Graduate
Management
Admission
Council®

# CAT Security:
# A Practitioner's Perspective

**Fanmin Guo**
**Graduate Management Admission Council**

*Keynote Address Presented June 7, 2007*

## Abstract

The security of computer adaptive testing (CAT) programs is a continued concern of measurement practitioners, because security breaches might destroy the validity of the test scores. This paper focuses on the factors that might impact the security of a CAT program, including item types, test length, item selection algorithm and content balance, exposure control, rotation of items, length of testing window, test center capacity, and the operational cost. The purpose of this paper is to reveal the possible impact of these factors on CAT security and to discuss how they should be handled in the design, administration, and cost considerations of a secure CAT program from the perspective of a measurement practitioner.

## Acknowledgment

**Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.**

## Copyright © 2007 by the Authors

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

## Citation

**Guo, F. (2007). CAT Security: A practitioner's perspective. In D. J. Weiss (Ed.),**
***Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.***
**Retrieved [date] from** www.psych.umn.edu/psylabs/CATCentral/

## Author Contact

**Fanmin Guo, Graduate Management Admission Council,**
**1600 Tysons Blvd, Suite 1400, McLean, VA 22102, USA**
**Email: fguo@gmac.com**

# CAT Security: A Practitioner's Perspective

The security of computer adaptive testing (CAT) programs is a continued concern of measurement practitioners, because security breaches of CAT programs might invalidate the scores for their intended use. For example, schools will not be able to use the scores from admission tests as important evidence for valid selection of candidates for admissions, because the scores will no longer rank-order candidate correctly since some candidates have scores higher than what they should be. All practitioners in high-stake testing programs, CAT or paper-and-pencil tests (PPT), take test security very seriously. I shared the ideas of this keynote speech with a friend, who has more than fifteen years' experience with high-stake, large-scale CAT programs. His comments are "the topic isn't the most glamorous, but the realities of not thinking about them in advance can be severe."

It is often too late to address the security concerns of a CAT program when it is operational, because many of the measures for securing a CAT program have to be considered and incorporated into the design of a CAT program. This paper focuses on factors that might impact CAT security and how they should be handled in the design and the administration of a CAT program from the perspective of a measurement practitioner. I will group these factors into three categories: test design, test administration, and cost.

Before addressing these factors in detail, I would like to call our attention to the relationship of test validity and reliability in the CAT context. Not all the measures for increasing CAT security discussed in the literature and conference papers will increase the reliability of CAT test scores. In fact, a lot of them become feasible at the sacrifice of score reliability or score precision to some extent. It is important that a balance among all the measures for addressing both CAT security and score reliability be achieved in the design and implementation of a CAT program. The ultimate goal is to have a secure CAT program with reasonable reliability of its scores. I do not have a ready-to-use design that will guarantee an absolutely secure and reliable CAT. Rather, I will share my concerns of the factors that might impact CAT security and give recommendations and possible solutions from my own experience with CAT programs.

## CAT Security and Test Design

Decisions on the design of a CAT program might have direct impact on how secure the CAT program might be. A well-designed program with security as one of the design goals makes far more security measures available in the operation of the program. In this section, I will discuss the factors that need to be considered in the design for a secure CAT. These factors are item type, test length, item selection algorithm, and exposure control.

### Item Type

All item types are not equally secure. Some are easier to remember and easier to share among examinees after the item is administered. Let's consider three examples. The correct answer is marked with an *.

*Example 1:* An antonym item

succinct
(A) wordy *
(B) oblique
(C) bright
(D) industrious

This item type is very easy to remember, to recreate, and to share among examinees. It takes only two words, "succinct and wordy," for examinees to share the item.

*Example 2:* An analogy item

conglomerate : rock
(A) fish : water
(B) scissors : paper
(C) train : aircraft
(D) sofa : furniture*

This item type is more difficult to remember than Example 1. Examinees have to remember two pairs of words to share the item. However, it is still easy.

*Example 3:* A critical reasoning item from The Official Guide for GAMT® Review, (11th Edition).

In order to increase profits during a prolonged slowdown in sales, the largest manufacturers of automobiles in the United States have instituted record-setting price increases on all their models. The manufacturers believe that this strategy will succeed, even though it is inconsistent with the normal relationship between price and demand.

The manufacturers' plan to increase profits relies on which of the following assumptions?

(A) Automobile manufactures will, of necessity, raise prices whenever they introduce a new model.
(B) The smaller automobile manufactures will continue to take away a large percentage of business from the largest manufacturers.
(C) The increased profit made on cars sold will more then compensate for any decline in sales caused by the price increases.*
(D) New safety restraints that will soon become mandatory for all new cars sill not be very costly for manufacturers to install.
(E) Low financing and extended warranties will attract many price-conscious consumers.

This item type is much more difficult to remember than the first two examples, since it requires reading and understanding the content. It is almost impossible to share the text from memory.

With the computer as a delivery medium for CAT, new item types that take the advantage of computer interfaces can be far more secure than their paper-and-pencil counterparts because they can be very difficult to remember, to recreate, or to share among examinees. When designing a new CAT program or redesigning an existing program, it is important to make use of the item types that are more secure and measure what is to be measured better and more efficiently.

## Test Length

It is common knowledge that a CAT is more efficient, therefore shorter, than its linear counterpart. The reason is that CAT will select and administer items with appropriate difficulty for each examinee. It does not waste items by avoiding items that are too difficult or too easy for an individual examinee. However, an overly short CAT will create some problems in item use and test security.

Test length is related to score accuracy. Under the item response theory (IRT) framework, score reliability and standard error of measurement (SEM) are linked to test information. Higher test information leads to lower SEM or higher reliability. Higher test information can be achieved in two different ways. For similar items, longer tests tend to have higher score reliability, or smaller SEM than shorter tests. For the same test length, items with higher item information will lead to higher test information. If $I(\theta)_T$ is test information, $I(\theta)_I$ is item information, and *n* is test length in number of items, we have the following.

$$I(\theta)_T = \sum_n I(\theta)_I \tag{1}$$

$$I(\theta)_T = \frac{\sum_n I(\theta)_I}{n} \times n \tag{2}$$

$$I(\theta)_T = \overline{[I(\theta)_I]} \times n \tag{3}$$

From Equation 3, we know that test information is the average item information times the number of items.

For a CAT, $I(\theta)_T$ can be the target test information that will guarantee acceptable score reliability; $\overline{I(\theta)_I}$ is the average quality of the test items, and *n* is the number of items in the CAT . For the same score reliability (test information), decrease in *n* has to be compensated by the increase in the average item information. That is, a shorter test has to consist of better (more informative) items to maintain the same level of score reliability. If the length of a CAT is too short, the best items will be selected and administered very often and less informative items will be used less often or even rarely. From the perspective of CAT security, over-exposing a small number of items to a large number of examinees can be a serious threat, because examinees who have taken the test might share the items with those who will take test later and these examinees might benefit from pre-knowledge of these items. A reasonable test length is crucial for a CAT to allow balanced use of items of different information in order to reduce the threat of security breaches while maintaining reasonable score reliability.

## Item Selection Algorithm

A CAT item selection algorithm is designed to achieve two goals. One is to implement the test specification so that all the CATs are assembled to the specifications of content representation. The other task is to select the item that has the most favorable psychometric properties, such as item difficulty and item information, to be administered to a examinee based on his or her ability. Some item selection algorithms handle the content specifications and item statistical properties simultaneously. One example of such an algorithm is the weighted deviation model (WDM) by Stocking & Swanson (1991). It has been used in large-scale CAT programs. Other algorithms separate content balancing from selecting items with the most favorable psychometric properties. A "bin-structure" content balancing method (Steffen, 2005) is a promising example of these algorithms. The basic idea is very simple. For a fixed-length CAT test with $n$ items, a similar number of "bins" of items are constructed by partitioning the item bank in such a way that each bin represents a combination of sub-content areas and that a test assembled by randomly selecting one item from each bin satisfies all the constraints set by the test specifications. Selection of items with the most favorable psychometric properties is implemented within each bin. No matter what algorithm might be used for a CAT program, the definition of "most favorable psychometric property" has serious consequences on the security of a CAT program.

Some define the most favorable items as the items with the highest information at the examinee's current estimated trait level ($\theta$). Others define them as items with a $b$ parameter closest to the current estimated $\theta$. Depending on the IRT models used for a CAT program, the two definitions might impact test security differently. In case of the one-parameter IRT model, there is no difference, since all items with the same location/difficulty ($b$) parameter have same information at $b$. Therefore, selecting items by information has the same result as selecting by $b$ parameters. It is quite a different story if a two- or three-parameter IRT model is used. In this case, items with the same $b$ parameters but different $a$ (discrimination) and/or $c$ (pseudo-guessing) parameters do not have the same information at $b$. Therefore, the maximum information based selection algorithm always selects the most informative items and ignores items with similar difficulty but lower information. On the other hand, $b$ parameter based algorithms will treat items with the same $b$ parameters as equally favorable, although some might provide higher information.

A maximum information algorithm tends to select high $a$ parameter items far more often than those with low $a$ parameters. It will achieve a higher score precision level than the $b$-based algorithm because the resultant test information tends to be higher. However, it causes two problems in test security. The first is the repeated use of high information items, making them more vulnerable to become compromised than other items. At the same time, items with less information are used less. In some extreme cases, some low information items will never be used at all. This is a scenario where the use of the items in the bank would be unbalanced and the total number of items available for CAT operations would be greatly reduced. The second problem is that the algorithm might make new item development for CAT so expensive that no testing company could afford such a program. The only way to reduce the cost is to limit the total number of new items. That would be a disaster to test security. This problem would arise in the fact that the average quality of items selected by these algorithms is inevitably higher than the average quality of the item in the item bank, because low information items will be less used. It is not easy for item writers to write only items with high $a$ parameters. There seems to be a limit

to the percentage of high *a* parameter items among all items they write. Therefore the maximum information algorithm does not only waste item development cost, but also it might accelerate compromise of high information items.

In contrast, algorithms selecting items on difficulty tend to have a more balanced selection of items, although it takes more items to achieve the same score precession level due to lower test information when compared with the other algorithm. However, if the test length is reasonable, this should not be a problem. Note that CATs with difficulty based selection algorithms remain more efficient than their linear counterparts because they adapt to the ability levels of the examinees.

For security reasons, try to use an item selection algorithm for balanced use of items, avoid algorithms that only select high *a* value items (using "the cream" only), avoid using only better-than-average items and not using average and below-average items, and achieve test efficiency by selecting items with appropriate difficulties.

### Exposure Control

The purpose of controlling item exposure in a CAT program is to prevent items that have favorable psychometric or content characteristics from being over-exposed because of security concerns. Sympson and Hetter (1985) presented a probabilistic exposure control method that imposes exposure control for the items in a CAT pool based on their probabilities of being administered given being selected. This method allows control of the exposure rate of the items in the pool at a specified level. Stocking & Lewis (1998) extended Sympson and Hetter's method to controlling not only the overall exposure but also the exposure rate of items for examinees with similar ability. Both methods require that the exposure control parameters be estimated through simulations before a CAT bank can be used. If the simulations are not set properly, or the assumptions for the simulations are not reasonably met, resultant exposure control parameters might not be accurate. In such situations, exposure control might fail for some items.

There is a group of simple exposure control methods that add a random component to the item selection methods, such as the method by Kingsbury and Zara (1989). They share the strategy of randomly selecting one from several most favorable items. This method can be used with programs with simple content constraints and an item selection algorithm that resulst in more or less balanced item use. It is less effective in protecting the most favorable items. For testing programs that employ complicated content constraints to implement complex content specifications, the Sympson and Hetter and the Stocking and Lewis methods might be better choices.

## CAT Security and Test Administration

### Rotation of Items (Item Pools

Using the entire item bank of a testing program is probably most psychometrically desirable in assembling adaptive tests for individual examinees, because the item selection algorithm will select the best items from all available items and will build the best adaptive tests possible for that program. If a testing program had an extremely large item bank with hundreds of thousands of usable items, the concerns of test security would be minimal since the probability of an individual examinee obtaining previous knowledge about an item is very, very small. Even if he or she had gained knowledge of some of the items in the bank, he or she might not see those compromised items at all in their CAT. However, no testing program can afford to have such a

luxury and most testing programs have much limited sizes of operational item banks. Therefore, using an entire item bank in CAT operations could be detrimental for a CAT program from the perspective of test security, because all the available items will be exposed to examinees continuously throughout the testing window. The current practice in addressing the security concerns is to partition the item bank into smaller units of items (called "item pools") for operational use. Rotating item pools in and out of operation becomes a tool for deploying different items at different times during the testing window. The focus of this paper is not on how to partition the item bank and form optimized pools, but on the factors of the pools related to test security.

One factor is the size of a CAT pool. If everything is equal, larger pools will lead to lower exposure rates of the items in the pool, which helps test security. However, if the size of the item bank is fixed, larger CAT pools mean few pools. The size of CAT pools is also directly related to the complexity of the content constraints. Tests with more and difficult-to-meet constraints tend to require larger pools and tests with simple content structures work with small pools. Longer tests also require larger pools than their shorter counterparts. Each CAT program needs special research in order to define an optimal pool size. From the perspective of security, larger pools and more pools tend to reduce the risk.

Another factor that might impact test security is the frequency of rotating items by pools. More frequent rotation of CAT pools will help shorten the duration of the use of individual pools, making more items available for the operations in a given period. That will make it less likely to "steal" the items and to benefit from pre-knowledge of the compromised items. Again, due to the limit of the total number of items available for a testing program, the number of CAT pools that can be assembled is not unlimited. Rapid rotation of CAT pools also means shorter time spans between reuse of the pools. Reuse of intact pools encourages organized efforts to steal items and makes it more likely to benefit from pre-knowledge of items.

It is especially vulnerable to rotate pools according to chronological patterns, such as once or twice a month, because item thieves might figure out the pattern and take advantage of it. Therefore items (pools) should be rotated as often as the number of pools allows and the rotation should follow some random pattern.

The other factor that has impact on test security is the reuse of items. It is neither possible nor economical to operate a CAT program without reuse of items. Most CAT programs, including high-stake programs, have to reuse items in order to operate a sustainable testing program. The common practice is to shelf some high exposure items for a period of time before bringing them back to operations. Here are some recommendations: Reuse as few items as possible to minimize the impact on security; reuse items that have been administered to fewer examinees first; reuse items that have been shelved for a longer period; and try to shelf used items as long as possible.

### Length of Test Window

The length of test windows is related to the accessibility of the test. For PPTs, a test window is an administration of the test, which is usually the same time as required to finish the test. For a CAT, if the test volume is very small and there are enough test seats for all examinees to take the tests assembled from a single pool, a test window can be defined similarly as that for a PPT. However, for large volume CAT programs, the length of test windows is defined quite differently. Most CATs are administered in computer test centers where physical and computer settings are standardized. It is often impossible to use a single CAT pool and schedule all

examinees in a single administration because of the limit of available test seats, since it is extremely costly to maintain computer test centers. The common practice is to build small computer centers with a limited number of test seats and spread the test volume across multiple days or months. Thanks to the fact that examinees do not receive the same items even if their tests are assembled from the same CAT pool, CAT pools can be in operation for multiple test administrations. Therefore, the length of a CAT test window can last several days or even several months. By rotating pools in and out, some CAT programs are available 24 hours a day and seven days a week all year round. This is called continuous testing. Others are limited to specific days or weeks of a month.

Continuous testing is the best for examinees because they might schedule and take the test whenever they need to. It also provides for the best access to a CAT. I have received very positive feedback from examinees who had to have their GMAT® scores in the schools they were applying to in two weeks in order to be considered for admission or for financial aid applications. They were happy that they were able to schedule and take the test within one or two days and have their scores in schools in such a short time. Schools love continuous testing, too, because candidates can take the GMAT at any time of a year.

However, from a perspective of test security, continuous testing presents the most risk because items are continuously exposed to theft as well as examinees. This is worsened by the fact that most CAT programs have to reuse items, which encourages cheaters to try to gain benefit from pre-knowledge of items. An extremely large item bank and a steady supply of pretested new items are needed in order to maintain a continuous CAT program. That makes such a program very, very expensive to maintain.

Non-continuous CAT programs are relatively more secure but limit the access of examinees to the test. While it might save money in item supplies, it might cost more in testing seats because it requires more seats than continuous testing.

### Seat Capacity

Seat capacity plays a very important role in the operation of a CAT program. Unlike PPT tests, seat capacity has direct impact on the cost and security of CAT programs. As computer test centers are expensive, it is economical to maintain just enough test seats for the test volume in a defined time period. However, limited seat capacity might cause concerns in test security.

Large seat capacity allows testing more examinees simultaneously. This is very favorable in test security for the following reasons.

1. Test windows can be shortened without concerns about losing examinees due to an inadequate number of test seats.
2. Fewer CAT pools are needed in a given period of time since each pool will accommodate more examinees.
3. More efficient use of items because each item will be given to more examinees.
4. Fewer CAT pools and more efficient use of items also mean less reuse of items when non-overlapping pools are not possible.

In contrast, inadequate seat capacity can be detrimental to a CAT program because it might have the following consequences.

1. The test window for a program has to be excessively long to accommodate all examinees, causing concerns about item theft and invalid test scores due to pre-knowledge of live items.
2. In order to keep the test secure in case of long test windows, more CAT pools are developed and deployed.
3. It is less efficient use of the items since each items is given to fewer examinees within a given period of time.

A recent example of failure of a program due to inadequate seat capacity is the planned new GRE test, although it would not have been a CAT. ETS canceled the plan and cited the concern of inadequate test seats as the reason.

## CAT Security and Cost

CAT is expensive. Programs should consider the cost of going CAT carefully before making the leap. The cost of maintaining a CAT programs incurs in three major types. One is the cost of test development. This is largely the development of new items, the assembly of CAT pools, and the packaging of components (pools, item selection and exposure control algorithms, scoring algorithms, and so on) for operational delivery. The most costly on the list is the development of new items because CAT programs require far more items than its PPT counterpart for security reasons. Another type of cost is test delivery through computers and secured networks, both of which are housed in computer test centers. Compute test centers are far more expansive to maintain then the test sites for PPT administrations. The third type of cost is post-administration analyses and score reporting. This is similar cost for both CAT and PPT test programs. In this section, I will discuss the cost of new item supply and cost of test seats because both exert direct impact on test security.

### Cost for New Item Supply

Most people understand that CAT is shorter then its PPT counterpart and assume that a CAT program requires a smaller supply of new items. Quite to the contrary, CAT programs tend to require more items than their PPT counterparts for security reasons, especially a continuous testing CAT.

CAT programs with continuous testing are favored by both examinees and schools, but they require an enormous supply of new items to minimize item reuse for security reasons. Keeping a steady supply of new items is crucial for a secure CAT. However, the development of new items is usually one of the most costly budget lines in operating a test program. Continuous testing is the most expensive configuration of CAT programs. Research efforts have been devoted to writing algorithms to automatically generate items with certain known statistical and content characteristics but so far they are not yet ready for CAT operational use. In short, the cost factor needs to be considered before switching a program from PPT to CAT. Be aware of the high cost when you plan for a continuous CAT program.

Item selection algorithm also interacts with the amount of item supply in a CAT program. Some algorithms dictate larger item supply than others. For example, algorithms based on maximum information only choose items with higher information and ignore lower-information items when higher-information items are available. As a consequence, only a portion of the item bank is used in operations and other items are never seen by examinees even if they are included

in CAT pools. Can we request that item writers write only high information items? We can try, but chances are very low that they will be able to do that. The reality is that our item writers do not have the knowledge or skills to write items with certain statistical properties. Given that, we will continue to receive items with varied information from item writers. If an item selection algorithm chooses to use high information items only, others will be sitting in the item bank and wasted. To maintain a secure CAT, we are forced to spend a large amount of money on writing more and more items, some of which will go into operations and others will simply be deposited in our item bank.

CAT tests that are too short might also cause unbalanced use of items as discussed above. They rely on fewer high information items to achieve a given level of score precision and ignore low or even medium information items. Therefore, short CAT programs also increase the cost of item supply in order to maintain score precision and test security.

From the perspective of test security, continuous testing and unbalanced item use drives up operational cost. If funding for item development is inadequate, the security of the CAT program might suffer.

### Cost of Test Delivery

Another major item on the operational cost list is CAT test delivery. The majority of it goes into the maintenance of computer test centers because they are much more expensive than the desk-and-chair-plus-#2-pencil PPT test sites. The total number of available seats also interacts with the cost and the security of a CAT program.

In order to maximize the use of the items in a CAT pool, the more seats the better during a test window. However, those seats will be wasted during the period the test is not administered. It makes the most business sense to build just enough seats to host the annual volume of a test. That total number of seats is the annual volume divided by 365 if only one test session is possible per day or the annual volume divided by 365 and then by 2 if two test sessions are possible per day. The corresponding test window is continuous testing and test centers are in operation all year round to their full capacity. In this sense, continuous testing is the most efficient delivery method although it requires the largest item supply for security reasons.

Generally speaking, a limit in the number of test seats might have a negative impact on CAT test security. It reduces the item use efficiency within a test window because few examinees take the test, causing more administrations to be available to test the same number of examinees. More administrations of the test means more CAT pools. There are two consequences, depending on the CAT operations. If non-overlapping pools are used across administrations, a limit in the number of seats forces the programs to develop more items, thus driving up the item development cost. If items are reused, it forces a high rate of item reuse, which is a big concern for test security.

## Summary

I have discussed the factors that might impact test security of CAT programs from the perspective of measurement practitioners. They are item types, test length, item selection algorithm and content balance, exposure control, rotation of items, length of testing windows, test center capacity, and the operational costs. A balance among all the factors is crucial for a secure CAT program that yields reliable and valid test scores with a controllable cost. It is

important to address security concerns in the designing phase of a CAT program because most of the factors interact with each other. Failure to balancing these factors could have serious problems in maintaining a CAT program.

## References


Graduate Management Admission Council® (GMAC®). (2005). The Official Guide for GMAT® Review, 11th edition.

Kingsbury, G.G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2(4),* 359–375.

Steffen, M. (2005). *Substantive considerations in item bin definition*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23,* 57– 75.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17,* 277–292.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association*. San Diego, CA: Navy Personnel Research and Development Center, 973–977.