Behav Res (2012) 44:742–752
DOI 10.3758/s13428-011-0178-5

Comparing single-pool and multiple-pool designs regarding test security in computerized testing

Jinming Zhang · Hua-Hua Chang · Qing Yi

Published online: 5 January 2012

Abstract This article compares the use of single- and multiple-item pools with respect to test security against item sharing among some examinees in computerized testing. A simulation study was conducted to make a comparison among different pool designs using the item selection method of maximum item information with the Sympson–Hetter exposure control and content balance. The results from the simulation study indicate that two-pool designs have a better degree of resistance to item sharing than do the single-pool design in terms of measurement precision in ability estimation. This article further characterizes the conditions under which employing a multiple-pool design is better than using a single, whole pool in terms of minimizing the number of compromised items encountered by examinees under a randomized item selection method. Although no current computerized testing program endorses the randomized item selection method, the results derived in this study can shed some light on item pool designs regarding test security for all item selection algorithms, especially those that try to equalize or balance item exposure rates by employing a randomized item selection method locally, such as the *a*-stratified-with-*b*-blocking method.

J. Zhang (✉)
Department of Educational Psychology,
University of Illinois at Urbana-Champaign,
236A Education Building 1310 S Sixth Street,
Champaign, Illinois 61820, USA
e-mail: jmzhang@illinois.edu

H.-H. Chang
Department of Educational Psychology,
University of Illinois at Urbana-Champaign,
236B Education Building 1310 S Sixth Street,
Champaign, Illinois 61820, USA

Q. Yi
Pearson Assessments,
San Antonio, Texas

Keywords Computerized testing · Adaptive testing · Item response theory · Statistics

Computer-based testing (CBT), including computerized adaptive testing (CAT) and computer-delivered nonadaptive testing, makes it possible for an educational or psychological test to be administered to small groups of examinees at frequent time intervals, which is referred to as *continuous testing*. A continuous test is preferred by examinees because of the flexibility it provides in scheduling to take the test. However, continuous testing results in 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 rapid growth of Internet applications all over the world has made such item-sharing activities easier. In one scenario, some test takers organize a special interest group and share test item information with each other through the Internet. Such a group of examinees is called an *examinee collaboration network* (ECN; Luecht, 1998). The tactic of memorizing and sharing test item information will inflate test scores for some examinees, and consequently will hurt the reliability and validity of the test (Davey & Nering, 2002; Guo, Tay, & Drasgow, 2009; Yi, Zhang, & Chang, 2008). The damage of such cheating could be so severe that a testing company might have to suspend its continuous-testing program (Honan, 1995; Steinberg, 2002; Wheeler, 2002). Without effective measures, such cheating activities could significantly undermine the credibility of any continuous-delivery test. Hence, 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 testing, a very large item pool is needed 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, the 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 of 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 a continuously delivered test, especially a CAT (see Mills & Steffen, 2000; Stocking, 1994; Stocking & Lewis, 1995, 1998; Sympson & Hetter, 1985; Way, 1998). A quantity related 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 (Chen, Ankenmann, & Spray, 2003; Way, 1998). Test overlap rates have been generalized to deal with the problem of large-scale item sharing (Chang & Zhang, 2002, 2003). For a fixed item selection algorithm, some other mechanism besides item exposure control must be employed in order to attain even better levels of test security.

Using multiple item pools is regarded as a viable strategy for enhancing test security for CBT/CAT (Ariel, Veldkamp, & van der Linden, 2004; Davey & Nering, 2002; Mills & Steffen, 2000). Suppose that several item pools can be formed from a given set of items. By periodically rotating item pools in and out of use, a multiple-pool approach can in theory maintain a high degree of test security for high-stakes CBT. However, it is unclear whether the use of multiple item pools really helps test security. Note that multiple pools are the subpools of the single (whole) pool. If an examinee who memorizes item information from one subpool is instead administered items from another subpool, the impact of collusion should be dramatically reduced. On the other hand, if the examinee is administered items from a subpool he/she has partially memorized, the percentage of compromised items that the examinee runs into should be much higher than would be the case using the single pool, because using multiple item pools reduces the size of each individual pool. Obviously, a fair comparison about test security between the single-pool approach and the multiple-pool approach is needed.

Different item selection methods have different degrees of resistance to test security breaches. For any specific item selection method, such as the weighted deviation method (Stocking & Swanson, 1993; Swanson & Stocking, 1993), the *a*-stratified method (Chang & Ying, 1999; Yi & Chang, 2003), and constrained CAT with shadow tests (van der Linden, 2000), a comparison of test security between the single-pool approach and the multiple-pool approach can be conducted by a simulation study. Below, we discuss a simulation study using a popular CAT item selection method, the maximum item information method (Lord, 1980) with the Sympson–Hetter item exposure control (Hetter & Sympson, 1997; Sympson & Hetter, 1985), to find out whether a two-pool design outperforms a single-pool design in terms of test security. The results from the simulation study indicated that all two-pool designs considered in the study had a better degree of resistance to item sharing than did the single-pool design in terms of both measurement precision in ability estimation and the number of compromised items encountered by examinees. It is possible that the results may be sensitive to the item selection method being used. Thus, a theoretical investigation is needed. Usually, the probabilistic scheme of an operational item selection algorithm is very complicated. It is extremely hard, if not impossible, to analytically compare the use of multiple pools with a single pool in terms of test security against item sharing when an operational item selection algorithm is used. In this article, we theoretically compare the single- and multiple-pool designs with respect to test security using a randomization item selection method and characterize the conditions under which employing multiple pools is better than using a single, whole pool in terms of minimizing the number of compromised items encountered by an examinee. A mathematical reason for using a randomized item selection method is that probability theory can easily be applied so that theoretical results and formulae can be obtained. Although no current CAT program endorses the randomized item selection method, the results derived in this study can shed some light on item pool designs regarding test security for all item selection algorithms, especially those that try to equalize or balance item exposure rates by employing a randomized item selection method locally, such as the *a*-stratified-with-*b*-blocking method (Chang, Qian, & Ying, 2001) and the match-ability-with-difficulty method (Hulin, Drasgow, & Parsons, 1983). These results also provide guidelines to practitioners in item pool designs.

Theoretical framework

Single- and multiple-pool designs

Suppose there is a set of N items for a continuously delivered test (e.g., a CAT). A single-pool design uses all of the items as a single pool, while a multiple-pool design constructs J item pools (more precisely, subpools) with N_j items in pool j ($1 \le j \le J$ and $J \ge 2$) and rotates the pools in and out of use. Here we presume that these items can form J item pools, each of which satisfies all of the constraints (e.g., content balance) that are required by the test; otherwise, a comparison between the two designs is out of the question. In a multiple-pool design, an item that appears in more than one pool is called a *common item*; otherwise, it is a *unique item*. Denote m_{jk} as the number of common items between pool j and pool k for $1 \le j < k \le J$. When $J = 2$, $m_{12} = N_1 + N_2 - N$. When $m_{12} = 0$, the two pools are mutually exclusive.

Let Q_j be the relative 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 an 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. In this scenario, a randomly selected examinee should have the same chance of being administered any one of the pools; that is, $Q_1 = \cdots = Q_J = 1/J$.

Compromised items in an item pool

In this article, an item is called a *compromised item* to an examinee if he or she has some preknowledge about this item. For example, if an item has been discussed or posted on a Web site, then it is a compromised item to the group of examinees who have visited that Web site. After an item pool has been used for a certain time period or after some number of examinees have taken the test, some items might be compromised by an ECN. Clearly, the number of compromised items, though unknown, is directly related to or determined by the number of professional test takers or the number of active members of an ECN. Let $n(t)$ be the number of compromised items at time t (i.e., after t examinees have taken the test or after the item pool has been used for a period of time t) in the case in which a single pool is used. Let $n_j(t)$ be the number of compromised items in pool j at time t in the case in which multiple pools are employed. Then, $r(t) = n(t)/N$ is the proportion of compromised items in the single pool at time t, and $r_j(t) = n_j(t)/N_j$ is 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. Let $p_{jk}(t)$ be the proportion of compromised items among the common items between pools j and k. Then, $m_{jk}p_{jk}(t)$ is the number of compromised common items between pools j and k. In a two-pool design, for example, $m_{12}p_{12}(t)$ is the number of compromised common items between the two pools, and the total number of compromised items is $n_1(t) + n_2(t) - m_{12}p_{12}(t)$.

Let

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

$X_i(t)$ is a random variable. Its randomness comes from the uncertainty about compromised items in a pool or pools and the item selection method used in the test. $\sum_{i=1}^{L} X_i(t)$ is the number of compromised items administered to an examinee at time t, where L is the test length.

Let $P_J(i \mid t)$ be the probability mass function of $X_i(t)$ in a J-pool design:

$$P_1(i|t) = \text{Prob}[\, X_i(t) = 1 | \text{single} - \text{pool design}],$$

and

$$P_J(i|t) = \text{Prob}[\, X_i(t) = 1 | J - \text{pool design}] \text{ for } J > 1.$$

$P_J(i \mid t)$ is the probability that the ith item that an examinee gets is a compromised item in a J-pool design. If $P_1(i \mid t) > P_J(i \mid t)$ 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 $P_1(i \mid t) < P_J(i \mid t)$. The summation $\sum_{i=1}^{L} P_J(i|t) = E_J\left[\sum_{i=1}^{L} X_i(t)\right]$ is the *expected number* of compromised items administered to an examinee at time t in a J-pool design. Although its value is typically unknown, this expected number can be used as a theoretical criterion for judging test security with different designs: the smaller the number, the better. The notation used in this article is summarized in Appendix A for the reader's convenience.

Simulation study

A simulation study was conducted to compare multiple-pool designs with a single-pool design in CAT. The item selection method considered in this article is the maximum item information method (Lord, 1980) with the Sympson–Hetter exposure control (Hetter & Sympson, 1997; Sympson & Hetter, 1985) and content balance. The maximum item exposure rate was set at 0.2.

A set of 720 items from a real, large-scale achievement test was used to form a single pool, denoted as S720, in this study. These items were calibrated using three-parameter logistic models. There are three content areas in the test, and the percentages of items in the three content areas are 40%, 30%, and 30%, respectively. These 720 items were also used to construct three different two-pool designs: (1) a no-overlap two-pool design in which each subpool contained 360 items, (2) a two-pool design with 60 common items between the two subpools, (i.e., each subpool containing 390 items, with 330 unique and 60 common items), and (3) a two-pool design with 120 common items between the two subpools (i.e., each subpool containing 420 items, with 300 unique and 120 common items). Those two-pool designs are denoted as T360, T390, and T420, respectively. Each subpool pair was constructed using the matched random subset method (Gulliksen 1950) so that the paired subpools within a two-pool design were content-balanced; roughly 40% of the items were from the first content area,

30% from the second content area, and 30% from the third content area. The paired subpools within a two-pool design are also similar in their distributions of item characteristics, such as item difficulty level. Table 1 contains descriptive information about the item parameters for the item pools in the four designs considered in this simulation study. Figure 1 displays the test characteristic curves (TCCs) of the paired subpools for the three two-pool designs. As is shown in Fig. 1, the two TCCs of the paired subpools in each of the two-pool designs are very close to each other.

The test length in this simulation study was 40 items. A content control procedure based on a modified multinomial model (Yi & Chang, 2003) was implemented such that each simulated test consisted of about 40% of the items from Content Area 1 and 30% of the items from each of the other two content areas.

In a simulation study, one may manipulate both the increment of the probability of answering a compromised item correctly and the number of compromised items (which is directly related to the number of professional test takers or the number of active members of an ECN). Both factors have virtually equivalent measurement consequences, and the same degree of severity can be achieved if one factor is fixed and the other increases. Thus, one may control just one factor while keeping the other fixed. Recall that the purpose of our simulation study was to compare different item pool designs. This comparison is fair as long as the probability of correct responses to compromised items is set to be equal across both designs. Therefore, we simply set this probability to 1 in our simulation. It is clear that if there are no compromised items, the single-pool design will produce more accurate measurement results than will the multiple-pool design for all reasonable item selection algorithms, because the former has more items to select from than the latter. If the degree of item leakage is light, the same conclusion can be expected. We seek to examine whether or not the conclusion also holds when item pools are heavily compromised.

There were two different scenarios involving item pools in our study: either without or with compromised items. The former case was used as a reference in this simulation study. In the latter scenario, the probability of an item becoming compromised was proportional to the corresponding item exposure rate obtained in the former scenario: the higher the item exposure rate, the greater the probability. In this simulation study, an unequal-probability-without-replacement sampling method (see Hartley & Rao, 1962) was adopted to select the compromised items. Sampling with unequal

Table 1 Descriptive statistics for item parameters of item pools under different item pool designs

Pools	*N*	Parameter	Mean	*SD*	Minimum	Maximum
Single	720	*a*	1.070	0.361	0.193	2.685
		b	0.066	1.089	−4.999	2.475
		c	0.182	0.082	0.032	0.500
Two						
A_1	360	*a*	1.076	0.372	0.271	2.685
		b	0.065	1.159	−4.999	2.475
		c	0.186	0.079	0.035	0.500
B_1	360	*a*	1.064	0.349	0.193	2.557
		b	0.066	1.016	−2.904	2.089
		c	0.179	0.085	0.032	0.500
Two						
A_2	390	*a*	1.081	0.375	0.271	2.685
		b	0.058	1.147	−4.999	2.475
		c	0.185	0.080	0.035	0.500
B_2	390	*a*	1.072	0.354	0.193	2.557
		b	0.057	1.046	−3.427	2.444
		c	0.177	0.084	0.032	0.500
Two						
A_3	420	*a*	1.086	0.378	0.271	2.685
		b	0.049	1.133	−4.999	2.475
		c	0.183	0.079	0.035	0.500
B_3	420	*a*	1.077	0.358	0.193	2.557
		b	0.051	1.043	−3.427	2.444
		c	0.177	0.084	0.032	0.500

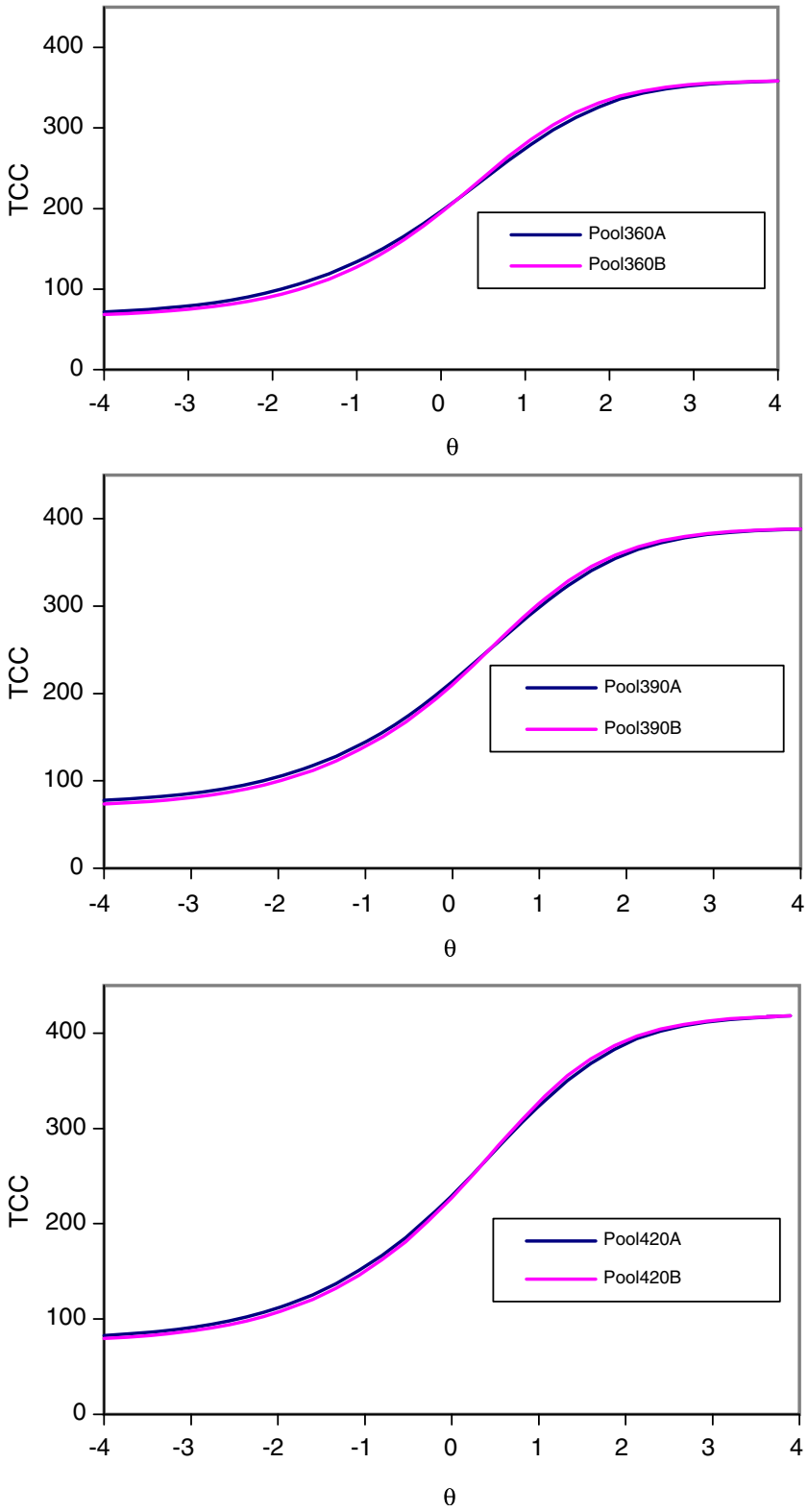

Fig. 1 Test characteristic curves for the paired subpools

probabilities and without replacement means to draw n (e.g., 150) out of N (e.g., 720) items in such a way that the probability for item i to be selected is proportional to its "size" r_i (i.e., item exposure rate). In a two-pool design, a common item compromised in one subpool is also regarded as compromised in the other subpool but is counted as only one compromised item. As was discussed in the Theoretical Framework section, the number of compromised items in a

pool increases with the number of times the pool is used. In order to make the contrast evident, we simulated a snapshot in which there were 150 compromised items. Two examples of such a situation are the cases in which 10 professional test takers each memorize 15 items or 30 active members of an ECN each post 5 items. A set of 150 compromised items was stochastically obtained for each of the four pool designs separately. Because item exposure rates are not the same in CAT with different designs, the compromised items in one design may be different from those in another design.

The number of examinees is 10,000 for each of the designs. Thus, in a two-pool design, 5,000 examinees use one pool, and the other 5,000 examinees use the other pool. The true abilities of the examinees were generated from a standard normal distribution. The expected a posteriori method with the standard normal distribution determined a priori was initially used to obtain a provisional estimate of an examinee's ability until at least five items had been administered and both correct and incorrect responses appeared. Afterward, the maximum likelihood estimation method was used to estimate ability. Note that a compromised item would affect the provisional ability estimate of an examinee, and thus change the subsequent item selection. Ten replications were performed for each of the combinations considered in the simulation, which is equivalent to having a total of 100,000 simulated examinees in the study.

The effectiveness of using a multiple-pool versus a single-pool design is evaluated in terms of test security control and measurement precision. The indices used to measure test security are the observed item exposure rate and the average number of compromised items that examinees encounter (for the second scenario only). The correlation between the estimated and true abilities, the overall bias, and the root-mean squared error (RMSE) of ability estimates are the indices for evaluating measurement precision. The bias and RMSE of ability estimates are calculated as follows:

$$\text{Bias} = \frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\theta}_k - \theta_k\right)$$

and

$$\text{RMSE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\widehat{\theta}_k - \theta_k\right)^2},$$

where K is the total number of examinees (K = 10,000 in this simulation), θ_k is the true ability of the kth examinee, and $\widehat{\theta}_k$ is its estimate.

The numbers (and percentages) of items in different ranges of exposure rates across all four designs in the two scenarios are displayed in Table 2. The third column (r = 0.0) of Table 2 shows the number of items that were never used. Nearly 50% of the items in the single-pool design and about 12%–22% of the items in the two-pool designs were not administered to any examinees. The single-pool design is worse than the two-pool designs with respect to item usage. It is well known that the smaller a pool is, the better the item usage will be for the item selection algorithm using maximum item information. Although the maximum item exposure rate was set at 0.2, the observed item exposure rates of some items could exceed the prespecified maximum level, due to the probabilistic nature of the Sympson–Hetter exposure control procedure. The last column of Table 2 shows the numbers (and percentages) of items with exposure rates exceeding the prespecified maximum level. The percentage is approximately doubled in the two-pool designs as compared to that in the single-pool design. In this respect, the single-pool design outperforms the two-pool designs. The patterns of the distributions of item exposure rates are quite similar to each other between the two scenarios, with or without compromised items, within any pool design. This means that the presence of compromised items did not noticeably change the distribution of the item exposure rates.

Table 3 presents the overall measurement precision in terms of bias, RMSE, and the correlation between the estimated and true abilities under different item pool designs. In addition, the last column of Table 3 shows the average number of compromised items an examinee encounters when there are 150 compromised items in a design. As

Table 2 Numbers (and percentages) of items in different ranges of exposure rates (r) across item pool designs in the two scenarios

Scenario	Pool	$r = 0.0$	$0.0 < r \leq 0.1$	$0.1 < r \leq 0.2$	$r > 0.2$
1*	S720	341 (47.4%)	189 (26.3%)	124 (17.2%)	66 (9.2%)
	T360	103 (14.3%)	220 (30.6%)	248 (34.4%)	149 (20.7%)
	T390	128 (17.8%)	237 (32.9%)	229 (31.8%)	126 (17.5%)
	T420	157 (21.8%)	243 (33.8%)	175 (24.3%)	145 (20.1%)
2*	S720	354 (49.2%)	171 (23.8%)	123 (17.1%)	72 (10.0%)
	T360	85 (11.8%)	246 (34.2%)	242 (33.6%)	147 (20.4%)
	T390	116 (16.1%)	243 (33.8%)	231 (32.1%)	130 (18.1%)
	T420	131 (18.2%)	265 (36.8%)	196 (27.2%)	128 (17.8%)

*Scenarios: 1, without compromised items; 2, with 150 compromised items

Table 3 Overall measurement precision and average numbers of compromised items examinees encountered (m) under different item pool designs in two scenarios

Scenario	Pool	Bias	RMSE	$\rho_{\hat{\theta}\theta}$	m
1*	S720	−0.001	0.201	.981	NA
	T360	−0.001	0.236	.974	NA
	T390	−0.001	0.231	.975	NA
	T420	−0.002	0.224	.976	NA
2*	S720	1.010	1.226	.788	16.6
	T360	0.552	0.694	.913	12.1
	T390	0.653	0.804	.891	13.1
	T420	0.774	0.943	.862	14.6

* Scenarios: 1, without compromised items; 2, with 150 compromised items

expected, when there are no compromised items, designs with large item pools yield relatively small RMSEs and large correlations between estimated and true abilities. However, the improvement is not remarkable in the cases considered here. The largest relative improvement is less than 15% from T360 to S720.

When there are 150 compromised items out of a total of 720, the bias and RMSE increase drastically. For instance, the bias is 1.010 for S720, increased from the amount of −0.001 when there are no compromised items. The results demonstrate that the damage of item sharing can be very severe. However, a CAT test with two small pools has relatively strong immunity (or resistance) to test security breaches, in the sense that the increments of bias and RMSE are relatively small. For instance, the increments of bias and RMSE due to the presence of compromised items for T360 are about half of those for S720. Among two-pool designs, when there are 150 compromised items, T360 works better than T390 and T420, with less bias, RMSE, and average number of compromised items administered to examinees, and with a larger correlation between estimated and true abilities (see the data in Table 3). This may be caused by the following two factors. First, there are no common items in T360, while there are 60 and 120 common items in T390 and T420, respectively. Generally speaking, common items have relatively higher exposure rates than do unique items. Thus, they have a higher probability of becoming compromised. In general, a compromised common item has a larger impact on the number of compromised items administered to examinees than does a compromised unique item. Second, more items are never administered in T390 and T420 than in T360 (see the $r = 0.0$ column in Table 2). In summary, two-pool designs perform better than the single-pool design, whereas among two-pool designs, T360 outperforms T390 and T420.

Analytically comparing single-pool and multiple-pool designs

There has been little discussion of the theoretical difference between single-pool and multiple-pool designs regarding test security. This is mostly due to the fact that the consequence of a pool design on test security is strongly related to the CAT operational item selection algorithm, which typically has a very complex probabilistic scheme. There is no analytical result available regarding the difference between single- and multiple-pool designs. The problem is that such analytical results or formulae are very difficult, if not impossible, to derive. In this section, we derive some formulae for the comparison of single- and multiple-pool designs under a randomized item selection method, which guarantees that probability theory can readily be applied. The mathematical derivations may advance our knowledge about how to extend theoretical work in probability theory to applied fields. As was pointed out by Wainer (2000), a randomized item selection method equalizes item exposure rates, and hence yields the best test security as compared to all other item selection methods. It is of interest to learn the degrees of resistance the two designs have toward security breaches under the selection method with the best test security.

Result 1 Single-pool design versus two-pool design

1. The probability that the ith item an examinee gets is a compromised item is simply the proportion of compromised items in the single-pool design, whereas in a two-pool design, it is a linear combination of the percentages of the compromised items in the two pools; that is,

$$P_1(i|t) = r(t) \text{ and } P_2(i|t) = r_1(t)Q_1 + r_2(t)Q_2.$$

Thus, $X_1(t)$, . . . , $X_L(t)$ have the same Bernoulli distribution when one 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. When the two item pools have no common items and have the same usage (i.e., $Q_1 = Q_2 = 1/2$), then

$$P_1(i|t) - P_2(i|t) = \frac{N_1 - N_2}{2N_2}(r_1(t) - r(t)). \qquad (1)$$

Thus, $P_1(i \mid t) > P_2(i \mid t)$ 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., $N_1 = N_2$), $P_2(i \mid t) \equiv P_1(i \mid t)$ 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. When the two pools have m_{12} common items and have the same usage, then

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

where $p_{12}(t)$ is the percentage of compromised items among the common items. Furthermore, if the two pools have the same number of items (i.e., $N_1 = N_2$), 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,

$$P_2(i|t) > P_1(i|t) \text{ if } p_{12}(t) > r(t),$$
$$P_2(i|t) = P_1(i|t) \text{ if } p_{12}(t) = r(t),$$

and

$$P_2(i|t) < P_1(i|t) \text{ if } p_{12}(t) < r(t).$$

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.[1]

The results above can be generalized to a general multiple-pool design. Below, we only consider special multiple-pool cases in which there is no item that appears in more than two pools, for mathematical simplicity. This constraint is satisfied automatically for any two-pool design. Let m_{jk} be the number of common items of pool j and pool k for $1 \le j < k \le J$. Under the constraint, $\sum_{j=1}^{J-1}\sum_{k=j+1}^{J} m_{jk}$ is the total number of common items in a J-pool design. Then, $m_{jk}p_{jk}(t)$ is the number of compromised common items of pools j and 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

$$p^*(t) = \sum_{j=1}^{J-1}\sum_{k=j+1}^{J} m_{jk}p_{jk}(t) \Big/ \sum_{j=1}^{J-1}\sum_{k=j+1}^{J} m_{jk}$$

as the overall proportion of compromised common items at time t.

Result 2 Single-pool design versus multiple-pool design

1. Let

$$P_J(i|t) = \sum_{j=1}^{J} r_j(t)Q_j. \quad (3)$$

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

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

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

$$\sum_{i=1}^{L} P_J(i|t) = LP_J(1|t). \quad (5)$$

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

$$P_j(i|t) = \frac{n(t) + \sum_{j=1}^{J-1}\sum_{k=j+1}^{J} m_{jk}p_{jk}(t)}{N + \sum_{j=1}^{J-1}\sum_{k=j+1}^{J} m_{jk}}, \quad (6)$$

and

$$P_1(i|t) - P_J(i|t) = \frac{1}{N + \sum_{j=1}^{J-1}\sum_{k=j+1}^{J} m_{jk}} \times \sum_{j=1}^{J-1}\sum_{k=j+1}^{J} m_{jk}\big(r(t) - p_{jk}(t)\big). \quad (7)$$

Thus, $P_1(i \mid t) > P_J(i \mid t)$ 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, $P_J(i \mid t) \equiv P_1(i \mid t)$.

The proofs of these results are presented in Appendix B.

In reality, the sizes of a set of multiple pools are likely to be close to each other. Thus, $(N_1 - N_2)/(2N)$ should be very small. According to Eq. 1, $P_2(i) \approx P_1(i)$ if the two pools are mutually exclusive. In general, if multiple pools have the same size and are mutually exclusive, and if each examinee

[1] This would only occur if the common items were used less frequently than the others, which is unlikely.

has the same chance of being administered each of the multiple pools, then the multiple-pool approach has the same probability that a compromised item will be given to an examinee as the single-pool approach.

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 above results, it is worse, in terms of test security, to use multiple pools than to use a single pool in this case. Therefore, it is of no benefit with respect to test security to use multiple pools under a randomized item selection method.

Discussion

In this article, a simulation study was carried out to compare multiple-pool designs to a single-pool design using the item selection method of maximum item information with the Sympson–Hetter exposure control. The results indicated that two-pool designs outperform the single-pool design with respect to the degree of resistance to item sharing, under the condition that the single pool can be split into two subpools, each of which satisfies the item pool requirements (e.g., content balance) of a test. This study also analytically compared the use of multiple pools to the use of a single pool with respect to test security against item sharing under a randomized item selection method. The theoretical results show that, in general, simply using multiple pools instead of a single pool actually does not improve test security. The conclusion obtained from the simulation study employing a maximum item information selection method is quite different from what has been derived theoretically under randomized item selection. That is, different conclusions have been reached under different item selection methods. In a sense, the randomized item selection method and the maximum information method can be regarded as two approaches that are located at opposite ends of a continuum of possible item selection methods. The former has no adaptive feature at all, whereas the latter is completely adaptive. Any other CAT item selection method is located somewhere along that continuum. One can surmise that the benefit of employing a multiple-pool design may diminish as an item selection algorithm deviates from adaptability after trying to equalize item exposure rates. The results obtained in this study can shed some light on these operational CAT item selection methods and provide guidelines to practitioners in item pool designs if test security is of concern.

Appendix A: Notation

J	The number of item pools in a multiple-pool design
L	The test length
M	The number of new items
m_{jk}	The number of common items of pool j and pool k for $1 \le j < k \le J$
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 $j = 1, \ldots, J$
$P_J(i \mid t)$	The probability that the ith item of an examinee is a compromised item at time t if J item pools are used ($1 \le i \le L$ and $J \ge 1$)
$p_{jk}(t)$	The proportion of compromised items among common items of pools j and k for $1 \le j < k \le J$ at time t
$p^*(t)$	The overall proportion of compromised common items at time $p^*(t) = \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk} p_{jk}(t) / \sum_{j=1}^{J-1} \sum_{k=j+1}^{J} m_{jk}$
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 proportion of compromised items at time t, $r(t) = n(t)/N$
$r_j(t)$	The proportion 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 ($1 \le i \le L$)
S720	A single-pool design with 720 items
T360	A no-overlap two-pool design in which each subpool contains 360 items
T390	A two-pool design with 330 unique and 60 common items
T420	A two-pool design with 300 unique and 120 common items

Appendix B

Proof of Result 1

1. By the multiplication principle (see Hogg & Tanis, 1997),

$$P_1(i|t) = \frac{n(t) \times (N-1)!}{N!} = \frac{n(t)}{N} = r(t),$$

and the conditional probability that the ith item administered to an examinee is a compromised item given pool j

is $r_j(t) = n_j(t)/N_j$ for $j = 1, 2$. According to the total probability formula (see Bickel & Doksum, 1977, p. 440), we obtain $P_2(i \mid t) = r_1(t)Q_1 + r_2(t)Q_2$ for any fixed i.

2. In this case, $N = N_1 + N_2$. From the first part of Result 1,

$$P_1(i|t) - P_2(i|t) = \frac{n(t)}{N_1+N_2} - \frac{n_1(t)}{2N_1} - \frac{n(t)-n_1(t)}{2N_2} = \frac{(N_1-N_2)(n_1(t)N - n(t)N_1)}{2NN_1N_2} = \frac{N_1-N_2}{2N_2}(r_1(t)-r(t)).$$

3. The number of compromised common items is $m_{12}\,p_{12}(t)$. Thus, $n_2(t) = n(t) - n_1(t) + m_{12}\,p_{12}(t)$. By the first part of Result 1, Eq. 2 is obtained. Furthermore, if $N_1 = N_2$, then $2N_1 = N + m_{12}$. By Eq. 2,

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

and

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

Proof of Result 2

The proof for the multiple item pools is similar to the proof of Result 1. According to the total probability formula, Eqs. 3 and 4 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 $\sum_{j=1}^{J} n_j(t) = n(t) + \sum_{j=1}^{J-1}\sum_{k=j+1}^{J} m_{jk}p_{jk}(t)$, by Eq. 4,

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

and it is not difficult to verify Eq. 7.

References

Ariel, A., Veldkamp, B. P., & van der Linden, W. J. (2004). Constructing rotating item pools for constrained adaptive testing. *Journal of Educational Measurement, 41,* 345–359.

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

Chang, H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage computerized adaptive testing with *b* blocking. *Applied Psychological Measurement, 25,* 333–341.

Chang, H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23,* 211–222.

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: Erlbaum.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small-scale cheating. *International Journal of Testing, 9,* 283–309.

Hartley, H. O., & Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *Annals Mathematical Statistics, 33,* 350–374.

Hetter, R., & Sympson, B. (1997). Item exposure control in CAT-ASVAB. In W. Sands, B. Waters, & J. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141–144). Washington, DC: American Psychological Association.

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

Honan, W. H. (1995, January 4). Computer admissions test to be given less often. *New York Times*, p. A16.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Luecht, R. M. (1998, April). *A framework for exploring and controlling risks associated with test item exposure over time*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

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

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

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

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.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17,* 151–166.

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

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.),

Compuerized adaptive testing: Theory and practice (pp. 75–99). Dordrecht, The Netherlands: Kluwer.

Wainer, H. (2000). Rescuing computerized testing by breaking Zipf's law. *Journal of Educational and Behavioral Statistics, 25,* 203–224.

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*.

Yi, Q., & Chang, H. (2003). *a*-stratified multistage CAT design with content-blocking. *British Journal of Mathematical and Statistical Psychology, 56,* 359–378.

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.