Does Item Format Affect Test Security?

Kylie N. Gorney and James A. Wollack

Department of Educational Psychology, University of Wisconsin-Madison

Author Note

Kylie N. Gorney https://orcid.org/0000-0002-8924-0726

Abstract

Unlike the traditional multiple-choice (MC) format, the discrete-option multiple-choice (DOMC) format does not necessarily reveal all answer options to an examinee. The purpose of this research was to determine whether the reduced exposure of item content affects test security. To do so, participants were provided with study guides containing item information prior to taking a test. The test contained secure and compromised items, and it was divided equally into DOMC and MC halves. Results showed nearly identical score gains for both item formats when items were compromised. However, the DOMC cheating ability estimates displayed less error overall, and they were more highly correlated with the true ability estimates. Thus, for this particular study, the DOMC format did offer a slight advantage in the presence of item preknowledge.

Keywords: discrete-option multiple-choice, item preknowledge, test security

Does Item Format Affect Test Security?

Item preknowledge occurs when a source reveals information about test items to future examinees. The items for which information has been leaked are referred to as *compromised*, while the remaining items are said to be *secure*. As a result of item preknowledge, it is expected that examinees will score higher on the compromised items than what would have otherwise been anticipated, thereby decreasing the validity of their test scores and perhaps the test scores of countless others (Belov, 2016; Wang, Liu, & Hambleton, 2017). Despite the severity of this threat, item preknowledge continues to remain at large because it is often difficult to pin down the source. Over the years, it has been shown that teachers, students, test preparation companies, and websites could all serve as the source, and the amount of information divulged has ranged from small hints to a complete exposure of the test and all of its items (Kyle, 2002; Wollack & Maynes, 2011).

Often, a scenario is observed where a previous test-taker has served as the source. In this case, it is possible that they had retained the information by memorization or by using a camera as a recording device (Murphy & Holme, 2015). The former strategy, though more intensive, is nearly impossible to monitor, and the latter is becoming increasingly difficult to detect, as well, given rapidly developing technology. Therefore, assuming that it is impossible to prevent this type of theft from occurring, it may be sensible to focus on the more realistic goal of reducing the amount of theft that occurs, instead.

Though item exposure is unavoidable, some item formats may be more susceptible to compromise than others. For instance, each time a multiple-choice (MC) item is administered, every answer option is displayed, and the item is able to be harvested in its entirety. Though an examinee may be unaware of the correct answer at the time they are taking the test, they (or future examinees) may be able to determine it later, with the help of the internet or other resources.

One way to reduce item exposure is by using the discrete-option multiple-choice (DOMC) item format instead (Foster & Miller, 2009). DOMC items are similar to MC

items in that they possess a stem and set number of answer options. Typically, one of these options serves as the key and is marked as the correct answer, though it is possible to have multiple keys for a single item. The main difference, then, between the two formats is the way in which the answer options are presented. Rather than displaying all answer options simultaneously, a DOMC item displays its options sequentially, and in a random order. After each option has been displayed, the examinee must decide whether they believe it to be correct or incorrect by responding *Yes* or *No*. Options continue to be randomly presented, one after another, until all have been exhausted, or the item has been scored. An item is scored as correct if the examinee endorses the correct option and refutes all prior incorrect options. An item is scored as incorrect if the examinee endorses an incorrect option or refutes the correct one. As a further layer of security, after an item has been scored, it is common to present examinees with an additional option at a prespecified probability (often 0.5) so that information regarding the correctness of the previous option is not inadvertently revealed. Many testing programs choose not to score this extra option, though whether or not it is scored does not affect the test security in either case.

Foster and Miller (2009) piloted the DOMC item format by including it in four assessments that were administered to 109 students. Each assessment was comprised of two item sets: one containing DOMC items, and the other containing MC items. All items were displayed in both formats to allow for comparisons at the item-level as well as at the test-level. In most cases, the authors found that items displayed in the DOMC format produced lower p-values than when they were displayed in the MC format. This difference was also reflected in the near 10% decrease that was observed in DOMC test scores, suggesting that the DOMC items were more difficult than their MC counterparts. In part, this may be due to the reduced opportunity that individuals had to use valuable test-taking skills (e.g., process of elimination) when answering the DOMC items. Furthermore, one might also consider the confusion that may have occurred since participants had just been presented with a new item format for the first time. The authors also noted that the

average number of options seen for each DOMC item ranged from 1.87 to 3.46 (out of 5 total options). For the MC items, this number was always 5. Thus, they concluded that in addition to producing items that were more difficult, the DOMC item format provided the added benefit of revealing fewer answer options, on average, to each examinee.

Eckerly, Smith, and Sowles (2018) considered the advantages and disadvantages of showing a varying number of answer options to each examinee. They examined data collected from a certification program in which 635 examinees were randomly assigned to take one of two DOMC test forms. As predicted, they found that there was a positive relationship between test scores and total number of answer options seen. They attributed this to the fact that lower ability examinees were more likely to answer the earlier options incorrectly, thus leading to fewer answer options being presented overall. Interestingly, those who viewed a higher number of answer options did not run into problems of test speededness as they approached the end of the test. However, the authors noted that this may not hold true for all implementations of the DOMC format, and testing programs should carefully consider the time required for longer versions of the test when setting a time limit. In addition, they considered the role that key position plays in the difficulty of a DOMC item. They found that items displaying keys later in the sequence of answer options were more difficult, on average. Thus, they concluded that when generating test scores, an examinee's average key position should be taken into account. Else, those who happened to receive earlier key positions, on average, would possess an unfair advantage.

In response, Bolt, Kim, Wollack, Pan, Eckerly, and Sowles (2020) proposed the key location model for DOMC items. This model extends the two-parameter item response theory (IRT) model by adding three parameters: (1) the effect of key position across items for a particular examinee, (2) the effect of key position for a particular item across examinees, and (3) the (last) key position for a particular examinee and item. The authors found that these parameters were able to significantly explain the variability of the difference in difficulties between DOMC and MC items. Furthermore, they suggested the

use of a constrained randomization key schedule as a way of balancing the distribution of key positions across examinees. By doing so, they were able to greatly reduce the score variability that was related to key position effects.

Tiemann, Miller, Kingston, and Foster (2014) considered a different aspect of test security by analyzing the results of two types of simulated cheating. Their study was not the only of its kind to simulate preknowledge (e.g., Toton & Maynes, 2019); however, it was the first to do so for DOMC items. For the first type of cheating, *before* taking the test, source examinees were instructed to try to remember as much of the test content as possible, thereby simulating an examinee intent on using memorization to harvest items. In the second case, source examinees did not enter the test with the intent of memorizing the content. Rather, they were asked to recall information *after* having taken the test. This condition was designed to simulate an examinee who is paid to brain dump, or discuss their experience, following a test. Once the two groups had completed the test, they were told to prepare study guides containing as much item information as they could remember. New test-takers were then brought in and were randomly assigned to receive one of the written study guides. They were allowed to study for 30 minutes before starting the test, at which point they were asked to return the study guides. Results showed that new test-takers earned lower scores on the DOMC portion of the test, regardless of the study guide they received. That is, the same pattern that was observed in Foster and Miller (2009) prevailed even when the items were compromised. However, it should be noted that this study did not account for the difference in item format difficulties, more generally. Thus, one could ask whether the observed score differences were due to (1) the difference in item format difficulties, (2) the increased security offered by the DOMC format and its reduced exposure of answer options (which is only applicable to compromised items), or (3) some combination of both ideas.

Soon after, Willing, Ostapczuk, and Musch (2015) examined the difference in item format difficulties more closely by considering the effects of testwiseness. To do this, they

administered two sets of items: one containing cues that would increase the probability of a correct guess for a testwise test-taker, and one that did not contain such cues (see Millman, Bishop, & Ebel, 1965). All items were administered in the DOMC and MC formats, and the effects of item format were then compared. As expected, those who were administered DOMC items were less successful at using the provided cues than those who were administered MC items. The authors conducted this experiment on three separate groups of participants (having sample sizes of 48, 86, and 106), and they found this effect to be significant for each group. However, contrary to expectations, the authors did not find a significant difference in difficulty for items that did not contain cues. That is, participants were just as likely to answer an item correctly if it was presented in DOMC format as they were if it was presented in MC format as long as it did not contain any cues. The results of this research suggest that testwiseness may affect how the difficulties of the two formats should be interpreted. Thus, when comparing the effects of item format on item difficulty, the items being administered should first be examined to see whether they contain any testwiseness cues.

The present study compares the effects of the DOMC and MC item formats on compromised items. Because item compromise and item format are the primary variables of interest, all items were written without testwiseness cues, and the difference in item format difficulties is addressed by the use of anchor items. Hence, any conclusions may be attributed to the compromise status and format of each item. In addition, this study differs from those that preceded it by the way in which cheating is simulated. While Tiemann et al. (2014) relied purely on the source examinees' memories to create the study guides, this study assumes that the source examinees were wearing cameras as they took the test. Thus, there were no mistakes in the study guides that were produced, and entire items were able to be captured without error. Ultimately, the purpose of this study was to determine whether participants benefitted more from item preknowledge when DOMC items were administered versus when MC items were administered.

Method

Participants

The sample consisted of 150 University of Wisconsin-Madison (UW-Madison) students who were enrolled in at least one undergraduate human development course taught by the Department of Educational Psychology in the Spring 2020 or Fall 2020 semesters. In exchange for their participation, students were compensated with two research credits that could be used to satisfy a course requirement. In addition, all participants who followed the instructions and earned a test score in the top 50% of all test-takers in their given semester received a $40 cash prize. This incentive was designed to motivate them so that their performance could be captured under more realistic circumstances.

Design and Materials

In order to measure participants' understanding of human development, a 68-item test was created. Item content was intended to reflect material covered in all four human development courses from which students were recruited. Based on the courses' syllabi and assigned readings, a blueprint was designed which contained five themes that were common to each course (Table A1). Items were written in accordance with each theme such that the more frequently mentioned themes received more items. Moreover, items were phrased in a way so that they could be easily converted from the MC format to the DOMC format without any additional editing.

Each item was comprised of a stem and five answer options, only one of which was correct and was marked as the key. In addition to belonging to one of the five content themes, all items were grouped into one of six item sets, as well (Table A1). Sets 1 and 2, comprised of 10 items each, contained the anchor items. These items were always secure (i.e., no compromise was simulated), and all participants received them in the same format, regardless of the test form to which they were assigned. Specifically, items belonging to Set

1 were always displayed in the MC format, while items belonging to Set 2 were always displayed in the DOMC format.

Sets 3–6 contained 12 items each. These sets had the possibility of appearing in either the DOMC or MC formats, and they may or may not have been compromised, depending on the particular test form that was administered (Table A2). This enabled direct comparisons to be made at the item level, since each item was delivered under four different conditions. More specifically, this allowed conclusions to be drawn regarding how item scores and testing behaviors were affected by both item compromise and item format, which was the main purpose of this study.

In all forms, items on the first half of the test were not mixed with items on the second half so as not to confuse participants with alternating item formats. However, within each half, the items appeared in a random order, and each item's answer options were displayed in a random order, as well. This applied to both the DOMC and the MC test halves. Thus, it is unlikely that any two participants would have viewed the test in the exact same way, though they may have been assigned to receive the same form.

In order to provide participants with item preknowledge, each participant was supplied with one of 20 study guides prior to test administration. Each study guide contained information pertaining to two of the six administered item sets (i.e., the compromised item sets), and the specific information that was included was determined by the source examinees who created the study guide. All sources were students who took this test in Spring 2019. Half of the sources experienced the test entirely in the DOMC format, while the other half experienced it entirely in the MC format. Study guides were then created that contained screenshots of each item exactly as it was displayed to a particular source. As a result, the MC sources were able to capture complete items and all of their answer options, while the DOMC sources were only able to capture the item stems and the answer options that were presented to them. Therefore, some items that were captured by

the DOMC sources appeared on the study guides *without* the key as one of the listed answer options, simply because it had not been revealed to them.

The study guide for any given participant included information that was captured by one DOMC source and one MC source. For example, consider a participant assigned to Group A1 where Sets 3 and 4 were compromised. This participant received a study guide where an MC source had leaked information for items belonging to Set 3, and a DOMC source had leaked information for items in Set 4. They did not receive any information regarding the items in Sets 1, 2, 5, or 6 since these items were secure.

The test was delivered using the Secure Exam Interface by Caveon. This system was chosen due to its ability to handle both the DOMC and MC item formats. Furthermore, this was the system that had been used in Spring 2019 to administer the test to the sources.

Procedure

At the start of each semester, students had the opportunity to participate in exchange for two research credits. Because this study took place entirely online, participants were able to choose a time and location that were convenient for them. The only restrictions were that participants could not collaborate with one another, and no student could participate more than once.

After reading the instructions, participants were presented with one of the 20 study guides. They were instructed to review their assigned study guide for 50–60 minutes and use whatever means necessary (e.g., textbooks, the internet) to prepare for the upcoming test. They were informed that the amount of time they spent viewing the study guide would be monitored, and if they fell outside of the 50–60 minute range, they would not be eligible to receive one of the $40 cash prizes.

When the allotted time had passed, participants were informed that the use of any outside resources beyond this point would be considered a form of cheating and was not permitted. They were then given up to 70 minutes to complete the test. If more than 70

minutes had passed and a participant had not finished, then they were routed to the end of the test and were not given the opportunity to view or answer any of the remaining items.

Results

For each test form that was administered, there existed an opposite form in which the two test halves were presented in reverse-order (e.g., Forms A and C). To determine whether there was an order effect, multivariate analyses of variance (MANOVAs) were conducted on each of the four pairs of test forms: A1 and C1, A2 and C2, B1 and D1, and B2 and D2. For each comparison, test form served as the independent variable, and the raw scores obtained on Sets 1–6 served as the six dependent variables (Table 1). Raw scores, rather than IRT ability estimates, were used due to the small sample size.

Table 1

Test Statistics by Test Form

Test Form	*N*	Set 1 (10 pts)	Set 2 (10 pts)	Set 3 (12 pts)	Set 4 (12 pts)	Set 5 (12 pts)	Set 6 (12 pts)
A1	22	5.9	4.8	10.2	7.7	8.2	4.6
C1	19	5.7	4.8	8.7	7.8	7.5	4.7
A2	21	6.2	4.7	8.3	6.1	9.6	6.9
C2	15	6.1	4.5	9.1	6.7	10.3	7.8
B1	16	6.2	5.3	9.0	10.1	5.8	7.4
D1	20	6.3	5.4	8.5	10.7	5.4	7.6
B2	22	6.0	4.8	6.7	8.1	7.2	8.2
D2	15	6.1	5.8	7.3	8.3	6.8	8.8

Results indicated that Form A1 scores did not significantly differ from C1 scores, Wilks's $\Lambda = .78, F(6, 34) = 1.59, p = .18$, nor did Form A2 scores significantly differ from C2 scores, Wilks's $\Lambda = .87, F(6, 29) = 0.75, p = .61$. Likewise, Form B1 scores did not significantly differ from D1 scores, Wilks's $\Lambda = .89, F(6, 29) = 0.59, p = .73$, nor did Form B2 scores significantly differ from D2 scores, Wilks's $\Lambda = .86, F(6, 30) = 0.84, p = .55$. In other words, the order in which the test halves were presented did not significantly affect the item set scores.

Classical Statistics

Item Statistics

For each item, the p-value (i.e., average score), point-biserial correlation, and average response time (RT) were computed. Specifically, for items in Sets 1 and 2, these statistics were computed across all examinees, since each item was delivered securely and in the same format, regardless of the test form and group to which one was assigned. However, for items in Sets 3–6, these statistics were computed four times, since each item was administered under four separate conditions (secure DOMC, secure MC, compromised DOMC, compromised MC). Summary statistics for each condition are provided in Table 2, and the corresponding plots may be viewed in Figures 1 and 2.

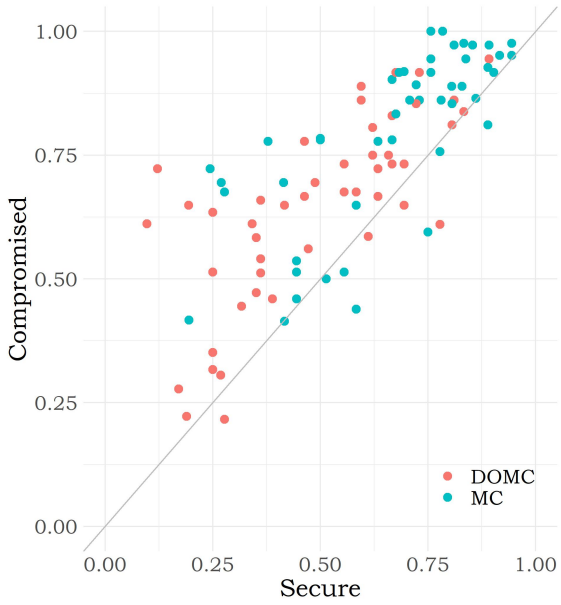
Table 2

Item Statistics

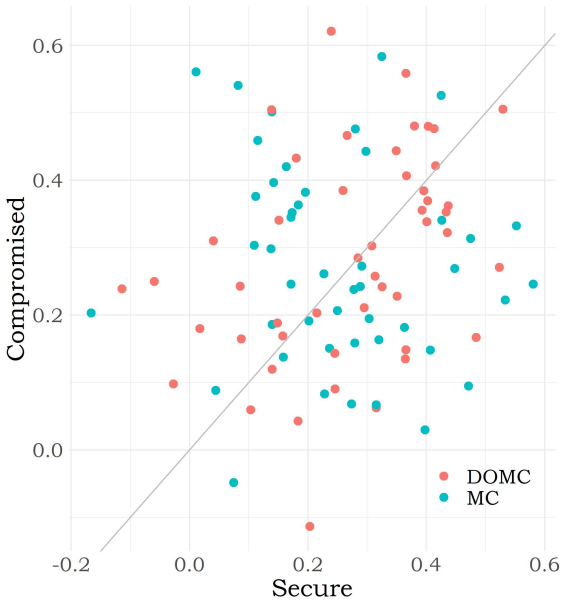
Condition	Number of Items	Item *p*-value	Item PB Correlation	Item RT (in Seconds)
Secure DOMC (Anchor)	10	0.50 (0.08)	0.26 (0.10)	19.55 (5.23)
Secure MC (Anchor)	10	0.61 (0.18)	0.29 (0.10)	24.51 (5.35)
Secure DOMC (Non-Anchor)	48	0.49 (0.21)	0.27 (0.15)	19.57 (4.80)
Secure MC (Non-Anchor)	48	0.67 (0.20)	0.25 (0.15)	24.07 (7.47)
Compromised DOMC	48	0.64 (0.19)	0.29 (0.15)	15.92 (3.43)
Compromised MC	48	0.80 (0.18)	0.27 (0.15)	12.70 (3.77)

Figure 1

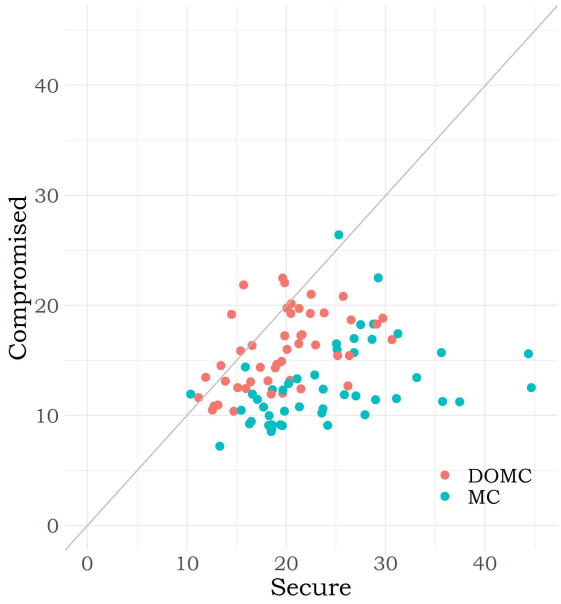
Item Statistics by Item Compromise Status



(a) *Item p-value*



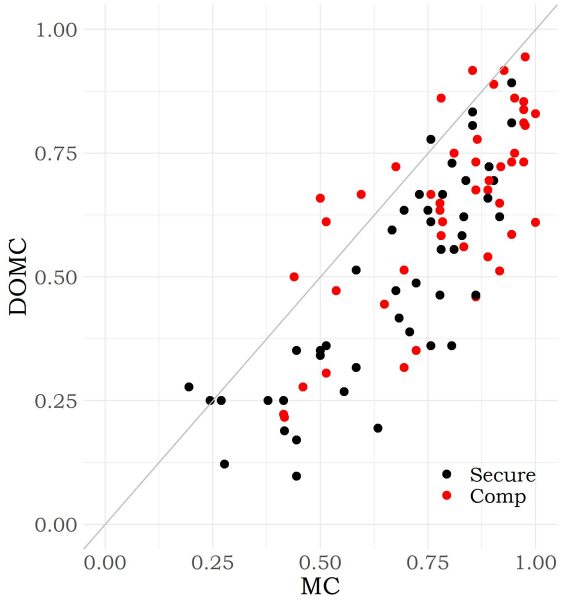
(b) *Item Point-Biserial Correlation*



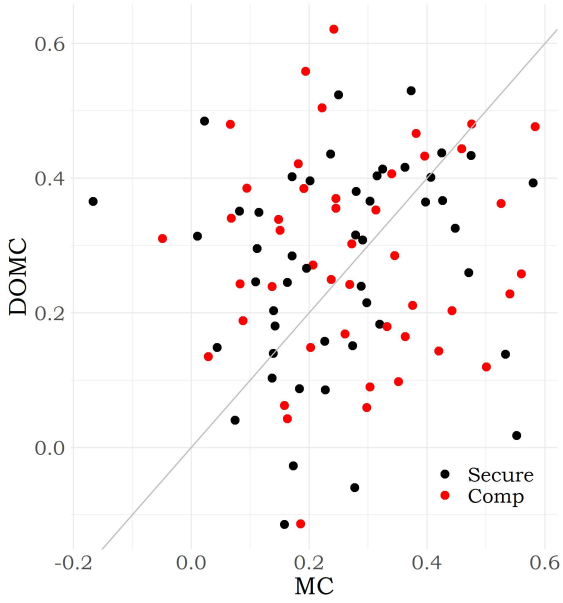
(c) *Item RT (in Seconds)*

Figure 2

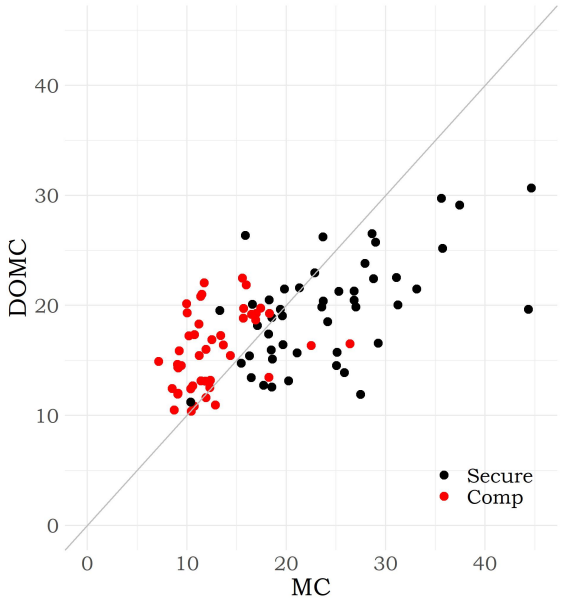
Item Statistics by Item Format



(a) *Item p-value*



(b) *Item Point-Biserial Correlation*



(c) *Item RT (in Seconds)*

Secure items tended to be more difficult than their compromised counterparts, and the difference in difficulties was similar for both the DOMC and MC formats. For the DOMC items, the secure p-values were .15 lower than the compromised p-values, on average, and the correlation between the secure and compromised p-values was .73. Meanwhile, for the MC items, the secure p-values were .13 lower than the compromised p-values, on average, and the correlation between the two sets of values was .74. The DOMC version of an item was almost always more difficult than the MC version, and this held true when the items were both secure and compromised. Specifically, for secure items, the DOMC p-values were .18 lower than the MC p-values, on average. Meanwhile, for compromised items, the DOMC p-values were .16 lower than the MC p-values, on average.

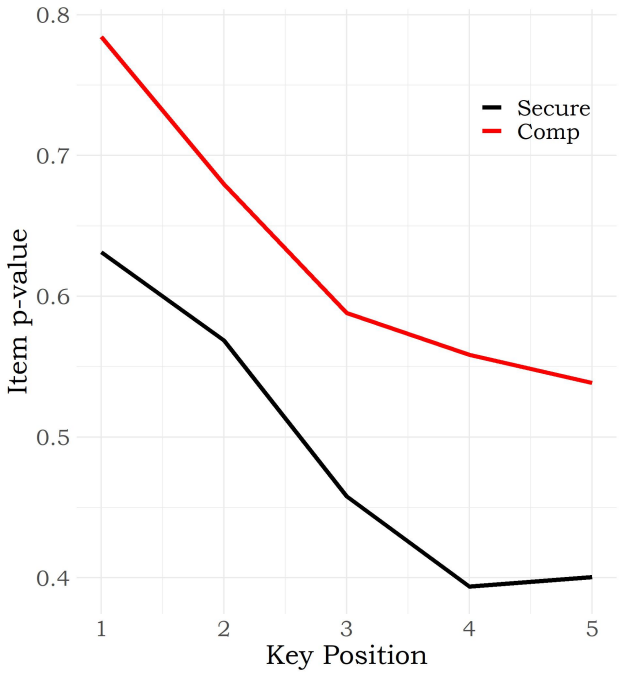
Across all four conditions, items displayed similar average point-biserial correlations. For the DOMC items in particular, the secure point-biserial correlations were .02 lower than the compromised point-biserial correlations, on average, and the correlation between the two sets of values was .37. Meanwhile, for the MC items, the secure point-biserial correlations were also .02 lower than the compromised point-biserial correlations, on average, though the correlation between them was considerably lower at $-.12$.

Secure RTs tended to be longer than compromised RTs, though this effect was less pronounced for items administered in the DOMC format. Specifically, within the DOMC item format, compromised items were answered 19% faster than secure items, and the correlation between the secure and compromised RTs was .48. Meanwhile, for the MC item format, the compromised items were answered 47% faster than the secure items, and the correlation between the two sets of RTs was .40. Interestingly, the item format which required more time varied depending on whether or not the item was compromised. For secure items, those that were administered in the DOMC format were answered 19% faster than those administered in the MC format, on average. However, for compromised items, the opposite was true: items in the MC format were answered 20% faster than items in the DOMC format, on average.

One primary advantage of the DOMC format is that responses may be examined at the option level. In particular, it may be of interest to consider the order in which the options are presented. Though MC items allow answer options to be displayed in a random order, they are, in fact, presented simultaneously. DOMC items, on the other hand, present options in a sequential order, thereby allowing two examinees to have vastly different experiences when answering the same item. For instance, if the key is displayed in the first position, an examinee would only be required to answer one option correctly to receive credit for the item. However, if the key is displayed in the fifth position, an examinee would need to answer all five options correctly before receiving credit for the item. Figure 3 displays the average score for DOMC items having each of the five key positions.

Figure 3

Item Statistics by Key Position (DOMC Items)



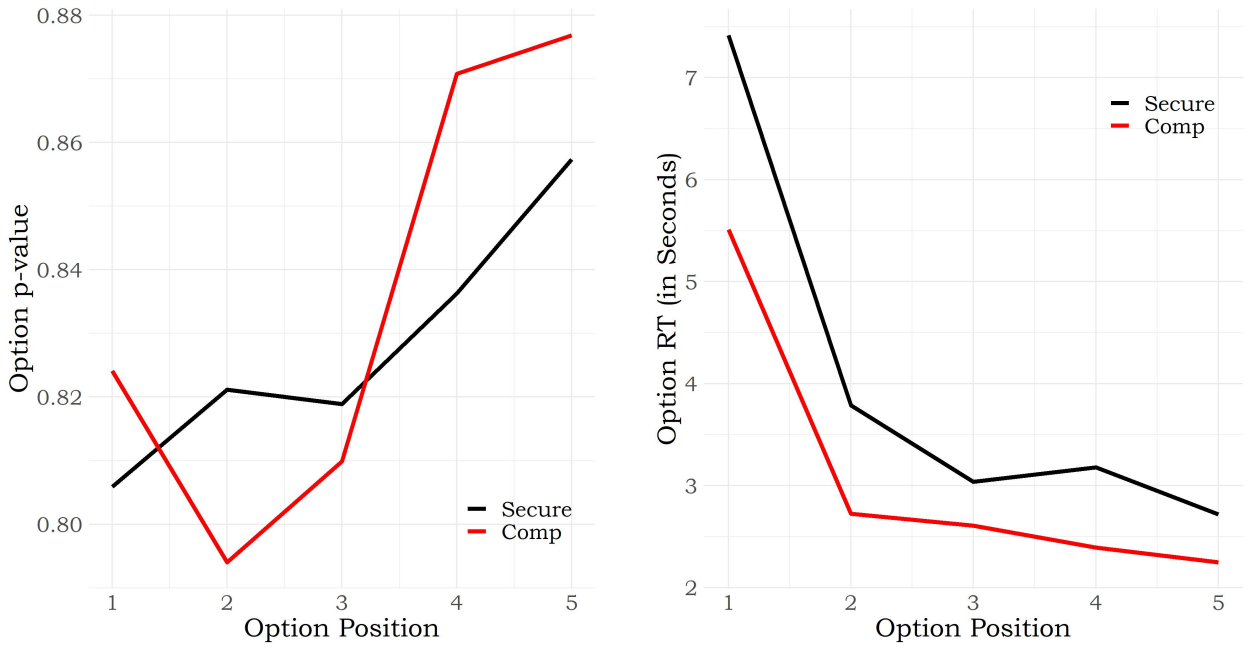
For both secure and compromised items, those with later key positions tended to be more difficult than those with earlier key positions. This effect was most noticeable when comparing items having key positions 1 and 2 or key positions 2 and 3. In both cases, the item *p*-values decreased by .09, on average, regardless of whether the items themselves were secure or compromised. However, whether the key was presented in position 4 or 5 seemed to have little to no impact on item score, suggesting that the effect of key position levels off after a certain point.

Option Statistics

As previously mentioned, each item was written with five options, and options were presented in a random order to each examinee. Thus, each option had the potential to assume one of five positions. For DOMC items in particular, two questions to consider are whether option position affects option score or option RT.

Figure 4

Option Statistics by Option Position (DOMC Items)



(a) *Option p-value*

(b) *Option RT*

Generally, the options administered in later positions were slightly easier than the options administered in earlier positions (Figure 4). This effect was similar for both secure and compromised options. Specifically, secure options administered in position 5 had an average p-value that was .05 higher than the average p-value for the options that were administered in position 1. Likewise, the compromised options administered in position 5 also had an average p-value that was .05 higher than the one observed for the options administered in position 1.

When considering option RT, the effect of option position was even more noticeable. The options presented in position 1 required the most time, while all subsequent options required considerably less time. Again, this effect was similar for both secure and compromised options. Generally, the options that were presented in position 2 were answered 50% faster than those that were presented in position 1. However, there was little difference between RTs for the options that were presented in positions 2 through 5, suggesting that this effect may also level off after a certain point.

Another question worth asking regarding DOMC items is whether option compromise (rather than item compromise) affects option score or option RT. In part, the answer depends on whether the option itself was a distractor or the key (Table 3).

Table 3

Option Statistics (DOMC Items)

Compromise Status	Key	Option p-value	Option RT (in Seconds)
Secure	No	0.83 (0.16)	5.13 (2.14)
Comp	No	0.81 (0.21)	3.75 (2.39)
Secure	Yes	0.70 (0.22)	4.84 (1.58)
Comp	Yes	0.86 (0.15)	3.18 (1.54)

Secure and compromised distractors were similar in difficulty, with secure distractors having an average p-value of .83 and compromised distractors having an average p-value of .81. Despite this similarity, a noticeable difference emerged with respect to RT. Compromised distractors were answered 27% faster than their secure counterparts, suggesting that participants did recognize the options, even though this additional knowledge may not have affected their responses. In contrast, option responses did seem to be affected when it was the key that had been revealed. More specifically, secure keys had an average p-value of .70 while compromised keys had an average p-value of .86, indicating a higher rate of endorsement and a noticeable improvement in performance. Additionally, compromised keys were answered 34% faster than secure keys, this being an even greater reduction in RT than had been observed for the compromised distractors. In general, these results could suggest that when item information was revealed, participants tended to focus more on memorizing the keys than the distractors.

Item Response Theory

In addition to the classical statistics, item and ability parameter estimates were generated using the Rasch model of item response theory (IRT). This model was chosen in part due to its simplicity and its small sample size requirement. Under the Rasch model, the probability of a correct response to item i for examinee j may be written as

$$\mathbb{P}(X_{ji} = 1|\theta_j) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)} \tag{1}$$

where θ_j represents the ability of examinee j, and b_i represents the difficulty of item i.

One assumption of this model is that of unidimensionality. In other words, there exists a single, latent trait that is collectively being measured by all of the items on the test. Furthermore, an examinee's location on this latent trait (i.e., their ability) is the sole indicator of their performance on each item, and no amount of outside information should affect their response. However, when examinees possess any amount of item preknowledge, this assumption no longer holds. One might say that there now exist two latent traits that

are responsible in determining a person's performance: their "true" ability and their "cheating" ability. Examinees are assumed to rely on their true ability when answering secure items and their cheating ability when answering compromised items. Because an item can never be both secure and compromised, an examinee will only rely on one of these two abilities when answering a given item.

In order to obtain parameter estimates, the initial calibration included only responses to the secure items. Since these were the items that had been captured under ideal circumstances, this allowed for the estimation of uncontaminated item parameter estimates. Moreover, items that were displayed in both the DOMC and MC formats (Sets 3–6) received two sets of parameter estimates: one for each format. Also, it should be noted that the inclusion of the 20 anchor items (Sets 1 and 2) enabled all item parameter estimates to exist on a common scale, regardless of item format.

Each participant received one true ability estimate that was based only on the secure items. To compute the cheating ability estimates, the uncontaminated item parameter estimates were fixed. Compromised items were then entered into the model, replacing their secure counterparts. Each participant received two cheating ability estimates: one based only on compromised DOMC items, and the other based only on compromised MC items. Thus, all participants received three ability estimates in total.

To compare cheating ability to true ability for both item formats, three criteria were assessed: the bias, the root mean squared difference (RMSD), and the correlation between estimates. Bias measures whether the cheating ability estimates tend to over- or underestimate the true ability estimates and is computed as the average difference across examinees. Thus, bias may be written as

$$\text{Bias} = \frac{1}{J}\sum_{j=1}^{J}(\hat{\theta}_j^{(c)} - \hat{\theta}_j^{(t)}) \tag{2}$$

where J is the total number of examinees, and $\hat{\theta}_j^{(c)}$ and $\hat{\theta}_j^{(t)}$ denote the cheating and true ability estimates, respectively, of examinee j. In contrast, the RMSD is concerned with the absolute difference that exists between the cheating and true ability estimates, meaning

that the direction of error is less important than the fact that the error exists at all.

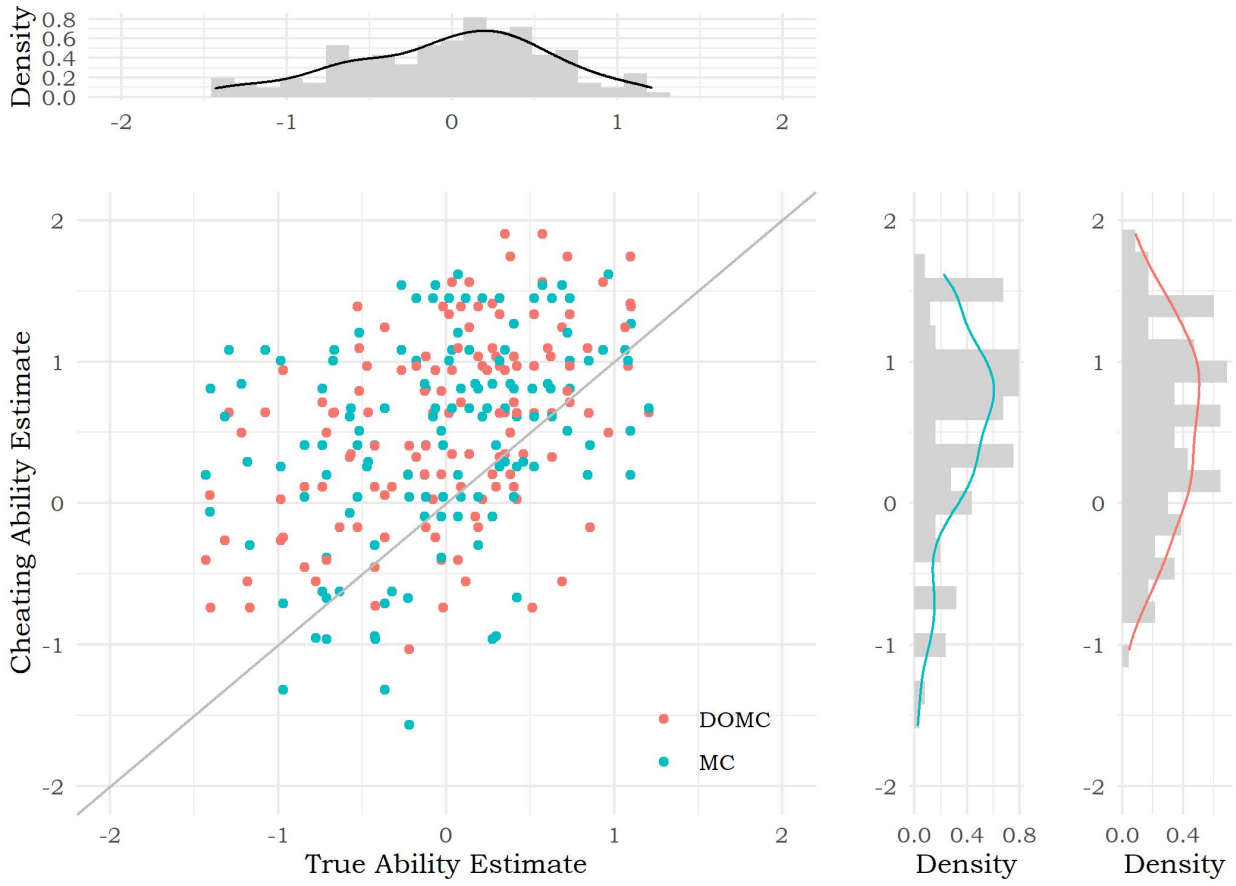
$$\text{RMSD} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\theta}_j^{(c)} - \hat{\theta}_j^{(t)})^2} \tag{3}$$

The third and final criterion was the correlation between the cheating and true ability estimates. Theoretically, the item format that is more secure would produce bias and RMSDs closer to zero, and positive correlations that are larger in magnitude. This would indicate that participants benefitted less from item preknowledge when that particular format was administered.

The distributions of the ability estimates may be viewed in Figure 5. For most participants, the cheating ability estimates were higher than the true ability estimates, and this was true for both the DOMC and MC item formats. In particular, the DOMC cheating ability estimates yielded an upward bias of 0.55, while the MC cheating ability estimates yielded an upward bias of 0.53. Moreover, the DOMC cheating ability estimates produced an RMSD of 0.72 and a correlation of 0.47. In contrast, the MC cheating ability estimates produced an RMSD of 0.84 and a correlation of 0.36. In summary, the two item formats yielded results that showed similar amounts of bias, but the DOMC format produced less error overall.

Figure 5

True and Cheating Ability Estimates



Discussion

Each test form was comprised of two test halves: a DOMC half and an MC half. The order in which the halves were presented did not appear to affect item scores, thereby allowing for a simple and straightforward interpretation of the item scores.

Classical Statistics

Item Statistics

It is no surprise that secure items were generally more difficult than compromised items. However, it is interesting to note that the change in difficulty was very similar for both the DOMC and MC formats. This suggests that participants benefitted equally from

item preknowledge, regardless of the item format that was administered. An additional finding was that the secure and compromised item p-values were closely related within both item formats. In other words, an easy secure item tended to be an even easier compromised item, while a difficult secure item tended to be a slightly less difficult compromised item. Furthermore, the DOMC version of an item was almost always more difficult than its MC counterpart. This result was expected as it agrees with previous research that has been conducted on both secure (e.g., Eckerly et al., 2018; Foster & Miller, 2009; Kingston, Tiemann, Miller, & Foster, 2012) and compromised items (Tiemann et al., 2014).

Across all conditions, the average point-biserial correlations were very similar. In fact, for both the DOMC and MC item formats, the average differences between the secure and compromised point-biserial correlations were nearly identical, suggesting that item format had little to no effect on this statistic. Kingston et al. (2012) drew similar conclusions when comparing secure DOMC items to secure MC items, though they did not consider the case of compromised items. Additionally, it should be noted that for the DOMC items specifically, the secure point-biserial correlations were positively correlated with the compromised point-biserial correlations. This suggests that the secure and compromised DOMC items behaved similarly in their measurement of the underlying construct. However, for the MC items, the secure point-biserial correlations were negatively correlated with the compromised point-biserial correlations, suggesting that the secure and compromised MC items may actually have been measuring somewhat different constructs.

For both item formats, the secure and compromised RTs were positively correlated. That is, the amount of time required to answer a secure item was somewhat indicative of the amount of time required to answer the same item when it was compromised. In addition, the secure items tended to require more time to answer than the compromised items, though precisely how much more time was needed depended on the item format. Specifically, the difference in RTs tended to be larger for MC items than for DOMC items. Therefore, a compromised MC item saw a greater reduction in RT than a compromised

DOMC item. One possible explanation could be that DOMC items require an examinee to read and respond to each answer option that appears on the screen in front of them. Thus, some cognitive energy must be devoted to each of the presented options. In contrast, when a compromised MC item is presented, an examinee need only search for what they know to be the correct answer. And, assuming that they know only one answer option is correct, they do not need to consider any alternatives beyond this point, thus resulting in a shorter RT. One implication of this result is that item preknowledge may actually be easier to detect when the MC format is used as opposed to the DOMC format if RTs are considered.

For DOMC items in particular, the effect of key position was examined. Generally, the DOMC items with later key positions were more difficult than the DOMC items with earlier key positions. This relationship held for both secure and compromised items, though in both cases, the effect appeared to level off after a certain point. Specifically, whether the key was displayed in position 4 or 5 seemed to make little difference, as items having both key positions were found to be similar in difficulty. To a certain extent, these results parallel those that were found by Eckerly et al. (2018). Though they specifically considered secure DOMC items, they also found that the effect of key position weakened as the key position itself increased. The reason for this could be that those of lower ability would likely have been eliminated earlier in the sequence of answer options. Thus, they would not have been given the chance to see or answer any of the later options. Those who did see the later options were likely of higher ability, and presumably, key position would have had less of an impact on their response.

Option Statistics

For this particular test, each option had the potential to assume one of five positions. For DOMC items specifically, it was wondered whether the position in which an option was presented would affect the option score or option RT. Ultimately, for both secure and compromised DOMC items, option position appeared to have a positive effect

on option score. That is, participants were more likely to answer options in later positions correctly than they were to answer options in earlier positions correctly. It seems reasonable to assume that the explanation used above would apply here, as well. Consider that the earlier options would have been answered by those of lower ability *and* those of higher ability. In contrast, it is likely that the later answer options would have only been seen by those of higher ability. As a result, options presented in the later positions would appear to be less difficult than those that were presented in the earlier positions.

Additionally, it seems as though option position affected option RT. Specifically, participants tended to spend the most time viewing the option that was presented in position 1, whereas all subsequent options were answered in considerably less time. A similar pattern was observed for both secure and compromised items, perhaps suggesting some form of item familiarity. In other words, the options that were administered in the later positions may have been answered more quickly because the participant had more time to consider the item and all it entailed. Therefore, not as much time was required to determine whether the option itself was correct or incorrect.

In addition to examining option position, the compromise status of each option was also considered more generally. It is interesting to note that the effect of option compromise varied depending on whether the option was a distractor or the key. When a distractor was compromised, the option score was relatively unaffected. That is, participants answered similarly to how they would have had the option not been revealed. Yet, a noticeable difference emerged with respect to the option RT. Specifically, the RT was considerably faster if the distractor had been compromised, suggesting that participants did recall having seen it before. However, when the key was compromised, differences were observed in both option RT *and* option score. Not only were the RTs considerably faster for compromised keys, but participants were also more likely to endorse a key that they had seen before as opposed to one that they had not. Therefore, when attempting to determine whether a distractor has been compromised, it may be useful to

examine the option RT. However, when attempting to determine whether a key has been compromised, both option RT and option score may be worth investigating.

Item Response Theory

In addition to examining the raw scores and RTs, another way to detect item preknowledge is by using a modelling approach, such as IRT. By comparing each participant's true ability estimate to their cheating ability estimate, it becomes easy to identify those who performed better on the compromised items than what would have otherwise been anticipated. As expected, participants' cheating ability estimates generally exceeded their true ability estimates. Both the DOMC and MC formats yielded similar amounts of bias, suggesting that reviewing the study guides led to similar increases in performance, regardless of item format. However, the DOMC format also produced a lower RMSD and a higher correlation with the true ability estimates, thereby offering a slight advantage over the MC format. In other words, the DOMC cheating ability estimates displayed less error overall, and they were more closely related to the true ability estimates.

Conclusion

In the past, the DOMC item format has been described as a mechanism by which a testing program might increase its security. In fact, few studies have examined the validity of this statement, and those that did failed to distinguish between the effects of item compromise and item format, more generally. The purpose of this study was to determine whether the DOMC format is indeed superior to the MC format in combatting item preknowledge.

Participants were allowed to view study guides before taking a DOMC and MC test, and their resulting scores were then examined. The IRT results showed nearly identical score gains for both formats when items were compromised, and the classical item p-values yielded similar results, as well. However, the DOMC format did produce cheating ability estimates that displayed less error overall, and these estimates were more highly correlated

with the true ability estimates. Thus, it appears as though the DOMC format did offer a slight advantage over the MC format in the presence of item preknowledge.

In general, the results of this study provide several insights regarding the process by which examinees obtain preknowledge from harvested items. Prior research has studied similar behavior when MC items were administered, but one advantage of the DOMC item format is that responses may be examined at the option level. Interestingly, participants seemed to benefit the most when the key had been compromised. This suggests that participants were more focused on memorizing the keys than the distractors, though this required them to first identify that the key was, in fact, the correct option. It seems reasonable to believe that this pattern would carry over to MC items, as well. However, because the key is always revealed when an MC item is administered, this could be seen as one disadvantage of the MC item format.

This information, as well as many of the other findings that were gleaned from this study, could prove useful when conducting simulation studies that incorporate item preknowledge. For example, it would be wise to acknowledge that item compromise led to a greater reduction in RT for MC items than for DOMC items. Consequently, RT behaviors that are flagged as suspicious in one format may not necessarily be flagged in the other. Additionally, for the DOMC items in particular, it may be useful to incorporate option-level information, such as key position. By taking this information into account, researchers may be able to design better simulations that lead to the development of more effective preknowledge prevention and detection methods.

Limitations

As is often the case, this study was affected by a series of limitations. First, though several efforts were made in an attempt to increase participants' motivation, this was, in fact, a low-stakes test. As long as the participants answered the required questions, they were able to receive research credit, regardless of how they actually performed.

Furthermore, participants were only given a 50–60 minute window during which they could study. In practice, examinees would likely have much more time to study the material and memorize the test content if they so desired.

Second, participants were not directly monitored during the study period or as they took the test. Because the instructions were not overly prescriptive as to how the study time should be used, participants could have engaged in many different behaviors (e.g., scouring textbooks, browsing the internet, talking with classmates) during this time. Additionally, though process data was able to reveal whether participants left the testing window, there was no way of observing their behavior outside of the particular device that was being used to take the test. Thus, they could have accessed outside notes or even used another device while answering the items. Moreover, due to the $40 incentive, participants may have been more inclined to cheat on this test in particular. However, it should be noted that because this test was timed, participants would have had to balance the time spent searching for answers with the time required to read and respond to each item. In part, this may have been able to reduce the aforementioned threat.

Third, recall that participants were recruited from a small subset of students who attended a single university. In addition to limiting the generalizability of the results, the use of a small sample has the potential to affect the IRT parameter estimates. Though the Rasch model is known for its ability to handle small sample sizes, this is still a limitation worth mentioning, as larger sample sizes are almost always desired.

Fourth, though this study acknowledges that key position may affect the difficulty of a DOMC item, it does not attempt to balance key position across examinees or the item sets. Thus, for any one examinee, the average key position for the compromised items may have been considerably higher or lower than the average key position for the secure items that were administered. Though the key position of each item was left to random chance, constraining this feature may have been useful when drawing comparisons between the secure and compromised item sets.

Future Research

Additional research is needed to determine whether these findings are applicable to other situations, as well. For example, it would be useful to conduct real-data studies that examine different populations and different tests. It would also be interesting to see whether these findings hold in high-stakes environments where examinee motivation is less of a concern.

In addition, existing preknowledge prevention and detection methods should be examined to see how they perform when DOMC items are used. Moreover, simulation studies could be conducted to develop new methods that are specifically designed to handle DOMC items. Such methods might differ from the existing methods by taking advantage of the option-level information that DOMC items are able to provide.

References

Belov, D. I. (2016). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement*, *40*(2), 83–97. https://doi.org/10.1177/0146621615603327

Bolt, D. M., Kim, N., Wollack, J., Pan, Y., Eckerly, C., & Sowles, J. (2020). A psychometric model for discrete-option multiple-choice items. *Applied Psychological Measurement*, *44*(1), 33–48. https://doi.org/10.1177/0146621619835499

Eckerly, C., Smith, R., & Sowles, J. (2018). Fairness concerns of discrete option multiple choice items. *Practical Assessment, Research & Evaluation*, *23*, Article 16. https://doi.org/10.7275/chaw-y360

Foster, D., & Miller, H. L. (2009). A new format for multiple-choice testing: Discrete-Option Multiple-Choice. Results from early studies. *Psychology Science Quarterly*, *51*(4), 355–369.

Kingston, N. M., Tiemann, G. C., Miller, H. L., & Foster, D. (2012). An analysis of the discrete-option multiple-choice item type. *Psychological Test and Assessment Modeling*, *54*(1), 3–19.

Kyle, T. (2002, August 9). Cheating scandal rocks GRE, ETS. *The Dartmouth*. https://thedartmouth.com/article/2002/08/cheating-scandal-rocks-gre-ets/

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of testwiseness. *Educational and Psychological Measurement*, *25*(3), 707–726. https://doi.org/10.1177/001316446502500304

Murphy, K. L., & Holme, T. A. (2015). What might cell phone-based cheating on tests mean for chemistry education? *Journal of Chemistry Education*, *92*(9), 1431–1432. https://doi.org/10.1021/acs.jchemed.5b00637

Tiemann, G., Miller, H., Kingston, N., & Foster, D. (2014, October 1–2). *Protecting item content via the discrete-option multiple-choice item format* [Oral presentation]. Conference on Test Security, Iowa City, IA.

Toton, S. L., & Maynes, D. D. (2019). Detecting examinees with pre-knowledge in experimental data using conditional scaling of response times. *Frontiers in Education*, *4*, Article 49. https://doi.org/10.3389/feduc.2019.00049

Wang, X., Liu, Y., & Hambleton, R. K. (2017). Detecting item preknowledge using a predictive checking method. *Applied Psychological Measurement*, *41*(4), 243–263. https://doi.org/10.1177/0146621616687285

Willing, S., Ostapczuk, M., & Musch, J. (2015). Do sequentially-presented answer options prevent the use of testwiseness cues on continuing medical education tests? *Advances in Health Sciences Education*, *20*, 247–263. https://doi.org/10.1007/s10459-014-9528-2

Wollack, J. A., & Maynes, D. (2011, April 9–11). *Detection of test collusion using item response data* [Paper presentation]. National Council on Measurement in Education Annual Meeting, New Orleans, LA, United States.

Appendix

Test Construction

Table A1

Test Blueprint

Theme	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Total
Biological Development	1, 2	11, 12	21, 22	33, 34	45, 46	57, 58	18%
Cognitive Development	3	13	23, 24	35	47, 48	59	12%
Emotional and Social Development	4, 5	14, 15	25, 26, 27	36, 37, 38	49, 50, 51	60, 61, 62	24%
Prenatal Development and Birth	6	16	28	39, 40	52	63, 64	12%
Theoretical Perspectives	7, 8, 9, 10	17, 18, 19, 20	29, 30, 31, 32	41, 42, 43, 44	53, 54, 55, 56	65, 66, 67, 68	35%

Table A2

Test Forms

	Half 1		Half 2	
Test Form	Item Format	Item Sets	Item Format	Item Sets
A1	MC	1, 3, 5	DOMC	2, 4, 6
A2	MC	1, 3, 5	DOMC	2, 4, 6
B1	MC	1, 4, 6	DOMC	2, 3, 5
B2	MC	1, 4, 6	DOMC	2, 3, 5
C1	DOMC	2, 4, 6	MC	1, 3, 5
C2	DOMC	2, 4, 6	MC	1, 3, 5
D1	DOMC	2, 3, 5	MC	1, 4, 6
D2	DOMC	2, 3, 5	MC	1, 4, 6

Note. Secure item sets are indicated using black text, while compromised item sets are in red.