Graduate Management Admission Council®

Guess What? Score Differences with Rapid Replies Versus Omissions on a Computerized Adaptive Test

Eileen Talento-Miller
and
Fanmin Guo
Graduate Management Admission Council

Presented at the Realities of CAT Paper Session, June 2, 2009

Abstract

One of the strategies for completing standardized tests is to guess if unsure of the answer or if running out of time. In the case of a computerized adaptive test (CAT), the effect of guessing on any single item would depend on the individual's ability estimate and the item's parameters. Rapid guessing as time expires migh have extreme effects depending on current estimates or may result in only slight changes. The current study used data from an operational CAT to examine differences in scaled scores in three cases: (1) if the score was calculated before guessing; (2) if the score was calculated based on completing the last few items by guessing; and (3) if the score was calculated before guessing but adjusted with a penalty for omissions. Using these data, the relative benefits and possible penalties of guessing or leaving questions blank were explored for different ability levels. Although the results tended to favor guessing as a strategy, the degree of difference varied based on section content, number of items involved, and estimated ability of the examinee. Future research could include more definitive methods for determining random guessing and examine guessing at different positions within the test rather than merely at the end. Ultimately, the advice for candidates remains the same for a CAT as it would for other tests: Time management is important to allow ample opportunity to give thought to every question.

Acknowledgment

The views and opinions expressed in this paper are those of the authors and do not necessarily reflect those of the Graduate Management Admission Council®. Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Talento-Miller, E. & Guo, F. (2009). Guess What? Score differences with rapid replies versus omissions on a computerized adaptive test. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Eileen Talento-Miller, Graduate Management Admission Council®, 1600 Tysons Boulevard, Suite 1400, McLean, VA 22102, U.S.A. Email: talento-miller@gmac.com.

Guess What? Score Differences With Rapid Replies Versus Omissions on a Computerized Adaptive Test

Guessing is a standard test-taking strategy presented to examinees taking a multiple-choice assessment. This strategy provides an opportunity to have an item counted correct even when the examinee has insufficient knowledge of the subject matter. If test scores are based simply on the number of questions answered correctly, then a random guess increases the chance of a higher score. Formula scoring introduces an adjustment, in the attempt to recapture ability estimation undiluted by chance responses. With formula scoring, guessing is only advisable when the choices are not completely random, but based instead on partial knowledge (Angoff & Schrader, 1984; Lord, 1975; Lord & Novick, 1968). The probability of improving one's score and the expected amount of score difference are mathematically predictable for completely random guesses with either number-right correct or formula scoring. Due to differences in test administration and scoring of a computerized adaptive test (CAT), however, predicting the effect of guessing on a test score is considerably more challenging.

In CAT, items administered to each examinee are selected dynamically during the test and as a result, each examinee will see a unique set of items. Items are chosen based on their characteristics and on the examinee's then current ability estimate. For an examinee with a high ability estimate, test items will typically have higher difficulty values than those administered to examinees with low ability estimates. This raises the question: Will a purely random response from an examinee with a high ability estimate have the same effect on his or her score as a similar response from an examinee with a low ability estimate? In other words, does guessing on an easy item have the same effect as guessing on a difficult item?

With formula scoring, guessing is only advisable with partial knowledge; otherwise, leaving the question blank might be the wiser option. In CAT, are there identifiable instances when omitting the item would be a wiser choice than randomly guessing? Omitting items in the middle of the test is generally not allowed, since it would affect the computer program's ability to select subsequent items. Toward the end of the test, however, with the time limit rapidly expiring, examinees would have to decide whether to leave the questions blank or respond rapidly at random. Knowledge of the likely differences in score could help to inform examinees trying to maximize their score.

The current research examines guessing and omission behaviors and their effects on scores from actual administrations of an operational CAT. The probability of observing score differences when items at the end of the test are omitted or guessed are based on examinee behaviors. Furthermore, differences were examined for both verbal and quantitative sections to determine whether results were consistent across section types.

Related Literature

Existing research on the effect of guessing on scores from CATs is quite limited. The question of whether to guess or omit items is based on the amount of time pressure an examinee might feel toward the end of the test. The study by Bridgeman and Cline (2004) showed how time pressure affects the ability estimation of different examinees on the last six items of a CAT. Different time pressure groups were described, with the most extreme group having approximately two minutes left for the set of items remaining. Information was not available on

whether examinees omitted or guessed each of the items remaining, nor was it available on the effect at each item, but only on the cumulative effect from all six items based on time remaining. One might argue that with up to 20 seconds left for each of the items, random guesses might be supplemented by answering some of the questions with partial knowledge. The study did not differentiate among effects of random guessing behavior, solution behavior (guessing with partial knowledge), or omitting items and did not provide item-level feedback for each of the last items.

Although not directly addressing the effect of guessing on scores, a study by Schmitt, Walker, and Sass (2006) reported on the effects of random guesses and omissions on ability estimates. The authors found that the negative bias on ability estimates depended on the percentage of items involved and the item response theory model used. Because bias is related to ability estimate, this study does little to help examinees form expectations in terms of score changes. Furthermore, the results described the effects of not reaching 20% to 30% of items, which might be more than would concern a typical time-pressured examinee who might be concerned only with the last three to four items. The use of simulated data in this study might also be considered a limitation, since these data might not capture all the variations observed in practice.

Another study using simulated data was conducted by Mills and Steffen (2000). This research offered specifics in terms of the effects of random responses and omissions on scores in a simulation of the quantitative section of the Graduate Record Examination. Score differences were estimated based on number of items affected and by examinees' ability level. Because this study examined only the quantitative section of the test, it is unknown whether the effects would generalize to behaviors on the verbal section of the test. Peculiar differences related to the verbal section, such as having sets of items associated with a reading comprehension passage, rather than having all independent items, might result in differences in both guessing behavior and score differences.

Based on the previous research described, there are opportunities to expand what is known about guessing on CATs. The study by Bridgeman and Cline (2004) described time pressure on an operational test, but offered no descriptions of observed guessing or omission behavior and the resulting effects of those behaviors. The study by Schmitt, et al. (2006) addressed bias in ability based on speededness of a test, necessitating different strategies to be used at the end of the test. The results were combined across groups of items and bias was given in terms of θ estimates. Because of the general and technical nature of the results, examinees might have difficulty translating results into guidance for practice. In this study, omitted items were simulated by counting all remaining items as incorrect. The problem with this practice is that in the simulation, items will continue to be selected based on the responses. In reality, when a test is ended, there will be no record of items that would be chosen if the test were to continue. For this reason, the treatment of omissions as incorrect might introduce differences in scoring versus what is observed in practice. The Mills and Steffen (2000) research directly compared the effects of guessing, omissions, and incorrect responses in scoring, with differences described at each item level. Like the Schmitt, et al. (2006) study, however, this research was based on simulated data. Constraints placed on item selection and the items available in the bank used for the operational test are some of the reasons why these simulations might not reflect practice.

Research is needed on rapid guessing during an operational test and the possible score differences if items are not reached. Furthermore, because previous research only examined

limited content areas, additional studies are needed to determine directly whether differences exist in scoring of quantitative and verbal sections. Placing the results in context of the test would inform examinees of the likely effects of their end-of-test behaviors and help them choose the strategy that gives them the best chance for a higher score.

Method

Instrument

The Graduate Management Admission Test® (GMAT®) is a CAT used as part of the admission process for graduate business programs around the world. The test contains three separately timed sections: analytical writing, quantitative, and verbal. Only the latter two use the CAT multiple-choice format. For each of these two sections, candidates have 75 minutes to complete all the items in that section. There are 37 questions in the quantitative section and 41 questions in the verbal section. Examinees are not permitted to skip items during the administration of the test and final scores are adjusted if the section is incomplete. Section scores range from 0 to 60; both sections have a standard error of measurement of approximately 3 points.

Data

For this study, data were collected from operational administrations of the GMAT exam over an eight-month period. Data included item scores, time spent overall and on each item, and relative score at each item position in the test. Cases were removed from the dataset if they tested under any special conditions or if scores were not reportable for any reason. Also, following the precedent established in the Bridgeman and Cline (2004) study, any cases that spent fewer than 20 of the 75 minutes allotted were removed from the analyses for that section, with the assumption that these candidates did not approach the section with serious effort. The final dataset analyzed included more than 135,000 cases for both the quantitative and verbal sections.

Defining Guessing

Since the data were from actual examinee behaviors rather than simulated, it was necessary to determine how to characterize guessing behavior as opposed to solution behavior. A number of studies have been conducted on how to differentiate guessing behavior from solution behavior. The identification of rapid guesses in the current study followed a combination of procedures suggested in various studies (DeMars, 2007; Kong, Bhola, & Wise, 2005; Schnipke & Scrams, 1999; Swygert, 2003) that used graphical displays and proportions of correct answers to determine a time threshold for rapid guessing. For each item position, a histogram of time spent on correct items was examined. At the lower end of the graph, an inflection point where more solution behavior appeared to be starting was approximated. Based on the visual information, a time value was chosen and the proportion of correct responses at or below that value was compared to 20% correct, the expected value for a random guess with five answer options. Based on these analyses, it was determined that 10 seconds was an appropriate threshold to define rapid guessing for verbal items and 7 seconds was used as the threshold for quantitative items.

Analyses

For each section, descriptive information was collected on the prevalence of guessing and omitting items at the end of the test. Examinees were only characterized as guessing a certain

number of items at the end if those guesses were consecutive. For instance, if an examinee spent 4 seconds, 12 seconds, 3 seconds, and 5 seconds on the last four items in the quantitative section, that examinee was classified as guessing only the last two items, since only the string of consecutive guesses at the end of the test counted toward the classification, and the 12 seconds spent on the third to last item did not meet the threshold for a rapid guess.

Descriptive statistics were compiled for those who were classified as guessing at least one item and up to five items at the end of the test. These examinees were then compared to the group of examinees who did not respond to the same number of items in each section. This initial comparison provided information on the prevalence of the two strategies in dealing with time pressure during the operational administrations of each section.

To compare the effect of guessing versus omissions directly, a single-group design was used. This comparison was designed to separate the effect of the strategy used at the end of the test from other qualities of the examinee, while still using real data. Out of the available candidates who finished the test, only those who guessed at least one item toward the end of the test were selected and included in these analyses.

The observed score (G) was the score the examinee received based on guessing behavior. An additional score (O) was calculated based on the examinee's ability estimate prior to the string of consecutive guesses with the remaining items calculated as omissions. The differences between the omission score and the guessing score (O – G) were calculated for those who consecutively guessed the last one to five items. Proportions of examinees observing score differences up to a standard error in either direction provide a sense of the effects of the two strategies. In addition, the expected value of the difference is given for each number of items and for each section.

Because candidates of different ability levels might have items with radically different difficulty levels at the end of the test, the impact of guessing versus omitting was calculated for those of low, medium, or high ability. Examinees were categorized based on their ability estimate prior to the consecutive guesses at the end of the test. The low ability and high ability groups contained approximately 30% of the examinees at each end, and the middle group contained the approximately 40% remaining. Relative score differences were examined again, paying particular attention to values of a standard error or more, and calculated expected values by number of items.

Results

The first set of results describes the scores of examinees who chose to guess toward the end of the test versus those who omitted the last few items. Table 1 shows the number of examinees who guessed at or omitted each of the last five items. These tables also summarize statistics for examinees' observed scores for the verbal and quantitative sections. Based on the relative frequencies, examinees were considerably more likely to guess rapidly at the end of each section than to omit the items, with the differences greater in the verbal section. The only exception was for five items in the quantitative section, which more examinees omitted rather than guessed; however, the relative frequencies were low for both, as these two groups combined still represented less than 1% of the total examinees.

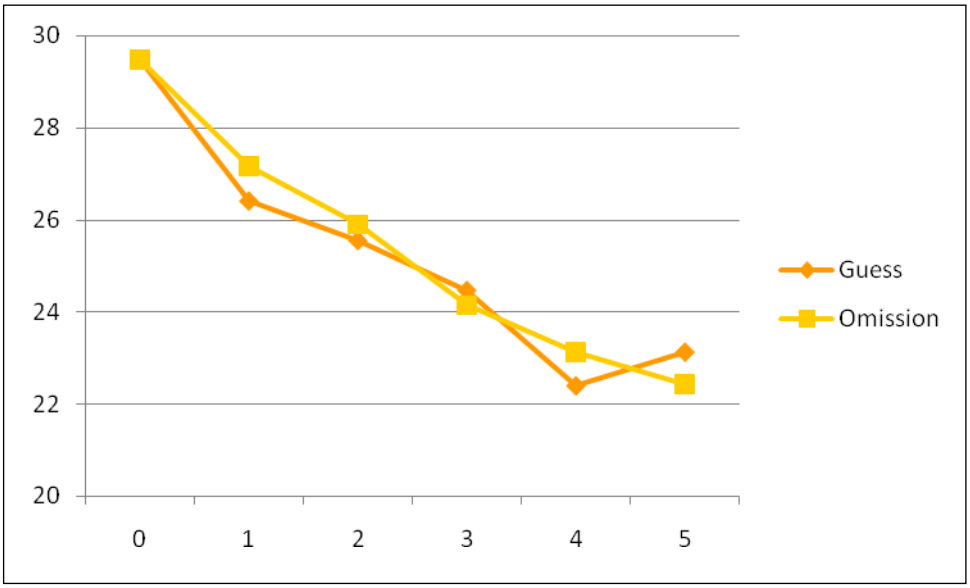
Table 1. Mean and Standard Deviation (SD) of Observed Verbal and Quantitative Scores by Number of Guesses and Omissions

Section and No. of Items	Guessed			Omitted		
	N	Mean	SD	*N*	Mean	SD
Verbal						
0	106,078	29.47	8.53	106,078	29.47	8.53
1	4,530	26.40	8.47	812	27.15	8.04
2	1,815	25.54	8.15	526	25.89	7.95
3	1,595	24.46	8.29	342	24.14	8.06
4	705	22.39	8.14	210	23.12	6.57
5	261	23.12	8.72	202	22.42	6.74
Quantitative						
0	102,989	35.68	10.90	102,989	35.68	10.90
1	5,447	34.71	9.89	2,504	36.39	9.36
2	2,486	34.30	9.34	1,285	35.80	9.27
3	1,198	34.63	9.20	805	34.00	8.92
4	693	34.27	9.51	444	33.11	8.93
5	420	33.96	9.15	665	32.93	8.30

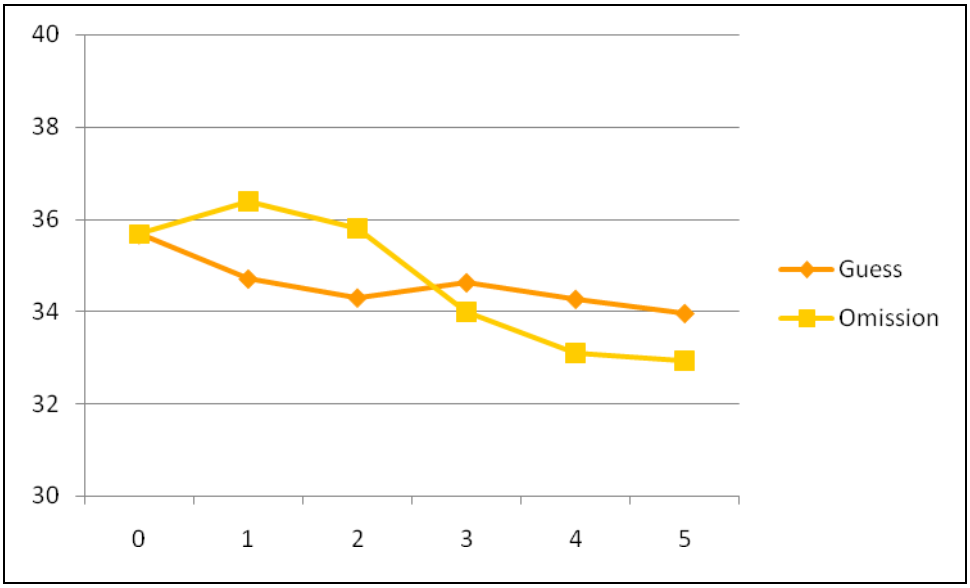
The graphical presentations of the data shown in Figure 1 suggest that there might be differences in scores by section. For instance, it appears that the choice of omitting or guessing made little difference in average observed scores for examinees on the verbal section. On the quantitative section, however, those who omitted up to two items had average scores slightly higher than those who guessed items. Beyond two items, however, the group of guessers had the higher average scores. The largest difference was fewer than two points and would be considered a small effect size at less than a fifth of a standard deviation. Although these results provide an interesting observation of the operational GMAT data, differences in scores might not necessarily be related simply to guessing or omitting. Clearly, sample sizes differ between the two groups. There might be other relevant differences that existed between the two groups that could affect the observation of an effect between guessing and omitting items. The single-group design allows direct comparison of scores and attribution of effect to omit or guess strategies.

Figure 1. Average Observed Scores for Guesses Versus Omissions

a. Verbal



b. Quantitative



For each examinee who guessed the last item up to five consecutive guesses at the end, the observed guessing scores were subtracted from the scores they would have received had they omitted rather than guessed the remaining items. Positive differences indicated higher scores from omitting and negative values indicated an advantage from guessing. Table 2 shows the proportion of examinees with a magnitude of score differences up to or exceeding 3 points. The data seemed to follow conventional wisdom that it is better to guess than leave questions blank, but there were still a number of cases that benefited more from omitting items.

Table 2. Proportion Observing Score Difference Magnitudes for O – G

Section and No. of Items	Amount of Scale Score Difference						
	≤-3	-2	-1	0	+1	+2	≥+3
Verbal							
1	0.001	0.043	0.179	0.634	0.139	0.003	0.001
2	0.021	0.088	0.194	0.417	0.247	0.027	0.006
3	0.044	0.098	0.218	0.324	0.208	0.083	0.024
4	0.067	0.132	0.174	0.305	0.206	0.085	0.031
5	0.161	0.111	0.211	0.245	0.138	0.084	0.050
Quantitative							
1	0.003	0.050	0.213	0.570	0.132	0.010	0.004
2	0.030	0.121	0.251	0.387	0.160	0.038	0.013
3	0.094	0.201	0.260	0.259	0.124	0.043	0.019
4	0.237	0.185	0.209	0.193	0.107	0.048	0.022
5	0.374	0.214	0.138	0.157	0.069	0.026	0.021

When the candidate guessed five consecutive items at the end, even though the guessing score met or exceeded a standard error for 16% of the candidates, the omission score was a standard error or more higher for 5% of the candidates for the verbal section. The difference was much more straightforward in the quantitative section. For five items, more than a third of the examinees had a guessing score 3 or more points higher than their omission score, and only 2% had an omission score 3 or more points above their guessing score.

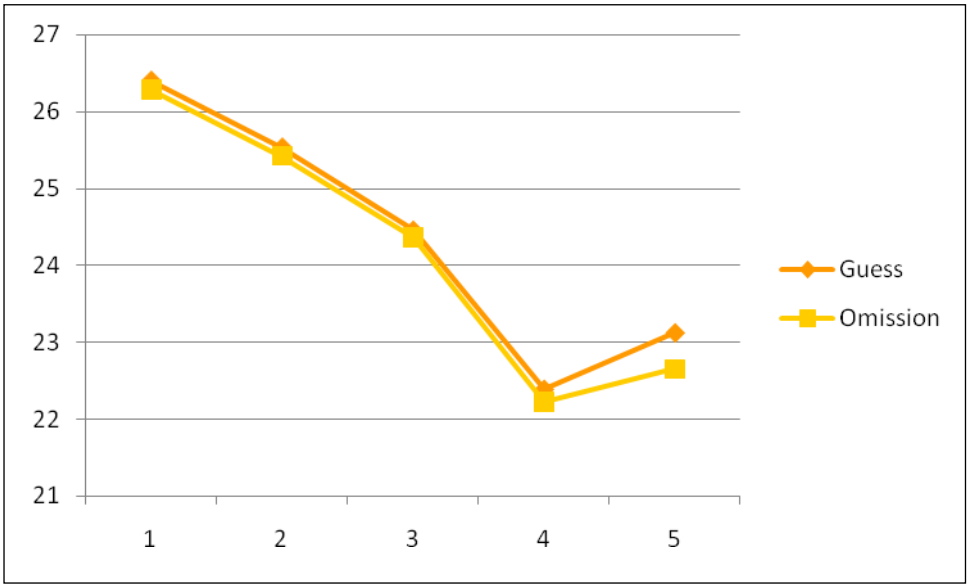
Table 3 shows the mean value of the difference at each number of items for both sections. The expected value favored guessing across all sections and numbers of items, although the difference appeared relatively constant for many of the verbal cases while increasing rapidly for the quantitative items. Figure 2 illustrates these differences. Average verbal scores for those who guessed five consecutive items were around 22 to 23, which approximately correspond to the 27th and 29th percentiles in the distribution of GMAT scores. For the quantitative section, the scores around 32 to 34 are also near the bottom of the distribution for that section, falling around the 34th and 40th percentiles.

Table 3. Mean and SD of Scale Score Differences by Section

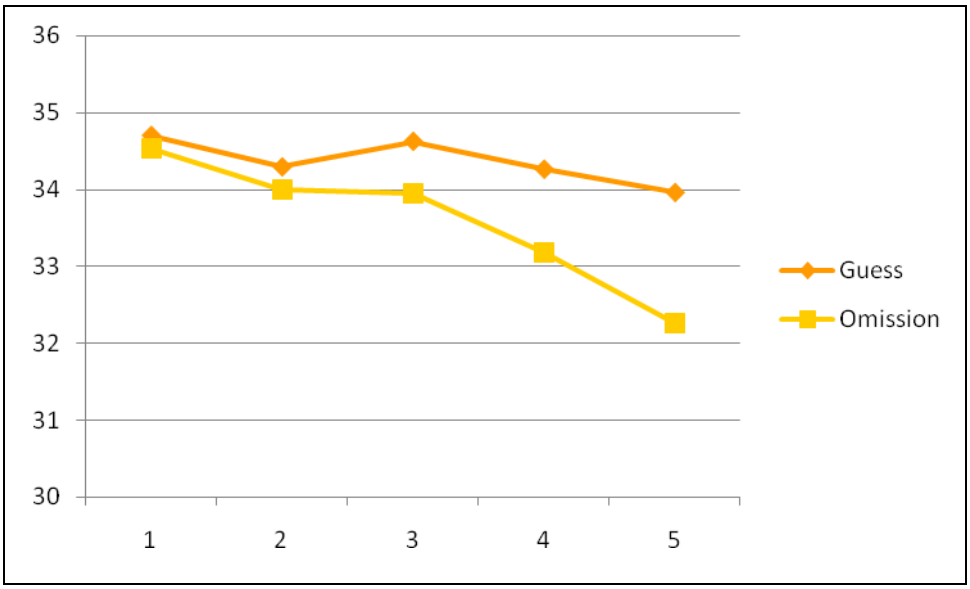
No. of Items	Verbal			Quantitative		
	N	Mean	SD	*N*	Mean	SD
1	4,530	-0.119	0.725	5,447	-0.170	0.879
2	1,815	-0.112	1.095	2,486	-0.295	1.291
3	1,595	-0.097	1.382	1,198	-0.674	1.556
4	705	-0.164	1.575	693	-1.078	1.810
5	261	-0.471	2.067	420	-1.698	1.986

Figure 2. Average Scores by Number of Items

a. Verbal



b. Quantitative



The score differences between guessing or omitting items might differ based on estimated ability level. Tables 4 and 5 show the differences in proportions by ability group for the two sections. The results by ability group shed more light on the previous analyses. For verbal, there were several cases that had significantly higher scores when omitting versus guessing, with fewer observing similar differences in quantitative scores. These higher scores appeared to be most likely in the low ability group for both sections. In fact, only a single case from the high

ability group in either section benefited by 3 or more points by omitting items as opposed to guessing them.

Table 4. Proportion Observing Differences in Verbal Scores for O − G by Ability Group

No. of Items	*N*	Amount of Scale Score Difference						
		≤-3	-2	-1	0	+1	+2	≥+3
Low Ability Group								
1	1,673	0.002	0.016	0.143	0.596	0.233	0.008	0.003
2	668	0.007	0.030	0.138	0.398	0.371	0.042	0.013
3	639	0.006	0.034	0.150	0.362	0.279	0.138	0.038
4	330	0.012	0.045	0.115	0.379	0.288	0.112	0.048
5	112	0.018	0.036	0.170	0.286	0.241	0.152	0.089
Medium Ability Group								
1	1,901	0.002	0.061	0.158	0.667	0.112	0.001	0.000
2	814	0.025	0.111	0.184	0.425	0.231	0.023	0.001
3	670	0.058	0.101	0.230	0.309	0.210	0.070	0.021
4	287	0.080	0.164	0.230	0.265	0.164	0.077	0.021
5	104	0.221	0.154	0.240	0.240	0.077	0.038	0.029
High Ability Group								
1	956	0.000	0.054	0.281	0.636	0.027	0.001	0.000
2	333	0.039	0.150	0.333	0.432	0.039	0.006	0.000
3	286	0.094	0.231	0.343	0.276	0.045	0.007	0.003
4	88	0.227	0.352	0.216	0.159	0.034	0.011	0.000
5	45	0.356	0.200	0.244	0.156	0.022	0.022	0.000

Table 5. Proportion Observing Differences in Quantitative Scores for O – G by Ability Group

No. of Items	*N*	Amount of Scale Score Difference						
		≤-3	-2	-1	0	+1	+2	≥+3
Low Ability Group								
1	1,635	0.005	0.046	0.120	0.546	0.245	0.027	0.011
2	691	0.014	0.054	0.191	0.314	0.289	0.097	0.041
3	304	0.023	0.095	0.158	0.326	0.220	0.109	0.069
4	174	0.080	0.086	0.190	0.264	0.216	0.109	0.057
5	110	0.118	0.194	0.145	0.282	0.145	0.055	0.055
Medium Ability Group								
1	2,471	0.002	0.055	0.206	0.613	0.119	0.005	0.001
2	1,228	0.037	0.113	0.215	0.454	0.156	0.021	0.004
3	580	0.090	0.179	0.269	0.290	0.138	0.031	0.003
4	341	0.179	0.170	0.240	0.246	0.109	0.041	0.015
5	213	0.343	0.244	0.169	0.150	0.061	0.023	0.009
High Ability Group								
1	1,341	0.001	0.045	0.413	0.521	0.019	0.000	0.000
2	567	0.034	0.220	0.402	0.333	0.009	0.002	0.000
3	314	0.172	0.344	0.341	0.137	0.006	0.000	0.000
4	178	0.500	0.309	0.169	0.022	0.000	0.000	0.000
5	97	0.732	0.175	0.062	0.031	0.000	0.000	0.000

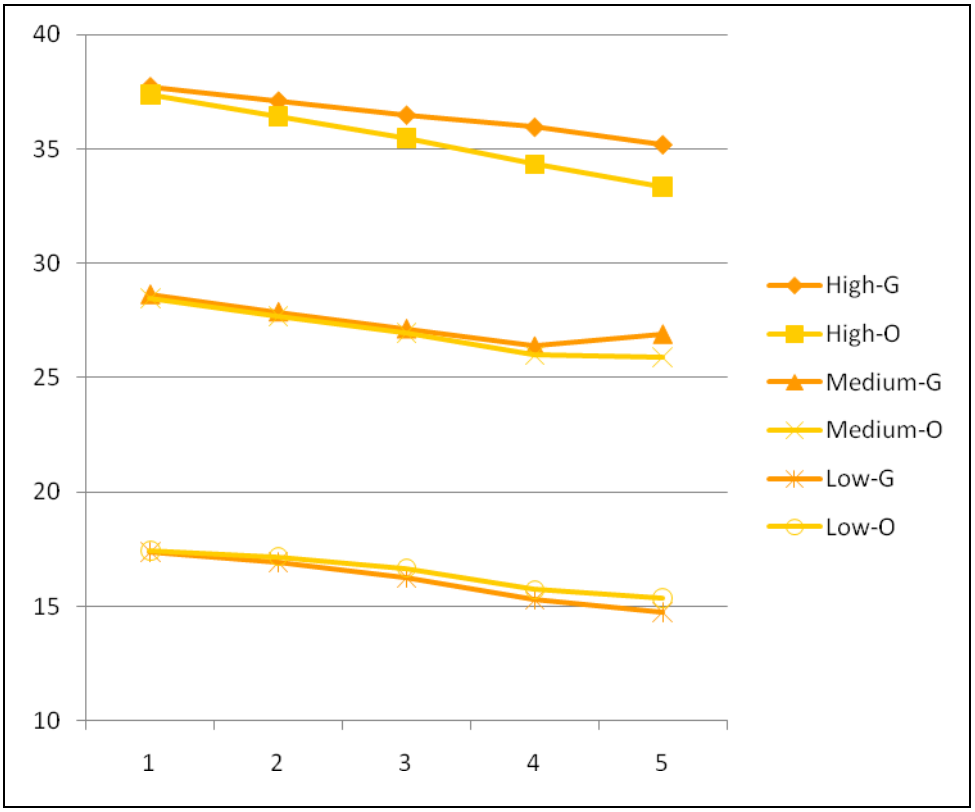
For more information, mean differences by ability group are listed in Table 6. The positive differences shown for most of the low group averages suggest that omitting the items tended to result in better scores for this group, although the differences were very small. Large differences can be found for the high ability group in the quantitative section, where the average difference favored guessing by more than the standard error of 3 points for the five-item group. Figure 3 shows the average guessing and omission scores for each ability group. The most striking results are in the quantitative section and occurred for the high ability group, with omitting items resulting in rapidly decreasing scores.

Table 6. Mean (SD) of O – G Differences by Ability Group

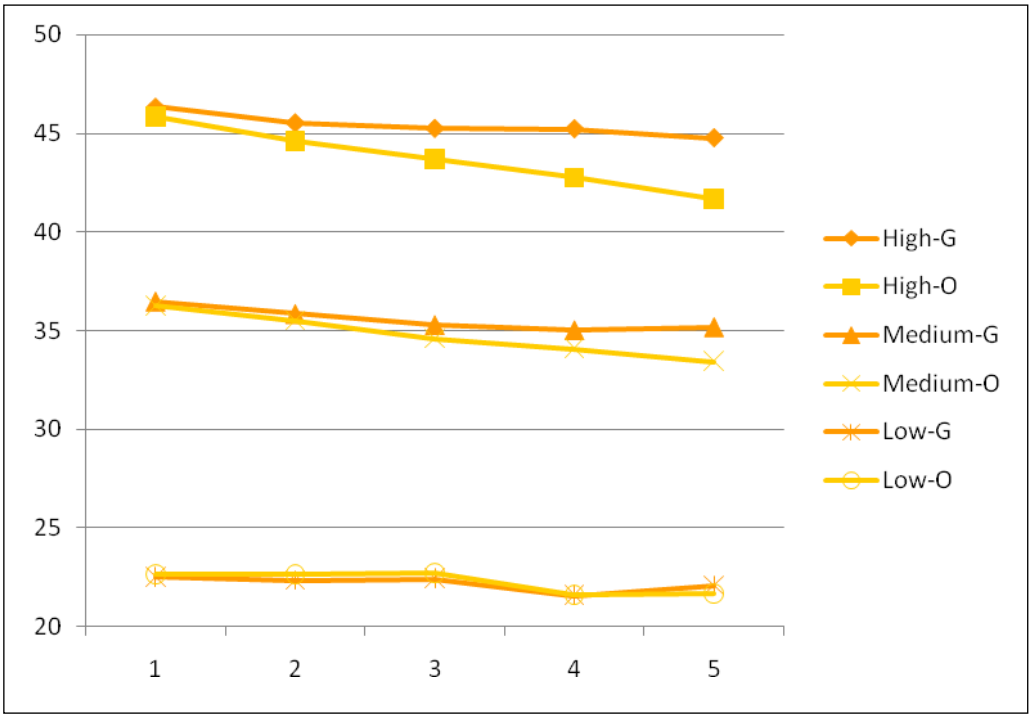
No. of Items	Low		Medium		High	
	Verbal	Quant	Verbal	Quant	Verbal	Quant
1	0.078 (0.744)	0.122 (1.118)	-0.172 (0.708)	-0.190 (0.745)	-0.361 (0.633)	-0.488 (0.622)
2	0.209 (1.030)	0.317 (1.597)	-0.198 (1.070)	-0.344 (1.093)	-0.709 (0.955)	-0.935 (0.868)
3	0.430 (1.241)	0.293 (1.819)	-0.203 (1.381)	-0.698 (1.340)	-1.028 (1.114)	-1.567 (1.010)
4	0.442 (1.309)	0.046 (1.880)	-0.411 (1.566)	-0.915 (1.589)	-1.636 (1.323)	-2.489 (1.101)
5	0.598 (1.891)	-0.373 (2.063)	-1.019 (1.907)	-1.737 (1.750)	-1.867 (1.604)	-3.113 (1.257)

Figure 3. Average Scores by Ability Group and Strategy

a. Verbal



b. Quantitative



Discussion

The current research using operational data from a CAT showed many interesting findings. The distribution of differences (O – G) can be examined to determine whether there is an effect of guessing versus omitting. With no effect, one would expect differences to be distributed symmetrically around zero, with differences reflecting only measurement error. Skew in either direction suggests an effect for the end-of-test strategy, with more positive values (negative skew) favoring omitting items, and more negative values (positive skew) favoring the guessing strategy. The distributions in the current research show skew in favor of scores for guessing over omitting items. The amount of skew varied by section, however.

Previous research has examined only quantitative and analytic sections. The current study shows results can be quite different for the verbal section. Distributions of differences show only slight skew in favor of guessing. Few notable differences in verbal scores between omitting items and guessing items suggested that neither strategy presented much of an advantage. If an examinee found herself with only a minute remaining to answer the last four items of the verbal section, it would be to her benefit to spend time trying to answer at least one of the remaining questions with thought while feeling confident that leaving the remaining items blank would not affect the score much differently than random responding.

The situation differed considerably in the quantitative case, where omitting items showed greater differences. The skew of the distributions was quite noticeable, becoming more obvious as the number of items involved increased. The examinee has only a 2% chance of having a significantly higher score by omitting 3 to 5 items compared to guessing. This probability

disappears if one is in the high ability group. Not a single examinee in the high ability group received 3 or more points by omitting items in the quantitative section. In contrast, in the same section, more than 80 individuals in the low ability group would have scored 3 or more points higher if they had omitted the remainder of their items as opposed to guessing. If the examinee has an idea of his/her relative ability prior to test administration, then the choice of omitting or guessing in the quantitative section is slightly better informed. Those with lower relative ability should try to answer the questions with thought in the time they have remaining and leave the rest blank. Those who have higher relative ability should not consider leaving questions blank under any circumstances.

The effect of omitting items differed by section, with greater differences being observed in the quantitative section. One factor to consider is the difference in length of the two sections. Since the verbal section has more items (41 compared with 37 in the quantitative section), then omitting three to four items would be expected to have a lesser effect. It is less clear why omitting items in the verbal section would show such a consistent advantage for the low ability group, indicating a need for further study.

The guessing advantage for the high ability group was much more dramatic and not entirely surprising. Certainly, high scorers have farther to fall when we apply an adjustment for unanswered items. By definition, some of the random responses to the items will be consistent with the ability level of the examinees and will provide an estimate that is more consistent with their observed responses to that point. Studies of error across the ability groups might provide more insight into these phenomena.

There are several limitations to the current study. The prevalence of guessing during the test can only be estimated using the available data of timing and accuracy. Because each examinee received different items, the threshold might have been too high for some types of items and too low for others. For instance, it is possible that 10 seconds would have been sufficient time to answer a sentence-correction question of moderate difficulty, but that might not have been enough time just to read through a difficult critical reasoning question. The majority of the analyses involved a single-group design and included only those who guessed at the end of the test. This design was intended to remove variations in groups of people who choose a random guessing strategy over an omission strategy. Because only the guessers were included, their behaviors and score differences might not generalize to the larger testing population. Also, the study only included consecutive guessing behavior at the end of the test sections. We did not address the prevalence of, and effects of, random guessing at other points of the test. Future research might be able to identify guessing behavior explicitly through examinee feedback and could examine the use of time-based strategies throughout the test.

Although the results provide examinees some guidance for maximizing their scores, it is important to reiterate that the differences observed were quite small for the majority of cases. The vast majority of cases observed no difference between the guessing and omission score. Across ability levels and content sections, when only one item was at stake, the proportion observing a difference of 0 ranged from more than half to two-thirds of the examinees. The impact of the differences, however, will likely be perceived quite differently for test developers versus examinees. In terms of standard error, a difference of one point is a small effect, but to the examinee it might appear to make all the difference.

As one might expect, the data favor guessing in most cases, although the distribution of

differences implies that the result is by no means clear cut. Although highly improbable, even a single guess might result in a standard error difference either for or against the candidate. The results by ability level offer a bit more information about when these cases might occur. The question of whether it is best to guess or to omit depends on how many test items are left, on one's relative ability estimate up to that point, and on the specific section of the test being considered. Ultimately, the advice for candidates remains the same for a CAT as it is for a paper-and-pencil test: Answer the questions to the best of your ability within a reasonable amount of time, to allow ample opportunity to give thought to every question.

References

Angoff, W. & Schrader, W. (1984). A study of hypotheses basic to the use of rights and formula scores. *Journal of Educational Measurement, 21*, 1–17.

Bridgeman, B. & Cline, F. (2004). Effects of differentially time-consuming tests on computer-adaptive test scores. *Journal of Educational Measurement, 41,* 137–148.

DeMars, C. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12,* 23–45.

Kong, X., Bhola, D., & Wise, S. (2005, April). *Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Lord, F. (1975). Formula scoring and number right scoring. *Journal of Educational Measurement, 12,* 7–11.

Lord, F. & Novick, M. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Mills, C. & Steffen, M. (2000). The GRE computer adaptive test: Operational issues. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice.* (pp. 75–99). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Schmitt, T., Walker, C., & Sass, D. (2006, August). *The effects of test speededness on computer adaptive ability estimation.* Paper presented at the annual convention of the American Psychological Association, New Orleans, LA.

Schnipke, D. & Scrams, D. (1999). *Exploring issues of test taker behavior: Insights gained from response-time analyses.* Law School Admission Council Computerized Testing Report 98–09. Newton, PA: Law School Admission Council.

Swygert, K. (2003). *The relationship of item-level response times with test-taker and item variables in an operational CAT environment.* Law School Admission Council Computerized Testing Report 98–10. Newton, PA: Law School Admission Council.