Graduate Management Admission Council®

Quantifying the Impact of Compromised Items in CAT

Fanmin Guo

Graduate Management Admission Council

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

Abstract

If a few test items should become compromised, their impact on test scores would not be constant across different computerized adaptive testing (CAT) programs. Each CAT program is unique in a wide range of factors, such as the complexity of test specification, the characteristics of CAT item banks, item selection algorithm, item exposure control, item response theory model, pretest strategy, and scoring method. All of these factors interact with the impact of compromised items on test scores. As a result, evaluating the impact of compromised items in a CAT program is a challenging task. The current study approached the problem from a unique perspective and used a new method of simulation to quantify the impact of compromised items on the Graduate Management Admission Test® (GMAT®) CAT. Although the results are specific to the GMAT CAT program, the simulation method employed applies to any CAT program.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2009 by the Author

All rights reserved. Permission is granted for non-commercial use.

Citation

Guo, F. (2009). Quantifying the impact of compromised items in CAT. 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

**Fanmin Guo, 1600 Tysons Boulevard, #1400,
McLean, VA 22102, U.S.A. Email: fguo@gmac.com**

Quantifying the Impact of Compromised Items in CAT[1]

If a few test items should become compromised, their impact on test scores would not be constant across different computerized adaptive testing (CAT) programs. The impact might be more serious in some CAT programs than others. Like its paper-and-pencil counterpart, a CAT program might give a higher score to an examinee as a result of his or her pre-knowledge of the answers to the compromised items. Unlike its paper-and-pencil counterpart, a CAT program assembles a test for each examinee adaptively. If some compromised items are answered correctly due to pre-knowledge of the answers, the CAT algorithm might recover the true ability through the subsequent item selections, depending on the location of compromised items and the number of them. Each CAT program is unique in a wide range of factors, such as the complexity of test specification, the characteristics of CAT item banks, item selection algorithm, item exposure control, item response theory (IRT) model, pretest strategy, and scoring method. All of these factors interact with the impact of compromised items on test scores. As a result, evaluating the impact of compromised items in a CAT program is a challenging task.

Several simulation studies have been reported on the impact of compromised items but for different purposes. Steffen and Mills (1999) reported an unpublished study in Mills and Steffen (2000) on the impact of item overlap between CAT item banks for the Graduate Record Exam (GRE) program. Their study focused on the comparison of the advantage of using two overlapping CAT banks over one. Assuming some or all of the overlapping items become compromised, they simulated and reported the percent of examinees seeing these items and average score gain by examinee ability groups. They indicated that the impact on scores was defined as the difference between simulees' true ability and the simulated performance with all or some compromised items answered correctly.

Yi, Zhang, and Chang (2008) investigated the damage of two types of item theft to CAT programs that employed two different item selection algorithms. They first identified the stolen items with a simulation. Then they simulated and compared simulees' true ability (θ) and estimated ability ($\hat{\theta}$) with answers to stolen items set as correct answers.

The current study approached the problem from a different perspective and used a different method of simulation to quantify the impact of compromised items on the Graduate Management Admission Test® (GMAT®) CAT. Although the results are specific to the GMAT CAT program, the simulation method employed applies to any CAT programs.

Method

GMAT CAT

The GMAT is an admission test for applicants to post-graduate management education programs. It has three sections: quantitative, verbal, and analytical writing. The quantitative and verbal sections are both CATs. There are 37 items in the quantitative section and 41 items in the verbal section, both with pretest items embedded but not contributing to the score calculations. Four scores are reported. The quantitative and verbal scores are reported on the scale of 0 – 60 with an increment of 1. A total score based on the combined performance of quantitative and

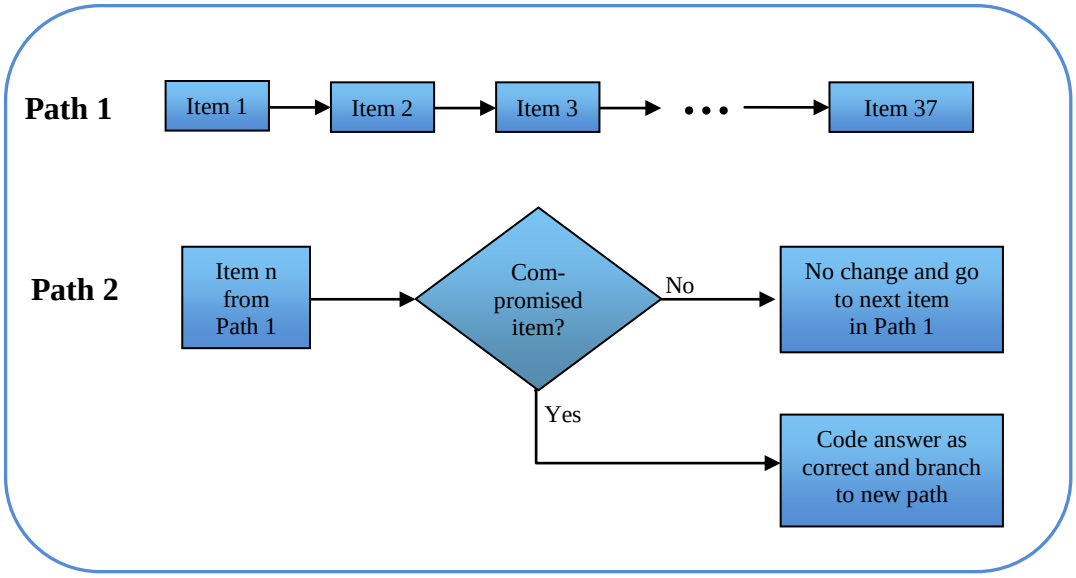
[1] A pilot simulation study under the same title was presented at the 2009 National Council on Measurement in Education Meeting, San Diego, US.

verbal sections is reported on a scale of 200 – 800 with increments of 10. The standard error of measurement is about 3.0 for the quantitative and verbal scores and is about 30 for the total score. An analytical writing score is also reported on the scale of 0 – 6 with an increment of 0.5.

Simulation

This study was a simulation study with a method focusing on the impact on individual examinees. A flowchart of the simulation process is presented in Figure 1.

Figure 1. Flowchart of the Simulation Process



The simulation study was run in two paths. The first path was a conventional simulation under a no-compromised-items condition. Then some items were selected as compromised items. The second path followed the selected items and response patterns in the first path for each simulee until a "compromised" item was "administered." Then the answer to this item was reset to a correct answer to simulate the "security breach." After that, the algorithm branched to selecting new items based on the "breached" interim $\hat{\theta}$. All the answers to subsequent "compromised" items were set as correct answers. Since each simulee had two scores from the two separate paths, this method allowed estimating the range of score gains due to the compromised items seen by each individual. The purpose of the two-path simulation method was to quantify the impact of compromised items as well as its interaction with the item selection method and other CAT operational configurations.

The simulation method implemented differed from those in the previously mentioned studies. A true θ was used as a simulee's base ability under the unbreached performance in those studies, but the base ability of a simulee was estimated from Path 1 in this study. The items selected and answer patterns were also fixed between the two paths for each simulee until a compromised item was selected. The two-path method was designed to address the concern that estimator bias (Weisstein, 2009) might influence the magnitude of the impact of the compromised items on an individual's scores.

In operational CAT programs, important assumptions have been made about the item and person parameters. Item parameter estimates ($\hat{a}$, $\hat{b}$, and $\hat{c}$) and person ability estimates ($\hat{\theta}$) are

assumed to be close to their true values and therefore used as if they were true parameters in both item calibrations from pretest data and in the estimation of examinees' ability. These assumptions are valid for group averages. However, cautions are needed when individual items or individual examinees are the interest of investigation. Table 1 presents a summary of the differences between the true θ values and the base θ estimates from Path 1for the GMAT verbal and quantitative sections by ability group. The difference can be broken down into bias and measurement error. The mean of the differences is the bias, since measurement errors are random and should sum or average to 0.

Table 1. Differences Between True and Estimated θ

	Ability Group	Mean	25^{th} Percentile	Median	75^{th} Percentile
Verbal					
	Low	.03	-.22	.03	.30
	Average	.01	-.23	.00	.24
	High	.03	-.24	.00	.28
Quantitative					
	Low	-.02	-.28	.01	.27
	Average	.00	-.21	.00	.21
	High	.02	-.20	.03	.25

In Table 1, the mean differences (bias) are all small for all groups in both sections. However, they are closer to zero in the average ability groups (0.01 for Verbal and 0 for Quantitative scores) than they are in the other ability groups (0.03, 0.03, –0.02, and 0.02). This is true for both the verbal and quantitative scores, indicating that the influence of bias on the impact is not uniform across the ability scale. They are larger at the two ends but smaller in the middle of the scale. Table 1 also shows that at least 50% of these differences are larger than 0.2 in magnitude since all the 25^{th} percentiles are ≤ -0.2 and all the 75^{th} percentiles are ≥ 0.2. A difference of 0.2 is surely to impact the magnitude of the impact of compromised items on individuals' scores. In this study, the impact of compromised items was defined as the difference between the two θ estimates from the two paths. The selected items and response patterns were fixed as much as possible between the two paths of simulation in order to keep the bias and measurement error as similar as possible between the two θ estimates.

From the examinees who took the GMAT in January 2007, 5,000 examinees were randomly selected and their quantitative and verbal θ estimates were used as the true θs in this simulation study. Items with their scaled item parameters were also the same items used in that period. The configurations in the simulation were identical to those used in the real GMAT exam. They included the test specification, CAT bank configuration, item selection algorithm, item exposure control, IRT model, pretest strategy, scoring method, and others. These configurations are not described here because the purpose of this study was not to make the results generalizable to other CAT programs and the author believes that each program has to do such a study with its own CAT operational configurations because of the complexity and uniqueness of each program. The contribution of this study is the simulation method that applies to any CAT program, in addition to the results for serving GMAC, GMAT examinees, and business schools that use

GMAT scores in the admissions processes.

In this study, the number of compromised items was set at five. This was selected based on our analyses of examinee test records from those who discussed GMAT items on the ScoreTop website. GMAC won the lawsuit against ScoreTop and confiscated its database with users' identification information. Although SoreTop users claimed that they had seen more compromised items, our analyses showed that most of them saw only one or two compromised items and rarely three. A pilot study (Guo, 2009) for this project showed that this was comparable to a situation where some examinees might have gained pre-knowledge of five items.

In order to evaluate the differential impact of compromised items of different difficulty on examinees with different ability, three levels of item difficulty and three ability groups of examinees were built into the simulation study. The items used in this study were sorted into three levels (easy, medium, and hard) of similar numbers based on their *b* parameters. The 5,000 simulees were also sorted into three ability groups (low, average, and high) based on their true θ, each group having approximately the same number of simulees. Crossing these two factors resulted in a 3×3 matrix with nine cells, each representing a combination of one item difficulty level and one ability group. For each cell, five items were randomly sampled from their corresponding difficulty level without replacement and treated as "compromised" items in the second path of the simulations.

For stable estimates of the impact, ten simulations were conducted for each cell. These simulations were performed separately for the GMAT quantitative and verbal sections.

Analyses

The two paths of simulations produced two scaled scores, converted from the two θ estimates for each simulee. The impact of the compromised items could be quantified as the gain between the two scaled scores for those who had received at least one compromised item. The gain also reflected all possible interactions between the impact of the compromised items and the influence by the CAT configuration, such as item selection and scoring methods. The percent of simulees seeing the compromised items were calculated using the simulation results. The percent of simulees gaining *n* score points and seeing *m* compromised items are reported in the next section.

The simulation performed was the worst scenario in which every simulee had pre-knowledge of five compromised items. In other words, 100% of the test population gained the pre-knowledge of five items. This is rarely the case in real CAT operations. The question asked more often is what would be the impact if a certain percent of the test population (10%, for example) have gained pre-knowledge of five items. The design in this study also enabled calculations for a smaller percent of the test population gaining pre-knowledge of the compromised items. The calculations and two examples are also given in the next section.

Results

Number of Compromised Items Seen

Ten simulations were run in each of the nine combinations of simulee ability groups and item difficulty levels. For each run, five items were randomly selected without replacement as compromised in the second paths. Tables 2 and 3 summarize the percent of simulees receiving

compromised items by ability groups and item difficulty levels. The percent in each cell is calculated after pooling the results from all ten simulations in that cell.

Table 2. Percent of Simulees Seeing Compromised Verbal Items

Examinee Ability	Item Difficulty	Number of Compromised Verbal Items Seen			
		0	1	2	≥ 3
Low	Easy	66.8%	28.1%	4.7%	0.5%
	Medium	75.6%	21.3%	2.9%	0.2%
	Hard	88.0%	11.1%	0.8%	0.1%
	Any item	**76.8%**	**20.2%**	**2.8%**	**0.3%**
Average	Easy	82.4%	16.4%	1.1%	0.1%
	Medium	74.2%	22.6%	3.0%	0.1%
	Hard	74.8%	21.6%	3.2%	0.4%
	Any item	**77.1%**	**20.2%**	**2.4%**	**0.2%**
High	Easy	86.5%	12.9%	0.6%	0.0%
	Medium	87.8%	11.6%	0.6%	0.0%
	Hard	54.7%	35.4%	8.6%	1.3%
	Any item	**76.3%**	**20.0%**	**3.3%**	**0.4%**
All examinees		**76.8%**	**20.1%**	**2.8%**	**0.3%**

The columns in Table 2 are the percent of simulees seeing 0, 1, 2, or 3 and more compromised verbal items. They are presented by item difficulty levels for each ability group. The percentages in boldface are also presented across the three difficulty levels for each ability group and across all ability groups. In order to show patterns of interactions between simulees' ability and the difficulty of compromised items, blue horizontal bars are also graphed in each cell. They are scaled within each column, so comparing the length of the bars is meaningful only within columns but *not* within rows. Some patterns are obvious.

A larger percent of low ability simulees (88%) did not see any of the five compromised hard items than that of those simulees (66.8%) not seeing any of the five easy items. A smaller percent of high ability simulees (54.7%) did not see any of the five hard items than that of those simulees (86.5%) not seeing any of the five easy items. As a result of the adaptive testing, GMAT verbal examinees had a larger chance not seeing the compromised items when these items were not appropriate for their ability. This is true of the low ability examinees with hard items and high ability examinees with easy items.

That relationship was reversed when they saw at least one of compromised item. Of the low ability simulees, 28.1% saw one easy item and 11.1% saw one hard item. On the contrary, 12.9% of the high ability simulees saw one easy item and 35.4% of them saw one hard item. Similar patterns emerged for those who saw two compromised items. The percentages were 4.7%, 0.8%, 0.6%, and 8.6%. The patterns repeated for those who saw three and more compromised items. The percentages were 0.5%, 0.1%, <0.1%, and 1.3%. GMAT verbal examinees had a larger chance to see the compromised items when these items were appropriate for their ability.

If any five items were compromised regardless of their difficulty, the percentages of simulees seeing zero, one, two, or three and more items were similar in each of the ability groups as well as across the ability groups. 76.8% did not see any compromised items; 20.1% saw one; 2.8%

saw two; and 0.3% saw three or more items.

Table 3, for the Quantitative items, has the same structure as Table 2. All the patterns for the GMAT verbal section observed in Table 2 also show in Table 3 for the GMAT quantitative section. The only differences are in the percentages in boldface for the ability groups and across ability groups, regardless of item difficulty. 74.9% (76.8% for verbal) did not see any compromised items; 21.6% (20.1 % for verbal) saw 1; 3.2% (2.8 % for verbal) saw 2; and 0.3% (0.3 %) saw three and more items. It seems that a slightly smaller percent of simulees did not see any compromised items in the quantitative as compared to the verbal section, but a slightly larger percent of simulees saw one or two items. This is not a surprise, because there were 37 items in the quantitative section but 41 items in the verbal section of the GMAT CAT Exam. Five compromised items consist of a larger proportion of total test items in quantitative than in the verbal sections. It is expected that more examinees will see the compromised items if a larger portion of the test items are compromised.

Table 3. Percent of Simulees Seeing Compromised Quantitative Items

Examinee Ability	Item Difficulty	Number of Compromised Quantitative Items Seen			
		0	1	2	≥ 3
Low	Easy	63.9%	30.1%	5.5%	0.6%
	Medium	74.0%	22.3%	3.4%	0.4%
	Hard	87.2%	12.1%	0.7%	0.0%
	Any item	**75.0%**	**21.5%**	**3.2%**	**0.3%**
Average	Easy	82.9%	15.6%	1.4%	0.0%
	Medium	66.8%	27.9%	4.9%	0.4%
	Hard	74.6%	22.0%	3.1%	0.3%
	Any item	**74.8%**	**21.8%**	**3.1%**	**0.2%**
High	Easy	87.7%	11.8%	0.5%	0.0%
	Medium	80.7%	18.0%	1.3%	0.1%
	Hard	56.1%	34.7%	8.1%	1.1%
	Any item	**74.8%**	**21.5%**	**3.3%**	**0.4%**
All examinees		**74.9%**	**21.6%**	**3.2%**	**0.3%**

Impact on Verbal Score

Ten simulations were run in each of the nine combinations of simulee ability groups and item difficulty levels. Two paths were run for each simulee. The first run resulted in a scaled score serving as the basis as the estimated ability without the impact of compromised items. The second path simulated the impact of the five compromised items. Another scaled score was calculated at the end. The score gain was calculated for each simulee by subtracting the base score in Path 1 from the score with impact in Path 2. Table 4 summarizes the verbal score gain attributable to five compromised items by ability group and item difficulty level. The percent of simulees in each cell was calculated after pooling the results from all the simulations in that cell. Again, blue horizontal bars that show the patterns are also graphed in each cell. They are scaled within each column, so comparing the length of the bars is meaningful only within columns but *not* within rows.

Table 4. Verbal Score Gain by Examinee Ability and Item Difficulty

Examinee Ability	Item Difficulty	Gain on Verbal Score				
		≤ -1	0	1	2	≥ 3
Low	Easy	0.3%	91.2%	5.3%	2.1%	1.1%
	Medium	0.6%	93.0%	3.4%	1.8%	1.1%
	Hard	0.6%	97.4%	0.8%	0.6%	0.6%
Average	Easy	0.1%	98.5%	0.5%	0.4%	0.4%
	Medium	0.5%	94.5%	2.0%	1.8%	1.3%
	Hard	0.7%	91.8%	3.4%	2.2%	1.9%
High	Easy	0.0%	99.7%	0.2%	0.1%	0.0%
	Medium	0.1%	99.3%	0.3%	0.2%	0.2%
	Hard	0.4%	87.9%	6.3%	3.7%	1.7%

The columns in Table 4 are the percentages of simulees losing one and more point, gaining no points, one point, two points, or three and more points. By design, simulees who did not receive any compromised items had the same estimated scores in both paths of simulations. Any negative gains were the real observed differences from those who did receive compromised items. In an adaptive test, pre-knowledge of items might cause examinees' scores to decrease, because the CAT algorithm tends to correct inflated interim ability estimates due to a breached item by selecting more difficult items until the interim ability estimate stabilizes around the true ability. For some simulees, the correction might lead to a lower score than their base ability. However, the percent of simulees showing negative gain scores is very small in Table 4.

The patterns of score gain are similar to the patterns of number of compromised items seen by examinees. A larger percent (97.4%) of low ability simulees did not gain any points when they had pre-knowledge of hard items than they did (91.2%) on the easy items regardless of whether they saw the compromised items or not. The opposite was true of high ability simulees, with 87.9% vs. 99.7%. Out of the low ability simulees, 5.3% gained one score point on easy items but only 0.8% gained the same amount on hard items. However, for the high ability simulees, 0.2% gained one score point on easy items but 6.3% had the same gain on hard items. Similar patterns showed for the groups of simulees with two or three and more score point gains. The corresponding percentages are 2.1%, 0.6%, 0.1%, and 3.7% for the two-point gain group and 1.1%, 0.6%, less than 0.1%, and 1.7% for the three-and-more-point gain group. These patterns show that examinees tended to gain more when the compromised items were of appropriate difficulty to their ability.

Table 5 highlights the impact of five compromised verbal items by ability groups. Note that, by design, those who did not see any compromised items would have no score changes. 0.5% of low ability simulees, 0.4% of average ability simulees, and 0.2% of high ability simulees had lower scores than their base scores after compromised items were administered to them. This pattern seems to indicate that low ability examinees might be more susceptible to the correction due to the CAT adaptive algorithm.

Table 5. Verbal Score Gain by Examinee Ability

Examinee Ability	Gain on Verbal Score				
	≤ -1	0	1	2	≥ 3
Low	0.5%	93.9%	3.2%	1.5%	0.9%
Average	0.4%	94.9%	2.0%	1.5%	1.2%
High	0.2%	95.6%	2.3%	1.3%	0.6%
All	**0.4%**	**94.8%**	**2.5%**	**1.4%**	**0.9%**

Of all simulees, 93.9% of low ability simulees, 94.9% of average ability simulees, and 95.6% of high ability simulees did not gain any score points. High ability examinees seemed to be less likely to have score gains although they had a similar probability of seeing the compromised items (Table 2). High ability simulees also showed smaller gains than the low ability simulees: 2.3% vs. 3.2% gaining one point, 1.3% vs. 1.5% gaining two points, and 0.6% vs. 0.9% gaining three points or more.

Across all ability groups, or for the whole test population, 94.8% did not gain any points; 2.5% gained one point; 1.4% gained two points; 0.9% gained three or more points; and 0.4% received lower scores, assuming they had pre-knowledge of five compromised items. Note that the standard error of measurement (SEM) for GMAT verbal scores is 3 points. Score gains of only 0.9% of the examinees were equal to or larger than one SEM.

Impact on Quantitative Score

Using the same methods, score gains were calculated for the GMAT quantitative section and are summarized in Tables 6 and 7. Both are in the same format as Tables 4 and 5.

Table 6. Quantitative Score Gain by Examinee Ability and Item Difficulty

Examinee Ability	Item Difficulty	Gain on Quantitative Score				
		≤ -1	0	1	2	≥ 3
Low	Easy	0.9%	88.9%	3.9%	3.3%	3.0%
	Medium	1.1%	91.4%	3.0%	2.1%	2.4%
	Hard	0.5%	97.9%	0.6%	0.4%	0.7%
Average	Easy	0.3%	98.7%	0.4%	0.3%	0.3%
	Medium	0.7%	92.0%	3.1%	2.2%	2.0%
	Hard	0.8%	92.7%	2.9%	1.8%	1.8%
High	Easy	0.1%	99.8%	0.1%	0.0%	0.0%
	Medium	0.1%	98.8%	0.7%	0.2%	0.2%
	Hard	0.4%	92.2%	5.4%	1.4%	0.7%

Table 6 has an identical structure to that of Table 4. All the patterns observed for the verbal scores in Table 4 also show in Table 6.

1. A small percent of simulees showed negative gains when compromised items were administered to them.

2. A larger percent of low ability simulees did not gain any points when they had pre-knowledge of hard items compared to their gains on the easy items, regardless of whether they saw the compromised items or not. The opposite is true of high ability simulees.
3. Low ability simulees gained more on easy items and high ability simulees gained more on hard items. These patterns show that examinees tended to gain more when the compromised items are of appropriate difficulty for their ability.

Table 7. Quantitative Score Gain by Examinee Ability

Examinee Ability	Gain on Quantitative Score				
	≤ -1	0	1	2	≥ 3
Low	0.8%	92.7%	2.5%	1.9%	2.0%
Average	0.6%	94.5%	2.1%	1.4%	1.4%
High	0.2%	96.9%	2.1%	0.5%	0.3%
All	**0.5%**	**94.7%**	**2.2%**	**1.3%**	**1.2%**

Table 7 summarizes the quantitative score gains by examinee ability groups. An obvious pattern in Table 7 shows that low ability simulees had larger percentages in both positive and negative score gains than the average ability simulees, whose percentages were larger than those of the high ability simulees. For those who did not gain at all, low ability simulees seemed to have a smaller percent than that of the average ability simulees, which is in turn smaller than that of the high ability simulees. These patterns indicate that high ability examinees benefited less than their low ability counterparts. The reason might be that high ability examinees would answer some of the compromised items correctly even if they did not have pre-knowledge of them. Therefore, the impact of some compromised items was zero in the second path of the simulation.

Across all ability groups, or for the whole test population, 94.7% did not gain any points; 2.2% gained one point; 1.3% gained two points; 1.2% gained 3 and more points; and 0.5% received lower scores, assuming they had pre-knowledge of five compromised items. Note that the SEM for GMAT quantitative scores is about 3 points. Score gains of only 0.5% of the examinees were equal to or larger than one SEM.

Impact on Total Score

In addition to the quantitative and verbal scores, a GMAT total score is also reported. It is not a linear transformation of the sum of the two section scores. It is on a different scale (200 – 800 with increments of 10) based on the combined performance of the two sections. The SEM for the GMAT total score is about 30 points.

The five compromised verbal or quantitative items will impact the total scores as well as the section scores. In this simulation study, the effect of the five compromised verbal items on the total score was calculated as the combined performance of Path 1 of the quantitative section and Path 2 of the verbal section, and the impact of the five compromised quantitative items on the total score was calculated as the combined performance of Path 2 of the quantitative section and Path 1 of the verbal section for each simulee. The impact was calculated as the difference between the affected scores and the base scores from Path 1s for both sections. Table 8

highlights the impacts.

Table 8 Impact on Total Score by Compromised Verbal or Quantitative Items

Items	Gain on Total Score				
	≤ -10	0	10	20	≥ 30
Verbal	0.5%	95.0%	3.0%	1.0%	0.5%
Quant.	0.6%	95.0%	2.8%	1.0%	0.5%

Table 8 shows that the impact on total scores was similar between the five verbal or quantitative compromised items. About 95% simulees had no score changes at all; about 0.5% or 0.6% lost 10 or more points; about 3% or 2.8% gained 10 points; about 1% gained 20 points; and about 0.5% gained 30 or more points. As the SEM of GMAT total score is 30 point, about half a percent of the examinees gained one SEM or more.

If Not All Examinees Had Pre-Knowledge of the Five Compromised Items

Test companies often want to know what the impact would be if only a certain percent of the test population gained the pre-knowledge of the five compromised items. This impact can be calculated from the simulation results.

Assume that 10% of the test population gained pre-knowledge of 5 compromised items. What are the percentages of examinees seeing the items? Table 2 was used as an example for calculating the cell percentages and the results are presented in Table 9. For all the cells in the columns for one, two, and three or more items seen, the corresponding cell percentages in Table 2 were multiplied by 0.1. For the cells in the column of zero-items seen, the percentages are 100%—the sums of the three cell percentages in the same row. That is

$$\tilde{p}_{s\neq 0} = p_{s\neq 0} \times x \qquad (1)$$

and

$$\tilde{p}_{s=0} = 100\% - \sum_{s\neq 0} \tilde{p}_s \,, \qquad (2)$$

where $\tilde{p}$ is the cell percent for the 10% breaching scenario in Table 9, p is the cell percent of the worst scenario (100% breaching) in Table 2, s indexes the number of compromised items seen (zero, one, two, and three or more) within each row, and x is the percent of examinees who gained the pre-knowledge of compromised items; x equals 0.1 in this example. Table 9 is the percent of simulees seeing the verbal compromised items if only 10% of the test population gained pre-knowledge of five compromised items by examinee ability and item difficulty.

Table 9. Impact of 10% Examinees Gaining Pre-Knowledge: Number of Items Seen

Examinee Ability	Item Difficulty	Number of Compromised Verbal Items Seen			
		0	1	2	≥ 3
Low	Easy	96.67%	2.81%	0.47%	0.05%
	Medium	97.56%	2.13%	0.29%	0.02%
	Hard	98.80%	1.11%	0.08%	0.01%
	Any item	**97.68%**	**2.02%**	**0.28%**	**0.03%**
Average	Easy	98.24%	1.64%	0.11%	0.01%
	Medium	97.43%	2.26%	0.30%	0.01%
	Hard	97.48%	2.16%	0.32%	0.04%
	Any item	**97.72%**	**2.02%**	**0.24%**	**0.02%**
High	Easy	98.65%	1.29%	0.06%	0.00%
	Medium	98.78%	1.16%	0.06%	0.00%
	Hard	95.47%	3.54%	0.86%	0.13%
	Any item	**97.63%**	**2.00%**	**0.33%**	**0.04%**
All examinees		**97.68%**	**2.01%**	**0.28%**	**0.03%**

The same method also applies to the tables of score gains, with *s* indexing the score gain categories (≤–1, 0, 1, 2, ≥3) within each row. Table 5 can be used as an example for calculating Table 10. For the cells in the columns of ≤–1, 1, 2, ≥3, multiply the corresponding cell percentages in Table 5. For the cells in the column of 0-point gain, the cell percentages are 100%—the sums of the four percentages in the same rows. Table 10 presents the percent of examinees gaining score points if only 10% of the examinees gained pre-knowledge of five verbal items.

Table 10. Impact of 10% of Examinees Gaining Pre-Knowledge: Points of Score Gains

Examinee Ability	Gain on Verbal Score				
	≤ -1	0	1	2	≥ 3
Low	0.05%	99.39%	0.32%	0.15%	0.09%
Average	0.04%	99.49%	0.20%	0.15%	0.12%
High	0.02%	99.56%	0.23%	0.13%	0.06%
All	0.04%	99.48%	0.25%	0.14%	0.09%

If 10% of the examinees gained pre-knowledge of 5 GMAT verbal items, the impacts are:

1. About 97.7% of the test population will not see any of the compromised items.
2. About 2% will see one item in their verbal section.
3. About 0.3% will see two items.
4. About 0.03% will see three or more items.
5. About 99.5% of the test population will gain no points on the verbal scale score.
6. About 0.25% will gain one point.
7. About 0.14% will gain two points.

8. About 0.09% will gain three points or more.
9. About 0.04% will get lower scores.

Discussion and Conclusions

In this study, a two-path simulation method was used to investigate the impact of five compromised items in the GMAT CAT exam. The impact was estimated for easy, medium, and hard compromised items on low, average, and high ability examinees, as well as items across difficulty levels on examinees of all ability groups. The following conclusions can be drawn:

1. With five verbal or quantitative items compromised, about 21% will see one item; about 3% will see two items and about 0.3% will see three or more items. About 77% of verbal examinees and 75% of quantitative examinees will not see any of the compromised items.
2. For the verbal section, about 95% of examinees will have no score gains at all; 2.5% will have a one point gain; 1.4% will gain two points; about 1% will gain three points or more. However, 0.4% will have lower scores.
3. For the quantitative section, about 95% examinees will have no score gain; 2.2% will have a one point gain; 1.3% will gain two points; about 1.2% will gain three points or more. However, 0.5% of them will have lower scores.
4. The impact on GMAT total score shows the same patterns but with increments of 10 points instead of 1.
5. Examinees have a higher chance of seeing the compromised items if the items are appropriate for their ability. This is true for both GMAT quantitative and verbal sections.
6. Examinees have a higher chance of score gains when the compromised items are appropriate for their ability for both GMAT sections.

The above results are reported assuming that the items have been compromised to such an extent that all examinees have gained pre-knowledge. If the security breaching is not as bad and only a certain percentage of examinees have gained pre-knowledge of five items, the impact will be smaller and can be calculated with the method described.

Interpretations of the results can be made from two perspectives: that of a test company and of examinees. To a testing company, this is an assessment of the risk of five compromised items in terms of score change. Although about 5% of the examinees had score gains, only about 1% had gains of one SEM or more for either verbal or quantitative sections. If the test breach is less serious, the impact is even smaller.

For the examinees, a one-point gain is a gain whether it is 0.1 or 1 SEM. Interpretations can be made for individual examinees. For example, if an examinee gains pre-knowledge of five items, he or she will have a 76% chance of not seeing any of them, a 21% chance of seeing one, a 3% chance of seeing two, and a 0.3% chance of seeing three or more of the items in his or her test. He or she also has a 95% chance of not gaining any score points at all, 2.5% or 2.2% chance of gaining one point, 1.4% or 1.3% chance of gaining two points, or a 0.9% or 1.2% chance of gaining three or more points on verbal or quantitative scores. However, there is also a small chance that his or her score will be lower because of the compromised items seen on the test. Of course, if he or she gains pre-knowledge of test items of appropriate difficulty to his or her ability, the chances of seeing them and gaining score points are larger. An examinee would ask himself or herself whether it is worth risking the consequences of being caught cheating on the GMAT test, given the small magnitude of score gains.

This study intended to showcase the simulation and analysis methods. Although the results from this study are not generalizable to other CAT programs, the methods presented here apply to any CAT program. The author believes that each CAT program is so unique in test configuration that each one should do such a study to quantify the impact for its own program.

References

Guo, F. (2009). *Quantifying impact of compromised items in CAT*. Paper presented at the 2009 National Council on Measurement in Education Meeting, San Diego CA.

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

Steffen, M., & Mills, C. (1999). *An investigation of item overlap and security risks in an operational CAT environment*. Unpublished manuscripts.

Weisstein, Eric W. (2009). Estimator bias. In *MathWorld--A Wolfram Web Resourc*e. Retrieved May 19, 2009 from http://mathworld.wolfram.com/EstimatorBias.html

Yi, Q., Zhang, J., & Chang, H. (2008). Severity of organized item theft in computerized adaptive testing: A simulation study. *Applied Psychological Measurement, 32*, 543-558.