GEFAD / GUJGEF 39(2): 1113-1134(2019)

Investigating Consequences of Using Item Pre-knowledge in Computerized Multistage Testing *

Bireye Uyarlanmış Çok Aşamalı Testlerde Madde Ön Bilgisinin Test Sonuçlarına Etkisinin Araştırılması

Halil İbrahim SARI[1]

[1]*Kilis 7 Aralik University, Department of Educational Science, Measurement and Evaluation in Education Program. hisari87@gmail.com*

Makalenin Geliş Tarihi: 08.03.2019 ***Yayına Kabul Tarihi: 02.05.2019***

ABSTRACT

The goal of this study is to determine the effects of test cheating in a scenario where test-takers use item pre-knowledge in the c-MST, and to urge practitioners to take additional precautions to increase test security. In order to investigate the statistical consequences of item pre-knowledge use in the c-MST, three different cheating scenarios were created, in addition to the baseline condition (e.g., no pre-knowledge usage). The findings were compared under 30-item and 60-item test length conditions with 1-3-3 c-MST panel design. A total of thirty cheaters were generated from a normal distribution, and EAP was used as an ability estimation method. The findings were discussed with the evaluation criteria of mean bias, root mean square error, correlation between true and estimated thetas, conditional absolute bias, and conditional root mean square. It was found that using item pre-knowledge severely affected the estimated thetas, and as the number of compromised items increased, the results got worse. It was concluded that item sharing and/or test cheating seriously damage the test scores, test usage, and score interpretations.

Keywords: *Computerized multistage testing, Test cheating, Item pre-knowledge*

ÖZ

Bu çalışmanın amacı, bireye uyarlanmış çok aşamalı (BUÇAT) testi alan bireylerin madde ön bilgisini kullandıkları durumlarda yetenek seviyelerinin nasıl etkilendiğini ortaya çıkarmak ve bu durumun meydana getirmiş olduğu sonuçlar konusunda testi düzenleyenleri test güvenliğini arttırmak için ek önlemler almaya teşvik etmektir. BUÇAT'ta madde ön bilgi kullanımının istatistiksel sonuçlarını araştırmak için, null durumuna (madde ön bilginin kullanılmadığı) ek olarak üç farklı madde hırsızlığı senaryosu simülasyonla üretilmiştir. Bulgular, 30 maddelik ve 60 maddelik test uzunluğu koşullarında 1-3-3 BUÇAT panel tasarımı ile karşılaştırılmıştır. Madde hırsızlığı yapan 30 bireyin yetenek seviyeleri normal dağılımla üretilmiştir. Bireylerin ara ve final

* **Alıntılama:** Sarı, H.İ. (2019). Investigating consequences of using ıtem pre-knowledge in computerized multistage testing*. Gazi Üniversitesi Eğitim Fakültesi Dergisi, 39*(2), 1113-1134.

yetenek seviyeleri beklenen sonsal dağılım (EAP) ile hesaplanmıştır. Simülasyon sonuçları iki farklı istatistik grubuyla değerlendirilmiştir: (a) genel sonuçlar ve (b) koşullu sonuçlar. Genel istatistikler için, ortalama yanlılık (mean bias), ortalama kareler hatası (RMSE) ve hesaplanan ve doğru yetenek seviyeleri arasındaki korelasyon hesaplanmıştır. Bulgulara göre madde ön bilginin kullanılmasının öğrenci yetenek seviyelerini ciddi şekilde etkilediği ve risk altındaki (test sonrasında paylaşılan maddeler) maddelerin sayısının artmasıyla sonuçların daha da kötüleştiği görülmüştür. Madde paylaşımının ve / veya test hırsızlığının test puanlarına, test kullanımına ve puan yorumlarına ciddi şekilde zarar verdiği sonucuna varılmıştır.

Anahtar Sözcükler: *Bilgisayarlı çok aşamalı test, Test hilesi, Madde ön bilgisi*

INTRODUCTION

There are three main test administration models used to measure student success in the area of education and psychology. A linear test and/or static test is administrated on paper, and item order and item number do not change during the test. Computerized adaptive testing selects items one-by-one, according to one's current ability estimate (Weiss & Kingsbury, 1984). And computerized multistage testing is another type of adaptive testing that selects a group of items called modules, based on one's current ability level (Luecht & Nungester, 1998). All three types of test administration models are designed to measure success accurately and precisely. It has been evidentially proven that the latter two achieve this goal much better than linear tests (Luecht & Sireci, 2011).

High accuracy in measurement is not the only advantage of adaptive tests compared to linear tests; others include lower test length, quick scoring, and test security (Weiss & Kingsbury, 1984). However, test security in adaptive tests is not perfect, as researchers have noted (Guo, Tay, & Drasgow, 2009; Segall, 2004). This is because test items are selected from an item pool and some of the items in the pool, e.g., items that have not reached a pre-specified exposure rate, are reused in future test administrations. The associated risk of these questions being shared among test takers thus negatively impacts the validity and reliability of the test (Segall, 2004). Therefore, issues with test security in adaptive testing must be addressed, and alternative ways of making tests should be explored.

Adaptive tests are administrated on computers via the internet around the world. For example, the Graduate Record Examinations, an ETS test, is administered at more than 1,000 test centers in more than 160 counties[†], and more than 1,600,000 test takers took it after it switched from CAT to c-MST-revised-GRE (from 2012 and 2015).[‡] These numbers obviously make the test vulnerable to cheating. In fact, ETS's past experience clearly supports the claim that test items are easily shared, especially in Asian counties. For example, ETS suspended the CAT version of the GRE due to test fraud that occurred in four Asian countries—China, Hong Kong, Taiwan, and South Korea—on August 6, 2002, switching to a paper version only. Due to an abnormally large occurrence of high scores on GRE, these countries have been designated red flag countries by ETS Board (Ewing, 2002). Apparently, frequent test administration without replenishing the item banks or rotating test materials in and out of active use may incentivize collaboration to cheat. The first test taker does not necessarily cheat on the test, but this does not guarantee that test takers who have received previously administered items and take the test later do not use their item pre-knowledge during the test.

Another well-known case of test fraud occurred in 2008 at the Kaplan Test Preparation Company. In this case, a group of employees working at Kaplan repeatedly took the GRE, with each employee attempting to memorize some items, so that group could collectively steal the whole item bank. ETS filed a lawsuit against Kaplan Test Company charging a violation of test security, although Kaplan defended itself by saying that the theft was intentionally instigated by the company to show it is possible to steal items in the bank and encourage ETS to take actions to protect the item bank. This experiment cost Kaplan $150,000 in fines. For further readings about test cheating, one can refer to reports released by the U.S. Department of Justice[§].

The potential impact of item sharing and/or theft have been investigated in relation to both linear tests and CAT (see Guo, Tay, & Drasgow, 2009; Segall, 2004; Zopluoglu &

[†]https://www.ets.org/gre/revised_general/register/
[‡]https://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2015.pdf
[§]https://www.justice.gov/opa/pr/fifteen-chinese-nationals-charged-fraud-scheme

Davenport, 2012) but a limited number of studies have examined these issues in the context of computerized multistage testing. Because the c-MST is a new trend in the area of educational and psychological measurement, interest in it has recently increased, especially after GRE switched from CAT to c-MST (Yan, von Davier, & Lewis, 2014). Therefore, this study investigates some statistical consequences of using item pre-knowledge and/or test cheating in computerized multistage testing and discusses some potential ways to prevent cheating on the c-MST. The purpose of this study is to show how severe the impacts would be if test takers were to use item pre-knowledge during administration of the c-MST. We aim to emphasize and empirically support the notion that item theft and item sharing can seriously damage the test scores, test usage, and score interpretations, especially in large scale test administrations. We also aim to draw researchers, test developers, and test users' attention to the issues of test security in computerized multistage testing, because its use is rapidly increasing year by year (e.g., the Massachusetts Adult Proficiency Test, r-GRE, Law School Admission Council, Certified Public Accountants (CPA) Examination).

CONCEPTUAL FRAMEWORK

Item response theory (IRT) is a strong (e.g., in terms of restrictive nature of the local independence assumptions needing to hold item by item) statistical theory used to describe the expected probability of a particular response pattern to an item conditional on the latent trait levels ($P(\theta)$) (Baker, 1992). Of the different IRT models used to calculate this probability, the three-parameter model (3PL) (Birnbaum, 1968) is most widely used. The 3PL model defines the conditional probability of a correct response on item i for person p (X_{ip}=1) as

$$P\left(X_{ip}=1\middle|\theta_p\right)=c_i+\frac{(1-c_i)}{1+\exp\left[-a_i(\theta_p-b_i)\right]} \quad (1)$$

where b_i is the difficulty parameter, a_i is the item discrimination, c_i is the pseudo-guessing parameter for item i, and θ_p is the latent trait score for person p. The assumption in Equation 1 is that the person did not use item pre-knowledge (e.g., the person did not

memorize the item), and the probability of getting the item correct is accounted for by the underlying latent trait only. In other words, it denotes the probability of answering the item without item pre-knowledge. If the person has item pre-knowledge, the overall probability of getting the item correct (e.g., final probability) increases. As discussed and presented in McLeod, Lewis, and Thissen (2003), this final probability is defined as

$$P\left(X_{ip}=1\middle|\theta_p\right)_{overall}=P(m_i)+\left(\left(1-P(m_i)\right)*P\left(X_{ip}=1\middle|\theta_p\right)\right) \tag{2}$$

where $P\left(X_{ip}=1\middle|\theta_p\right)_{overall}$ is the overall probability of getting the item correct, $P(m_i)$ is the probability of using item pre-knowledge for item *i*, and $P\left(X_{ip}=1\middle|\theta_p\right)$ is the probability of answering the same item without item pre-knowledge. In the case of the 3PL item pre-knowledge model, which is the focus of this study, the overall probability of getting the item correct defined in Equation 2 can be expressed as

$$P\left(X_{ip}=1\middle|\theta_p\right)_{overall}=\left(P(m_i)+c_i-c_i\,P(m_i)\right)+\frac{1-\left(P(m_i)+c_i-c_i\,P(m_i)\right)}{1+\exp\left[-a_i(\theta_p-b_i)\right]}, \tag{3}$$

This equation can be simplified as

$$P\left(X_{ip}=1\middle|\theta_p\right)_{overall}=c_i'+\frac{1-c_i'}{1+\exp\left[-a_i(\theta_p-b_i)\right]}, \tag{4}$$

where c_i' is the new guessing parameter and equals to

$$c_i'=P(m_i)+\ c_i-c_iP(m_i) \tag{5}$$

The key component in the item pre-knowledge model (Equations 2 and 3) is the probability of having item pre-knowledge, $P(m)$. This is because if the person solves the item without pre-knowledge (e.g., $P(m)=0$), Equation 2 turns out Equation 1, and if the test taker definitively knows the correct answer for an item (e.g., $P(m)=1$), the overall ability in Equation 2 equals to 1. For example, if the $P(m)$ is equal to .80 for an item, this means that there is a 80% chance of a test taker using item pre-knowledge when

answering this item. Assuming that difficult items are shared with other test takers after the exam, the $P(m)$ is going to be

$$P(m) = \frac{1}{1+\exp(1-b)} \quad (6)$$

(McLeod, Lewis, & Thissen, 2003). This indicates that as the item difficulty increases, the probability of memorizing the item—and of others then receiving this knowledge after the test— increases. In other words, more-difficult items are more-often solved by using pre-knowledge. One can refer to McLeod et al. (2003) for more descriptive information about the conceptual framework for the item pre-knowledge model.

METHOD

Design Overview

In this study, we first manipulated a null condition where no item pre-knowledge usage is allowed, and responses to the test items are assumed to be correlated with innate abilities and knowledge only. This null condition is called Case 1. However, in reality, it is impossible to know how a test taker used item pre-knowledge in the test. Thus, in this study, we created three possible test cheating scenarios, called Case 2, Case 3 and Case 4. We manipulated the assumed item pre-knowledge in different ways for each case. All four scenarios were tested under 1-3-3 c-MST panel design with two levels of test length, 30 items and 60 items. All manipulated conditions were fully crossed with one another. This resulted in a total of eight scenarios (two test lengths x four cheating models). For each condition, 100 iterations were performed. 30 cheaters were generated from a normal distribution, $N(0, 1)$. The theta values that represent examinees were re-generated for each replication, but for better comparability, the same theta values were used for each of the eight scenarios. The whole simulation process was completed in RStudio version 0.99.903 (R Development Core Team, 2009–2016). Both the fixed and varied study conditions are detailed in the following sections.

Fixed Conditions

In this study, the item parameters were based on a real ASVAB military test used in Armstrong, Jones, Li, and Wu (1996). As in the original item bank, our simulated item bank had 450 multiple-choice items from four different content areas. The item parameters and number of items for each content area are provided in Table 1. The test length in the ASVAB was 30, and the distributions across the content areas were 10, 11, 4, and 5 for Case 1, Case 2, Case 3, and Case 4, respectively. For the 30-item test length condition we used the same target numbers for each content area, while for the 60-item condition we doubled all corresponding numbers. It should be noted that within each panel structure, the number of items and content distributions for the modules at the same stage were the same.

Table 1. Item Parameters of Each Content Area in the Item Bank

Content Area	***a***		***b***		***c***	
(Number of items)	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Content 1 (*n*=150)	1.079	.40	-.467	1.179	.210	.09
Content 2 (*n*=165)	1.128	.43	-.154	1.033	.200	.10
Content 3 (*n*=60)	1.092	.53	-.025	.815	.203	.08
Content 4 (*n*=75)	1.237	.38	-.014	.678	.162	.08

As already explained, a total of 30 cheaters were generated from a normal distribution. The 3PL model (Birnbaum, 1968) was used to generate item responses for non-compromised items (see Equation 1). The 3PL item pre-knowledge model (McLeod, Lewis, & Thissen, 2003) was used to generate item responses for the compromised items (see Equation 3). The expected a posteriori (EAP) (Bock & Mislevy, 1982) with a prior distribution of $N(0, 1)$ was used for both provisional and final ability estimates across all simulation conditions. Again, all study conditions were tested with a 1-3-3 c-MST design, which is one of the most commonly used panel designs in the literature (see Schnipke & Reese 1999; Zenisky, 2004). The details of the panel design are presented in later sections. For all conditions, the maximum Fisher information method (Lord 1980; Thissen &

Mislevy 2000) was used as the routing method. That computer algorithm calculates an examinee's ability level based on previously administered module(s) and then selects the module that best matches his/her current ability estimate (see Weissman, Belov, & Armstrong, 2007 for more technical details).

Varied Conditions

It is impossible to know how a test taker used item pre-knowledge when taking a test. Thus, in addition to the null condition, we manipulated three different item pre-knowledge usage conditions. In Case 2, it is assumed that test takers know the correct answers for some of the items (i.e., the probability of getting them was 1). In this condition, all items in stage two were chosen as compromised items. This means that each test cheater cheated on the same *number* of items, but depending on the module they received in stage two (e.g., easy, medium, or hard), the items they cheated on were not necessarily the same for all test takers. In this condition, depending on the total test length, the number of compromised items varied—10 in the 30-item test and 20 in the 60-item test, respectively. This is because there were 10 and 20 items in stage two modules in the 30 item and 60 item test length conditions, respectively.

In Case 3, it was assumed that cheaters memorized more-difficult items, and the degree of the probability of a test taker having pre-knowledge was correlated with item difficulty. The compromised questions were selected from stage two and stage three items. Again, this does not mean that all stage two and three items were necessarily solved by using pre-knowledge, but that the probability of a test taker having item pre-knowledge was set as lower for the easy items and higher for difficult items. The probability of a test taker having item pre-knowledge for those items was generated by Equation 6.

In Case 4, the probability of item pre-knowledge was generated from a uniform distribution, with minimum value 0 and maximum value of 1. This manipulated condition is somewhat similar to random-strings type of test cheating (Wollack, Cohen, & Serlin, 2001). Similar to in Case 3, the compromised items were selected from stage two and stage three items—60 total stage two and three items in the 30-item condition and 120

total stage two and three items in the 60-item condition. Again, this does not mean all stage two and three items were solved by item pre-knowledge. Since the probability of pre-knowledge ranged from 0 to 1, it was negligibly low for some of the items. Thus, the stated number of compromised items in both test length conditions were the maximum number of items that could have been solved with item pre-knowledge.

The purpose of manipulating these conditions was to explore the potential damage of item pre-knowledge use on ability estimates under different test cheating scenarios. Our ultimate goal is not to compare the findings of Cases 2, 3, and 4 with one another, but to compare them with the results of Case 1 (e.g., null condition or baseline condition).

Test Assembly

All study conditions were tested under a 1-3-3 c-MST panel design with three non-overlapping essentially parallel panels generated from a simulated item bank (see Table 1). The 1-3-3 c-MST panel design had one routing module in stage one, and two easy, three medium, and two hard modules in stages two and three. Regardless of the test length condition, there were an equal number of items in all modules. This means that in each module in any stage, there were 10 and 20 potentially compromised items in the 30-item and 60-item test length conditions, respectively. The multiple panel design was used in an attempt to hold the maximum panel, module, and item exposure rates at 0.33. After the panels were built, thirty cheaters were randomly assigned, ten per panel. The IBM CPLEX program (ILOG, Inc, 2006) was used to build the panels and modules: First items were clustered into different modules, then modules were randomly assigned to the panels. The bottom-up strategy was used to create panels, which means that modules at the same difficulty level were exchangeable across the panels.

The automated test assembly finds a solution to maximize the IRT information function at a fixed theta point; denote θ_0 as the fixed theta point. We first define a binary decision variable, x_i, (e.g., x_i=0 means item *i* is not selected from the item bank, x_i=1 means item *i* is selected from the item bank). The information function we want maximize is

$$I(\theta_0) = \sum_{i=1}^{N} I(\theta_0, \xi_i) x_i \quad (7)$$

where ξ_i represents the item parameters of item *i* (e.g., *a*, *b*, *c* parameters). As in the original bank, our simulated item bank had items from four content areas (e.g., C_1, C_2, C_3, and C_4), and the target distributions across the four content areas were 10, 11, 4, and 5 items, respectively. The automated test assembly for the 30-item test length condition was modeled to maximize

$$\sum_{i=1}^{N} I(\theta_0, \xi_i) x_i, \tag{8}$$

subject to

$$\sum_{i \in C1}^{N} x_i = 10, \tag{9}$$

$$\sum_{i \in C2}^{N} x_i = 11, \tag{10}$$

$$\sum_{i \in C3}^{N} x_i = 4, \tag{11}$$

$$\sum_{i \in C4}^{N} x_i = 5, \tag{12}$$

$$\sum_{i=1}^{N} x_i = 30, \tag{13}$$

and

$$x_i \in (0,1),\ i = 1, \ldots . N, \tag{14}$$

which puts constraints on C_1, C_2, C_3, and C_4, the total test length, and the range of decision variables, respectively. The test assembly models under 60-item condition were modeled similarly.

As in Diao and van der Linden (2011), when building 1-3-3 c-MST panel design, the three fixed theta scores were chosen as θ_1=-1, θ_2=0, and θ_3=1, which represent the target information functions for easy, medium, and hard modules, respectively. In the panel design, the items in the routing modules were chosen from medium difficulty items (e.g., items that maximize information function at theta point of 0). After modules were built, they were randomly assigned to the panels.

Evaluation Criteria

The results of the simulation were evaluated with two sets of statistics: (a) overall results and (b) conditional results as evaluated in similar studies (see Zenisky, 2004). For overall statistics, mean bias, root mean squared error (RMSE), and the correlation between estimated and true theta ($\rho_{\hat{\theta}\theta}$) were computed from the simulation results. Mean bias was calculated as

$$\bar{e} = \frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)}{N} \quad (15)$$

RMSE was calculated as

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)^2}{N}} \quad (16)$$

The correlation between true and estimated theta values was calculated as

$$\rho_{\widehat{\theta_j},\theta_j} = \frac{cov\,(\hat{\theta}_j,\theta_j)}{\sigma_{\hat{\theta}_j}\sigma_{\theta_j}}, \quad (17)$$

In any particular condition, each overall statistic was calculated separately for each iteration across the 30 examinees, and then averaged across 100 replications. For conditional results, conditional absolute bias and conditional root mean squared error were calculated between θ =-2 and θ=2, with the width of the θ interval at 0.1 (e.g., over 41 theta values).

RESULTS

Overall Results

The results of mean biases, root mean square errors, and correlations between estimated and true theta values across the four cases under two test length conditions are provided in Table 2. In terms of mean bias and root mean square error, as expected, both outcomes were lowest in condition Case 1 (e.g., null condition) regardless of test length, and the outcomes decreased as the test length increased. Compared to Case 1, mean bias and root

mean square error were very high in the other three cases, and both outcomes were the worst in Case 2 and Case 4. This was more likely due to the fact that Case 3 was manipulated so that the probability of solving difficult items by cheating increased, and probability of solving easy items by using item pre-knowledge was lower. As the test length increased, both outcomes increased, and this was due to the fact that as the test length increased, the number of compromised items increased in the three different test cheating scenarios (Cases 2, 3, and 4). In terms of correlations between true and estimated theta values, Case 1 resulted in higher estimates, but they were not much lower in the three different test cheating conditions. Also, increasing test length did not meaningfully affect the correlation estimates in all conditions. Overall, the main finding was that regardless of how cheating occurred, item pre-knowledge use severely impacted the outcomes, and this impact was even not comparable with the case where test cheating did not happen.

Table 2. Results of Overall Outcomes

	Mean Bias		**RMSE**		**Correlation**	
Case	30-item	60-item	30-item	60-item	30-item	60-item
Case 1	.08	.04	.32	.27	.98	.99
Case 2	.59	.70	.79	.82	.96	.96
Case 3	.22	.70	.45	.82	.97	.96
Case 4	.59	.70	.80	.83	.96	.96

Conditional Results

The results of conditional absolute biases across the four cases under two test length conditions were given in Figure 1. As expected, regardless of test length, Case 1 resulted in the lowest absolute biases and root mean square errors across the estimated theta values. Aligned with the mean bias and root mean square errors discussed above, increasing test length made the conditional results more stable (i.e., the fluctuations were more stable) and improved both conditional absolute bias and root mean square error. The conditional results were worse in the three test cheating conditions. Since the number of compromised items increased with the increase in test length, the findings were worse in the 60-item test length condition for all three cheating scenarios. Under Case 2, both

conditional error estimates (absolute bias and root mean square error) were very high across all estimated theta values (from -2 through 2). This was due to the fact that, regardless of the size of the true theta values, all test takers cheated on an equal number of items. Under Case 3, since only the higher-difficulty question were solved with item pre-knowledge, the conditional results were worse for the cheaters with low theta values, because the cheaters with high theta values ultimately were going to solve at least some of the compromised items without using item pre-knowledge.

DISCUSSSION AND LIMITATIONS

Test security, especially in high-stake standardized tests, is a fear for all test developers and test users, and violation of it is the biggest barrier to valid test use and reliable test scores (Yi, Zhang, & Chang, 2006). The main purpose of this study was to show the potential consequences of using item pre-knowledge on the estimated theta scores in a computerized multistage testing administration. This study does not aim to increase practitioners' fears about test security, but instead to encourage them to take additional precautions to increase test security in a computerized multistage testing administration. There is a consensus that adaptive tests increase test security, because test and/or item overlap is lower in adaptive tests and everyone works on his/her own pace (Luecht, &Sireci, 2011). In c-MST, this is basically done by specifying module exposure rate (i.e., by creating multiple panels). The modules that reach the maximum exposure rate, and the items within those modules, are no longer used in future administrations. Regardless of c-MST panel structure and number of panels, only routing modules (and so items within routing modules) are seen by all examinees assigned to that panel. Therefore, only routing modules reach the maximum exposure rate. The subsequent modules within the panels are used by fewer number examinees and thus also potentially used in future test administrations (Luecht, &Sireci, 2011). This creates the possibility that the items in those modules might be seen by some of the same examinees again in future exams, and that these test takers might solve them by using item pre-knowledge. This, of course, does not mean they are cheaters, but receiving the same item twice will increase the probability

of their solving items correctly, jeopardizing test validity and test fairness (Segall, 2004). Another possible test fraud is when test takers share test items after the exam with other, future test takers. In this case, the test taker does not necessarily cheat on another test administration, but others might cheat on those items if they receive them. This might seem unlikely. However, when the number of shared items increases, and if test takers deliberately memorize items (Segall, 2004), we see it is not impossible. Kaplan's employees memorized over 200 items in 2008 (Foster, 2013). A Chinese website (www.scoretop.com) was shut down for storing and posting Graduate Management Admissions Test (GMAT) items after each test administration. These recent examples and many other instances show that, unfortunately, organized item theft is possible in adaptive tests.

In order to explore the consequences of using item pre-knowledge on theta scores in a c-MST administration, we manipulated three test cheating scenarios. We also ran a baseline condition (i.e., null condition) to be able to compare the results yielded under these three scenarios with it. The study showed that when test cheating occurred during the test, depending on the test length condition, mean bias got worse—up to seven times on the 30-item test and seventeen times in the 60-item test (see mean biases in Cases 1 and 4 in Table 2). In terms of root mean square error, the results were up to two or three times worse than the baseline condition. These findings illustrates that test cheating might destroy the greatest advantage of computerized multistage testing, high measurement accuracy on the theta estimates. It is necessary not to forget the potential impact of cheating on pre-tested items (e.g., seeded items). As discussed by Meijer (1996), test cheating causes aberrant response patterns, harming item and person fit statistics (Meijer, 1996) and diminishing the accuracy of testing the properties of seeded items.

Another potential impact of test cheating is the module usage rates (i.e., how many people receive a module in a stage). In the case of no item pre-knowledge usage (Case 1), after receiving the routing module, the module usage rates for easy, medium, and hard modules in stages one and two should be roughly equal. In the case of item pre-knowledge usage (Cases 2, 3, and 4), depending on where the cheating occurred, the module usage rates

for the subsequent modules will be affected. The percentages of module usage rates across all study conditions are provided in Table 3. As hypothesized, regardless of the test length, the module usage in stage two in Cases 2, 3, and 4 were roughly equal with the rates in Case 1 (baseline condition), because no test cheating occurred in the routing module. However, since the test cheating occurred in stage two, the module usage rates and thus item usage rates in stage three changed. As can be inferred from Table 3, the module usage percentages for the difficult modules evidently increased, and this increase was more obvious under the 60-item test length condition.

Table 3. Results of Module Usage Rates as Percentages (%)

Module	**Case 1**		**Case 2**		**Case 3**		**Case 4**	
	30-item	60-item	30-item	60-item	30-item	60-item	30-item	60-item
Stage 2 Easy	43	42	44	41	43	41	43	42
Stage 2 Medium	14	15	14	15	14	15	14	14
Stage 2 Hard	43	43	42	44	43	44	43	44
Stage 3 Easy	46	46	33	23	45	23	40	38
Stage 3 Medium	12	11	15	20	10	20	11	12
Stage 3 Hard	42	43	52	57	45	57	49	50

It is important to note that this artificial increase in item usage rates in difficult modules also decreases the chance of those items being used in future administrations.

This study showed there are devastating consequences of test cheating on the estimated theta scores, and this should be strictly obstructed. As discussed in Foster (2013), several actions can be taken to prevent this. These include extending the time between test administrations, creating more panels, setting more conservative exposure rates, monitoring websites to prevent organized item theft, protecting databases from hackers, increasing test security in testing centers, preventing communication devices during the administration of tests, using multiple item pools, increasing items in the pools, randomizing the answer key by using discrete-option multiple-choice items (yes/no answer options), avoiding very long tests, and banning repeatedly retaking the test.

This study has several weaknesses. First, its design is too simple. This is because that our purpose was to show some statistical consequences of having item pre-knowledge in c-

MST. Thus, we were more focused on manipulating more cheating scenarios than other factors (e.g., different MST designs). Second, in similar studies in the literature, researchers usually generate thousands of examinees, and randomly select test cheaters from those generated examinees (see Wollack, Cohen, & Serlin, 2001). We intentionally generated the cheaters only because the purpose was to see the impact of test cheating on theta estimates for cheaters only. Further studies may want to generate thousands of examinees and select cheaters from those, and seek the same impact on overall results across the whole test taking population. Third, in this study we used a 1-3-3 panel design only. This was because it is one of the most popular and commonly preferred c-MST design in the literature (see Schnipke & Reese 1999; Zenisky, 2004). A future study may want to conduct the same study by using different panel designs (e.g., 1-2-3, 1-3-2) including two stage designs (e.g., 1-2, 1-3, 1-4 etc) and test how using item pre-knowledge affected the studied outcomes when the panel design varied. Lastly, even though the recent examples given in this study demonstrate that our manipulated cheating conditions may not be entirely fictitious, one can consider them the worst possible scenarios. It is also possible that some of them have never happened in a real-world high-stakes test administration. A future study could use more realistic conditions (e.g., fewer compromised items).

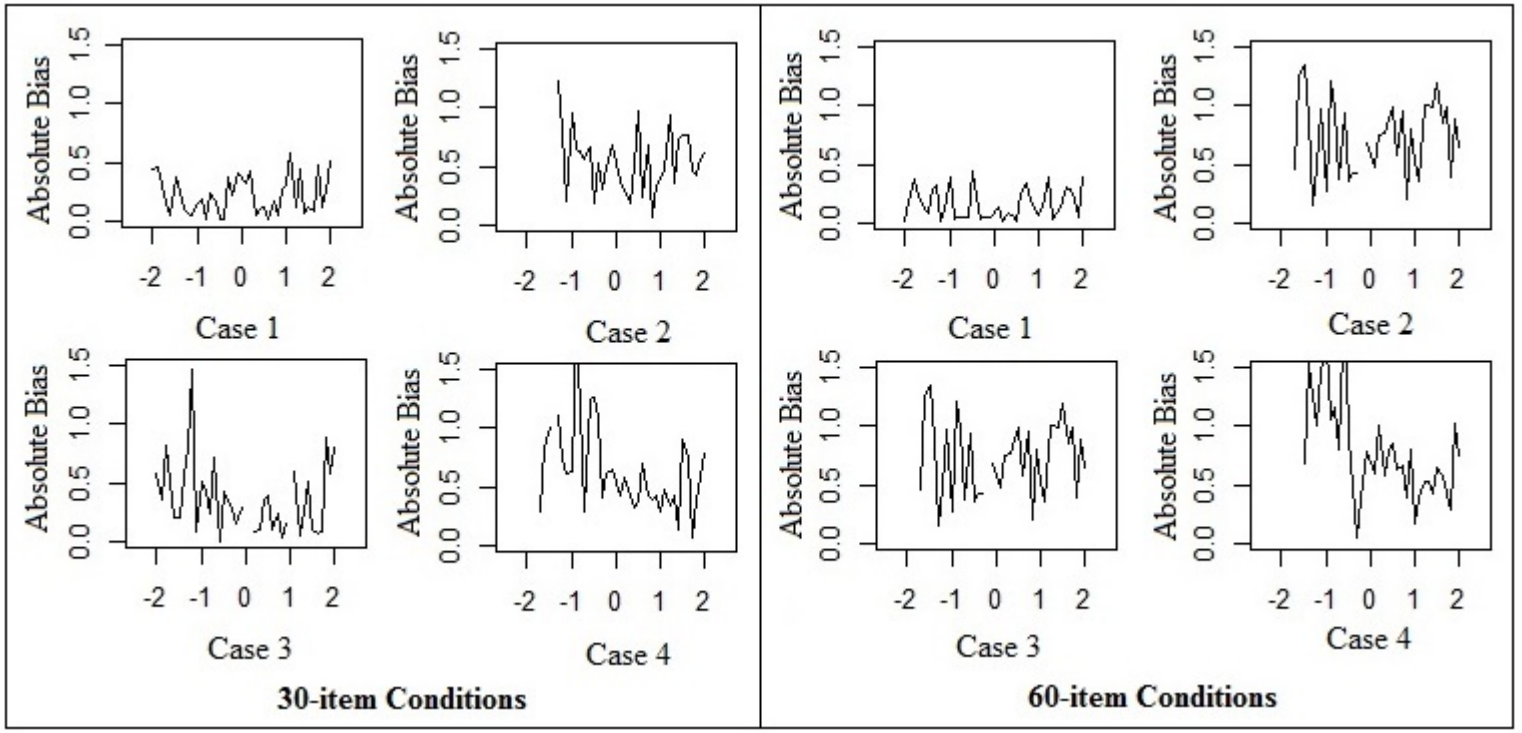

Figure 1. Conditional absolute biases across all study conditions.

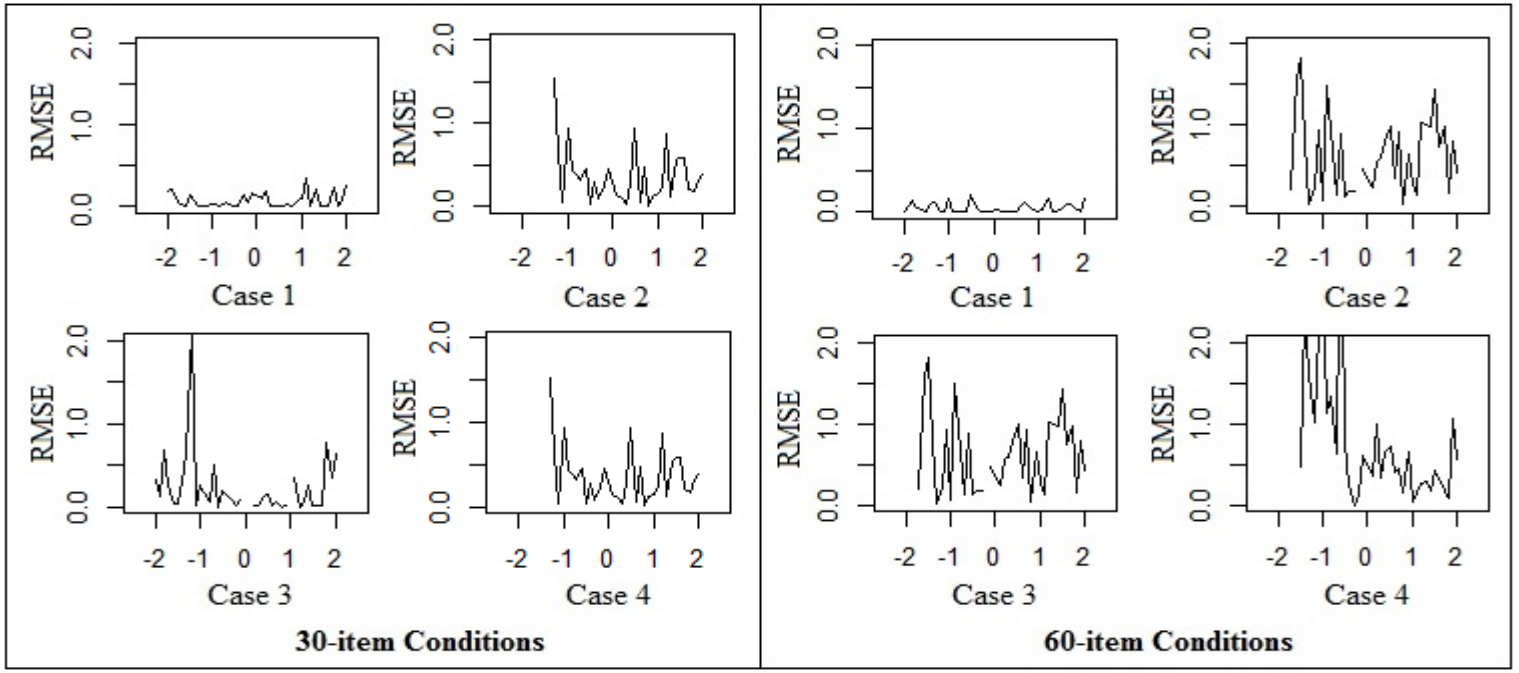

Figure 2. Conditional root mean square errors across all study conditions.

REFERENCES

Armstrong, R. D., Jones, D. H., Li, X., & Wu, L. (1996). A study of a network-flow algorithm and a noncorrecting algorithm for test assembly. *Applied Psychological Measurement*, *20*(1), 89-98.

Baker, F. (1992). *Item response theory*. New York, NY: Markel Dekker, INC.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick. In *statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bock, R. D., &Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied psychological measurement*, *6*(4), 431–444.

Diao, Q., & van der Linden, W. J. (2011). Automated test assembly using lp_solve version 5.5 in R. *Applied Psychological Measurement*, *35*(5) 398–409.

Foster, D. (2013). Security issues in technology-based testing. *Handbook of test security*, 39-83.

Guo, J., Tay, L., &Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small-scale cheating. *International Journal of Testing*, *9*(4), 283-309.

ILOG. (2006). ILOG CPLEX 10.0 [User's manual]. Paris, France: ILOG SA.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Luecht, R. M., &Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*(3), 229–249.

Luecht, R. M. &Sireci, S. G. (2011). A review of models for computer-based testing. Research Report RR-2011–12. New York: The College Board.

McLeod, L., Lewis, C., &Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*(2), 121–137.

Meijer, R. R. (1996). Person-Fit research: An introduction. *Applied Measurement in Education*, *9*, 3-8.

Schnipke, D. L., & Reese, L. M. (1999). A Comparison [of] Testlet-Based Test Designs for Computerized Adaptive Testing. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.

Segall, D. O. (2004). A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *29*(4), 439–460.

Team, R. (2016). RStudio: integrated development for R. *RStudio, Inc., Boston, MA.* Retrieved from http://www.rstudio.com.

Thissen, D., &Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101–133). Hillsdale, NJ: Lawrence Erlbaum.

Weiss, D. J., & Kingsbury, G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–375.

Weissman, A., Belov, D.I., & Armstrong, R.D. (2007). Information-based versus number-correct routing in multistage classification tests. (Research Report RR-07–05). Newtown, PA: Law School Admissions Council.

Wollack, J. A., Cohen, A. S., & Serlin, R. C. (2001). Defining error rates and power for detecting answer copying. *Applied Psychological Measurement*, *25*(4), 385-404.

Yan, D., von Davier, A. A., & Lewis, C. (Eds.). (2014). *Computerized multistage testing: Theory and applications*. CRC Press.

Yi, Q., Zhang, J., & Chang, H. H. (2006). Severity of Organized Item Theft in Computerized Adaptive Testing: An Empirical Study. *ETS Research Report Series*, *2006*(2), i-25.

Zenisky, A. L. (2004). *Evaluating the effects of several multi-stage testing design variables on selected psychometric outcomes for certification and licensure assessment* (Order No. 3136800).

Zopluoglu, C**.,** & Davenport, E.C. (2012). The empirical power and type ı error rates of the gbt and omega ındices in detecting answer copying on multiple-choice tests. *Educational and Psychological Measurement*, *72(6)*, 975–1000.

GENİŞ ÖZET

Amaç

Bu çalışmanın amacı, bilgisayarda bireye uyarlanmış çok aşamalı testi alan kişilerin madde ön bilgisi olduğunda, bunun kişilerin yetenek seviyelerine olan etkisini incelemektir. Çalışma özellikle büyük ölçekli test uygulamalarında madde hırsızlığı veya madde paylaşımının test puanlarına, test kullanımına ve puan yorumlarına ciddi şekilde zarar verebileceğini vurgulamayı ve deneysel olarak kanıtlamayı hedeflemektedir. Ayrıca bilgisayarlı çok aşamalı testlerde araştırmacılar, test geliştiriciler ve test kullanıcılarının test güvenliği konusundaki dikkatini çekmek çalışmanın amaçlarından biridir çünkü bilgisayar üzerinde yapılan uygulamaların kullanımı her geçen yıl hızla artmaktadır.

Yöntem

Çalışmada öncelikli olarak madde ön bilgisinin hiç kullanılmadığı yani null durumu yapay verilerle üretilmiştir. Bu durum çalışmada durum 1 olarak adlandırılmıştır. Testi alan kişilerin madde ön bilgisini nasıl kullandıklarını tam olarak bilmek mümkün olmadığı için 3 farklı durum daha oluşturulmuştur. Durum 2'de madde hırsızlığı yapanların önceden bildikleri maddeleri doğru cevaplama ihtimalleri 1 olacak şekilde veri üretilmiştir. Durum 3'te madde hırsızlığı yapanların daha zor maddeleri bildikleri varsayılmış ve buna göre veri üretilmiştir. Durum 4'te ise madde hırsızlığı yapanların madde ön bilgisini kullanma ihtimalleri normal dağılımla değişmiştir. Madde hırsızlığı yapan 30 kişinin yetenek seviyeleri normal dağılımla üretilmiş, üretilen madde havuzunun parametreleri Tablo 1'de verilmiştir. Testi alanların verileri ise 3 parametreli madde tepki kuramı ile üretilmiştir. Çalışmada 1-3-3 panel dizaynına sahip bilgisayarda bireye uyarlanmış çok aşamalı test kullanılmış, toplam test uzunluğu 30 ve 60 olacak şekilde, bireylerin madde ön bilgisini farklı durumlarda kullandıklarında yetenek seviyelerinin nasıl değiştiğine bakılmıştır. Bu koşulları manipüle etmenin amacı, farklı test hilesi senaryoları altındaki yetenek tahminlerinde madde ön bilgi kullanımının potansiyel zararını araştırmaktır. Nihai hedef, 2, 3 ve 4 numaralı durumların bulgularını birbirleriyle karşılaştırmak değil, onları 1 numaralı (madde ön bilgisinin kullanılmadığı) durumun sonuçları ile karşılaştırmaktır. Simülasyon sonuçları iki farklı istatistik grubuyla değerlendirilmiştir: (a) genel sonuçlar ve (b) yetenek seviyesine bağlı sonuçlar. Genel istatistikler için, ortalama yanlılık (mean bias), ortalama kareler hatası (RMSE) ve hesaplanan ve doğru yetenek seviyeleri arasındaki korelasyon hesaplanmıştır.

Bulgular

Ortalama yanlılık, ortalama kareler hatası, ve gerçek ve hesaplanan yetenek seviyeleri arasındaki korelasyon değerleri Tablo 2'de verilmiştir. Ortalama yanlılık ve kareler hatası açısından 3 durumda durum 1'den kötü çıkmıştır. Bu durumlar arasında ise

özellikle durum 2 ve 3'te oldukça kötü sonuçlar elde edilmiştir. Bunun yanı sıra test uzunluğu arttıkça daha yüksek değerler elde edilmiş ve sonuçlar kötüleşmiştir. Bunun temel nedeni test uzunluğu arttıkça çalınan veya paylaşılan maddelerin sayısının da artmasıdır. Ancak durum 1'de ve diğer durumlarda elde edilen korelasyon değerleri arasında çok fazla değişiklik olmadığı görülmüştür. Farklı yetenek seviyelerine bağlı ortalama yanlılık sonuçları Şekil 1'de gösterilmiştir. Farklı yetenek seviyelerine bağlı ortalama kareler hatası sonuçları Şekil 2'de gösterilmiştir. Genel sonuçlara paralel şekilde, test uzunluğunun artması, sonuçları daha stabil hâle getirmiş, farklı yetenek seviyelerindeki yanlılık ve ortalama kareler hatası miktarları birbirine yaklaşmıştır. Bununla birlikte durum 2'de sonuçların tüm yetenek seviyeleri için oldukça kötü olduğu görülmüştür. Bunun nedeni muhtemelen tüm yetenek seviyelerindeki kişilerinde benzer sayıda maddeyi önceden bilerek sınava girmesinden kaynaklanmaktadır.

Tartışma ve Kısıtlamalar

Test güvenliği, özellikle yüksek öneme sahip testlerde, tüm test geliştiricileri ve test kullanıcıları için bir korkudur ve bunun ihlal edilmesi geçerli test kullanımının ve güvenilir test puanlarının önündeki en büyük engeldir. Bu çalışmanın temel amacı, bilgisayar ortamında bireye uyarlanmış çok aşamalı bir test uygulamasında madde ön bilgisinin kullanılmasının veya test maddelerinin daha önceden bilinmesinin test puanlarına olan etkisini göstermektir. Bu çalışma, uygulayıcıların test güvenliğine ilişkin korkularını artırmayı değil, çok aşamalı bir test uygulamasında test güvenliğini arttırmak için ek önlemler almalarını teşvik etmeyi amaçlamaktadır.

Bu çalışma test maddelerinin önceden bilinmesi durumunda bilgisayarlı çok aşamalı testlerin en büyük avantajını tahrip edebileceğini, yetenek kestirimlerinin suni bir şekilde çok yüksek çıkabileceğini göstermiştir. BUÇAT uygulamalarında başlangıç modülü tüm bireyler tarafından ortak olarak alınmakta ancak diğer modüllerin alınma sayıları değişebilmektedir. BUÇAT'ta madde ön bilgisi kullanılması durumunda, bireyler artık gerçek yetenek seviyelerinin çok üzerinde modüller alacaklardır. Bu durumda da modüllerin kullanım oranları olması gerektiğinden farklı çıkacaktır. Buna dair sonuçlar da çalışmada verilmiş olup, madde ön bilgisine sahip olmanın sadece mevcut sınavı ve testi alanların yetenek seviyelerini değiştirmekle kalmayacağı aynı zamanda sonraki sınavlarda kullanılabilecek maddeleri de etkileyebileceği gösterilmiştir.