GRADUATE MANAGEMENT ADMISSION COUNCIL

Guess Again: The Effect of Correct Guesses on Scores in an Operational CAT Program

Eileen Talento-Miller, Kyung T. Han, Fanmin Guo

GMAC® Research Reports • RR-11-04 • March 14, 2011

Abstract

Previous theoretical research shows that recovery of ability estimate on computerized adaptive tests after initial correct responses depends on factors such as the item selection method and test length. Because it would be difficult to evaluate all factors present in an operational setting, data from an actual live testing program and examinees were used to simulate the effect of strings of correct responses in different positions on the exam. Differences interacted relative to string length, position, section, and examinee ability. Findings refuted the myth that the beginning items are the most important. The greatest limitation of the study is that the use of specific constraints for an operational program yields results that are not generalizable. Future research may include looking at incorrect responses and evaluating how increased time pressure at the end of the test would affect any benefit accrued from better performance at the beginning of the test.

Introduction

It is said that the first few items of a computerized adaptive test (CAT) are the most important. The conception of CAT administration is that the test begins with a medium difficulty item and a correct response then leads to a more difficult item. Answer the more difficult item incorrectly and the next item would be harder than the first, but easier than the second. By this logic, successive correct answers would lead the examinee to the top levels of the ability scale, whereas wrong answers would not mean as much since the examinee already correctly answered the lower difficulty items. This simplistic view of CAT methodology therefore leads some to believe that items at the beginning of the test are somehow weighted more heavily in estimating scores than are other items throughout the test.

The reality of CAT administration and scoring is far more complex. Research shows that the influence of responses to initial items can depend on the parameters of the test, such as length and item selection method. For instance, studies by Guyer (2008) and Chang and Ying (2008) both suggested that missing items at the beginning of the test makes it difficult to recover to the true theta. The proclamations regarding correct responses at the beginning are not as dire, with the overestimation reduced when the test is of sufficient length. The design of the simulations in both studies does not allow for generalization into operational CAT programs due to the limited number of variables studied. For instance, Chang and Ying (2008) examined differences based on two different item selection methods. The study by Rulison and Loken (2009) also examined different item selection methods as well as different item parameter models. As with the previous studies, the effects of misfitting correct responses in initial positions did not present as much bias as the effects of misfitting incorrect responses. Rulison and Loken concluded that the belief that the beginning of the test is most important in CAT is partially supported when item selection is based on maximum information with a three-parameter model and an idealized item pool. Under different conditions, or if constraints are added, how would theta

estimation be affected? Indeed, the effect on theta estimation would have different effects on final score depending on the scale. How then do particular factors affect, not just theta, but final score estimation? Small differences in theta may not be meaningful if the values translate to the same scaled score.

The simplistic view of CAT administration presented above does not take into account the numerous operational factors that will affect how item responses influence the score. The previous research shows that item selection method, test length, and item response theory model will affect the influence of initial items. Another factor to consider would be whether the operational program includes unscored experimental or pretest items within the test. If an unscored item appears in one of the early item positions, then getting the item right or wrong would have no effect on the subsequent score, regardless of previous or subsequent item responses. Other considerations for operational tests that would interact include methods for initial, interim, and final ability estimation, content constraints, item exposure constraints, pool characteristics, treatment of omitted items, and conversion to score scale (Stocking, 1997; van der Linden & Pashley, 2000; Wise & Kingsbury, 2000). With all the permutations possible, it would be impractical to create simulations to determine the contribution of each factor to the influence of initial items on score. Therefore, the most practical way to address the question would be to use information from an operational program. Although the results would not generalize among programs, the research could develop a method that applies to other CAT programs.

Some may argue that it is important to answer the first few items on a CAT correctly; others may counter with the fact that inflated ability estimates in the beginning still have the rest of the exam to compensate. If that were the case, then what would be the effect of a string of correct responses in the middle of the test or at the end of the test? Research by Talento-Miller and Guo (2009), conducted in an operational setting, shows that the effect of guessing at the end of the CAT varies by section and by ability. Because the differences are based on actual responses, however, the specific effects expected from a string of all correct (or all incorrect responses) cannot be addressed. The purpose of this research is to investigate the interaction between a string of correct responses and their positions on an operational CAT. Specifically, this study will quantify score differences relative to strings of correct responses at the beginning, in the middle, and at the end of a CAT. Emphasis is placed on correct responses since lucky guesses from lower ability examinees are more likely than careless mistakes from higher ability examinees. The string of correct responses represents the extreme benefit that could be expected from ability and luck. Previous research suggests that recovery of true ability is more likely with misfitting correct responses (Guyer, 2008), but with the qualifications that the test must be of sufficient length and with an idealized item pool, without regard to the many other constraints present in an operational exam. This research extends the previous study by providing a practical example to achieve a more realistic view of any effects, using operational constraints and scaled scores instead of theta estimates. The comparison to other positions in the test also provides a relative sense of whether the beginning is most important, or whether lucky guesses have similar effects wherever they occur.

Methodology

For this practical example, information was gathered from an operational CAT that is used to help with admissions decisions for higher education. Two separate CAT sections were examined: one that measures verbal reasoning ability and the other that measures quantitative reasoning ability. Both sections are multiple-choice, have embedded pretest items, and strict time limits that impose penalties for omitting items at the end of the test. The two sections differ in relevant ways, however, necessitating a separate examination of each. For instance, there are more items in the verbal section, with 41 items compared to the 37 items in the quantitative section. Furthermore, the verbal section contains sets of items related to reading passages, which means the rules for item selection will differ from the quantitative section.

Based on this CAT program, information was gathered from item pools and respondents to simulate the effects of correct responses in specific item positions. Operational statistics were used from the item pool, affecting item selection and theta

estimation. Five thousand examinees were randomly drawn from the examinee database in an effort to ensure a representative sample that matches the examinee population. Their theta scores were used as their true ability in the simulations. A baseline condition simulated responses based only on the examinee's true ability. Additional simulations for each examinee involved setting all affected item responses to correct and simulating remaining responses based on the inflated interim ability estimates. Conditions simulated included strings of 3, 5, or 7 consecutive correct responses in the initial, middle, and final item positions. All simulations were conducted separately for verbal and quantitative sections. Final theta estimates for each condition were calculated and translated onto the scaled score unit. It should be noted that the score scale was a nonlinear transformation from the theta scale. The difference between the examinees' baseline scores and simulated scores for each condition were calculated and compared by ability level. Five ability levels were defined based on true total score (a combination of quantitative and verbal section performance on a 200 to 800 scale) so that each group had approximately 20 percent of cases and ability groups were consistent across verbal and quantitative trials. Group sizes differed from exactly 1,000 cases (20%) due to tied values. Table 1 shows the score range for each ability level and group sample sizes. Results are only reported for the lowest and highest ability groups in order to illustrate the extreme differences.

Table 1: Ability Levels

Ability Level	N	Total Score Range
1 (lowest)	1,053	250-470
2	1,144	480-540
3	983	550-590
4	932	600-650
5 (highest)	888	660-780

Results

Tables 2 and 3 and Figures 1 and 2 summarize the findings for the quantitative and verbal sections. As expected, score differences increase as the number of manipulated item responses increases. Score differences also show greater variability as the number of items with manipulated correct responses increases but generally less variability as position moves from the beginning of the test to the end.

Table 2: Mean (SD) Differences From Baseline for Quantitative Section

	3 items	5 items	7 items
Beginning	1.39 (2.58)	1.86 (2.93)	2.76 (3.54)
Middle	1.00 (1.61)	2.10 (2.31)	3.13 (2.94)
End	1.68 (1.54)	2.21 (1.88)	2.80 (2.18)

Table 3: Mean (SD) Differences From Baseline for Verbal Section

	3 items	5 items	7 items
Beginning	1.29 (2.24)	2.81 (3.13)	4.65 (3.97)
Middle	0.98 (1.28)	2.02 (1.93)	3.58 (2.40)
End	1.53 (1.29)	2.15 (1.55)	2.73 (2.73)

In the quantitative section, there appear to be few differences among the relative section positions. Evaluations of effect size show very small to small effects for all comparisons, with the exception of the middle position versus end position with three items with a medium effect (*Cohen's d* = –0.430), where correct answers at the end lead to a larger score. Differences in the verbal section are more noticeable, with the improvement in score increasing more rapidly for the beginning position compared to the other positions. In terms of effect size, the difference between the beginning and end position is more than half a standard deviation for the 7-item condition (*Cohen's d* = –0.576). Indeed, comparing similar conditions between the quantitative and verbal sections, there are clear differences only in the beginning position for 5-items (*Cohen's d* = –0.313) and 7-items (*Cohen's d* = –0.504), with higher differences observed with the verbal section. Further examination of differences between the two sections reveals that the inclusion of embedded pretest items may have influenced the results based on the rules imposed by the operational constraints in the respective sections.

Figure 1. Mean Differences for Quantitative Section

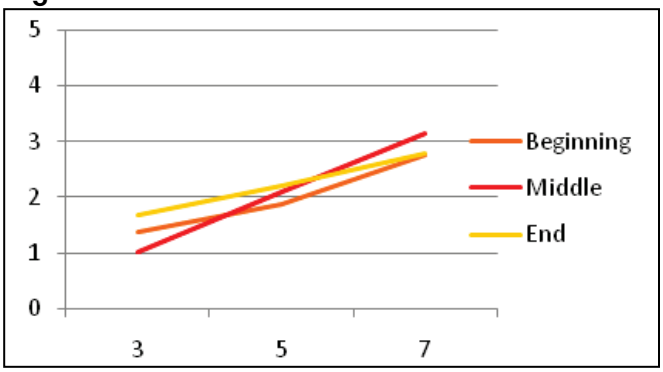
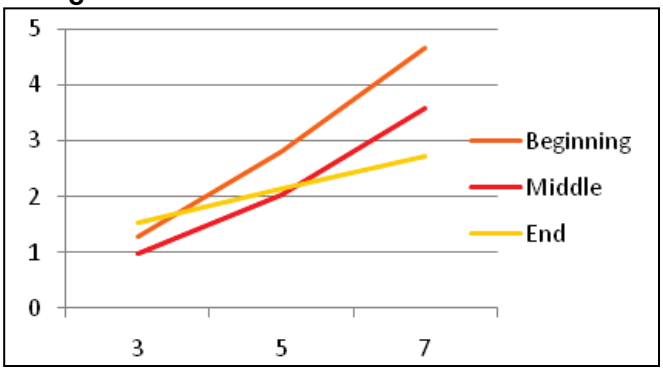


Figure 2. Mean Differences for Verbal Section



When examining the extreme ability levels, different patterns emerged. Figures 3 through 6 illustrate these differences. For the lowest ability group in the quantitative section, both beginning and end positions improved scores compared to the middle position for the 3-item condition, but position differences were minimal as the number of items increased. For the highest ability group in the quantitative section, results for position were consistent, with the end position resulting in higher increases by about a third to more than half a standard deviation over the beginning position, with smaller, yet consistent differences compared to the middle position. The lowest ability group had extreme improvements in the verbal section in the beginning position. The pattern of improvements for this ability group was similar to the pattern for all cases, although the 7-item condition was considerably higher. There appear to be no notable differences in improvements based on position of the correct items for the high ability group in the verbal section. Comparing the quantitative to the verbal section, the largest effect was observed in the 7-item beginning position for the lowest ability group (*Cohen's d* = –0.665). For the highest ability group, all comparisons showed greater improvement in the verbal section with small to moderate effects (*Cohen's d* = –0.071 to –0.514).

Figure 3. Quantitative Score Differences for Ability Group 1

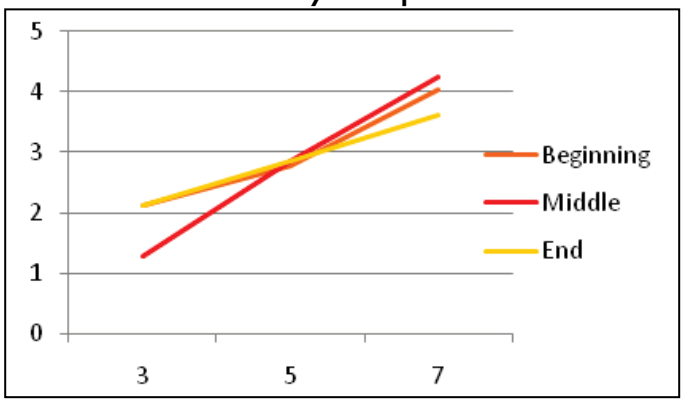


Figure 4. Quantitative Score Differences for Ability Group 5

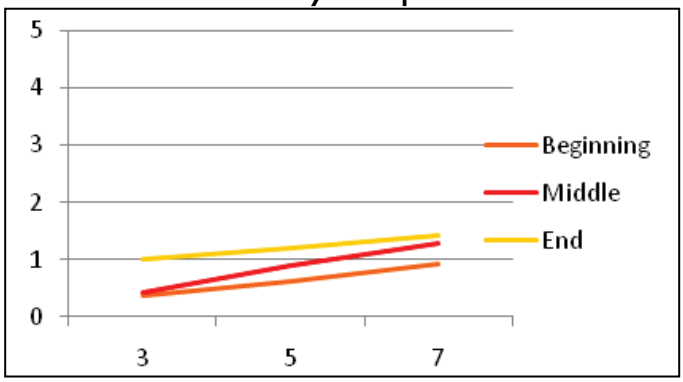


Figure 5. Verbal Score Differences for Ability Group 1

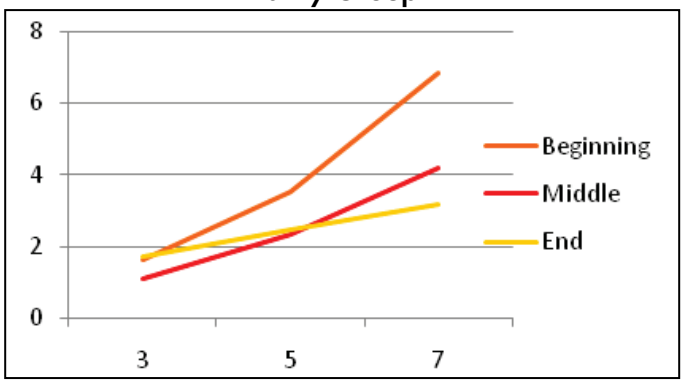


Figure 6. Verbal Score Differences for Ability Group 5

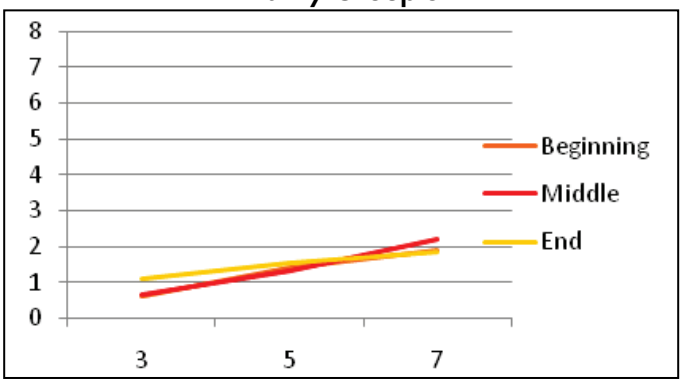


Table 4 provides a clearer picture of the comparison of position and section by including only median differences for the 5-item condition.

Table 4: Median Differences for the 5-Item Condition

Quantitative			
	All	Group 1	Group 5
Beginning	1	2	0
Middle	1	3	0
End	2	3	0
Verbal			
	All	Group 1	Group 5
Beginning	2	3	1
Middle	2	2	1
End	2	2	1

Relative to the standard error of the scaled scores, which is approximately three points for both sections, improvements overall are relatively small. As expected, greater improvement is observed for the lower ability group, although even these values are a standard error or less. The data appear to suggest that improvements from strings of correct responses are greater in the verbal section than in the quantitative section. The evidence refutes the notion that the beginning of the test is always the most important. The small differences that exist appear to interact with both section and ability.

Discussion

The current research contributes to existing information about CAT by examining the effect of correct responses by position. The use of operational information allows for the combination of factors that are frequently left out of theoretical simulations, such as content constraints and embedded pretest items. The focus on correct responses was based on the supposition that anyone can have a string of correct responses, but that an examinee with high ability is less likely to have a string of incorrect responses. The results of the study suggest that small strings of correct responses have little effect on final score. Interactions that were observed among position, string length, section, and ability counter the notion that the beginning of the test is always the most important.

Operational information was used to allow the simultaneous inclusion of several factors in the simulation, but the use of this information also means that conclusions are limited. Variations that were examined in previous studies, such as test length and item selection, were held constant in the current study. Closer examination of the operational factors in the study suggests that the inclusion of embedded pretest items in the simulation contributed to the observed results. Although the results may not generalize to other CAT programs, it is hoped that other programs will examine the effects of their own operational choices on scores to provide advice to candidates. One of the motivations for conducting this study was to show examinees why it is not advisable to spend extra time on the beginning of the test to the detriment of the remainder of the test. The current study assumed, however, that the rest of the section was completed at the examinee's ability level. To yield a better sense of the possible negative effects of spending more time on initial items, a model that includes penalties for time pressure at the end of the test would need to be constructed and evaluated. The current study focused on correct responses. A logical extension of this study for future research would involve examination of incorrect responses. As stated previously, a string of incorrect responses is highly unlikely at the beginning of a test for high-ability examinees; the likelihood may increase toward the end of the test with time pressure.

Examination of the issue through different positions of the test may provide additional insight.

Although CAT has been in use for decades, many examinees feel less than confident about their knowledge of CAT, mainly because of its complexity, and tend to rely on incomplete/incorrect information from unverified sources. One of the common CAT strategies suggested to test takers is to focus on initial items. This emphasis may actually deter examinees from showing the best of their ability in a timed test with penalties for not finishing. Spending more time on initial items means less than ideal time allotments for subsequent items and possibly not finishing the test. The results from the current study suggest that there is no one position which is most important. Examinees should understand that all items are important on a computerized adaptive test, not just the first few.

Contact Information

For questions or comments regarding study findings, methodology or data, please contact the GMAC Research and Development Department at research@gmac.com.

References

Chang, H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 7,* 441–450.

Guyer, R. (2008). *Effect of early misfit in computerized adaptive testing on the recovery of theta.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis, MN. Retrieved from http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/gu08001.pdf

Rulison, K., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33*, 83–101.

Stocking, M. (1997). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics, 21*, 365–389.

Talento-Miller, E., & Guo, F. (2009). *Guess what? Score differences with rapid replies versus omissions on a computerized adaptive test.* Paper presented at the GMAC Computerized Adaptive Testing Conference, June 1–3, 2009, Minneapolis, MN.

van der Linden, W., & Pashley, P. (2000). Item selection and ability estimation in adaptive testing. In W. van der Linden and C. Glas (Eds.), *Computerized adaptive testing: Theory and practice.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Wise, S., & Kingsbury, G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica, 21*, 135–155.