The Susceptibility of Performance Items to Exposure

Russell W. Smith

Alpine Testing Solutions

Presented at

The 9th Conference of the International Test Commission

San Sebastian Spain

July 4, 2014

Contact: Russell.Smith@alpinetesting.com

Introduction

Regardless of the specific terminology utilized to characterize unintended advantages — exposure, prior knowledge, cheating, item parameter drift — security concerns in information technology (IT) certification programs are pervasive. Items, forms, and even entire item banks are often available on the Internet within days of exam publication (Smith, 2005). As one way of minimizing security risks, Cizek (1999) suggests using "a variety of other assessment approaches" and notes that "performance assessments require students to actually demonstrate their knowledge or skill" (p. 168). The content domains and computer-based delivery mode of IT certification exams lends such exams to the development and administration of performance item types, often simulations or emulations. However, little is known about the susceptibility of these item types to exposure.

Sireci and Zenisky (2006) hypothesize that performance items are more memorable, and therefore more susceptible to exposure. However, Harmes and Wendt (2009) suggest that little empirical evidence exists to either support or refute Sireci and Zenisky's claim. They investigated the memorability of innovative items and concluded that "the surface characteristics of the innovative items may be more memorable than a multiple-choice item but the specific details were not" (p. 19). One counter argument of Sireci and Zenisky's (2006) claim is that performance items are at less risk of exposure because the test taker still needs to complete the task even if the test taker has prior knowledge of the task itself.

The ubiquitous cheating in IT certification threatens validity. As the use of performance items increases, it is particularly important to understand the nature of the differences between these items and more traditional selected response (SR) items. Additionally, it is key to understand the impact of exposure on item banks (Jones & Smith, 2005). It is also worthwhile to investigate ways to mitigate threats to validity. The purpose of this research is to investigate the susceptibility of performance items to exposure relative to SR items across multiple exams and certification programs. Specifically, performance and simulation items are compared to traditional SR items (e.g., single response multiple choice items) and less traditional SR items (e.g., hot spot graphic items where the correct answer is selected by clicking on an image) that may or may not have a more memorable format. These comparisons may result in a better understanding of the susceptibility of various item types to exposure, which could lead to improved processes that would strengthen the validity of the interpretations of the results of these examinations.

Methods

The data for this study are from seven IT certification exams within three different certification programs offered by three different test sponsors. Each exam is administered via computer, available on demand, and administered globally. Though the examinations may be translated to multiple languages, the analyses in this study are limited to the English language version. Selected examinations include a combination of performance items and SR items that were released concurrently as unscored pilot items.

For each examination, items were classified into one of the following groups: simulations, performance, selected response (SR), and selected response plus (SR+). Simulation items include a virtual or simulated hardware or software environment, are automatically scored, require candidates to perform a process, have logic built in, have multiple correct answers and solutions, and are all worth more than one point (though could be scored dichotomously or polytomously). Performance items are similar to simulation items in that they require candidates to perform a process, have logic built in, may have multiple correct answers and solutions, and are worth more than one point. However, these items are not associated with a simulated hardware or software environment. Selected response (SR) items are traditional multiple choice items, may include multiple select type items (e.g., choose 2, choose 3), are all worth one point, and have only one correct response. Selected response plus (SR+) items are all worth one point and have only one correct response but may include a graphic or hotspot graphic, drag and drop, or a matching list. While some may consider SR+ items to be “innovative,” these items are essentially a SR item with a different format and are quite common in IT certification. Each examination in this study includes some combination of these four item types.

Two different statistical approaches are used to evaluate the impact of exposure on each of the examinations and item types: the change in *p*-value and item measure displacement. In order to place dichotomous SR and SR+ and multi-point simulation and performance items for which the total number of points varied onto the same scale, the *p*-value was defined as the proportion of points available. The *p*-value change analysis included only items with a sufficient sample size that were initially administered with the performance items. The number of administrations was determined for each examination based on a reasonable balance between the number of administrations per item and a sufficient number of items of each type. The number of administrations per item was not consistent across all items within all examinations as not all items were administered to all candidates due to the administration of multiple forms and/or the random seeding of unscored items. The *p*-value, or proportion of points available for multi-point items, was calculated for the first 50 administrations of each item and the last 50 administrations, up to the maximum for each examination. Therefore, the change in *p*-value is based on counts for each item and does not reflect changes by date or number of total examination administrations.

The item measure displacement analysis consisted of running a Rasch calibration across all items on the first 200 administrations starting when the performance items were initially administered. The item parameters were anchored and another calibration was run across all additional administrations. The item measure displacement was calculated for each item. The displacement represents the difference between the anchored item measure and the measure for the same item in the subsequent administrations if all other parameters are held constant. A positive displacement shows items that become harder, relative to the other items; a negative displacement shows items that become easier, relative to the other items. Each of these analyses was conducted using Winsteps (Linacre, 2014). While each calibration included all items, the reported changes are only for the items that were initially administered at the same time as the performance items (including simulations, when present). The average displacement was calculated for each item type within each examination. Unlike the *p*-value change, the calibrations are based on incomplete matrices, reflect changes by date, and include all exam administrations.

Results

Table 1 summarizes the proportion of each examination composed of each item type, the total number of administrations used for the displacement analysis, and the number of administrations used per item for the *p*-value change analysis.

Table 1. Composition of examinations and counts of administrations

	Proportion of Item Types				**Counts of Exam Administrations**	
Exam	**Simulations**	**Performance**	**Selected Response**	**Selected Response +**	**Total n**	**Max n: *p*-value change**
A		2%	80%	18%	9,971	449
B		5%	78%	17%	5,507	283
C		6%	94%		13,156	1650
D	7%	1%	92%		12,492	400
E	6%	2%	92%		3,395	237
F	83%		17%		2,431	800
G	14%		86%		2,221	400

Figure 1 shows the change in *p*-value by item type, converted to a proportion of points and averaged, for examinations A – G. Overall, the figures demonstrate two main findings about simulation and performance items relative to SR items: 1) they tend to be more difficult and 2) their difficulty tends to be more stable than SR items. The following is a summary of the *p*-value change results by exam:

- Exam A shows fairly stable SR and SR+ items whereas performance items show a decrease in scores after 449 administrations.
- Exam B shows fairly stable performance items whereas the SR and SR+ items show an increase in scores.
- Exam C shows both SR and performance items increasing in score but SR items at a faster rate; note that this occurs after 1,650 administrations.
- Exam D, like Exam A, shows SR items as fairly stable whereas simulation and performance items show a decrease in score.
- Exam E shows reasonably stable simulation and performance items whereas SR items show an increase in average score.
- Exam F is the only exam where SR items are relatively easier than performance items, at least to start. Exam F also shows an increase in score for those items while the simulation items remain fairly stable. This is the only examination that shows an intersection between the changes in *p*-values. Exam F is also the examination with the highest proportion of simulation items (83%).
- Exam G shows an increase in both SR and simulation scores, though the SR item *p*-values change at a faster rate.

Figure 1. Change in *p*-value for examinations A – G

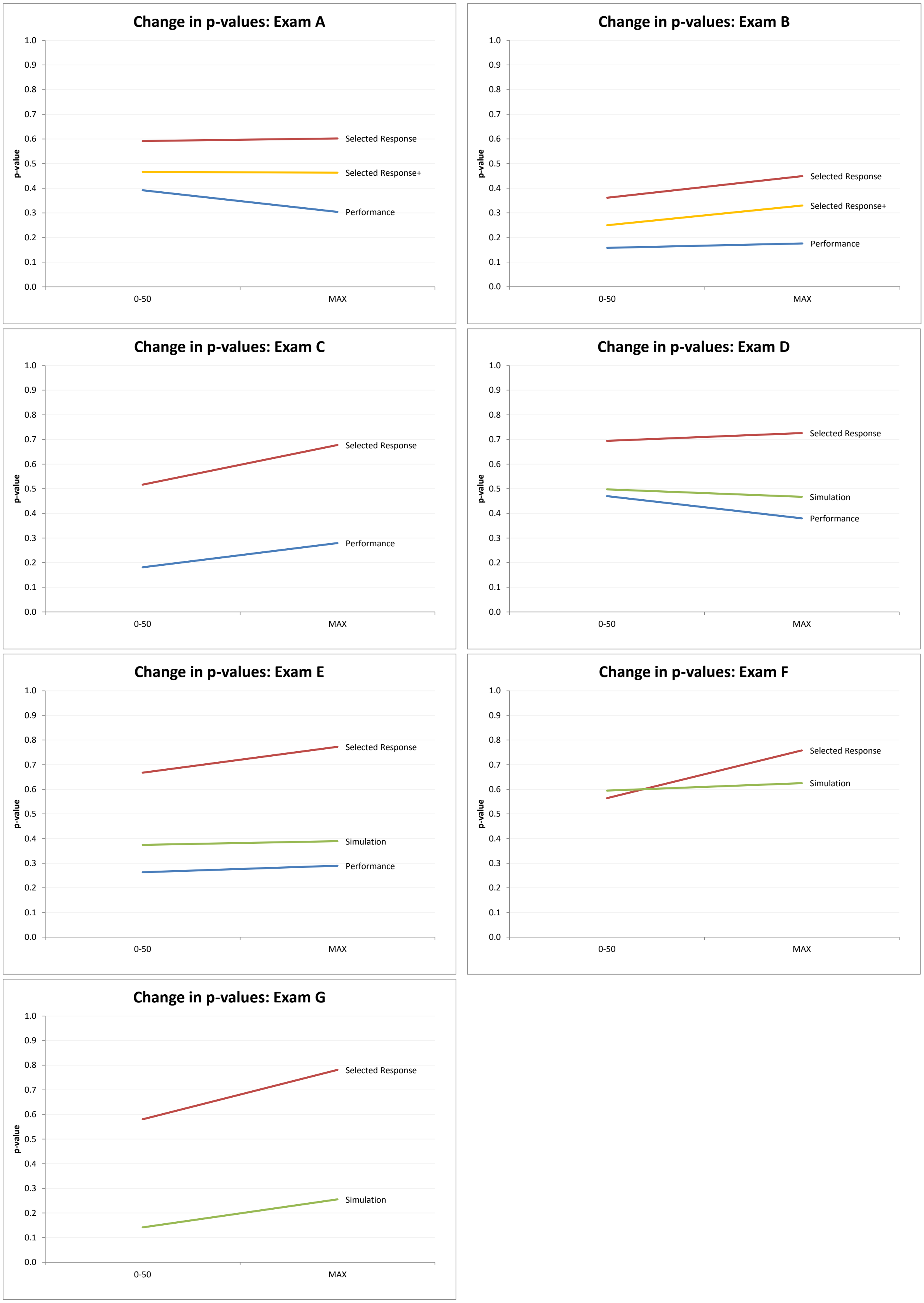
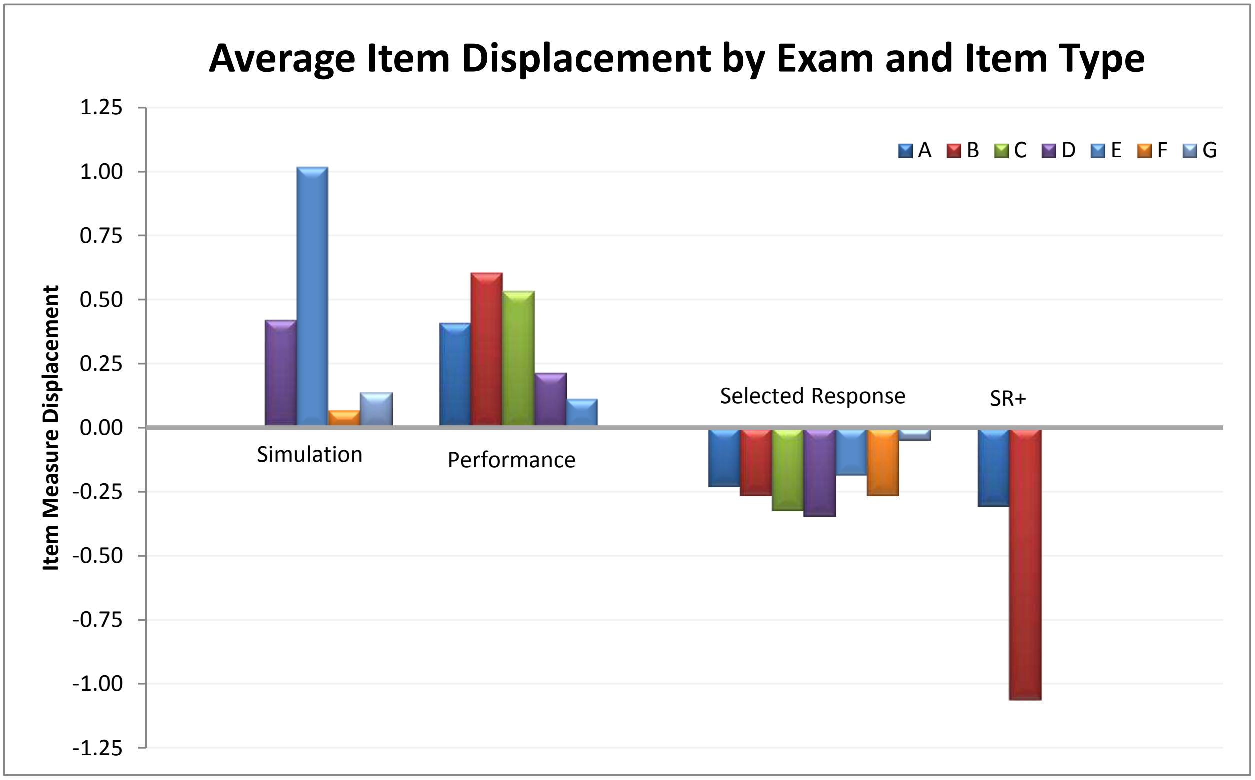


Figure 2 summarizes the results of the item measure displacement analyses by showing the average displacement by item type for each of the seven examinations. Across all seven examinations and on average, SR item displacements were negative—showing that SR items become relatively easier. The SR+ items on average, for the two exams that include them, were negative—showing that these items also become relatively easier. The performance items on average, for all five of the exams that include them, were positive—showing that they become relatively more difficult. The simulation items on average, for all four of the exams that include them, were positive—showing that they become relatively more difficult. The smallest change in the simulation items was for Exam F, consisting of 83% simulation items. The largest change was in the SR+ items in Exam B, consisting of 2% performance items, 80% SR items, and 8% SR+ items. The second largest average change was the simulation items on Exam E, consisting of 6% simulation items, 2% performance items, and 92% SR items.

Figure 2. Average item displacement by type for each exam



Conclusions and Discussion

A few key limitations of this study should be considered when reviewing the results, conclusions, and discussion. First, these exams are administered globally and are continuously available, which impacts the effects of exposure. Second, the sample of examinations in this study is limited to IT certification examinations. While IT certifications often lend themselves to performance and simulation items, the results of this study may not generalize to other domains. Third, the information regarding item types was gathered from multiple sources and multiple organizations, not from a single evaluation of the content. Therefore, there may be inherent inconsistency in some of the descriptions and understanding of the item types. In future research, more concrete operational definitions of the various item types should be established a priori.

The primary purpose of this study was to investigate the susceptibility of performance items to exposure, relative to SR items. Overall, the results indicated that performance and simulation items were more stable across these seven exams than their SR counterparts. This is not to say that simulation and performance items are not susceptible to exposure, only that they are *less* susceptible than SR items. Specifically, the displacement analyses showed that SR items became easier for all seven examinations and simulation and performance items became more difficult for each examination that included those item types. Further, the *p*-value change analysis showed that when SR *p*-values increased, the simulation and performance items were either essentially stable or their *p*-values also increased, but not as greatly.

For Exams A and D, SR item *p*-values items were relatively stable whereas the simulation and performance item *p*-values tended to decrease. One hypothesis that may explain this result is that sometimes high ability candidates take examinations more immediately when they are refreshed than lower ability candidates. Another observation was that the SR+ items tended to have the same *p*-value change as SR items, but their average displacement showed them becoming relatively easier than both performance and traditional SR items. For the two examinations that included SR+ items, those items were the most susceptible based on the results of the displacement analyses. It is possible that the factors that define these items as SR "plus," such as graphic interfaces, also makes them more memorable. A logical hypothesis is that items with a memorable component (e.g., a graphic), but only one correct response are at more risk of exposure. Items that have a memorable component, multiple possible correct answers, and require a candidate to complete a process (e.g., performance and simulation items) are at less risk. Given the limitations of the current study, these hypotheses are mostly conjecture, but certainly worth more investigation.

Finally, the differences between the stability of the different item types have important implications that could influence item calibrations, scaling, linking, and equating. Further, it seems reasonable that the development of simulation and performance items may help mitigate exposure issues, particularly those due to nefarious behavior. Therefore, the return on investment of their development may be improved due to the fact that the items may be useful for more administrations and/or a longer period of time than SR items. However, simulation and performance items are potentially time consuming, expensive to develop and maintain, and may not be appropriate for the purposes of some examinations. Therefore, despite indications of their decreased susceptibility to exposure, simulation and performance items should still only be developed when appropriate given the intended uses and interpretations of examination scores.

References

Cizek, G. J. (1999). *Cheating on tests: How to do it, detect it, and prevent it*. Mahway, NJ: Lawrence Erlbaum.

Harmes, J. C., & Wendt, A. (2009, Winter). Memorability of innovative items. *CLEAR Exam Review, 20*(1), 16–20.

Jones, P. E., & Smith, R. W. (April, 2006). *Item parameter drift in certification exams and its impact on pass-fail decision making*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Linacre, M. (2014). Winsteps (Version 3.75.1) [Computer software].

Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329–347). Mahwah, NJ: Lawrence Erlbaum Associates.

Smith, R. W. (2005, Summer). The impact of Internet sites on item exposure and item parameter drift. *CLEAR Exam Review, 16*(2) 12–15.