Research Report

The Effectiveness of Enhancing Test Security by Using Multiple Item Pools

Jinming Zhang

Hua-Hua Chang

Research & Development

September 2005
RR-05-19

The Effectiveness of Enhancing Test Security by Using Multiple Item Pools

Jinming Zhang
ETS, Princeton, NJ

Hua-Hua Chang
University of Illinois at Urbana-Champaign

September 2005

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This paper compares the use of multiple pools versus a single pool with respect to test security against large-scale item sharing among some examinees in a computer-based test, under the assumption that a randomized item selection method is used. It characterizes the conditions under which employing multiple pools is better than using a single whole pool in terms of minimizing the expected number of compromised items encountered by an examinee. The results obtained in this paper provide guidelines for constructing multiple pools optimally and evaluating the risks associated with some test designs that employ item reuse schemes. Finally, an adjusted-pool approach is proposed to achieve better test security than either the single-pool approach or the multiple-pool approach.

Key words: CBT, CAT, continuous-delivery test, multiple item pools, item sharing, item piracy, test security

1. Introduction

Computer-based testing (CBT), including computerized adaptive testing (CAT), makes it possible for educational and psychological tests to be administered frequently or continuously. A continuous-delivery test is preferred by examinees because of the flexibility it provides in scheduling to take the test. However, continuous testing causes constant item exposure that increases the risk of test item sharing; examinees who took tests earlier may share information with those who will take tests later. The tactic of memorizing and sharing test item information will inflate test scores for some examinees and consequently hurt other honest examinees.

A new problem that has emerged from current applications of CBT is large-scale coordinated item sharing activities (Steinberg, 2002; Wheeler, 2002; Davey & Nering, 2002). The rapid growth of Internet applications all over the world has created a new format for cheating, namely item piracy. That is, some test takers organize a special interest group and share test item information with each other through the Internet. As indicated by recent CBT test security incidents, filing lawsuits can hardly stop this kind of item sharing. Without effective measures, this piracy and sharing form of cheating could significantly undermine the credibility of any continuous-delivery test. Thus, test security becomes one of the major unsolved issues for continuous-delivery tests, especially for CAT tests that are used in making high-stakes decisions.

With continuously or close to continuously administered CBT, a very large item pool is needed so as to maintain test security by making sure that examinees who take the test later do not have an advantage over those tested earlier. In practice, available items for a test are always limited. To minimize the impact of possible item sharing, item exposure rates should be controlled. The exposure rate of an item is defined as the proportion of examinees who are administered the item among all the examinees taking the test in a specified time period. Given a set of items, an item exposure control mechanism is considered to be a major component in maintaining the security of CBT (see Mills & Steffen, 2000; Stocking, 1994; Stocking & Lewis, 1995, 1998; Sympson & Hetter, 1985; Way, 1998). A related quantity to the item exposure rate is the test overlap rate, which is defined as the average of the percentage of items shared by a pair of examinees across all such pairs (Way, 1998; Chen, Ankenmann, & Spray, 2003). Test overlap rates have been generalized to deal with the problem of large-scale item sharing (Chang & Zhang, 2002, 2003). When all item exposure rates are equalized, which would happen if a randomized item selection method were used, then the item exposure control that is achieved is better in terms of

test security than that achieved by any of the other item selection algorithms. For a fixed item selection algorithm, some other mechanism besides item exposure control must be employed to attain even better levels of test security.

Using multiple item pools is regarded as a viable strategy for enhancing test security for CBT (Mills & Steffen, 2000; Davey & Nering, 2002). The multiple-pool approach periodically rotates item pools in and out of use, trying to maintain a high degree of test security for CBT used for making high-stakes decisions. However, it is unclear whether the use of multiple item pools really helps test security. If an examinee who memorizes item information from one item pool is instead administered items from another item pool, the impact of collusion should be dramatically reduced. On the other hand, the examinee in this situation could be administered items from an already compromised item pool. Then, because using multiple item pools reduces the size of each item pool, the percentage of compromised items would be much higher than that in the case of a large single item pool. Thus a fair comparison about test security between the single-pool approach and the multiple-pool approach is needed. This paper investigates whether using multiple item pools instead of a single pool is effective in improving test security. The major criterion used for the comparison is the expected number of compromised items administered to an examinee after a single pool or multiple pools have been used for some time period. The smaller the expected number, the better the corresponding approach is.

A randomized item selection algorithm is assumed in this paper when comparing the multiple-pool approach with the single-pool approach. The reasons for such an assumption are that (a) probability theory can be easily applied under this assumption, and (b) item exposure rates are balanced (i.e., the best test security is achieved). This paper also discusses how to construct item pools optimally.

2. Main Results

Suppose there is a set of N items. The single-pool approach uses all items as a single pool, while the multiple-pool approach constructs J item pools with N_j items in pool j ($1 \leq j \leq J$ and $J \geq 2$) and rotates the pools in and out of use. An item that appears in more than one pool (more precisely, subpool) is called a *common item* when the multiple-pool approach is employed; otherwise it is a *unique item*. This paper considers special multiple-pool cases in which there is no item that appears in more than two pools. When $J = 2$, this constraint is satisfied

automatically. Let m_{jk} be the number of common items of pool j and pool k for $1 \leq j < k \leq J$. Under the constraint, $\sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}$ is the total number of common items in the J pools and $\sum_{j=1}^{J} N_j = N + \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}$. When $J = 2$, $m_{12} = N_1 + N_2 - N$. When $m_{12} = 0$, the two pools are mutually exclusive. Note that using the single pool may be regarded as employing two pools that are exactly the same (i.e., $m_{12} = N_1 = N_2 = N$).

Let Q_j be the frequency of usage of pool j. Note that $(Q_j,\ j = 1, \ldots, J)$ is the probability distribution of the random variable for the selected pool being administered to a randomly selected examinee when J pools are used. Typically, each pool will be used for the same amount of time or for the same total number of examinees in the case of multiple pools. Thus a randomly selected examinee should have the same chance of being administered any one of the pools; that is, $Q_1 = \cdots = Q_J = 1/J$.

After an item pool has been used for certain time period or after some number of examinees have taken the test, some items might be compromised (e.g., the information of these items has been shared by a group of examinees). Let $n(t)$ be the number of compromised items at time t (i.e., after t examinees have taken the test or after an item pool or multiple item pools have been used for a period of time t) in the case when a single pool is used. Let $n_j(t)$ be the number of compromised items in pool j at time t in the case when multiple pools are employed. Then, $r(t) = n(t)/N$ is the proportion of compromised items in the single pool at time t, and $r_j(t) = n_j(t)/N_j$ the proportion of compromised items in pool j for $j = 1, \ldots, J$. Note that the number of compromised items in an item pool increases as the time of use of the pool passes. Thus, $\{n(t)\}$ and $\{n_j(t)\}$ are monotone stochastic processes or time series. Let $p_{jk}(t)$ be the proportion of compromised items among common items of pool j and pool k. Then, $m_{jk}p_{jk}(t)$ is the number of compromised common items of pool j and pool k and $\sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}p_{jk}(t)$ is the total number of compromised common items between any two item pools. Denote $p^*(t) = \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}p_{jk}(t) / \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}$ as the overall proportion of compromised common items at time t.

Let

$$X_i(t) = \begin{cases} 1, & \text{if the } i\text{th item that is administered to an examinee is a compromised item;} \\ 0, & \text{otherwise.} \end{cases}$$

Let $P_J(i|t)$ be the probability that the ith item that an examinee gets is a compromised item when a single pool ($J = 1$) or J ($J > 1$) pools are used, where $1 \leq i \leq L$ and L is the test length.

That is,

$$P_1(i|t) = \text{Prob}(X_i(t) = 1 \mid \text{the single item pool is used}),$$

and

$$P_J(i|t) = \text{Prob}(X_i(t) = 1 \mid J \text{ item pools are used}) \text{ for } J > 1.$$

$\sum_{i=1}^{L} X_i(t)$ is the number of compromised items administered to an examinee at time t. Its expectation, $E_J[\sum_{i=1}^{L} X_i(t)] = \sum_{i=1}^{L} P_J(i|t)$, which is the *expected number* of compromised items administered to an examinee, is a key index used in this paper for test security. Here E_J is the operation of mathematical expectation given that one pool ($J = 1$) or multiple pools ($J > 1$) are used. This index will be used as a criterion for judging test security with different approaches in this paper. If $P_1(i|t) > P_J(i|t)$ for $J > 1$, then the multiple-pool approach is better than the single-pool approach at time t, while the single-pool approach outperforms the multiple-pool approach when $P_1(i|t) < P_J(i|t)$. Note that this paper always compares different approaches after the same amount of usage. The notation used in this section is summarized in the appendix for readers' convenience. In practice, only L, N, N_j, Q_j, and m_{jk} are known quantities, but $n(t)$, $n_j(t)$, $r(t)$, and $r_j(t)$ are unknown monotone stochastic processes increasing or nondecreasing with the time of use of the pools. In the rest of the paper, the time label is suppressed whenever there is no potential for confusion.

As discussed above, the number of compromised items is an unknown random variable or a time series. To predict its value, a statistical model must first be established. One simple and reasonable way to judge whether an item has been compromised is to resort to item exposure numbers, which is the number of examinees to whom the item has been administered in their tests. Note that an item may be leaked after it is administered to just one test taker who shares test item information with others, while an item that has been used by many honest examinees is not necessarily compromised. However, from a statistical point of view, the larger the item exposure number, the greater the chance is that the item is compromised. Thus, items with their exposure numbers exceeding a certain level should be retired. A simple and practical model for this is a linear additive probability model

$$\text{Prob}\{\text{An item is compromised}\} = \frac{\text{The exposure number of the item}}{C}, \tag{1}$$

where C is the item retirement age; that is, an item is retired if its exposure number reaches C. More realistic models may be established after a careful consideration of other factors involved

in an operational computer-based test. For example, a difficult item might be more likely to be compromised than an easy item. A detailed discussion about how to establish these kinds of models has to be case-based and will not be further pursued in this paper.

Based on the model in (1), one can calculate the expected number of compromised items in the pool(s) after t examinees have taken the test before any items are retired for both the single-pool approach and the multiple-pool approach. Because the model is additive, the probability that an item is compromised increases by $1/C$ every time the item is administered to an examinee. The total number of item exposures after t examinees have taken the test is tL whether a single item pool is used or multiple item pools are used. Thus, the expected number of compromised items in the J pools is the same (i.e., tL/C) as that in the single pool. This result holds with any linear additive probability models for an item to be compromised. In the rest of the paper, assume that the number of compromised items in the single pool is the same as the total number of compromised items in the multiple pools after the same pool usage, although the numbers are unknown. That is, $n(t) = \sum_{j=1}^{J} n_j(t) - \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk} p_{jk}(t)$ for any t, where $\sum_{j=1}^{J} n_j(t) - \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk} p_{jk}(t)$ is the total number of compromised items in the multiple pools under the constraint of no overlap among any three pools.

Below, an example with concrete numbers is discussed, followed by generalized results presented in a theorem. Assume there are 1,800 items available. Suppose an examinee has studied 360 items out of the 1,800 items after these items have been used for a while, either in the form of a single pool or multiple pools. Note that this example, which only presents a snapshot in time when there are 360 compromised items, is used here to better understand the generalized results presented in the theorem. If a single item pool is used, then, this examinee has a 20% chance that the ith item he or she gets is a compromised item (see the theorem below). That is, for any fixed i, $P_1(i) = 0.2$. Note that this probability is independent from i, the sequence or order of items the examinee gets.

If these 1,800 items can be used to construct two item pools with each pool (subpool) satisfying all measurement requirements of the test, then these two pools can rotate in and out of use. Assume that the sizes of these two item pools are the same (i.e., $N_1 = N_2$). There are two possible cases: one case where the two pools are mutually exclusive (i.e., $m_{12} = 0$) so there would be 900 items in each item pool, and the other case when there are m_{12} common items in the two pools (in this example, m_{12} is fixed to be 200 and hence there would be 1,000 items in each pool).

The number of compromised items in the first pool, n_1, is between 0 and 360, but is unknown.

If $m_{12} = 0$, then there are $360 - n_1$ compromised items in the other pool. The examinee who has studied 360 compromised items has a $n_1/900 = (n_1/9)\%$ chance that the ith item he or she gets is a compromised item if the first pool is administered to him or her, or a $(360 - n_1)/900 = ((360 - n_1)/9)\%$ chance if the second pool is used. Thus, according to the total probability formula (see Bickel & Doksum, 1977, p. 440), the chance that the ith item administered to an examinee is a compromised item is

$$P_2(i) = \frac{1}{2} \times \frac{n_1}{900} + \frac{1}{2} \times \frac{360 - n_1}{900} = 0.2$$

for the case without an overlap between two pools. Note that this chance is exactly the same as that in the single-pool scenario.

If the two pools have 200 common items, then the result depends on the number of compromised common items. Let us first look at three special situations: no common items are compromised (i.e., $p_{12} = 0.0$), the percentage of compromised common items is the same as the percentage of all compromised items (i.e., $p_{12} = 0.2$), or all common items are compromised (i.e., $p_{12} = 1$).

- If no common items are compromised, then for any fixed i,

$$P_2(i) = \frac{1}{2} \times \frac{n_1}{1000} + \frac{1}{2} \times \frac{360 - n_1}{1000} = 0.18.$$

  This probability is less than 0.2, corresponding to when a single item pool is used.

- If the percentage of compromised common items is the same as the overall percentage of compromised items, that is, 20%, then there are 40 compromised common items and the number of compromised items in the second pool is $n_2 = 360 - n_1 + 40$. Therefore, for any fixed i,

$$P_2(i) = \frac{1}{2} \times \frac{n_1}{1000} + \frac{1}{2} \times \frac{360 - n_1 + 40}{1000} = 0.2.$$

  This is exactly the same as that with a single item pool.

- If all common items are compromised (in this case n_1 should be at least 200), then for any fixed i,

$$P_2(i) = \frac{1}{2} \times \frac{n_1}{1000} + \frac{1}{2} \times \frac{360 - n_1 + 200}{1000} = 0.28.$$

This probability is much larger than the corresponding probability when using a single item pool.

Thus the probabilities are different under different situations. Generally, the number of compromised common items is $200p_{12}$ and the number of compromised items in the second pool is $n_2 = 360 - n_1 + 200p_{12}$. Therefore, for any fixed i,

$$P_2(i) = \frac{1}{2} \times \frac{n_1}{1000} + \frac{1}{2} \times \frac{360 - n_1 + 200p_{12}}{1000} = 0.18 + 0.1p_{12}.$$

Thus,

$$P_1(i) - P_2(i) = 0.1(0.2 - p_{12}).$$

If $p_{12} > 0.2$, then $P_2(i) > P_1(i)$; that is, the single-pool approach is better than the two-pool approach. When $p_{12} < 0.2$, the two-pool approach outperforms the single-pool approach. Since p_{12} is a time series, it is quite possible that one approach is better than the other for some period of time and is worse for some other period of time. The generalized results are presented in the following theorem.

Theorem. Assume that a randomized item selection method is used in the test. Then,

1. $P_1(i|t) = r$ and $P_2(i|t) = r_1(t)Q_1 + r_2(t)Q_2$ for any fixed i. Thus, $X_1(t), \ldots, X_L(t)$ have the same Bernoulli distribution when one item pool or two item pools are used. The expected number of compromised items encountered by an examinee at time t is

$$\sum_{i=1}^{L} P_J(i|t) = LP_J(1|t) \quad \text{for} \quad J = 1, 2.$$

2. If the two item pools have no common items and $Q_1 = Q_2 = 1/2$, then for any fixed i,

$$P_1(i|t) - P_2(i|t) = \frac{N_1 - N_2}{2N_2}(r_1(t) - r(t)). \tag{2}$$

Thus, $P_1(i|t) > P_2(i|t)$ if, and only if, $(r_1(t) - r(t))(N_1 - N_2) > 0$; that is, the two-pool approach is better than the single-pool approach at time t if, and only if, the proportion of compromised items in the larger subpool is greater than the overall proportion of compromised items at time t. When the sizes of two subpools are the same (i.e., $N_1 = N_2$), $P_2(i|t) \equiv P_1(i|t)$ for any t; that is, the two approaches are the same with respect to the expected number of compromised items administered to a randomly selected examinee.

3. If $Q_1 = Q_2 = 1/2$, the two pools have m_{12} common items, and among the common items the percentage of compromised items is $p_{12}(t)$, then for any fixed i,

$$P_2(i|t) = \frac{1}{2}(r_1(t) + r_2(t)) = \frac{1}{2}\left[\frac{n_1(t)}{N_1} + \frac{n(t) - n_1(t) + m_{12}p_{12}(t)}{N_2}\right]. \tag{3}$$

Further, if $N_1 = N_2$, then

$$P_2(i|t) = \frac{n + m_{12}p_{12}(t)}{N + m_{12}},$$

and

$$P_1(i|t) - P_2(i|t) = \frac{m_{12}}{N + m_{12}}(r(t) - p_{12}(t)).$$

Thus,

- $P_2(i|t) > P_1(i|t)$, if $p_{12}(t) > r(t)$,
- $P_2(i|t) = P_1(i|t)$, if $p_{12}(t) = r(t)$, and
- $P_2(i|t) < P_1(i|t)$, if $p_{12}(t) < r(t)$.

That is, the two-pool approach is better than the single-pool approach at time t if, and only if, the proportion of compromised common items is less than the overall proportion of compromised items at time t.

In general, suppose there are J item pools.

1.

$$P_J(i|t) = \sum_{j=1}^{J} r_j(t)Q_j. \tag{4}$$

If an examinee has the same chance of being administered each of the pools in the case of multiple pools, then

$$P_J(i|t) = \sum_{j=1}^{J} r_j(t)/J. \tag{5}$$

The expected number of compromised items encountered by an examinee at time t is

$$\sum_{i=1}^{L} P_J(i|t) = LP_J(1|t). \tag{6}$$

2. Suppose that the sizes of pools are the same in the case of multiple pools and an examinee has the same chance of being administered each of the pools. Then

$$P_J(i|t) = \frac{n(t) + \sum_{j=1}^{J-1}\sum_{k=j+1}^{J} m_{jk}p_{jk}(t)}{N + \sum_{j=1}^{J-1}\sum_{k=j+1}^{J} m_{jk}}, \tag{7}$$

and

$$P_1(i|t) - P_J(i|t) = \frac{1}{N + \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}} \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}(r(t) - p_{jk}(t)). \quad (8)$$

Thus, $P_1(i|t) > P_J(i|t)$ if, and only if, $r(t) > p^*(t)$. That is, the multiple-pool approach is better than the single-pool approach at time t if, and only if, the overall proportion of compromised common items is less than the overall proportion of compromised items at time t. If multiple pools are mutually exclusive, then for any t, $P_J(i|t) \equiv P_1(i|t)$.

In reality, the size of each of a set of multiple pools is likely to be close to each other. Thus, $(N_1 - N_2)/(2N)$ should be very small. According to (2), $P_2(i) \approx P_1(i)$ if the two pools are mutually exclusive. In general, when the sizes of pools are the same in the case of multiple pools and an examinee has the same chance of being administered each of the pools, the single-pool and the multiple-pool approaches are exactly the same in terms of the probability that a compromised item is given to an examinee if there is no common items.

When there are common items, the overall proportion of compromised common items is the key quantity in determining whether employing multiple pools is better than using a single pool. If all items are new at the beginning, the common items may get much higher exposure than unique items that belong to one pool only. Consequently, the overall proportion of compromised common items is expected to be larger than the overall proportion of compromised items. According to the theorem, it is worse in terms of test security to use multiple pools in this case than a single pool. Therefore, it is of no benefit with respect to test security to use the multiple pools unless there are other considerations. Some other possible applications of the theorem, including how to optimally construct item pools, will be discussed in the next section.

The proof of the theorem is presented below. The reader who is willing to accept all this without proof may skip the rest of this section and go to the next section. A basic result in the field of probability is needed to prove the theorem. This result can be found in a probability text from Fudan University (1979, p. 29). For the readers' convenience, this paper first states this result as a lemma and provides a brief proof and then presents the proof of the theorem.

Lemma. Suppose there are N balls with n red and $N - n$ green balls. Randomly and sequentially sample L balls without replacement. Then the probability that the ith sampled ball is red is n/N for any fixed i $(1 \leq i \leq L)$.

Proof of Lemma. First, continue to randomly sample the rest of the balls sequentially after sampling L balls. Let $P(i)$ be the probability that the ith sampled ball is red. The total number of possible outcomes in this experiment is $N!$. By the multiplication principle (see Hogg & Tanis, 1997), the number of outcomes that the ith sampled ball is red is $n \times (N-1)!$. Thus, for any fixed i,

$$P(i) = \frac{n \times (N-1)!}{N!} = \frac{n}{N}.$$

□

Proof of Theorem.

1. Regard compromised items as red balls and other items as green balls. By the lemma, $P_1(i) = n/N = r$, and the conditional probability that the ith item administered to an examinee is a compromised item given pool j is $r_j = n_j/N_j$ for $j = 1, 2$. According to the total probability formula, $P_2(i) = r_1 Q_1 + r_2 Q_2$ is obtained for any fixed i.

2. In this case, $N = N_1 + N_2$. From the first part of the theorem,

$$\begin{aligned} P_1(i) - P_2(i) &= \frac{n}{N_1 + N_2} - \frac{n_1}{2N_1} - \frac{n - n_1}{2N_2} \\ &= \frac{(N_1 - N_2)(n_1 N - n N_1)}{2N N_1 N_2} \\ &= \frac{N_1 - N_2}{2N_2}(r_1 - r). \end{aligned}$$

3. The number of compromised common items is $m_{12}p_{12}$. Thus, $n_2 = n - n_1 + m_{12}p_{12}$. By the first part of the theorem, (3) is obtained. Further, if $N_1 = N_2$, then $2N_1 = N + m_{12}$. By (3),

$$P_2(i) = \frac{n + m_{12}p_{12}}{2N_1} = \frac{n + m_{12}p_{12}}{N + m_{12}},$$

and

$$P_1(i) - P_2(i) = r - \frac{n + m_{12}p_{12}}{N + m_{12}} = \frac{m_{12}}{N + m_{12}}(r - p_{12}).$$

The proof for the multiple item pools is similar. According to the total probability formula, (4) and (5) can be obtained. When the sizes of item pools are the same, $JN_1 = N + \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}$. Since $\sum_{j=1}^{J} n_j = n + \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}p_{jk}$, by (5)

$$P_J(i) = \frac{\sum_{j=1}^{J} n_j}{JN_1} = \frac{n + \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}p_{jk}}{N + \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}},$$

and it is not difficult to verify (8). □

3. Examples of Applications

The theoretical results presented in the previous section can be applied to solve several issues related to test security with CBT. First, the theorem can be used directly to evaluate and compare the impact of item sharing when a single item pool or multiple item pools are used. Second, the theoretical results provide a guideline for constructing item pools optimally in terms of minimizing the expected number of compromised items administered to an examinee. Third, the results can be used to evaluate the risk of some CBT designs.

How to construct pools optimally. As pointed out in the last section, if all items are new at the beginning, it is better to use a single item pool than multiple item pools with overlap. After the single pool has been used for a while, multiple pools can be adopted. When compiling two item pools, new items and/or items with small item exposure numbers in the original pool should be put in both pools as common items. In this way, employing multiple pools is better than using the single pool in terms of minimizing the impact of possible item sharing among a group of examinees, according to the theorem. However, the exposure number of a common item usually increases more rapidly than that of a unique item. Thus the chance of a common item becoming a new compromised item is larger than that for an unique item. As time passes, the rate of compromised common items increases and likely becomes higher than the overall rate of compromised items. According to the theorem, it will be worse to employ multiple pools than the single pool at this point. Therefore multiple pools should be maintained (i.e., recompiled) periodically to keep the proportion of compromised common items small, by replacing high-exposure common items with low-exposure or new items. This mixed approach is called an *adjusted-pool* approach in this paper. According to the theorem, the adjusted-pool approach is better than the single-pool approach if item pool maintenance can be properly performed.

To evaluate the risk of item reuse scheme. Because items are expensive, some items will be reused. A typical case is that an item pool is suspended after it has been used for a while. Some items with extremely high item exposure numbers are retired, but others are possibly reused with new items later. Suppose that item pools, a single pool or two pools, are composed of $N - M$ used items and M new items. A practical issue is to evaluate the risk of this item reuse scheme. When evaluating the risk, it is better to consider the worst case scenario where all used items have been compromised by a group of test takers. Then, $P_1(1|t = 0) = (N - M)/N$ for the single pool, or $P_2(1|t = 0) = (N - M + m_{12}p_{12})/(N + m_{12})$ for two pools, where p_{12} is the percentage

of used items among common items. It is easy to verify that $P_2(i|t=0)$ is minimized when all and only new items are put in both pools as common items, that is, $p_{12}=0$ and $m_{12}=M$. Therefore, the optimal two-pool design is one in which one pool consists of all the new items and a half of used items and the other pool consists of all the new items and the other half of used items. If such a two-pool design is used, $P_2(i|t=0)=(N-M)/(N+M)$ and the expected number of compromised items encountered by a test taker who has studied all the used items is $L(N-M)/(N+M)$. Since the expected number of compromised items encountered by a test taker is $L(N-M)/N$ for the single pool, the difference of these expected numbers between using the single pool and the optimal two pools is

$$\frac{LM(N-M)}{N(N+M)}. \tag{9}$$

For example, if N=1,000, M=500, and L=60, the expected number of compromised items encountered by a test taker at the beginning of the testing program is 30 if the single pool is used, and it is 20 if the optimal two pools are used. Tables can be produced based on (9) for different L, M, and N for practical uses. Note that the price or the cost of the gain is that new items as common items will have higher usage than the other items. Thus new items will deteriorate quickly. When the mean raw score of used items is available, the average maximum gain in raw scores for the examinees who have studied the used items can be obtained. Further, the average maximum gain in reported scores due to the item reuse may also be calculated according to the raw score to reported score conversion table.

4. Discussion

This paper compares the use of multiple pools to use of a single pool with respect to test security against large-scale item sharing in a computer-based test. The comparison is made under the assumption that a randomized item selection method is used. The results show that, in general, simply using multiple pools instead of a single pool actually does not improve test security in terms of the expected number of compromised items encountered by a test taker. Based on the theoretical results, an adjusted-pool approach is proposed in this paper to achieve optimal test security. Usually, the probabilistic scheme of an operational item selection algorithm is very complicated. Thus it is hard to compare the use of multiple pools mathematically with a single pool in terms of test security when an operational item selection algorithm is used. To get actual

results for a given operational item selection algorithm, a Monte Carlo simulation study is needed.

The results obtained in this paper provide technical guidelines for solving many practical issues in CBT such as (a) evaluating the risk of employing an item reuse scheme, (b) optimally constructing multiple pools given a set of items to minimize the risk, and (c) estimating the possible gain in test security by using the adjusted-pool approach instead of the single-pool approach. Although different item selection algorithms may yield different average numbers of compromised items administered to examinees, the guidelines for optimally constructing item pools with respect to test security should still hold for any reasonable item selection algorithm. An item selection algorithm is considered to be reasonable if it tries to balance item exposure rates. The adjusted-pool approach will not improve test security if an item selection algorithm prefers to choose some items but never selects some other items in the pools. When applying the adjusted-pool approach to an operational CBT program, other factors, such as content requirements, will need to be considered simultaneously along with the test security issue.

To evaluate the impact of item sharing, the number of compromised items in a pool must be predicted first. A simple linear model for predicting the number of compromised items is proposed in this paper. Further research on this issue is needed.

References

Bickel, P. J., & Doksum, K. A. (1977). *Mathematical statistics.* San Francisco: Holden-day.

Chang, H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67*, 387-398.

Chang, H., & Zhang, J. (2003, April). *Assessing CAT security breaches by the item pooling index.* Paper presented at the Annual Meeting of National Council on Measurement in Education, Chicago, IL.

Chen, S., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40*, 129-145.

Davey, T., & Nering, N. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer, & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 165-191). Mahwah, NJ: Lawrence Erlbaum.

Fudan University. (1979). *Probability theory.* Beijing, China: People's Educational Press (in Chinese).

Hogg, R. V, & Tanis, E. A. (1997). *Probability and statistical inference* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

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

Steinberg, J. (2002, August 8]. Officials link foreign web sites to cheating on graduate admission exams. *The New York Times.*

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS RR-94-5). Princeton, NJ: ETS.

Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing* (ETS RR-95-25). Princeton, NJ: ETS.

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.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military*

Testing Association (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Way, W. D. (1998, Winter). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, 17-27.

Wheeler, D. L. (2002, August 7). ETS says GRE scores from China, South Korea, and Taiwan are suspect. *Chronicle of Higher Education*.

Appendix

Notation

J	The number of item pools in a multiple-pool approach
L	The test length
M	The number of new items
m_{jk}	The number of common items of pool j and pool k for $1 \leq j < k \leq J$
N	The total number of items
N_j	The number of items in item pool j for $j = 1, \ldots, J$
$n(t)$	The total number of compromised items at time t or after t examinees have taken the test
$n_j(t)$	The number of compromised items in pool j at time t for $j = 1, \ldots, J$
$P_J(i\|t)$	The probability that the ith item of an examinee is a compromised item at time t if J item pools are used ($1 \leq i \leq L$ and $J \geq 1$)
$p_{jk}(t)$	The proportional of compromised items among common items of pool j and pool k for $1 \leq j < k \leq J$ at time t
$p^*(t)$	The overall proportion of compromised common items at time t, $p^*(t) = \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk} p_{jk}(t) / \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}$
Q_j	The rate of usage of pool j in the case of multiple pools for $j = 1, \ldots, J$, and $Q_1 + \cdots + Q_J = 1$
$r(t)$	The overall proportional of compromised items at time t; $r(t) = n(t)/N$
$r_j(t)$	The proportional of compromised items in pool j at time t; $r_j(t) = n_j(t)/N_j$ for $j = 1, \ldots, J$
t	The number of examinees who have taken the test or the time period for which an item pool or multiple item pools have been used
$X_i(t)$	The indicator random variable if the ith item of an examinee is a compromised item at time t or after t examinees have taken the test ($1 \leq i \leq L$)