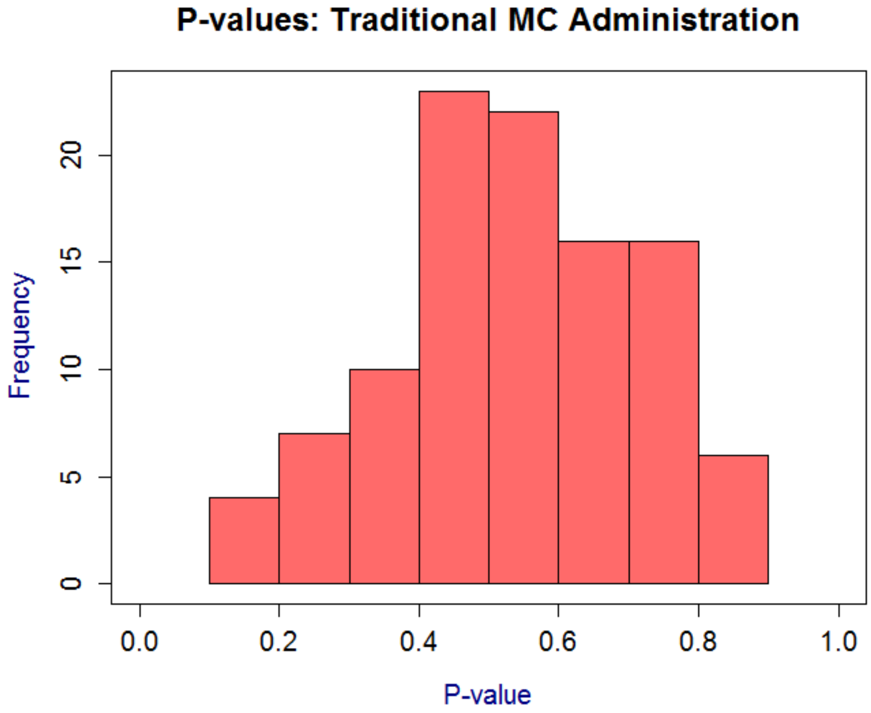
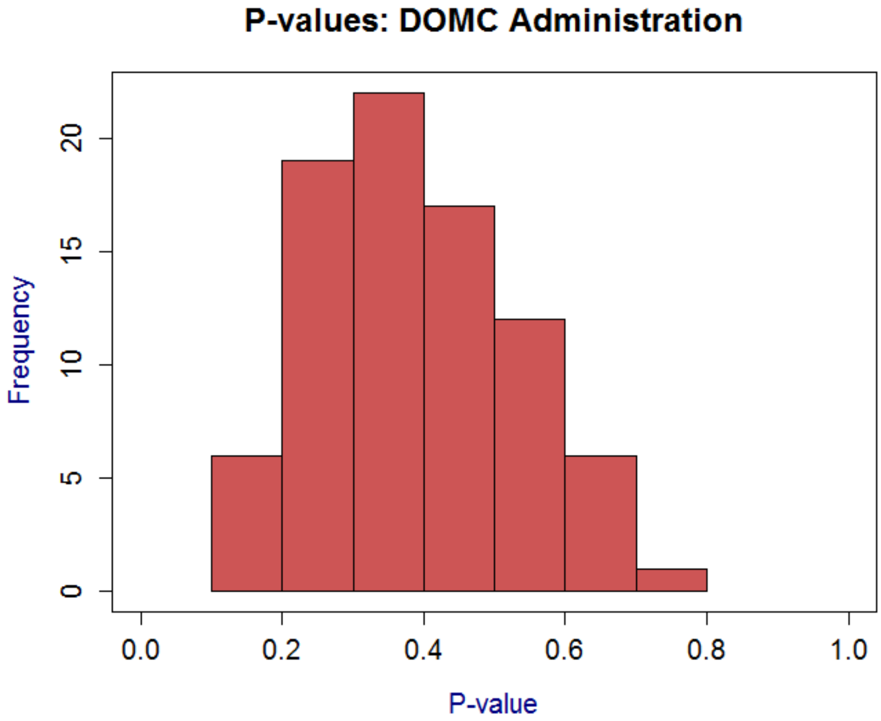
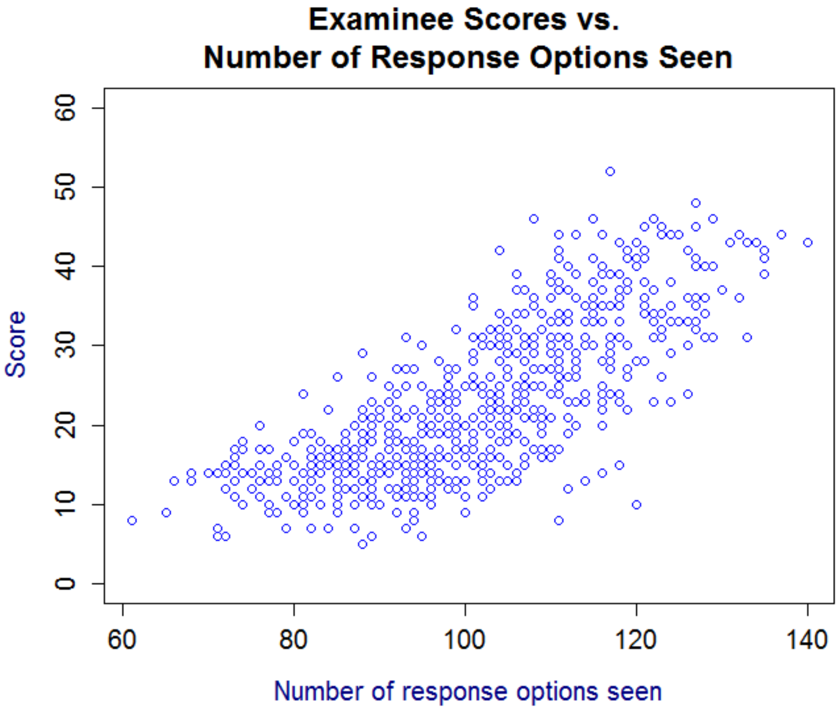


Figure 2



Because examinees responding to DOMC items see varying numbers of response options per item, we investigated the length of the exam for each examinee based on the number of response options seen. Figure 3 plots the number of response options seen versus the examinee total score on the exam. The number of response options seen does not include responses which 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 examinee scores and the number of response options seen. Lower ability examinees were more likely to answer items incorrectly earlier in the sequence of response options, terminating the further exposure of remaining response options. Thus, if low ability examinees were attempting to steal content, they would not be very successful at gaining access to a large portion of the response options. Each form of this exam consisted of 240 total response options across the 59 items, and the highest number of response options seen by any examinee was 140 (not including any response options shown after an item was scored).

Figure 3



While the varying number of response options based on examinee ability is an intended consequence of the DOMC item type, Figure 3 also shows that there was quite a bit of variability in number of response options seen for a given score. Thus, we investigated whether examinees who were administered higher numbers of response options were likely to run into time pressure at the end of the exam. Figures 4 and 5 address this question, showing examinee total score versus total time and examinee total time versus number of response options seen, respectively. Both of these figures indicate that examinees with higher scores and examinees who saw more response options did not seem to run into time pressure at the end of the exam. Very few examinees 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 exam. 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 examinees who effectively have longer exams have sufficient time to complete the exam without being speeded.

Figure 4

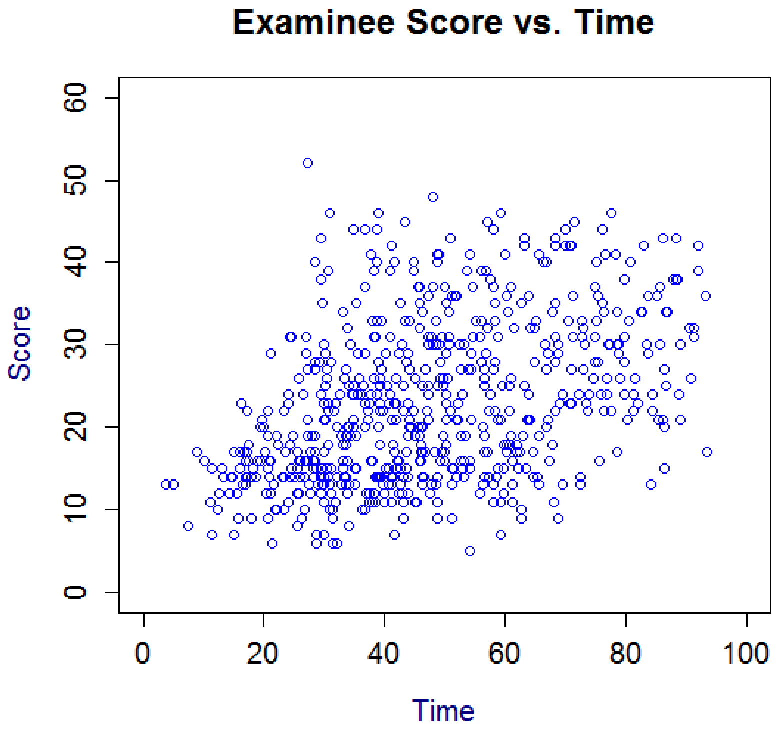
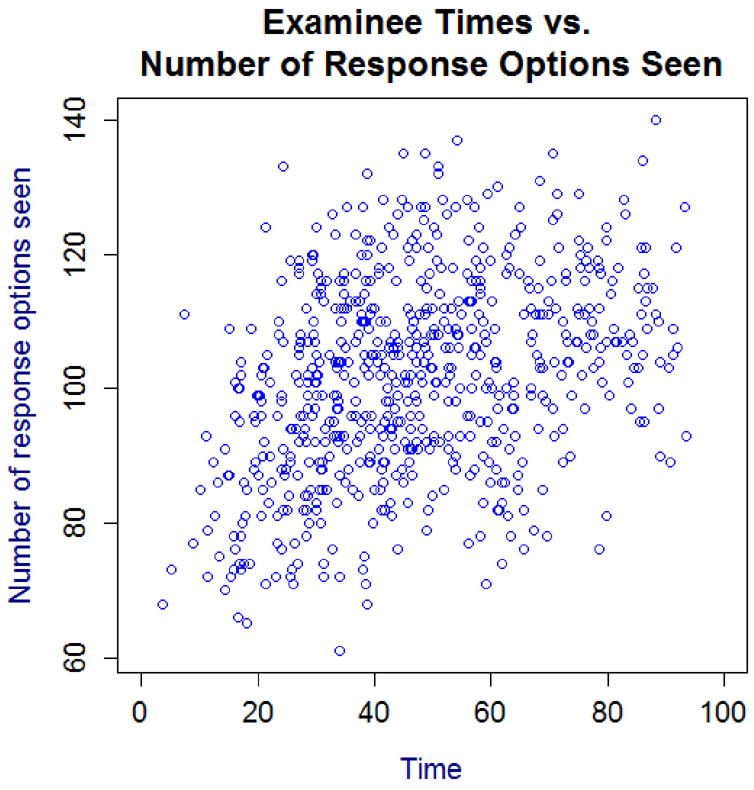


Figure 5



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, examinees 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, an examinee using true random guessing would answer the item correctly with a probability of 0.25. High ability examinees may be able to recognize the correct response right away, regardless of the attractiveness of the distractors. Examinees 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 examinee receives the response options. For example, if an examinee has no knowledge of the subject matter being assessed and is presented the correct option first, the examinee will answer the item correctly with a probability of 0.5 by using random guessing. However, if the examinee is administered the version of the item where the correct response would be presented last, the examinee would answer the item correctly with a probability of $0.50^4 = 0.0625$. Presumably, examinees 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. 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 examinees could be unfairly disadvantaged when items are administered in DOMC format.

To investigate the effects of differing response option orders on item statistics, we recoded each DOMC item which 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 examinee before the item was scored. For response option presentations in which the key is not shown in the first position, different permutations of distractors may be seen by examinees before the key is presented to them. However, we still based our recoding only on the key position for two reasons: 1) we hypothesized that key position would likely have a larger effect on any potential differences in item statistics than differences in distractors seen due to the differing role of guessing in obtaining a correct response for varying key positions, and 2) 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 examinee by 216 item response matrix for analysis purposes. We performed analysis to evaluate both classical and Rasch item statistics on the recoded items (Rasch, 1960). Because no examinee responded to more than one recoded item originating from the same DOMC item, the local independence assumption was not violated in this analysis.

Figure 6 shows the sample size for each of the recoded items used in the analysis. Sample sizes ranged from 65 to 185 for each of the 216 recoded items. The cluster of smaller sample sizes shown in the histogram represents recoded items which were on one form, and the cluster of larger sample sizes represents recoded items which were on both forms.

Figure 6

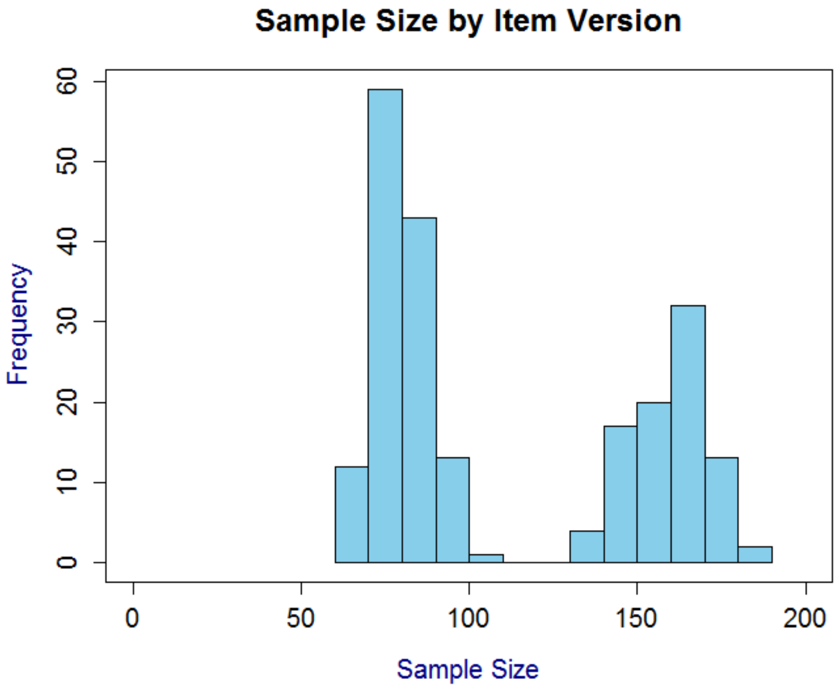


Figure 7 shows *p*-values for each of the recoded items grouped by key position. As key position increases, *p*-values generally decrease. Average *p*-values for recoded items with key position 1, 2, 3, and 4 were 0.64, 0.48, 0.35, and 0.29, respectively (shown in bold red on Figure 7). Similarly, Figure 8 shows estimated item difficulty parameters from the Rasch model for each of the recoded items grouped by key position. As key position increases, estimated item 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 8). These results are consistent with our hypothesis that the role of guessing in examinee responses differs depending on the key position, where guessing plays a larger role in the probability of answering the item correctly with lower key positions.

Figure 7

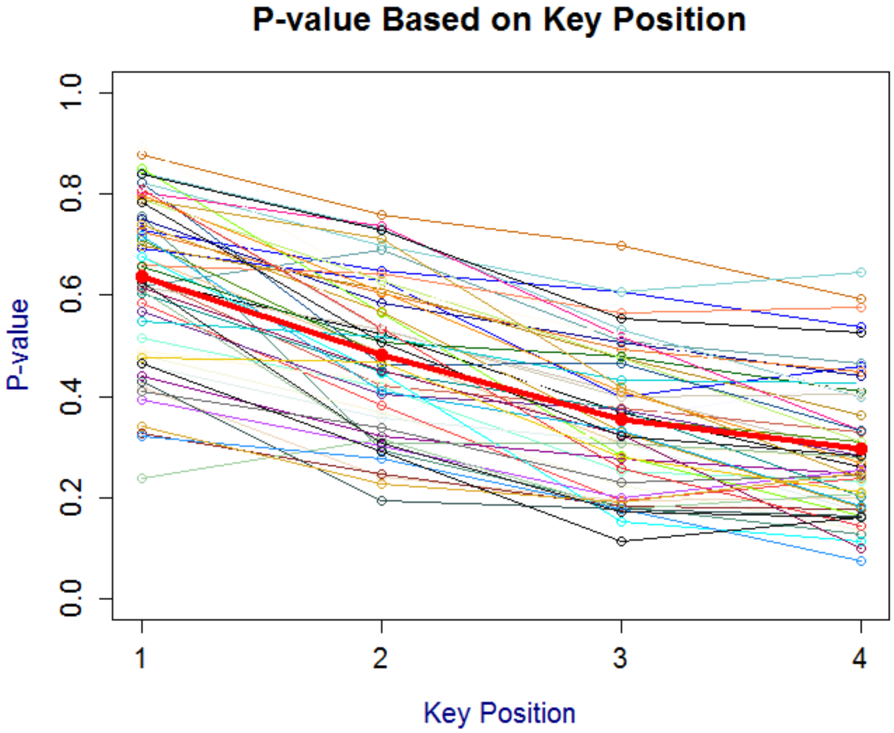
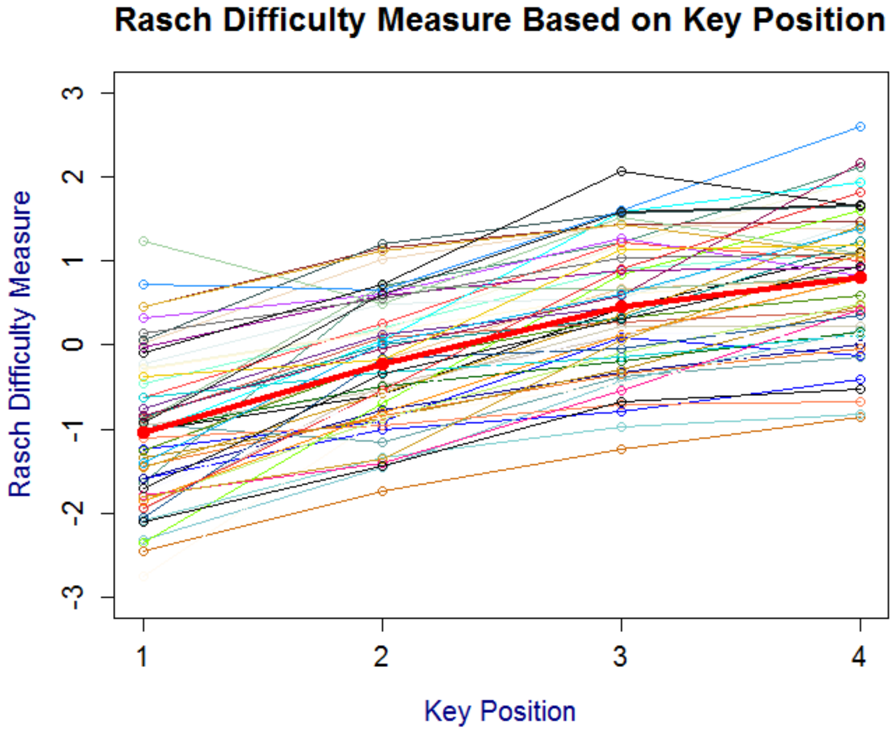


Figure 8



Because key position clearly has an influence on the difficulty of individual items, we also examined how these differences in item difficulty manifested themselves at the total test level for individual examinees. Figure 9 shows the distribution of average item difficulty for the complete subset of recoded items each examinee was administered. The different subsets of recoded items administered to examinees can be conceptualized as representing different forms, and for the purposes of this analysis, we will refer to these differing subsets as different forms. 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 examinee. To visualize how these differences in average form difficulty affect the test characteristic curves, Figure 10 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 measure (i.e., theta), the minimum and maximum example forms differ by six points at a Rasch measure of zero. Thus, if examinees of equal ability at a Rasch measure of zero were administered the hardest and easiest form of the exam, 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.

Figure 9

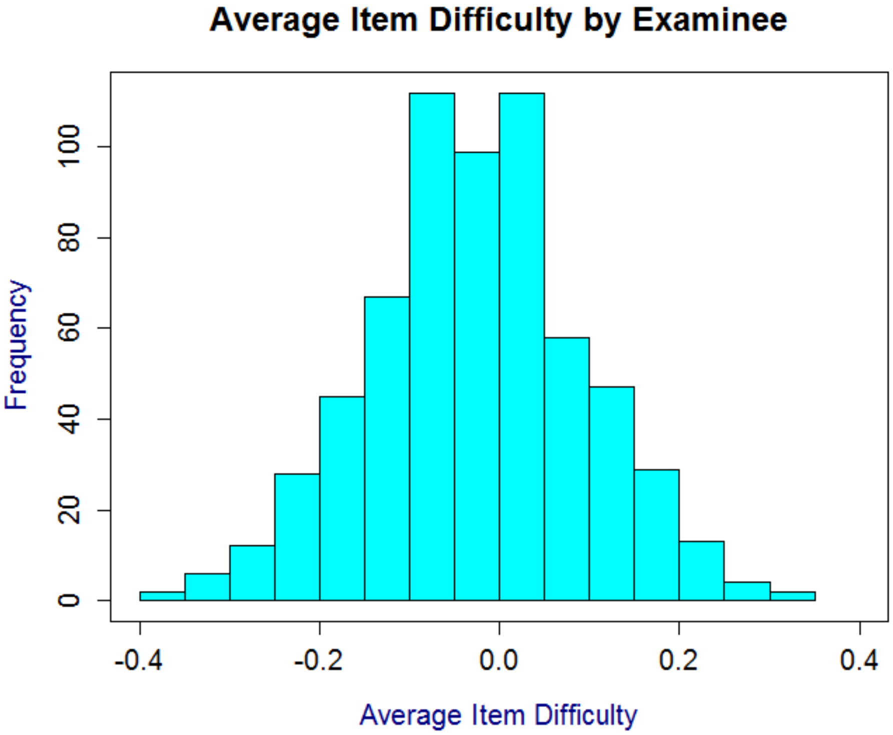
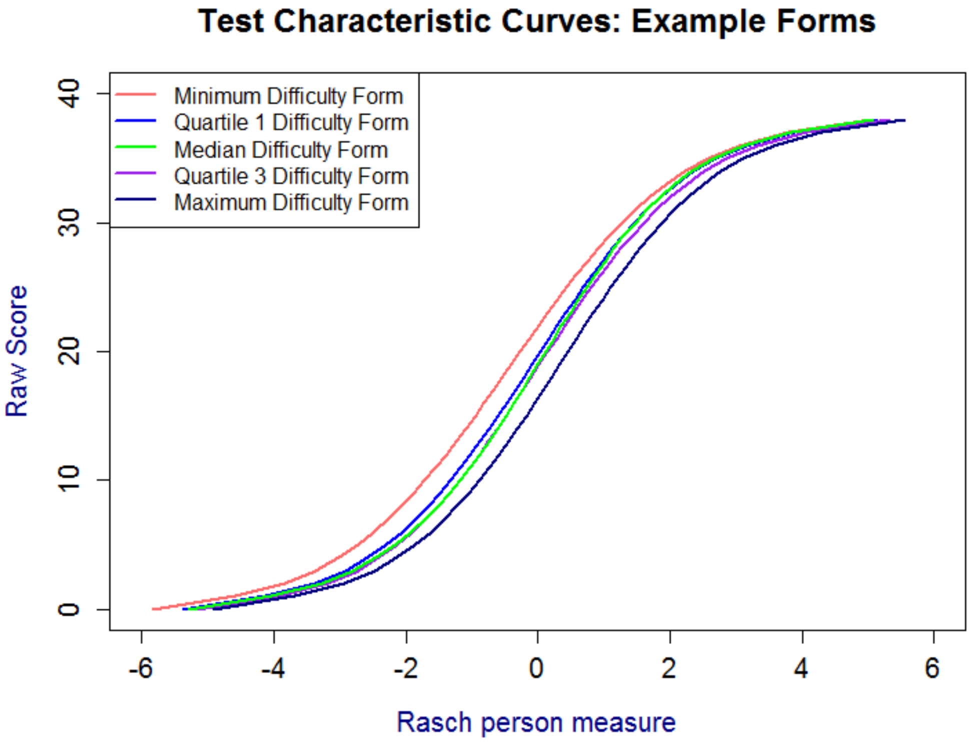
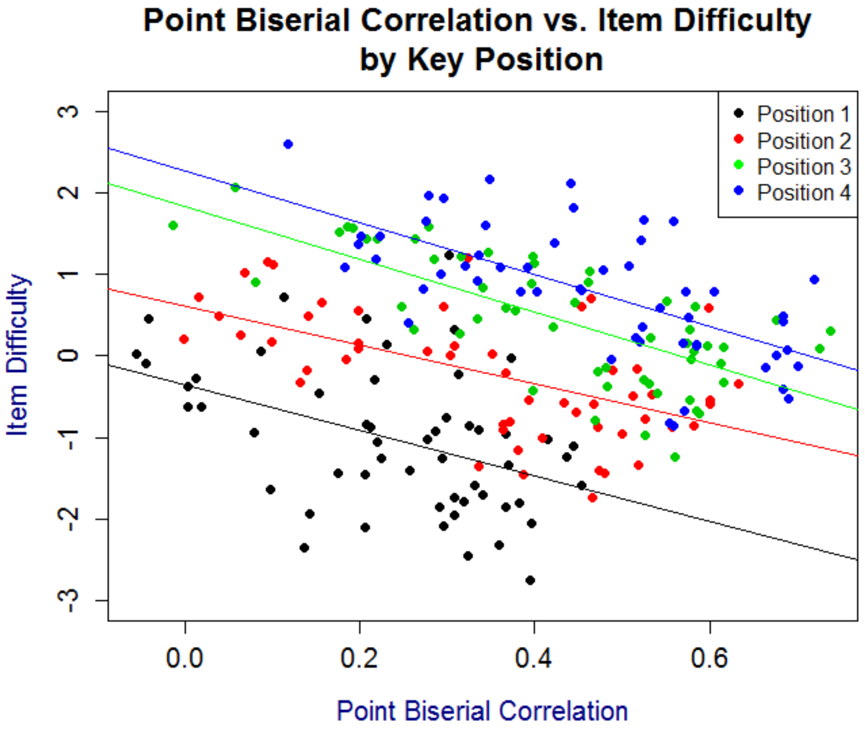


Figure 10



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 11 plots point-biserial correlations versus item difficulty for each of the items recoded based on key position. 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 indicates 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.

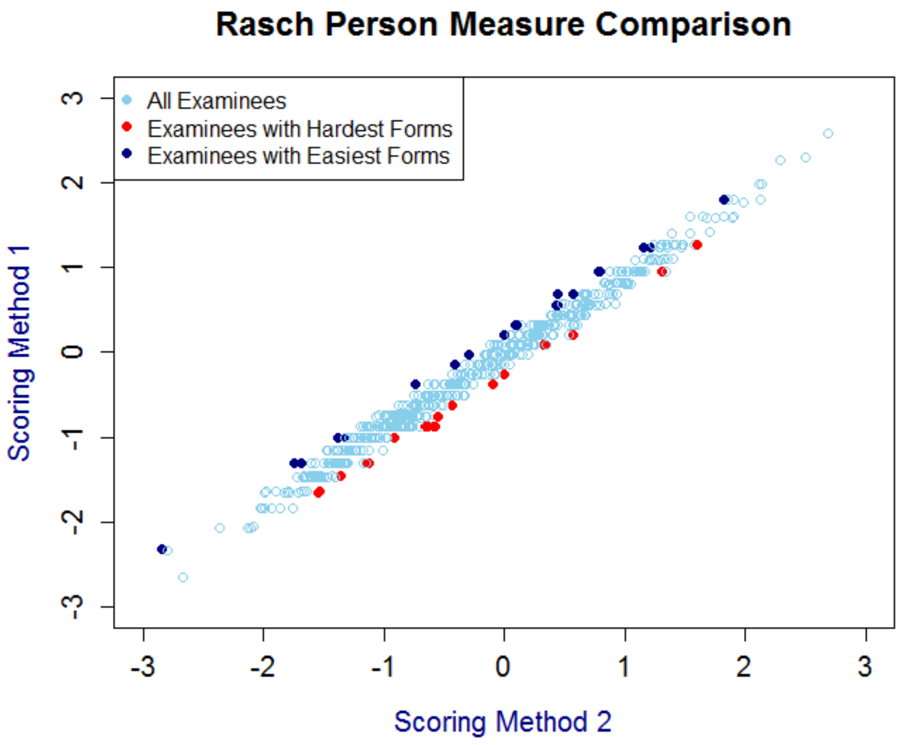
Figure 11



Cronbach's alpha is a common measure used to estimate reliability of exam scores, and can be calculated with the formula $\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_C^2}\right)$ 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 calculated 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 examinees would vary. Thus, we calculated 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 examinees. Cronbach's alpha values 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 exam-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 examinee measures. Figure 12 plots Rasch person measures (i.e., theta measures) using two different scoring methods. Scoring Method 1 estimates item difficulties for the 59 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 recording described above to estimate item difficulties for the four variations based on key position for each of the 59 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 examinees are advantaged or disadvantaged due to the particular combination of DOMC item variations they were administered. Examinees 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 examinees who were administered easier forms and underestimates the ability of examinees who were administered harder forms.

Figure 12



Conclusion

The DOMC item type was introduced to protect exam content from theft by presenting different subsets of response options to examinees, thereby creating unique testing experiences for each examinee where not all of the response options are revealed. The validity of exam scores depends heavily on exam content remaining secure, so efforts to reduce exposure and item theft can contribute to the overall health of an exam program. However, it remains necessary to ensure that the testing experience is fair to examinees. We have shown that item difficulty and discrimination varied substantially for the DOMC items in this dataset depending on key position, leading examinees 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. We recommend that testing programs choosing to administer exams in the DOMC item format take additional steps to control for differences in difficulty and discrimination due to the varying key positions in the items. One possible strategy could be including constraints for response presentation order in the DOMC algorithm, ensuring that examinees receive the same number of items with the key in each of the respective positions. This strategy would likely mitigate but not completely remove the differences in form difficulty for examinees. 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 scoring examinees based on Rasch person measures rather than raw scores. However, additional research should be conducted to determine whether this approach is sufficient to control for differences in difficulties between forms. Due to an insufficient sample size, we did not analyze whether differing distractor presentations affect item characteristics in addition to differing key positions. Lastly, it may be possible to model changes in item performance based on response presentation order and use those models to score examinees probabilistically, thereby mitigating fairness issues introduced by changing item difficulty based on key and distractor order.

References

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

Foster, D. F. & Miller, H. L., Jr. (2009). A new format for multiple-choice testing: Discreet option multiple-choice. Results from early studies. *Psychology Science Quarterly, 51*(4), 355-369.

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

Rasch, G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*. (Copenhagen: 1960) Chicago, IL: MESA Press, 1992