Practical Assessment, Research, and Evaluation

Volume 25 Article 3

2020

The Effect of Repeat Exposure to Simulation Based Items

Xiaodan Tang
UNiversity of Illinois at Chicago

Matthew Schultz
American Institute of Certified Public Accountants (AICPA)

Recommended Citation

Tang, Xiaodan and Schultz, Matthew (2020) "The Effect of Repeat Exposure to Simulation Based Items," *Practical Assessment, Research, and Evaluation*: Vol. 25 , Article 3.
DOI: https://doi.org/10.7275/bsxr-m225
Available at: https://scholarworks.umass.edu/pare/vol25/iss1/3

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 25 Number 3, January 2020 ISSN 1531-7714

The Effect of Repeat Exposure to Simulation Based Items

Xiaodan Tang, *University of Illinois at Chicago*

Matthew Schultz, *American Institute of Certified Professional Accountants (AICPA)*

This study aims to examine the potential impacts on repeat examinees' performance by reusing simulation-based items in a high-stakes standardized assessment. We examined change patterns of item scores, ability estimate, score pattern change, response time and compared the performance of repeat examinees who have received repeat items and those who haven't. Results suggest that there are limited benefits from encountering the same items. The practical implications to licensing/certification assessments are discussed.

Repeated item exposure has long been of concern, especially in licensing or certification test organizations, where exams are offered on a repeat basis. Ideally, examinees would receive a different set of items each time they take the test. However, this is not always possible, especially for tests containing performance assessment items or technology-enhanced innovative item types, which can be complicated as well as expensive to develop, resulting in limited item inventories for the following reasons. First, in high-stakes large-scale testing, the examinee volume is usually large so that the number of repeaters might be high, especially for exams with relatively low passing rates. For performance assessment items, there might be limited scenarios or resources that can be used as item prompts or stems. For this reason, it is expensive and time-consuming to develop such performance assessment items and hard to maintain an item bank large enough to deliver different or unique items to all repeaters. Finally, when using a panel assembly approach to construct test forms, in addition to item exposure constraints, psychometricians also need to balance many other constraints to yield optimal panels.

Taken together, it is common and inevitable to have the same items delivered to repeat examinees. Importantly however, it may raise some concerns regarding test validity and fairness. It is intuitive to assume that examinees encountering the same items might benefit from an unfair advantage by remembering some items at their first attempt and then searching for answers to prepare for their subsequent attempts. In this case, they would be more likely to answer items correctly, thus causing a test validity issue to the extent their score on the repeat exam is inflated due not to increased knowledge of the underlying construct, but rather due to their ability to remember the previously seen or repeat items. This unfair score advantage may result in inflated ability estimation and false positives in terms of pass-fail decisions. Therefore, it is crucial to identify any unfair score advantages for repeat examinees so that test developers can address any issues by expanding item inventories, modifying test specifications, etc. The purpose of this study is to examine whether prior exposure to certain simulation items or performance items has any impact on subsequent/repeat performance in the context of high-stakes standardized testing.

Brief Review of the Literature and Research

This section reviews and unpacks previous research on the impact of prior item exposure on repeat examinee performance in terms of score gains, response time, and score change patterns. As has been demonstrated, prior item exposure may pose a threat on test fairness and validity, which is seen as a big concern to test organizations. However, this threat is not warranted in most of the studies reviewed. Despite a frequently seen score increase in repeat examinations for the studies reviewed, this score increase is generally not an unfair advantage brought by repeatedly encountering the same test or the same items.

In a series of studies on the impact of reusing the same test for repeaters, Raymond, Neustel, and Anderson (2007; 2009) compared the effects of administering the identical and parallel exam forms. Although examinees receiving the identical exam form during their repeat attempt obtained higher scores than their initial scores, these score gains were indistinguishable from those receiving a parallel exam form of different items from their initial exam. Based on these results, the researchers claimed that it might not be necessary to concerned about unfair score advantages if test organizations or licensure boards plan to administer the same exam form for repeat examinees.

To examine score gains for repeaters' subsequent attempts in more detail, Chavez, Swygert, Peitzman, and Raymond (2013) applied a locally weighted scatterplot smoothing (LOESS) technique and a piecewise regression to illustrate whether such score gains can be explained by the within-session score increase. They discovered a score increase over the first few items for both single-take and repeat examinees, indicating a temporary warm-up effect within each attempt. Further, they revealed that the across-attempt score increase was more likely due to true ability improvement rather than the warm-up effect.

Raymond, Neustel, and Anderson (2007; 2009) further examined another indicator of retesting effect, response time, which might unfold examinees' responding behaviors toward items they encountered before. The researchers compared total testing time between examinees' first and second exam attempts and between identical and parallel forms and then found that repeat examinees on the identical form had shorter total response time than those who received the parallel form. In other words, repeat examinees tended to more quickly respond to items they came across on their initial attempt. Later, in another study examining retesting effect, Feinberg, Raymond, and Haist (2015) found that shorter response time was associated with incorrect-correct response pattern presented on both reused and new items, which spoke to the contention that the shorter response time was less likely due to the repeated exposure.

Based on another strategy to investigate the retesting effect, Hertz (2003) conducted a Rasch analysis of item parameters and person ability estimates and disclosed no large differences among four test administrations of the same exam. This result suggested that examinees did not benefit from receiving the same test content information on their repeat attempts. To have a closer look at the validity of the repeat test for multiple-take examinees, Raymond, Kahraman, Swygert, and Balog (2011) found that the criterion validity of test scores improved for repeat examinees on their second attempt by correlating their scores with other related exams and comparing confirmatory factor analysis results across subdomains between single-take and repeat examinees. In other words, the repeat test score would more accurately reflect an examinee's true proficiency.

The above research looked at the cases of re-administering the entire test for repeaters. In some licensing or certification exams, it is more common to only repeat some items rather than to administer the identical form, which provides an impetus for researchers to evaluate the impact of the repeat use of some items rather than entire forms. For example, Wood (2009) explored the reuse effect by randomly mixing some reused items with new items. The results revealed that repeat examinees achieved similar score increases on both reused and new items. Similarly, Wagner-Menghin, Preusche, and Schmidts (2013) examined the effect of reusing some items based on the Rasch modeling analysis by comparing item difficulties across different examinee samples and concluded that exam quality would not worsen when a low ratio of randomly selected items was reused. Contrary to the above findings based on the large-scale standardized testing, Joncas, St-Onge, Bourque, and Farand (2018) examined the impact of reusing some

items in classroom assessments. They claimed that reusing an item several times within a short time may pose a threat on the exam quality.

The reviewed literature collectively concludes with an argument that repeaters' score gains on subsequent attempts were less likely due to memorizing the items but more likely due to true ability improvement and possibly the existence of a warm-up effect in some situations. However, most of these studies investigated exams composed of only multiple-choice items. Rather than being exposed to a long exam of many multiple-choice items requiring a short response time for each item, examinees during a performance assessment need to work on the same item for a longer time, which leads to a longer exposure to each item. Additionally, they are presented with fewer questions compared to the length of a multiple-choice test. In this case, repeat examines would have longer time per item to memorize if they intend to do so.

Some studies have investigated the repeat exposure effect of performance assessment items. In concert with previous studies on multiple-choice exams, there was limited advantage for repeat examinees due to prior exposure of the same performance assessment items. For example, Boulet, McKinley, Whelan, and Hambleton (2003) analyzed repeat examinee performance based on an exam composed of scenario-based items with a response time of around 15 minutes for each item. Each item described a scenario in which examinees need to gather data, perform analysis, and take notes in order to successfully answer the questions. Repeat examinees received higher scores on repeat attempts, but this score gain was not attributable to encountering the same items. The reason is that a higher score gain was found for new items than previously seen items. Similarly, Rambler and Schultz (2017) found that repeat examinees who passed a high-stakes exam at their second attempt typically had an increase in performance on simulation-based items, however their score increase at the second attempt was not due to repeat exposure. In another study conducted by Swygert, Balog, and Jobe (2010) regarding a high-stakes performance assessment, they observed repeat score gains for both examinees who received reused items and those who received all new items. Moshinsky, Ziegler, and Gafni (2017) echoed the same findings based on a high-stakes non-cognitive test delivering multiple mini-interviews. Driven by another technique, Raymond, Swygert, and Kahraman (2012) compared score consistency of a performance assessment for repeat examinees among subdomains and found that for low-performing repeat examinees, their repeat scores were more consistent across subdomains than their initial scores. This finding bolstered the test validity of repeating a performance assessment.

Some studies examined the relationship between repeat examine performance and time lags between attempts. For example, Wilson (1987) observed that examinees' repeat scores increased, and longer time lags between two attempts were associated with larger score gains. This association spoke to a potential true ability improvement or practice effect as examinees may spend time studying the knowledge. Similarly, Geving, Webb, and Davis (2005) noted the pattern that the number of days between attempts was positively related to score gains of repeat examinees.

Response pattern, as an indicator of examinee behaviors, may also convey information on score changes of repeat examinees. Wood (2009) found that although repeat examinees tended to choose the same response option on subsequent attempts, the proportion of wrong to right pattern was lower than other patterns. Wood (2009) stated that the score increases on both reused and new items reflected a true ability improvement or practice effect rather than memory effect, and further explained that examinees' potential stress or poor testing strategies involved in their initial attempt might be alleviated on their second attempt.

In general, there was limited evidence supporting an unfair score advantage repeat examinees may receive when they encounter the same items or the same test at their repeat attempts. In addition, most of the abovementioned exams were certification or licensure exams with a relatively high pass rate of at least 70% for first-time examinees (e.g., Feinberg et al., 2015; Raymond & Luciw-Dubas, 2010). Thus, there is a need to enrich the literature by focusing on licensure and certification exams with relatively low pass rates. Additionally, most previous studies investigated the effect of reusing some multiple-choice items and the effect of repeating the entire performance assessment. As such, there is a lack of research on the effect of repeating some performance assessment items for a

test with a low pass rate composed of both multiple-choice questions and performance assessment items. This combination has become a common test structure in the wake of the popularity in innovative item types or technology-enhanced item types. An analysis of performance of repeat examinees on reused items and a comparison with those who received all new items would hold substantial promise for filling this research gap by informing the impact of repeat item exposure. To achieve this purpose, this study intends to examine the effect of repeating some of the performance assessment items in the context of high-stakes standardized testing.

Data and Analysis

The data of this study come from a high-stakes licensing exam comprised of four sections. Each section is independently delivered and scored. Examinees can take the exam for each section once within a testing window which lasts for a quarter of each year. This study investigates the exam data for 26407 examinees who have taken and repeated during four testing windows starting from the second quarter of 2017 to the first quarter of 2018. The specifications of test blueprints call for delivery to both content specifications as well as skill levels. The skill framework is based on the revised Bloom's Taxonomy of Educational Objectives (Krathwohl, 2002) including remembering and understanding, application, analysis, and evaluation. The test items are calibrated, and item responses are scored based on a 3-PL IRT model. It should be noted that because this exam is a pass/fail exam, only individuals who fail at first attempt make a subsequent one.

The structure of this exam contains both multiple-choice items (MCQs) and simulation-based items. This study will focus on the repeat exposure of simulation-based items, which are typically condensed case studies that test real life, work-related situations. They typically require examinees' capacity to process data by software, apply domain knowledge to solve problems, and/or use the provided literature to answer questions. Also, they allow examinees to demonstrate their knowledge and skills by generating responses to questions rather than simply selecting the correct answer. Some response options are open-ended, and others are selected from a drop-down list. Each simulation-based item usually has 6-8 questions scored dichotomously. As such, each item has a score range of 6-8 points from the perspective of binary scoring. When repeat examinees take the exam again, they might receive the same simulation-based items they have encountered before. Since simulation-based items belong to a type of performance-based items, examinees might work on the same item for a longer period of time due to their complexity, sometimes as much as 20 minutes per item. Examinees are also presented with fewer items in total (compared to MCQs), and thus their exposure to each item is comparatively long. Due to these features, there is a need to examine whether repeat examinees may receive an unfair advantage when they encounter the same simulation-based items on a subsequent attempt.

Among all the simulation-based items administered during the four testing windows, we find 125 reused items, which had been exposed to the same examinee at least twice. Note that any one examinee may see zero, one, or more than one simulation-based items repeated on a subsequent attempt. Table 1 specifies the frequency of repeat items examinees have encountered. The samples of this study are labeled as follows: B1 denotes multiple-take examinees with no repeat simulation-based items (N = 10274); B2 denotes multiple-take examinees with repeat simulation-based items (N = 16133).

Table 1. The frequency of repeat simulation-based items each examinee encountered across the four sections

# of repeat simulation-based items	Number of examinees			
	Section 1.	Section 2.	Section 3.	Section 4.
1	2644	621	2998	2625
2	3016	203	1185	1875
3	407	4	334	763
4	122	0	120	333
5	32	0	24	103
6	5	0	5	39
7	1	0	1	12

Results

Repeat item score change.

A set of paired samples t-tests were performed to examine the repeat item score changes for examinees between their initial and subsequent attempts. The results revealed positive and statistically significant differences for B2 group examinees between their initial and repeat scores of the same simulation-based items as shown in Table 2. Specifically, if examinees attempted their repeat exam during the subsequent exam window right after their first attempt, they tended to have larger score point increase ($\Delta = 0.516$) for each repeat simulation-based item than when the second attempt was after two windows ($\Delta = 0.489$) and three windows ($\Delta = 0.485$) based on the score metric of a simulation-based item ranging from 6-8 points. In other words, score gains of the reused items faded a little when the time delay between attempts became longer. The effect size corresponding to each score change had small effect (i.e., around .2), suggesting this difference may not be practically meaningful. However, the direct comparison of item scores failed to account for item parameters (i.e., item difficulty, item discrimination, guessing) of the 3-PL IRT modeling so that the different score gains might be due to diverse item parameters among items. Hence, we further compared the ability estimate changes in the following section.

Table 2. Score changes of repeat simulation-based items between B2 repeat examinees' first attempt and their subsequent attempts

Time interval	Score change				Cohen's d
	M	SD	t	p	
One window	.516	1.478	37.493	<.01	.326
Two windows	.489	1.512	35.367	<.01	.305
Three windows	.485	1.502	19.088	<.01	.323

Note. The score of each item is ranged 6-8 points.

Ability estimate change.

We compared ability estimates of reused simulation-based items, new simulation-based items, and all MCQ items for B2 group examinees over their first and subsequent attempts to unfold whether their ability estimates showed consistent patterns with their score gains. The results (see in Table 3) showed that the differences between two attempts for reused, new simulation-based items, and MCQ items were all positive and statistically significant. That is, the ability estimates of not only reused simulation-based items but also new simulation-based items increased. Moreover, incongruently with the score changes, the ability estimates of reused simulation-based items increased as time lags between attempts became longer. It suggests that examinees tended to perform better on the reused items if they waited for longer time to conduct their second attempt. MCQ ability change was consistent with the increasing ability change pattern. In contrast, the ability change of new simulation-based items decreased as the time interval went longer. In terms of effect size, the ability estimate changes showed moderate effect. These results may

Table 3. A comparison of ability estimate changes of B2 repeat examinees between first and subsequent attempts

Time interval	Reused simulation-based item ability change			New simulation-based item ability change			MCQ ability change		
	M(SD)	t	Cohen's d	M(SD)	t	Cohen's d	M(SD)	t	Cohen's d
One window	.249(.799)	16.025*	.312	.200(.903)	11.329*	.222	.240(.554)	22.124*	.432
Two windows	.421(.760)	52.595*	.553	.357(.847)	40.032*	.422	.361(.610)	56.215*	.592
Three windows	.480(.779)	46.373*	.617	.387(.885)	32.851*	.434	.447(.642)	52.220*	.695

Note. *p<.01

bolster the argument that the potential benefit brought by seeing the same items is probably not remembering the contents and studying for particular items (i.e., memory effect) after repeat examinees' first attempt, as it should fade away over time, but having exposed to a previously unfamiliar type of item and contents and then learning to respond better at their second attempt (i.e., practice effect). In other words, score gains may be due to a general improvement in the mastery of the content knowledge after repeat examinees make more practice within the time interval of attempts.

Response time.

We further examined the amount of time repeat examinees spent answering the new and repeat items. The results showed that examinees tended to spend more time on their second attempt for the same items (see in Table 4). This small difference (8 second difference out of 900 seconds for an item on average) would be more likely to be statistically significant due to the large sample size, but it may have fewer practical implications. With regard to the time spent on all repeat items in the case of encountering more than one repeat item at the later attempts, repeat examinees tended to spend less time than their first attempts. That being said, examinees spent slightly longer time for each repeat simulation-based item; but when considering the time they spent on all repeat simulation-based items if they encountered more than one repeat item, it took them less time to complete all repeat simulation-based items than their first attempt. Given small to none effect sizes, these response time differences were less meaningful to support a strong relationship between response time and repeat exposure. Further, we discovered that repeat examinees tended to spend more time on new items and all simulation-based items at their second attempt.

Although there were different patterns on response time between first and subsequent attempts, response time might also be impacted by item order. In general, examinees might be speeded on the last few items. If reused items were administered at the end of a test panel, examinees might be speeded on these items due to a time limit. To examine whether the longer response time is related to item order and previous item exposure, we conducted a regression analysis by considering both item order and the number of attempts as predictors (see in Table 5). The results showed that after accounting for item order, the number of attempts no longer significantly predicted response time ($F = 384.377$, $p < 0.01$, $R^2 = .013$). As such, the response time of each reused item was not related to whether examinees encountered this item before.

Table 5. Regression results for a model specifying the response time of repeat items is only significantly related to item order rather than the number of attempts examinees conducted.

	Response time of repeat items			
Variables	Coeffi-cient	SE	T	p
Intercept	1118.109	23.507	47.565	.001
First attempt	-11.188	23.365	-.479	.632
Second attempt	4.324	23.366	.185	.853
Item order	-31.366	.925	-33.894	.001

Score change pattern.

In general, 47% examinees received lower scores on their subsequent attempts than their first attempt, and 30% kept the same scores as before. The score increase pattern occurred less frequently (23%) than the same score or score decrease pattern. It suggests that the repeat exposure of simulation-based items didn't benefit a majority of examinees as they rarely had score change from incorrect to correct. This result may serve as an indication of no occurrence of memory

Table 4. Response time spent on simulation-based items (in seconds) by B2 repeat examinees

	A single reused item			All reused items			New items			All simulation-based items		
	M(SD)	t	Cohen's d	M(SD)	t	Cohen's d	M(SD)	t	Cohen's d	M(SD)	t	Cohen's d
Response time change (in seconds)	8 (505)	2.780*	.017	-75 (947)	-12.884*	.281	405 (1439)	45.727*	.235	329 (1401)	38.226*	.079

Note. *p<.01.

effect for most of the repeat examinees. We further conducted ANOVA analysis and we found that there were significant differences in the score changes of reused simulation-based items across skill components ($F = 5.994$, $p < .01$) and item types ($F = 87.357$, $p < .01$). We than further conducted a pairwise post-hoc analysis. As shown in Table 6, in terms of skill components, items requiring the analysis skill defined by the Bloom's taxonomy were associated with higher score gains than those measuring application. It may indicate that examinees intend to focus more on content at the level of analysis when remediating their levels of knowledge after their first attempt.

Ability estimate change of B1 vs B2.

We conducted independent sample t-tests to compare the ability estimate changes of simulation-based items between B1 and B2 groups by each test section as shown in Table 7. We also did this contrast on MCQ items. It was found that, across the four sections, B2 group had a slightly higher increase of MCQ ability than B1 group, which was the same case for the simulation-based item ability. The consistent pattern of MCQ and simulation-based items may confirm that B2 group examinees have a general ability improvement shown throughout the exam. Further, although the ability change differences of both MCQ and simulation-based items for the two groups were statistically significant on most of the exam sections, effect sizes ranged from very small to none, which failed to speak to a strong size of ability gain for the B2 group who encountered repeat items. To support this argument with more evidence, we plotted the ability estimate changes between the two groups for MCQ and simulation-based items (see in Figure 1). The data points in the figure indicated that individual ability estimate changes mostly overlapped between the two groups across the four sections, which seems almost indistinguishable. In other words, the mean ability estimate changes were comparable between examinees encountering the same items and those receiving all new items. Therefore, prior exposure to simulation-based items did not yield any major unfair score or ability estimate advantage for the B2 group as they performed similarly as the B1 group across the four test sections.

Table 6. Pairwise post-hoc comparison results of score changes for B2 group across item types and the measured skills.

Pair	Score change mean difference	p
Application – Analysis	-.084	<.01
Data analysis items – References items	.420	<.01

Practical Implications

This study examines the impact of the repeat exposure of performance-assessment items on the performance of repeat examinees. In general, our results echoed the conclusions of previous research that repeat examinees benefit from encountering in a modest sense. Although this study has parallels to prior research, it also expands the current literature by examining the impact of the reuse of some performance assessment items in a high-stakes licensure exam with a relatively low pass rate. Specifically, repeat examinees were found to have increase in their scores of reused simulation-based items on their subsequent attempts. This increase can be accounted for by several reasons: random measurement error, regression to the mean due to test unreliability, construct-irrelevant factors including memory effects, test anxiety, practice effects, and true

Table 7. The differences of MCQ and simulation-based item ability change between B1 and B2 groups

Section	Differences of MCQ item ability change between B1 and B2 groups			Differences of simulation-based item ability change between B1 and B2 groups		
	M (SD)	t	Cohen's d	M (SD)	t	Cohen's d
1.	-.041 (.015)	-2.663*	.062	-.025 (.018)	-1.409	.033
2.	-.072 (.023)	-3.044*	.118	-.113 (.034)	-3.296*	.128
3.	.041 (.013)	3.032*	.070	.072 (.018)	3.975*	.092
4.	-.013 (.017)	-.786	.021	-.028 (.022)	-1.270	.034

Note: * p<.01

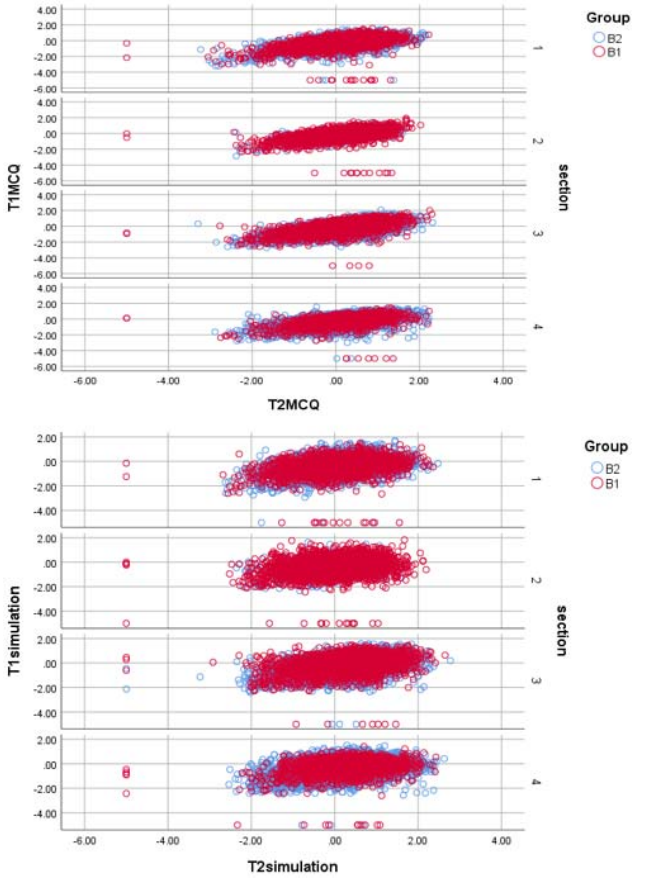

Figure 1. Upper: MCQ ability estimate change between B1 and B2. Bottom: Simulation-based item ability estimate change between B1 and B2.

ability increase. As high-stakes large-scale tests have been usually quality controlled by content developers and psychometricians, we were more concerned with memory effect rather than test unreliability issues. In other words, we would like to focus on whether repeat examinees' score increase is due to memory effect, which refers to memorizing and studying the particular item content that may produce the score increase on their subsequent attempts. To examine the existence of a memory effect, we conducted a set of analyses to study score changes as a function of repeat exposure. First, we observed score gains between examinees' first and second attempts on the reused simulation-based items. However, the analysis of reused item scores is not a direct reflection of examinees' ability as not every examinee received reused items with the same level of difficulty (as seen in item parameters). After undertaking a more direct comparison of ability estimates, we found that the ability estimates of all items (i.e., reused, new, MCQs) increased, suggesting a true ability improvement for repeat examinees rather than memory effect.

Second, as another indicator of the minimal likelihood of memory effects driving the observed results, the relationship between the time lag, which spans from the first exposure of items to the retest of the same items, and the score or ability increase might imply whether students use their memorized contents to search and remember the answers. The findings showed that the higher ability estimate increases were related to longer time lags. It implies that an immediate repeat attempt would not bring a better performance than a latter subsequent attempt, which again would potentially be indicative of practice effect and enhanced learning of the construct(s) rather than memory effect.

Third, in addition to the time interval between administrations, we looked more closely at how response time was impacted by prior item exposure. It was found that repeat examinees tended to spend more time at their second attempt on a single reused item but spend less time on all reused items and all items. However, this relationship was precluded after considering the item order. Response time was more related to item sequence than whether it was repeat examinees' first-time encounter.

Fourth, investigating score pattern changes from initial to subsequent attempts may contribute to understanding repeat examinees' knowledge levels at each attempt. We found that nearly half of repeat examinees who encountered repeat simulation-based items performed worse in terms of their score change patterns. The reason might be that examinees were feeling stressful when being challenged again by the items on their previously failed attempts. This finding suggests that prior simulation-based item exposure does not necessarily inflate repeat examinees' test scores, and it serves as evidence of no memory effect. Further, the analysis of score pattern change between item types and skills informs content developers and psychometricians of higher score gains on certain exam elements to further examine if it is necessary to check the issues of item or content overexposure or leakage.

Finally, the comparison of repeat examinees seeing some reused items and those seeing completely new items verifies the consensus that encountering the

same items will not yield prominent advantages over encountering all new items. It further supports that memory effect may rarely occur and the score and ability increase would be largely due to students' true ability improvement and greater familiarity or comfort and understanding of the item formats in question Since examinees may have personal and professional benefits from passing this high-stakes licensing exam and there is an exam fee for each attempt, it is reasonable to contend that examinees strive for enhancing their levels of knowledge and skills in an effort to pass the exam rather than memorizing some items not necessarily repeated on the next attempt.

Practically speaking, the results of this study may to some extent alleviate the concerns toward repeat exposure of some performance assessment items in high-stakes large-scale exams. Given that it is time consuming and expensive to develop performance assessment items and simulation-based items, the findings of this study help testing companies or organizations save these resources and support the strategy of administering the same performance assessment items to repeat examinees when necessary. Although the repeat exposure seems to have small impact on repeat examinee performance, it is still important to control item exposure, for instance, by increasing the mix of reused and new items. In the age of the Internet, test developers may assume that all items of any test context can be exposed. Hence, in order to maintain test validity, reliability, and fairness, it is important to carefully determine testing procedures contexts and refine retesting policies.

References

Boulet, J. R., McKinley, D. W., Whelan, G. P., & Hambleton, R. K. (2003). The effect of task exposure on repeat candidate scores in a high-stakes standardized patient assessment. *Teaching and Learning in Medicine, 15*(4), 227-232. doi:10.1207/S15328015TLM1504_02

Chavez, A. K., Swygert, K. A., Peitzman, S. J., & Raymond, M. R. (2013). Within-session score gains for repeat examinees on a standardized patient examination. *Academic Medicine, 88*(5), 688-692. doi:10.1097/ACM.0b013e31828af039

Feinberg, R. A., Raymond, M. R., & Haist, S. A. (2015). Repeat testing effects on credentialing exams: are repeaters misinformed or uninformed? *Educational Measurement: Issues and Practice, 34*(1), 34-39.

Geving, A. M., Webb, S., & Davis, B. (2005). Opportunities for repeat testing: Practice doesn't always make perfect. *Applied HRM Research, 10*(2), 47-56.

Hertz, N. R., & Chinn, R. N. (2003). *Effects of Item Exposure for Conventional Examinations in a Continuous Testing Environment.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME) Chicago, IL.

Joncas, S. X., St-Onge, C., Bourque, S., & Farand, P. (2018). Re-using questions in classroom-based assessment: An exploratory study at the undergraduate medical education level. *Perspectives on Medical Education, 7*(6), 373-378. doi:10.1007/s40037-018-0482-1

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4), 212-218.

Moshinsky, A., Ziegler, D., & Gafni, N. (2017). Multiple mini-mnterviews in the age of the internet: does preparation help applicants to medical school? *International Journal of Testing, 17*(3), 253-268. doi:10.1080/15305058.2016.1263638

Ramler, P., & Schultz, M. (April 2017). *Can subscore performance predict future test success?* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Antonio, TX.

Raymond, M., Kahraman, N., Swygert, K. A., & Balog, K. P. (2011). Evaluating construct equivalence and criterion-related validity for repeat examinees on a standardized patient examination. *Academic Medicine, 86*(10), 1253-1259. doi:10.1097/ACM.0b013e31822bc0a4

Raymond, M., & Luciw-Dubas, U. A. (2010). The second time around: accounting for retest effects on oral examinations. *Evaluation and the Health Professions, 33*(3), 386-403. doi:10.1177/0163278710374855

Raymond, M., Neustel, S., & Anderson, D. (2007). Retest effects on identical and parallel forms in certification and licensure testing. *Personnel Psychology, 60*(2), 367-396.

Raymond, M., Neustel, S., & Anderson, D. (2009). Same - form retest effects on credentialing examinations. *Educational Measurement: Issues and Practice, 28*(2), 19-27.

Raymond, M., Swygert, K., & Kahraman, N. (2012). Measurement precision for repeat examinees on a standardized patient examination. *Advances in Health Science Education Theory and Practice, 17*(3), 325-337. doi:10.1007/s10459-011-9309-0

Swygert, K. A., Balog, K. P., & Jobe, A. (2010). The impact of repeat information on examinee performance for a large-scale standardized-patient examination. *Academic Medicine, 85*(9), 1506-1510. doi:10.1097/ACM.0b013e3181eadb25

Wagner-Menghin, M., Preusche, I., & Schmidts, M. (2013). The effects of reusing written test items: a study using the rasch model. *ISRN Education, 2013*, 1-7. doi:10.1155/2013/585420

Wilson, K. M. (1987). Patterns of test taking and score change for examinees who repeat the test of english as a foreign language. *ETS Research Report Series, 1987*(1), i-68.

Wood, T. J. (2009). The effect of reused questions on repeat examinees. *Advances in Health Sciences Education Theory and Practice, 14*(4), 465-473. doi:10.1007/s10459-008-9129-z

Citation:

Tang, Xiaodan and Schultz, Matthew (2020). The Effect of Repeat Exposure to Simulation Based Items. *Practical Assessment, Research & Evaluation*, 25(3). Available online: https://scholarworks.umass.edu/pare/vol25/iss1/3/

Corresponding Author

Xiaodan Tang
Department of Educational Psychology
University of Illinois at Chicago
1040 W. Harrison St. (MC 147)
Chicago, IL 60607

email: xtang1322 [at] gmail.com>