Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 27 Number 15, August 2022 ISSN 1531-7714

Does Item Format Affect Test Security?[1]

Kylie Gorney, *University of Wisconsin-Madison*
James A. Wollack, *University of Wisconsin-Madison*

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 study was to determine whether the reduced exposure of item content affects test security. We conducted an experiment in which participants were allowed to view study guides prior to taking a test comprised of DOMC and MC items. Results showed that the DOMC format seems to offer a slight advantage over the MC format in the presence of item preknowledge.

Keywords: Discrete-option multiple-choice (DOMC), multiple-choice (MC), item format, item preknowledge, test security

Introduction

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 preknowledge, examinees are expected to answer the compromised items differently than otherwise anticipated, thereby decreasing the validity of their test scores and perhaps the test scores of countless others (Eckerly, 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 its items (e.g., Wollack & Fremer, 2013).

Often, a situation is observed where a previous test-taker has served as the source. In this case, it is possible they had retained the information by memorization or by using a camera as a recording device. The former strategy, though more intensive, is nearly impossible to monitor, and the latter is becoming increasingly difficult to detect given rapidly developing technology. That being said, although item exposure is unavoidable, some item formats may be more susceptible to compromise than others. For example, each time a multiple-choice (MC) item is administered, every option is displayed, and the item is able to be harvested in its entirety. Although examinees 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) format

[1] The authors would like to thank Caveon Test Security for providing access to the platform that was used to deliver the assessment. An earlier version of this paper was presented at the 2021 annual meeting of the National Council on Measurement in Education.

instead (Foster & Miller, 2009). DOMC items are similar to MC items in that they possess a stem and a set number of options. Typically, one of these options is marked as the correct answer and serves as the key, though it is possible to have multiple keys for a single item. The main difference between the two formats is the way in which the options are presented. Rather than displaying all options simultaneously, DOMC items display options sequentially, and in a random order. After each option has been displayed, an examinee must indicate whether they believe it to be correct or incorrect by responding *Yes* or *No*. After responding to an option, the examinee cannot go back to view it again. 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, or it is scored as incorrect if the examinee endorses an incorrect option or refutes the correct one. Because it is often possible to score an item prior to administering all of the options, the DOMC format tends to expose less content than the MC format, thus offering a potential security advantage. In fact, in situations where an examinee incorrectly endorses one of the distractors, the item can be scored before the key is even revealed.[2] As a further layer of security, examinees may be presented with an additional option after the item has been scored at a prespecified probability (often 0.5) so that information regarding the correctness of the previous option is not inadvertently revealed. Many testing programs that use the DOMC format elect not to score this extra option, though whether or not it is scored does not affect the security of the test either way.

Despite the theoretical security advantages of the DOMC format, to our knowledge, only one previous study has examined this claim. Tiemann et al. (2014) analyzed the results of two types of simulated cheating. For the first type of cheating, source examinees were instructed to remember as much of the test content as possible *before* taking the test, thereby simulating an examinee intent on using memorization to harvest items. For the second type of cheating, source examinees were asked to recall information *after* taking the test, thereby simulating 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 (i.e., beneficiary examinees) 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 had to return the study guides. Results showed that there were no significant differences between the test scores of the source examinees (who did not have preknowledge) and the beneficiary examinees (who did have preknowledge) for the DOMC and MC items. This suggests that the preknowledge effect was minimal, making it difficult to tell if the DOMC format provided the security advantage that was expected. In part, this weakened effect may have been due to the low-stakes nature of the test, therefore resulting in a lack of motivation in the source and beneficiary examinees. It is also important to note that individual person or item differences were not accounted for when reporting these results.

In this article, we extend the work of Tiemann et al. (2014) and attempt to overcome the limitations that were observed in previous research. Importantly, the design of our experiment allows us to account for individual person and item differences, and we address the issue of participant motivation in two ways. First, we simulate cheating as a scenario in which the items have been captured by camera. Because there are no mistakes in the study guides that are produced this way, source examinee motivation is no longer a concern. Second, we attempt to increase the motivation of beneficiary examinees by offering a monetary incentive for good performance. Combined, we believe that these efforts more closely parallel preknowledge as it would occur in a real testing situation. Ultimately, the purpose of this research is two-fold: (1) determine whether the DOMC format is more effective than the MC format in combatting item preknowledge, and (2)

[2] For this reason, there has been some debate as to whether or not it is fair to administer DOMC options in a random order (e.g., Bolt et al., 2018; Bolt et al., 2020; Eckerly et al., 2018). Previous research and the results of this article show that DOMC items tend to be more difficult when the key is administered in a later position. Although this does not directly affect the security of the test (which is the primary focus of this study), it is something practitioners should be aware of when deciding whether to implement the DOMC format.

investigate the statistical properties of DOMC items relative to MC items when preknowledge is present.

Method

Participants

The sample consisted of students from a large, midwestern university who were enrolled in an undergraduate human development course in the Spring 2020 or Fall 2020 semesters. Note that by exclusively recruiting participants from human development courses, we were able to identify a single content domain over which all should be familiar. In exchange for their participation, all students were compensated with research credits that could be used to satisfy a course requirement. In addition, those who scored in the top 50% received a $40 prize. This incentive, combined with the fact that the mock test was given shortly before finals, was designed to motivate participants so that their efforts would more closely parallel examinees using preknowledge in a real testing situation.

Because this study took place entirely online, students were not directly monitored as they took the test. However, process data was able to provide additional information regarding students' testing behaviors. After removing those who failed to follow the given instructions (i.e., they spent less than the required time reviewing the assigned study guide, or they left the testing window once the test had started), 150 participants remained and comprised the final sample. In addition, there were two instances in which a participant spent more than 10 minutes viewing and responding to a single item. These unusually long responses were treated as missing for all subsequent analyses.

Design and Materials

In order to measure participants' understanding of human development, a 68-item test was created. Item content reflected material that was covered in all four human development courses from which students were recruited. Items were carefully phrased so that they could easily be converted from the MC format to the DOMC format without any additional editing. In other words, the options were written so that they could be marked as correct or incorrect without having knowledge of any of the other options.

Each item consisted of a stem and five options, one of which was correct and was marked as the key. In addition, all items were grouped into one of six item sets (Table 1). Sets 1 and 2, comprising 10 items each, contained the anchor items that would be used to place all items onto a common metric. These items were always secure, and all participants received them in the same format, regardless of the test form to which they were assigned. Sets 3–6 comprised 12 items each. These items had the possibility of appearing in either the DOMC or MC format, and may or may not have been compromised, depending on the test form that was administered. Importantly, because each of these items was delivered under each of the four conditions, individual item differences were accounted for, thus allowing direct comparisons to be made at the item level.

In all test forms, items on the first half of the test were not mixed with items on the second half so as not

Table 1. Test Forms.

	Half 1		Half 2	
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 in plain text, while compromised item sets are in **bold**.

to confuse participants with alternating item formats. But within each half, the items appeared in a random order, and their options were displayed in a random order, as well. Thus, it is extremely 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.

To simulate preknowledge, each participant was supplied with 1 of 20 study guides. Each study guide contained information pertaining to two of the six item sets (i.e., the compromised item sets). 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 DOMC format, while the other half experienced it entirely in MC format. Study guides were then created that contained screenshots of the items exactly as they were displayed to a particular source. As a result, MC sources captured the complete items and all of their options, while DOMC sources were only able to capture the item stems and the options that were presented to them. Therefore, some of the items that were captured by the DOMC sources appeared on the study guides *without* the key being listed, simply because it had not been disclosed.

Each study guide included information that was captured by one DOMC source and one MC source. As an example, consider a participant assigned to Form A1 where Sets 3 and 4 were compromised (see Table 1). This participant received a study guide where an MC source had leaked information for the items in Set 3, and a DOMC source had leaked information for the items in Set 4. They did not receive any information regarding the items in Sets 1, 2, 5, or 6 since these item sets were secure.

Procedure

After reading the instructions, participants were presented with 1 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 spent viewing the study guide would be monitored, and if they fell outside the 50–60 minute range, they would not be eligible to receive one of the $40 prizes. When the allotted time had passed, participants were told 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, 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 and Discussion

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 between the test halves, 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. Raw scores, rather than equated scores or IRT ability estimates, were used due to the small sample sizes.

Results indicated that Form A1 scores did not significantly differ from C1 scores, Wilks's $\Lambda = .77, F(6, 33) = 1.68, p = .16$, 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 $\Lambda = .88, F(6, 28) = 0.66, p = .69$, nor did Form B2 scores significantly differ from D2 scores, Wilks's $\Lambda = .86, F(6, 30) = 0.84, p = .55$. Therefore, the order in which the test halves were presented did not significantly affect the item set scores, greatly simplifying all subsequent analyses.[3]

Reliability

Coefficient α (Cronbach, 1951) was computed as an internal estimate of reliability. Only Sets 1 and 2

[3] Additional analyses, which are not shown here, were conducted to see whether demographic variables affected test performance. For each variable (e.g., semester of data collection, gender, number of college credits earned), an independent t-test or an analysis of variance (ANOVA) was conducted to determine whether significant differences existed with respect to the secure scores. Notably, no significant differences were found between any of the groups.

were analyzed since they were delivered securely and in the same format for all examinees. For Set 1 (10 MC items), it was found that $\alpha = .49$, while for Set 2 (10 DOMC items), $\alpha = .43$. After applying the Spearman-Brown prophecy formula to project the reliability for a 68-item test, these values became $\alpha = .87$ and $\alpha = .84$, respectively. Thus, the MC anchor items returned a slightly higher reliability estimate than the DOMC anchor items.

Classical Item Statistics

For each item, the p-value (i.e., average score), point-biserial correlation, and average response time (RT) were computed. For the items in Sets 1 and 2, these statistics were computed once across all examinees. For the items in Sets 3–6, these statistics were computed four times, since each item was administered under four different conditions (secure DOMC, secure MC, compromised DOMC, compromised MC). Summary statistics are provided in Table 2, and item-level plots can be viewed in Figures 1 and 2.

Secure items tended to be more difficult than their compromised counterparts, and the average difference in difficulties was similar for both the DOMC and MC formats. This suggests that participants benefitted similarly from preknowledge regardless of the item format administered. Figure 1 further reveals that the secure and compromised item p-values were closely related. For DOMC items, the correlation between secure and compromised p-values was .73, while for MC items, the correlation was .74. Figure 2 shows that the DOMC version of an item was almost always more difficult than the MC version, and this was true regardless of whether the item was secure or compromised. 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 et al., 2012; Papenberg et al., 2017) and compromised items (Tiemann et al., 2014).

Across all four conditions, items displayed similar average point-biserial correlations. In fact, for both the DOMC and MC formats, the average differences between secure and compromised point-biserial

Table 2. Item Statistics.

		Item p-value		Item PB Correlation		Item RT (in Seconds)	
Item	n	Mean	SD	Mean	SD	Mean	SD
Secure DOMC (Anchor)	10	.50	.08	.26	.10	19.6	5.2
Secure MC (Anchor)	10	.61	.18	.29	.10	23.8	4.9
Secure DOMC (Non-Anchor)	48	.49	.21	.27	.15	19.6	4.8
Secure MC (Non-Anchor)	48	.67	.20	.25	.15	24.1	7.5
Compromised DOMC	48	.64	.19	.29	.15	15.8	3.4
Compromised MC	48	.80	.18	.27	.15	12.7	3.8

Figure 1. Item Statistics (Secure vs. Compromised).

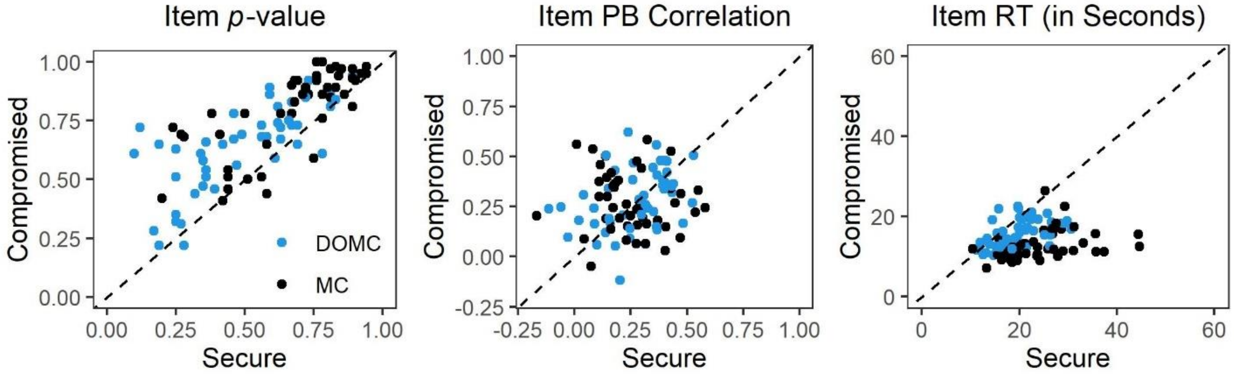
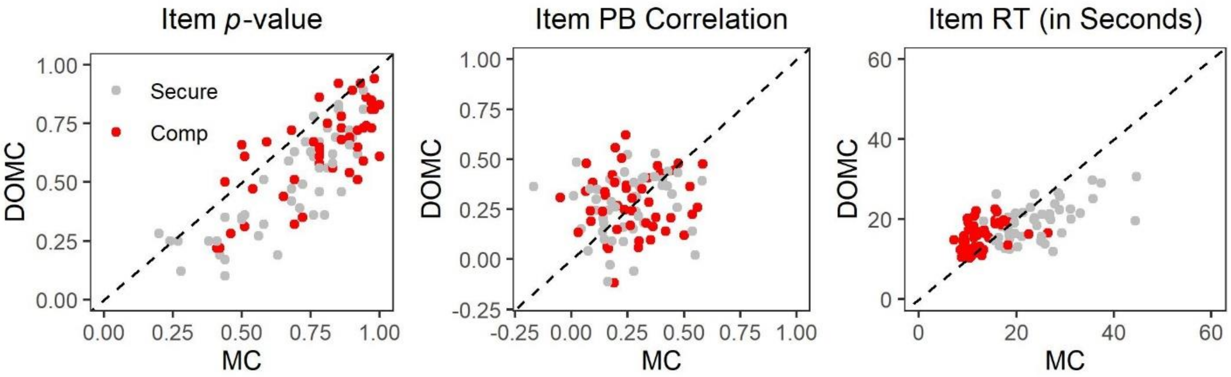


Figure 2. Item Statistics (MC vs. DOMC).



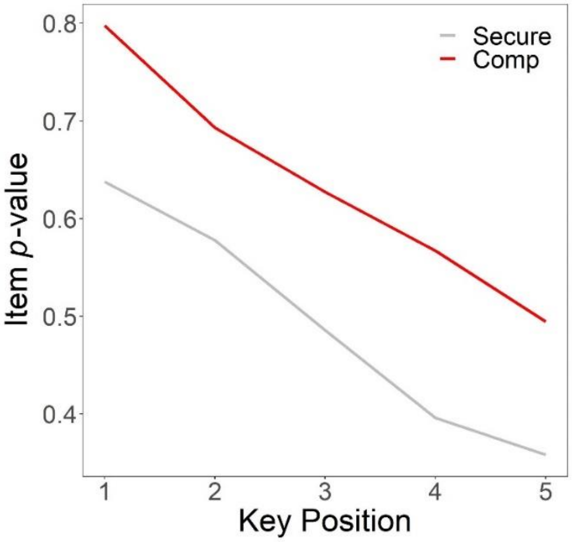
correlations were nearly identical, suggesting that item format had little to no effect on this statistic. Kingston et al. (2012) drew similar conclusions when comparing secure DOMC items to secure MC items, though they did not consider the case of compromised items. Figure 1 shows that for DOMC items specifically, the secure point-biserial correlations were positively correlated with the compromised point-biserial correlations ($r = .37$). This suggests that the secure and compromised DOMC items behaved similarly in their measurement of the underlying construct. In contrast, for MC items, the secure point-biserial correlations were negatively correlated with the compromised point-biserial correlations ($r = -.12$), suggesting that the secure and compromised MC items may have been measuring somewhat different constructs.

For both item formats, the secure and compromised RTs were positively correlated (DOMC $r = .48$, MC $r = .39$), as shown in Figure 1. This means that 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 compromised. In addition, secure items tended to require more time than compromised items, though precisely how much more time was needed depended on the item format. Table 1 reveals that the difference in RTs tended to be larger for MC items than for DOMC items. In other words, a compromised MC item saw a greater reduction in RT than a compromised DOMC item. One possible explanation for this could be that DOMC items require examinees to read and respond to each 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, examinees need only search for what they know to be the correct answer. And, assuming they know that only one option is correct, they need not consider any alternatives beyond this point, thus resulting in a shorter RT. One implication of this is that preknowledge may actually be easier to detect in the MC format than the DOMC format if RTs are able to be considered.

A major advantage of the DOMC format is that responses can be examined at the option level. In particular, it may be useful to consider the order in which the options are presented. Although MC items allow 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, thus allowing two examinees to have vastly different experiences when answering the same item. For example, if the key is displayed in the first position, an examinee is only required to answer one option correctly to receive credit for the item. But, if the key is displayed in the fifth position, an examinee must answer all five options correctly before receiving credit. Figure 3 displays the average score for DOMC items having each of the five key positions. In the event that the key was not displayed, the key position was randomly assigned a value amongst the remaining positions. For instance, if an examinee had incorrectly endorsed the first presented distractor, then the key position was randomly assigned a value between 2 and 5.

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 key positions 1 and 2 and key positions 2 and 3. However, whether the key was presented in position 4 or 5 seemed to have less of

Figure 3. DOMC Item Statistics by Key Position.



an impact on item score, suggesting that the effect of key position may diminish over time. To a certain extent, these results parallel those of Eckerly et al. (2018, p. 6). Although they only considered secure DOMC items, they also found that the effect of key position weakened as the key position itself increased. This could be explained by the fact that those of lower ability would likely have been eliminated earlier in the sequence of options. Therefore, 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 performance.

Classical Option Statistics

As mentioned previously, each item had a total of five options that were presented in a random order. Therefore, 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(a) reveals that options administered in later positions tended to be slightly easier than options administered in earlier positions. This effect was especially noticeable when the options had been compromised. It seems reasonable to assume that the explanation used above would apply here, as well. Consider that earlier options would have been answered by those of lower ability and those of higher ability. In contrast, later options may have only been answered by those of higher ability, making them appear less difficult overall.

For option RT, the option position effect was even more noticeable (Figure 4b). Participants spent the most time viewing the option that was presented in position 1, whereas all subsequent options were answered in considerably less time. This effect was similar for both secure and compromised options, suggesting some form of item familiarity. That is, later options may have been answered more quickly because the participant had more time to consider the item and all it entailed. As a result, not as much time was needed to determine whether the option itself was correct or incorrect.

Another question worth asking is whether option compromise affects option score or option RT. Table 3 reveals that the answer may depend on whether the option was a distractor or the key. See that when a distractor was compromised, option score was relatively unaffected. In other words, participants answered similarly to how they would have had the option not been disclosed. However, a noticeable difference emerged with respect to RT. Specifically, RTs were much shorter for compromised distractors than they were for secure distractors, suggesting that participants may have recalled having seen them before. Meanwhile, when the key was compromised, differences emerged with respect to both RT and score. Not only were RTs considerably shorter for the compromised keys, but participants were also more

Figure 4. DOMC Option Statistics by Option Position.

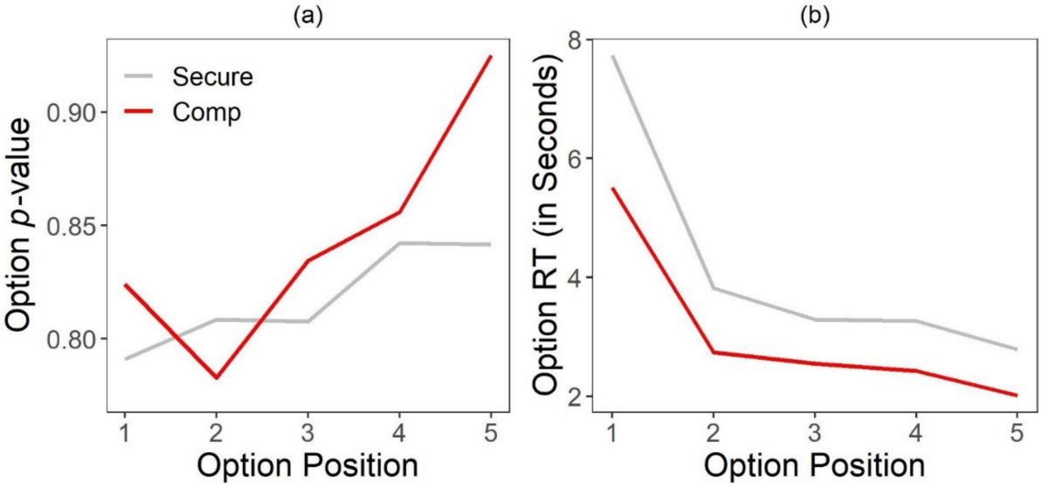


Table 3. DOMC Option Statistics.

	Option p-value		Option RT (in Seconds)	
Option	Mean	SD	Mean	SD
Secure Distractor	.83	.16	5.6	2.6
Compromised Distractor	.80	.23	4.0	2.9
Secure Key	.72	.22	5.0	1.8
Compromised Key	.87	.14	3.2	1.6

likely to endorse a key they had seen before as opposed to one they had not. Combined, these results suggest that when item content is disclosed, participants focus more on memorizing the keys than the distractors.

Item Response Theory

In addition to the classical statistics, IRT item and ability parameters were estimated using the Rasch model. This model was chosen because it was found to provide a significantly better fit than the more heavily parameterized 2PL and 3PL models. Under the Rasch model, the probability of examinee j answering item i correctly can be written as

$$P(X_{ji} = 1) = \frac{\exp{(\theta_j - b_i)}}{1 + \exp{(\theta_j - b_i)}}, \qquad (1)$$

where θ_j is the ability of examinee j, and b_i is 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 in the test. An examinee's location on this latent trait (i.e., their ability) is the sole indicator of their performance, and no amount of outside information should affect their responses. Yet, when examinees possess any amount of preknowledge, this assumption no longer holds. We could say that there now exist two latent traits that are responsible for 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 uncontaminated item parameter estimates, only the secure item responses were used. Furthermore, items that were displayed in both the DOMC and MC formats (Sets 3–6) received two sets of item parameter estimates: one for each format. Next, the item parameter estimates were treated as fixed, and each participant received three ability estimates: one true ability estimate (based only on the secure items) and two cheating ability estimates (one based only on the compromised DOMC items, and one based only on the compromised MC items).

To compare cheating ability to true ability, three criteria were assessed: bias, root mean squared difference (RMSD), and the correlation between estimates. Bias measures whether the cheating ability estimates tended to over- or under-estimate the true ability estimates and is computed as the average difference across examinees. This can be written as

$$Bias = \frac{1}{J}\sum_{j=1}^{J}(\hat{\theta}_{cj} - \hat{\theta}_{tj})\ , \tag{2}$$

where J is the total number of examinees, and $\hat{\theta}_{cj}$ and $\hat{\theta}_{tj}$ are the cheating and true ability estimates, respectively, of examinee j. In contrast, the RMSD is concerned with the absolute difference between estimates and can be written as

$$RMSD = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\theta}_{cj} - \hat{\theta}_{tj})^2}\ . \tag{3}$$

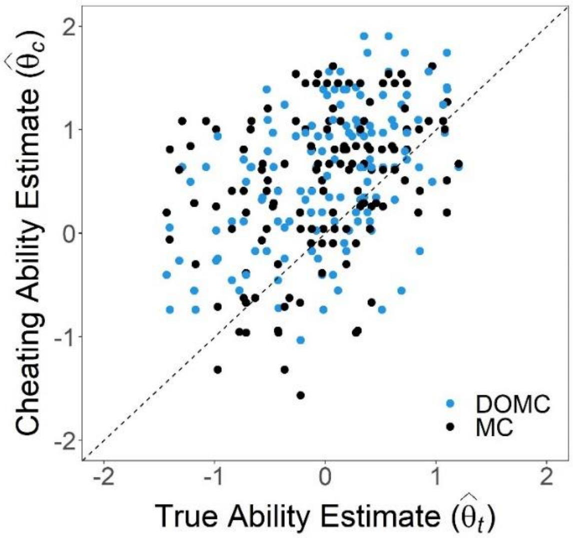
The final criterion was the correlation between the cheating and true ability estimates. Theoretically, the item format that is more secure should produce bias and RMSD values closer to 0, and a larger, positive correlation.

The ability estimates are shown in Figure 5. As expected, for most participants, the cheating ability estimates exceeded the true ability estimates. Specifically, 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. This suggests that reviewing the study guides led to similar increases in performance, regardless of item format. Notably, however, the DOMC cheating ability estimates yielded an RMSD of 0.85 and a correlation of 0.47, while the MC cheating ability estimates yielded an RMSD of 0.91 and a correlation of 0.36. In other words, the DOMC cheating ability estimates displayed less total error, and they were more closely related to the true ability estimates. Thus, the DOMC format seems to have offered a slight advantage in the presence of item preknowledge.

Conclusion

In the past, the DOMC format has been described as a mechanism by which a testing program might increase its security. Previous research has yielded inconclusive results (Tiemann et al., 2014), though such research may have been limited by the design of the experiment and a lack of motivation from the participants. The purpose of this study was to extend this work and address the following research questions: (1) Is the DOMC format more effective than the MC format in combatting item preknowledge? (2) How do the statistical properties of DOMC items compare to those of MC items when preknowledge is present?

Figure 5. Ability Estimates (True vs. Cheating).



To answer these questions, we conducted an experiment in which participants were allowed to view study guides prior to taking a test comprised of DOMC and MC items. Test scores were then examined using classical statistics and IRT. On average, the two item formats showed nearly identical score gains as a result of preknowledge. However, the DOMC cheating ability estimates displayed less total error, and they were more highly correlated with the true ability estimates. Therefore, to answer Research Question 1, it appears as though the DOMC format was slightly more effective than the MC format in combatting item preknowledge.

The answer to Research Question 2 is much broader and encompasses several interesting results. In general, we found that when examinees had preknowledge, DOMC items tended to be more difficult, similarly discriminating, and more time intensive than MC items. None of these results are surprising, although it is interesting to see that preknowledge affected the RTs of MC items more than it did the RTs of DOMC items. As mentioned earlier, this implies that preknowledge may be easier to detect in the MC format than the DOMC format if RTs are able to be incorporated into the analysis.

An additional contribution of this study is that it provides several insights regarding the process by which examinees obtain preknowledge from harvested items. Previous research has studied similar behavior when MC items were administered, but the analysis of DOMC items offers a unique perspective in that the responses can be examined at the option level. Interestingly, we found that participants seemed to benefit the most when the key was compromised. This suggests that participants were more focused on memorizing the key than the distractors, though this of course required them to identify that the key was, in fact, the correct option. It seems reasonable to believe that similar patterns would carry over to MC items, as well. Consequently, the fact that the key is always revealed when an MC item is administered could be seen as a major security disadvantage of the MC format.

In the interest of fairness, we would like to remind readers of the caveats associated with the DOMC format that could potentially outweigh any gains, security-related or otherwise (see, e.g., a discussion on the DOMC format's increased protection against the use of testwiseness cues in Papenberg et al., 2017). Most notable is the concern regarding the key position effect, where DOMC items having later key positions tend to be more difficult than DOMC items having earlier key positions (e.g., Bolt et al., 2018; Bolt et al., 2020; Eckerly et al., 2018). To our knowledge, available engines for delivering DOMC items do not yet offer a way of controlling this feature so as to ensure that all examinees are affected equally. Note that even if the average key position were constrained to be equal across all examinees (thus controlling the item-level variability), some examinees may still be more sensitive to the key position effect than others, which would manifest as person-level variability (Bolt et al., 2020; Kim et al., 2019). To address this issue, complex IRT models could be employed that account for such variance (e.g., Bolt et al., 2020). Alternatively, separate sets of item parameters could be estimated for each key position (Eckerly et al., 2018). Whether or not this additional effort is worthwhile, however, in exchange for the security benefits that the DOMC format has to offer is left to the discretion of individual testing programs.

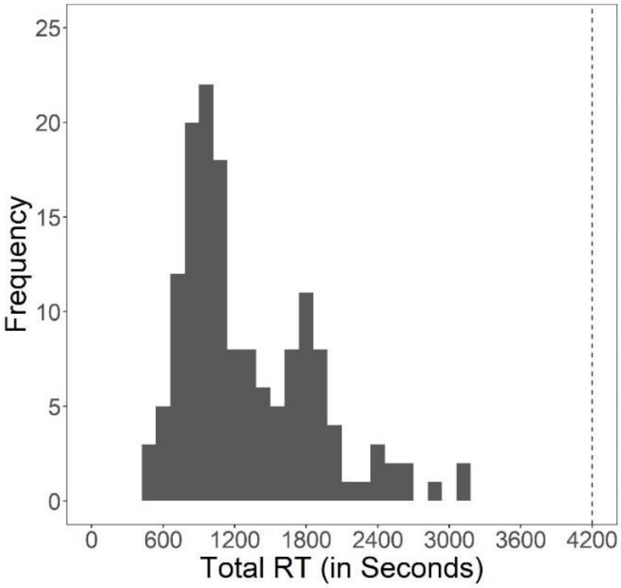
Limitations

As is often the case, this study was affected by a series of limitations. First, although several efforts were made in an attempt to increase participants' motivation, this was, in fact, a low-stakes test. As long as participants answered the required questions, they were able to receive research credit, regardless of how well they actually performed. Furthermore, participants were only given a limited window during which they could study. In practice, examinees would likely have had more time to study the materials if they so desired. However, we believe our results show that the participants were motivated, and that they engaged with the study guides, at least to a certain extent. If participants had not been motivated or had not engaged with the study guides, then we would expect to see similar scores and RTs on both the secure and compromised items. However, Figure 1 revealed that the compromised items were typically easier and were answered more quickly than their secure counterparts. A series of paired samples t-tests confirmed that these differences were statistically significant at the $\alpha = .05$

level,[4] suggesting that the study guides did, in fact, have an impact on testing behavior.

The second limitation is that participants were not directly monitored as they took the test. Although 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. Therefore, it is possible that participants could have accessed outside resources while answering the items and may have been particularly inclined to do so due to the $40 incentive. However, we believe this may not have been an issue for three reasons. (1) Recall that the study guides did not provide an answer key, and so participants were required to determine the correct answers to the compromised items on their own, perhaps by using the internet or other resources. If they had used similar strategies to cheat on the secure items during the live exam, then we would expect to see similar scores on both the secure and compromised items. However, as mentioned in the previous paragraph, statistically significant differences were observed between the secure item scores and the compromised item scores, suggesting that participants did not rely heavily (if at all) on outside resources once they had started the test. (2) Because the 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. It would be very difficult to do this for all 68 items while staying within the 70-minute time limit, and we further note that the majority of participants finished the test with ample time remaining (Figure 6). Not only does this mean that they had even less time to look up the test content, but it also suggests that they may not have been particularly driven to do so, especially given that the average score on the secure items was relatively low (see Table 2). (3) Even if participants had used outside resources to cheat during the live exam (i.e., *after* the test had already begun), any conclusions drawn from this study regarding preknowledge (defined as having item information *before* starting the test) still hold. In other words, the purpose of this study was to determine whether or not

Figure 6. Total Response Time Distribution.



[4] There was a significant difference between the secure DOMC item p-values and the compromised DOMC item p-values, $t(47) = -7.03, p < .05$, as well as a significant difference between the secure MC item p-values and the compromised MC item p-values, $t(47) = -6.33, p < .05$. There was also a significant difference between the average secure DOMC item RTs and the average compromised DOMC item RTs, $t(47) = 5.94, p < .05$, as well as a significant difference between the average secure MC item RTs and the average compromised MC item RTs, $t(47) = 11.44, p < .05$.

preknowledge of the study guides was differentially beneficial as a function of item format. Importantly, this question can still be answered regardless of whether students did or did not cheat during the live exam.

The third limitation is 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 IRT parameter estimates. Although 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 typically desired.

Fourth, in the interest of ensuring comparability across item formats, we intentionally capped each item at having five options. We believed this to be reasonable, since for many high-stakes credentialing and educational tests, there is a clear limit as to the number of high-quality, plausible options that can be constructed. However, in theory, the DOMC format could accommodate many more options, potentially including items with multiple keyed responses. It seems this would lead to improved security, though such a topic is left to explore in future research.

Future Research

Additional research is needed to determine whether these findings are applicable to other situations. For example, it would be useful to conduct more real-data studies to examine different populations and tests. It would also be interesting to see whether these findings hold in a high-stakes environment where motivation is less of a concern. In addition, existing preknowledge detection methods should be examined to see how they perform with DOMC items. Simulation studies could also be conducted to evaluate new preknowledge detection methods that are specifically designed to handle DOMC items. Such methods may differ from existing ones by taking advantage of the option-level information that DOMC items are able to provide.

References

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

Bolt, D. M., Lee, S., Wollack, J., Eckerly, C., & Sowles, J. (2018). Application of asymmetric IRT modeling to discrete-option multiple-choice test items. *Frontiers in Psychology*, *9*, Article 2175. https://doi.org/10.3389/fpsyg.2018.02175

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Eckerly, C. A. (2017). Detecting preknowledge and item compromise: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 101–123). Routledge.

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

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

Kim, N., Bolt, D. M., Wollack, J., Pan, Y., Eckerly, C., & Sowles, J. (2019). Modeling examinee heterogeneity in discrete option multiple choice items. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology: 83rd annual meeting of the Psychometric Society, New York, NY 2018* (pp. 383–392). Springer. https://doi.org/10.1007/978-3-030-01310-3_33

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*(1), 3–19.

Papenberg, M., Willing, S., & Musch, J. (2017). Sequentially presented response options prevent the use of testwiseness cues in multiple-choice testing. *Psychological Test and Assessment Modeling*, *59*(2), 245–266.

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

Wollack, J. A., & Fremer, J. J. (Eds.). (2013). *Handbook of test security*. Routledge.

Citation:
Gorney, K., & Wollack, J.A. (2022). Does item format affect test security? *Practical Assessment, Research, & Evaluation*, 27(15). Available online: https://scholarworks.umass.edu/pare/vol27/iss1/15/

Corresponding Author:
Kylie Gorney
University of Wisconsin - Madison
Madison, WI, USA

Email: kyliengorney [at] gmail.com