Graduate Management Admission Council®

Test Overlap Rate and Item Exposure Rate as Indicators of Test Security in CATs

Juan Ramón Barrada
Universidad Autónoma de Barcelona

Julio Olea, Vicente Ponsoda, and Francisco José Abad
Universidad Autónoma de Madrid

Presented at the Item Exposure Paper Session, June 3, 2009

Abstract

A computerized adaptive test (CAT) is considered more secure the lower the overestimation of the examinee's trait level due to item pre-knowledge. The common measures of test security have been the overlap rate between examinees and the distribution of item exposure rates. We explain that lower overlap rates or more homogeneous distributions of usage of the items might not lead to safer CATs. Instead of these variables, we show that the probability of item pre-knowledge of the first items administered and the overlap rate for high trait levels are better variables for assessing test security. This is illustrated in three different studies in which item bank disclosure is simulated. These studies compare the point Fisher information, the progressive method, and the alpha-stratified selection methods. The alpha-stratified method, the option among the three leading to the lowest overlap rate and most homogeneous item exposure distribution when there is no bank disclosure, is not the selection method offering the highest level of test security.

Acknowledgment

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

Citation

Barrada, J., Olea, J., Ponsada, V., & Abad, F. (2009). Test overlap rate and item exposure rate as indicators of test security in CATs. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Juan Ramón Barrada, Facultad de Psicología, Universidad Autónoma de Barcelona, 08193 Barcelona, Spain. E-mail: juanramon.barrada@uab.es

Test Overlap Rate and Item Exposure Rate as Indicators of Test Security in CAT

The risk of examinees receiving inflated trait estimates due to previous item knowledge is one danger in computerized adaptive testing (CAT; Chang, 2004). CAT allows continuous testing with an item bank that frequently is static over time. This characteristic of CAT makes it possible for future examinees to obtain information from previous examinees about the items they received. Test security is, thus, a major concern in some CATs (Davey & Nering, 2002).

A CAT is more secure the lower the benefit in trait level estimation that an examinee could obtain due to item pre-knowledge. Two main variables have been used as indicators of test security (Chang & Zhang, 2002). First, the pair-wise overlap rate, which is defined as the proportion of items that, on average, two examinees share (Way, 1998). It has been assumed that the greater the overlap rate the greater the trait level overestimation due to item bank disclosure. Second, the distribution of item exposure rates, under the assumption that CATs with more homogeneous rates will be more robust to item leakage (Chen, Ankenmann & Spray, 2003).

Limitation of Security Evaluation in CATs

Usually, the overlap rate and the distribution of the item exposure rates are obtained by means of studies where no examinee has item pre-knowledge. In these studies, the probability of a correct response is determined by a typical IRT model, where no parameter marking the presence or absence of pre-knowledge of each item is included. A small number of studies have simulated item sharing (Guo, Tay & Drasgow, 2009; McLeod & Lewis, 1999; McLeod, Lewis, & Thissen, 2003; Mills & Steffen, 2000; Segall, 2002; Stocking, Ward, & Potenza, 1998; Yi, Zhang, & Chang, 2008), although none of them have questioned the idea that lower overlap rates or more balanced usage of items lead to higher security.

When evaluating CAT security without including bank disclosure, we could be missing some important aspects that are present when there is item pre-knowledge. We will illustrate this with some examples where we will show that a lower overlap rate might not lead to higher test security. Consider a CAT with the following characteristics, intentionally simplistic: (1) item pre-knowledge means a probability of correct response equal to 1; (2) all the examinees start with the same estimated trait level ($\hat{\theta}_0$ is constant); and (3) different examinees with the same estimated θ receive the same item. Imagine two different scenarios, where the source is a previously tested examinee who shares the content of the items he/she received with a person who will be tested with the same CAT, the recipient:

1. A high trait level source of information (e. g., $\theta^S = +2$) and low trait level recipient (e. g., $\theta^R = -2$).This condition corresponds to a low overlap rate situation when there is no item disclosure (Way, 1998). The source would give a correct response to the item due to his ability. As source and recipient receive the same first item, the recipient would give a correct response to it, so $\hat{\theta}_1^R = \hat{\theta}_1^S$. As both examinees have the same estimated θ, they will receive the same item again and both will give correct responses, so $\hat{\theta}_2^R = \hat{\theta}_2^S$. This dynamic will be repeated until the source gives an incorrect response to an item, but the recipient gives a correct response. We will call *h*

this item position in the test. Then, $\hat{\theta}_h^R > \hat{\theta}_h^S$ and both $\hat{\theta}_h^R$ and *h* can be very high. As the recipient has an inflated θ, he/she will probably miss the next item, so $\hat{\theta}_{h+1}^R < \hat{\theta}_h^R$. The estimated θ of the recipient will start to decrease until his/her estimated θ is equal to some of the (high) estimated θ of the source, where more pre-known items will be presented. The expected benefit of item pre-knowledge in this condition would be high, although the overlap rate, when there is no bank disclosure, was small.

2. Both source and recipient have a low θ (e. g., $\theta^S = \theta^R = -2$). When there is no bank disclosure, this corresponds to a high overlap condition. Both examinees receive the same first item, but the source might incorrectly answer it, while the recipient will give a correct response. Thus, $\hat{\theta}_1^R > \hat{\theta}_1^S$. Probably, the recipient will give incorrect responses to the next items, slowly reducing his estimated θ (Rulison & Loken, 2009), until it is approximately equal to some of the (low) estimated θs of the source. Almost every time that the recipient responds correctly due to pre-knowledge and gets an inflated (although low) θ estimate, he/she will have no prior information in order to correctly answer the next item. Although the overlap rate when no item disclosure is present was high, a low benefit of item pre-knowledge is expected. The expected benefit of item pre-knowledge does not correspond to the overlap rate when there was no bank disclosure.

In fact, under the three characteristics of the CAT we have described, we could expect the following: (1) all examinees with very high θs would receive basically the same items (there is a clear route of items for obtaining a high estimated θ); and (2) as the first item is the same for all examinees, it is very easy for any examinee who has item pre-knowledge to be exposed to this route. This is the worst situation for a testing agency in the case of item bank disclosure.

Taking this into account, we consider that the overlap rate or the distribution of the exposure rates could be of limited value for evaluating CAT security. We propose that other variables could be more useful for evaluating CAT security: (1) the probability of recipient and sources sharing the first items administered; and (2) the overlap rate for high θs. Or, in other words, how many routes are available in a CAT in order to reach a high estimated θ and how easy is it to be exposed to one of these routes from the beginning of the test?

Importantly, the item selection methods with better performance in terms of overlap rate or distribution of exposure rates do not have to be those offering better results in terms of routes leading to high θ estimates.Thus, commonly used variables for assessing test security in CATs could be misleading. We will now present different item selection methods and show how they differ in these points.

Item Selection Methods in CATs

Point Fisher information method. The most commonly applied selection algorithm in CATs is the administration of the item offering maximum Fisher information at the estimated θ (Lord, 1980):

$$\arg\max_{i \in B_q} I_i(\hat{\theta}), \tag{1}$$

where $I_i(\hat{\theta})$ is the Fisher information of item *i* for the estimated trait level, $\hat{\theta}$ and B_q is the subset

of items belonging to the item bank that can be presented to the examinee, with q being the item position in the CAT. If no restriction is active, B_q consists of those items not presented to that examinee in the first $(q-1)$ items. The Fisher information function for the 3-parameter logistic model (Birnbaum, 1968) is calculated according to Equation 2:

$$I_i(\theta) = \frac{2.89 a_i^2 (1-c_i)}{\left(c_i + \exp\left[1.7 a_i (\theta - b_i)\right]\right)\left(1+\exp\left[-1.7 a_i (\theta - b_i)\right]\right)^2}, \quad (2)$$

where a_i is the discrimination parameter, b_i is the locating parameter and c_i is the pseudo-guessing parameter for item i.

This method, which we will call Point Fisher Information (PFI), leads to a high overlap rate and to a highly unbalanced distribution of exposure rates, with some items presented to almost all examinees and many that are never administered.

Alpha-stratified method. The alpha-stratified (AS) method (Chang & Ying, 1999) is the alternative method to PFI that has attracted the most attention in recent years. With this method, a much more balanced usage of items is obtained, leading to an overlap rate close to the minimum possible, although the cost for this is an increment in the measurement error compared to PFI.

In the AS method, prior to the administration of a test with a length of Q items, the bank is divided into K strata. In order to do so, the n items in the bank are ordered to have increasing a parameter values. The first n/K items in the bank are assigned to the first stratum, the second n/K to the second stratum, and so on. During the administration of the test, the first Q/K items will be selected from the first stratum, the second Q/K items will be selected from the items in the second stratum, and so on. The item selected is the item with the minimum difference in absolute value between $\hat{\theta}$ and its b parameter:

$$\arg\min_{i \in B_q} \left| \hat{\theta} - b_i \right|. \quad (3)$$

In this case, B_q consists of the intersection between the non-presented items and the items belonging to the active stratum for that item position in the test.

Progressive method. Another option for reducing the overlap rate, with negligible effects on accuracy in comparison with PFI, is to select items randomly at the beginning of the test and, as the test goes on, to increase the relevance of the Fisher information in the item selection (Barrada, Olea, Ponsoda, & Abad, 2008; Li & Schafer, 2005; Revuelta & Ponsoda, 1998). The progressive (PG) method (Revuelta & Ponsoda, 1998) uses this idea. In the PG method, the item selected is the one that maximizes the sum of two elements, one determined by Fisher information and the other by a random number:

$$\arg\max_{i \in B_q} \left[(1 - W_q) R_i + W_q I_i(\hat{\theta}) \right], \quad (4)$$

where W_q is the weight of item information in the selection criterion and R_i is a random number belonging to the interval [0, $\max_{i \in B_q} I_i(\hat{\theta})$].

Barrada et al. (2008) have proposed the following equation to relate W_q to q:

$$W_q = \begin{cases} 0 & if \quad q = 1 \\ \dfrac{\sum_{f=1}^{q} (f-1)^t}{\sum_{f=1}^{Q} (f-1)^t} & if \quad q \neq 1 \end{cases}. \qquad (5)$$

The *t* parameter marks the speed at which the weight of the random component is reduced and, thus, defines the improvement in the overlap rate and accuracy reduction, in comparison with PFI. With *t* equal to 1, it is possible to get an accuracy equivalent to that obtained with the PFI method, while improving item bank security.

Many other item selection methods have been proposed for improving bank security (Georgiadou, Triantafillou, & Economides, 2007). We will restrict our attention to the PFI, AS and PG methods. PFI is, currently, the common standard. AS is an option leading to a major reduction in overlap, with an increment in the measurement error. PG permits a more balanced usage of items, when compared with PFI, while maintaining the accuracy of the latter.

Hypotheses

Both PFI and AS have a negative characteristic for the scenario of bank disclosure: given the same B_q, two examinees with the same estimated θ will receive the same item. However, when the PG is applied, the same estimated θ can lead to different items administered, especially in the early part of the test, as item selection starts randomly.

The AS method has an overlap rate lower than the overlap obtained with the PG method and also a more homogeneous distribution of item exposure rates (Barrada, Olea, Ponsoda, & Abad, in press) .Therefore, one hypothesis, based on current standards for evaluating test security, would be that the AS method would be better in a condition of bank disclosure. However, we have shown that the PG method outperforms the AS method in the ease of incorporating some of the routes leading to a high estimated θ. Because of this, our expected results when simulating conditions of bank disclosure are that the PG method will outperform the AS method when there is bank disclosure.

SIMULATION STUDIES

To investigate the idea that a lower overlap rate or a more balanced usage of the items do not imply a higher resistance to item disclosure and that the other proposed variables are more adequate, we conducted a series of simulation studies comparing PFI, AS and PG item selection methods. In the first simulation, we present the condition of no item bank disclosure. This study will enable us to distinguish between the common expectation (AS, the method with lower overlap rate and more balanced usage of items, should be the method less affected by bank disclosure) and our expectation that it is more important to consider the number of routes for obtaining a high θ estimate and the ease of incorporating one of these routes (PG outperforms the other two item selection methods in this respect, so it should be the less affected by bank disclosure). The next three simulation studies are intended to evaluate the resistance of these item selection methods under different conditions of bank disclosure. They are complementary,

as they consider item bank disclosure from different perspectives. If similar results are obtained from them, this will increase the confidence in our conclusions. In Study 2, we analyze the impact of a different number of sources in overestimation of recipient θ, evaluating if the common indicators of bank security correctly order the different item selection methods in their resistance to bank disclosure. In Study 3, we study which θ level of sources is most useful for different θ levels of recipients. If overlap rate is a correct indicator of bank security, this θ level should be the one where maximum overlap was found in the condition of no disclosure. In Study 4, we analyze how the measurement error increases the longer the item bank has been in use and we introduce some more realistic conditions of bank disclosure.

Study 1: No Item Disclosure

Method

Item bank and test length. Ten item banks of 500 items were randomly generated. The distributions for the parameters were: $a \sim N(1.2, 0.25)$; $b \sim N(0, 1)$; $c \sim N(0.25, 0.02)$. Length of the test was set at 25 items.

Item selection methods. We evaluated three different item selection methods: PFI, AS and PG methods. In the AS method, the item bank was stratified into five strata and an equal number of items were presented from each stratum. For the PG method, the *t* parameter was fixed at 1 (Equation 5).

Restriction of maximum exposure rate. A common approach to improving bank security is to limit the maximum exposure rate (r^{max}) that no item should surpass. To do this, we used the Sympson-Hetter method (Sympson & Hetter, 1985), with r^{max} equal to 0.25.

Trait level of the simulees. We simulated two different conditions. In the first, where we obtained results for the overall population, the true θ level of the simulees was randomly sampled from a N(0, 1) distribution. For each item bank, 5,000 simulees were sampled. In the second condition, we were interested in the results conditional on θ levels. To do so, we simulated 1,000 examinees for nine different and equally spaced θ points, ranging from – 2 to 2.

Estimation/assignment of trait level. The starting $\hat{\theta}_0$ was randomly selected from the uniform interval (–0.5, 0.5). Maximum-likelihood estimation has no solution in real numbers when there is a constant response pattern, all correct or all incorrect responses. Thus, until there was at least one correct and one incorrect response, $\hat{\theta}$ was assigned using the method proposed by Dodd (1990): when all the responses were correct, $\hat{\theta}$ was increased by $(b_{\max} - \hat{\theta})/2$; if all the responses were incorrect, $\hat{\theta}$ was reduced by $(\hat{\theta} - b_{\min})/2$. After a mixed pattern of responses was obtained or when the test was finished, we applied maximum-likelihood estimation, with the restriction that $\hat{\theta}$ had to be in the interval [–4, 4].

Performance measures. Five dependent variables were used for the comparison among conditions:

1. RMSE, calculated according to Equation 6:

$$RMSE = \left(\sum_{g=1}^{v} \left(\hat{\theta}_g - \theta_g \right)^2 / v \right)^{1/2}, \quad (6)$$

where v is the number of simulees;

2. Distribution of item exposure rates ($r_{i,1..Q}$): the exposure rate of item i considering the whole test (from position 1 to position Q, with Q being the test length);

3. Overlap rate for the overall population. The common reported value of overlap rate is the pair-wise overlap rate, which provides information about the mean proportion of items shared by two examinees. If an item bank is disclosed, it would be possible for an examinee to gain item pre-knowledge from more than one source. Because of this, we will calculate the overlap rate with z different sources of information. This overlap rate was calculated according to Equation 7:

$$T^{z}_{1..Q,1..Q} = \frac{\sum_{i=1}^{n} r_{i,1..Q} \left[1 - \left(1 - r_{i,1..Q} \right)^z \right]}{Q}, \quad (7)$$

where n is the item bank size, z is the number of sources of item information and $T^{z}_{1..Q,1..Q}$ is the overlap rate considering z sources and the whole test. $T^{z}_{1..Q,1..Q}$ provides information about the proportion of items an examinee will face with item pre-knowledge, given z sources. When z is equal to 1, $T^{z}_{1..Q,1..Q}$ is the pair-wise overlap rate. We calculated this overlap rate with z values ranging from 1 to 5 sources.

4. Overlap rate between different θ levels. The mean overlap rate between an examinee with trait level equal to θ_1 and an examinee with trait level equal to θ_2 is

$$T^{z=1,\theta_1,\theta_2}_{1..Q,1..Q} = \frac{\sum_{i=1}^{n} r^{\theta_1}_{i,1..Q} r^{\theta_2}_{i,1..Q}}{Q}, \quad (8)$$

where $r^{\theta_1}_{i,1..Q}$ and $r^{\theta_2}_{i,1..Q}$ refer to the item exposure rate of item i for the two θ levels.

5. Probability that an examinee already tested could inform another examinee about the item content for each of the item positions. This probability is equal to the overlap rate, with z equal to 1, between a whole test and any single item position ($T^{z=1}_{1..Q,q..q}$):

$$T^{z=1}_{1..Q,q..q} = \sum_{i=1}^{n} r_{i,1..Q} r_{i,q..q}, \quad (9)$$

where $r_{i,q..q}$ is the exposure rate of item i in just the qth position of the test.

Some studies have reported the overlap rate conditional until an item position in the test ($T^{z=1}_{1..q,1..q}$; Barrada, Velkamp & Olea, 2009). This information, although useful, is not equivalent to the probability of an examinee informing about the item content conditional on item position in

the test. Imagine an item bank of 10 items and a test length of 10 items. Applying random item selection, the overlap between examinees until the qth item would be .1*q, but the probability of an examinee informing another about the any item he/she will receive is equal to 1.

The first variable measures accuracy. The second and third are the common indicators of test security. The fourth and fifth are those that we hypothesize can better measure test security.

Results

The results for the overall population in terms of overlap rate and RMSE can be seen in Table 1. As expected, the selection method with the greatest overlap rates was PFI. With the AS method, much lower overlap rates were obtained. The PG method was between these extremes. The order in overlap rates was the same for the three methods for any number of sources considered. While the PFI and PG methods offered the same RMSE, the accuracy was reduced with the AS method, as we obtained an RMSE 0.07 higher than with the other two methods.

Table 1. Overlap Rate and RMSE According to Item Selection Method With No Bank Disclosure

Method	$T^{1}_{1..Q,1..Q}$	$T^{2}_{1..Q,1..Q}$	$T^{3}_{1..Q,1..Q}$	$T^{4}_{1..Q,1..Q}$	$T^{5}_{1..Q,1..Q}$	RMSE
PFI	0.19	0.35	0.46	0.56	0.63	0.26
AS	0.07	0.14	0.20	0.26	0.31	0.33
PG	0.14	0.25	0.34	0.41	0.47	0.26

Figure 1 shows the exposure rates of the items considering the whole population. In accordance with the overlap data, the exposure rates for the PFI method were the most unbalanced, while those for the AS method were the most homogeneous. The distribution of the PG method was located between the PFI and AS methods. With the PFI method, up to 57% of the items had an exposure rate equal to 0. With the AS method, the proportion of unused items was 1%. For the PG method, there were no items that were never presented. The proportion of items with exposure rates over or near r^{max} (> 0.2) were 12% for the PFI method, 1% for the AS method, and 7% for the PG method. With the Sympson-Hetter method, some items had exposure rates slightly over r^{max} (van der Linden, 2003).

In Table 2, the overlap rates between examinees of the same and different θs are shown. The highest overlap rates were between examinees of the same θ. The higher overlap for examinees of the same θ can be found for examinees with extreme θs, as fewer items were available there. The greater the difference between the θs of the examinees, the lower the overlap (Way, 1998). The highest overlap rates corresponded to the PFI method. The PG method reduced the overlap. In general, the AS was the method with the lowest overlap. An exception can be seen in the overlap between examinees with $\theta = 2$, where it is higher for the AS method than for the PG method.

Figure 1. Exposure Rates of Items According to Item Selection Method With No Bank Disclosure and Items Ordered According to Their Exposure Rates

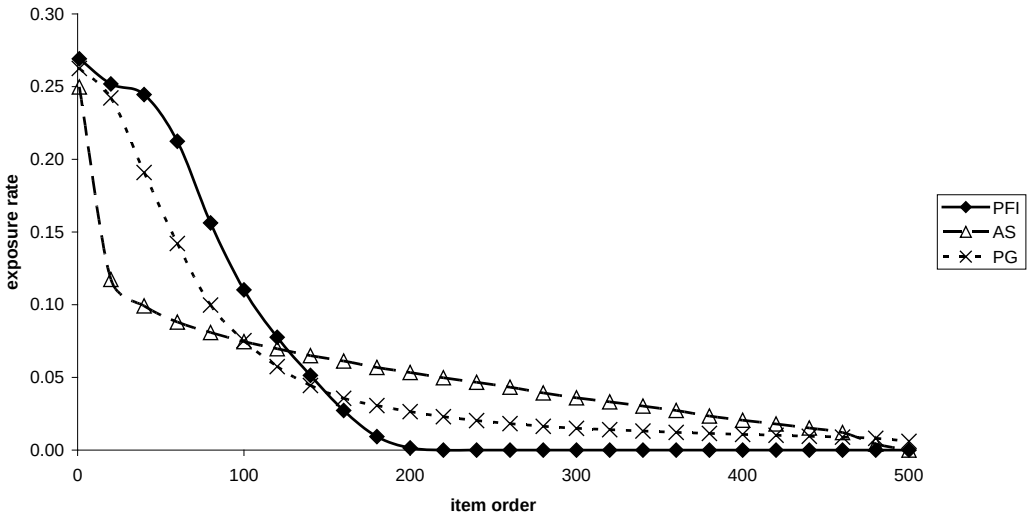


Figure 2 shows the probability of an examinee providing information about the item content at each item position. For the PFI method, the probability of an examinee giving information about the content of the items reduces as the test proceeds. The most interesting result is the pattern of results that is found when comparing the AS and PG methods. For the PG method, because of the randomness in item selection at the beginning of the test, the probability of receiving information about item content is lowest at the start of the test and increases with each new item. For the AS method, the probability of receiving information is higher at the start of the test and decreases as the test continues. After the sixth item, the probability of receiving information is lower for the AS method than for the PG method. Another interesting result is that, for the PG method the probability of receiving information about the content of an item at the end of the test is greater than for PFI. This is because with PG, the items that are most likely to be selected are not presented until the end of the test.

It has to be noted that the probabilities of Figure 2 hold while no examinee has item pre-knowledge or until the examinee reaches the item position where there is item pre-knowledge. With item pre-knowledge, as the probability of a correct response changes with respect to the simulation, these probabilities would be changed.

Discussion

This study performed simulations to evaluate three different item selection methods in terms of their item exposure control. According to Tables 1 and 2 and Figure 1, the selection method that seems to be preferred in order to maximize test security is the AS method, as it has lower overlap rates for the overall population of examinees, a more balanced distribution of the exposure rates and, in general, lower overlap rates conditional on θ levels.

The most generally accepted assumption would be that the AS method would be the method with the lowest inflation of the θ levels when there was item bank disclosure. However, our expectation was that the PG method would outperform the AS method in this condition. We have argued above that special attention should be paid to two other variables: (1) the number of different paths available for obtaining an estimate of a high θ level, and (2) the ease of

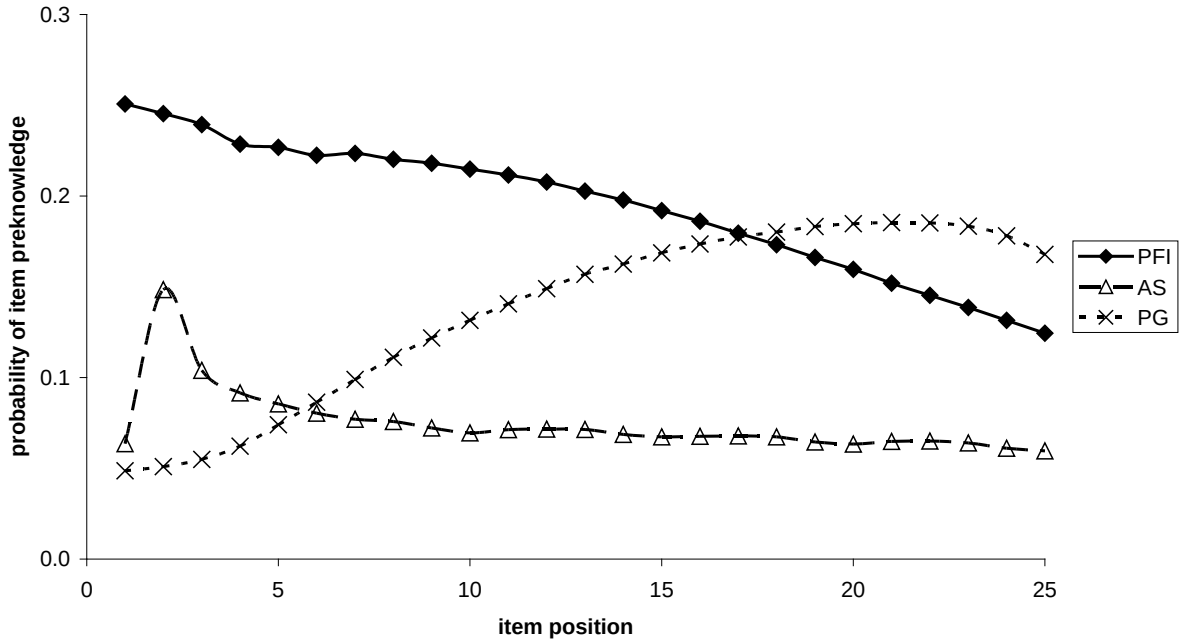
incorporating one of these paths from the beginning of the test. Table 2 and Figure 2 show that the risks for these two points are higher for the AS method than for the PG method. For the AS method, when compared with the PG method, the overlap rate conditional on $\theta = 2$ was greater and the probability of being informed about the item content at the beginning of the test was higher.

Table 2. Overlap Rate According to θ Levels and Item Selection Methods. . Bold-Faced Figures Correspond to θ Level Where the Overlap Rate Was Greater Given a Fixed Overlap Rate (Maximum by Row)

		θ level								
Method	θ	–2	–1.5	–1	–0.5	0	0.5	1	1.5	2
PFI	-2	**.61**								
	-1.5	.47	**.54**							
	-1	.22	.38	**.50**						
	-0.5	.10	.16	.32	**.43**					
	0	.06	.07	.13	.27	**.40**				
	0.5	.04	.05	.06	.11	.27	**.43**			
	1	.03	.03	.04	.06	.13	.31	**.50**		
	1.5	.03	.03	.03	.05	.08	.15	.37	**.59**	
	2	.02	.02	.03	.04	.06	.09	.20	.49	**.69**
AS	-2	**.44**								
	-1.5	.29	**.27**							
	-1	.12	.15	**.16**						
	-0.5	.05	.06	.10	**.11**					
	0	.03	.03	.05	.08	**.11**				
	0.5	.02	.02	.03	.05	.08	**.13**			
	1	.02	.02	.02	.03	.06	.11	**.19**		
	1.5	.02	.02	.02	.03	.05	.08	.18	**.31**	
	2	.01	.02	.02	.03	.04	.07	.15	.35	**.59**
PG	-2	**.44**								
	-1.5	.34	**.38**							
	-1	.17	.28	**.35**						
	-0.5	.07	.12	.23	**.31**					
	0	.03	.04	.09	.20	**.29**				
	0.5	.02	.02	.04	.08	.20	**.31**			
	1	.02	.02	.02	.03	.08	.23	**.36**		
	1.5	.02	.02	.02	.02	.04	.11	.27	**.41**	
	2	.02	.02	.02	.02	.03	.05	.14	.34	**.48**

To check the hypothesis that the common indicators of test security could be misleading and that a CAT using PG would be more robust to item disclosure than a CAT using AS, we conducted the following three studies.

Figure 2. Probability of Item Pre-Knowledge According to Item Position and Item Selection Method With No Bank Disclosure



Study 2: Effect of Disclosure According to the Number of Sources

Method

The method of this second study is similar to that of the first, with some exceptions. In this study, we investigated the effect on overestimation of the θ level when examinee-recipients take the exam after contacting *m* independent examinee-sources of items. The number of sources could be 1, 2, …,10. The process was: (1) to simulate the exam for *m* sources as standard CATs, (2) to do the union of the different sets of items and fix the probability of correct response to these items as equal to 1 for the recipient; and (3) to simulate the CAT for the recipient with these changed probabilities for some items. Sources and recipients were extracted from a standard normal distribution. For each of the 10 conditions (number of sources) 5,000 replications were simulated. Four dependent variables were recorded: (1) proportion of pre-known items in the bank (cardinal of the set formed by the union of items presented to the different sources divided by item bank size); (2) proportion of pre-known items in the test (cardinal of the set formed by intersection of the pre-known items in the bank and the items presented to a recipient divided by test length); (3) bias; and (4) RMSE. The bias was calculated according to Equation 10:

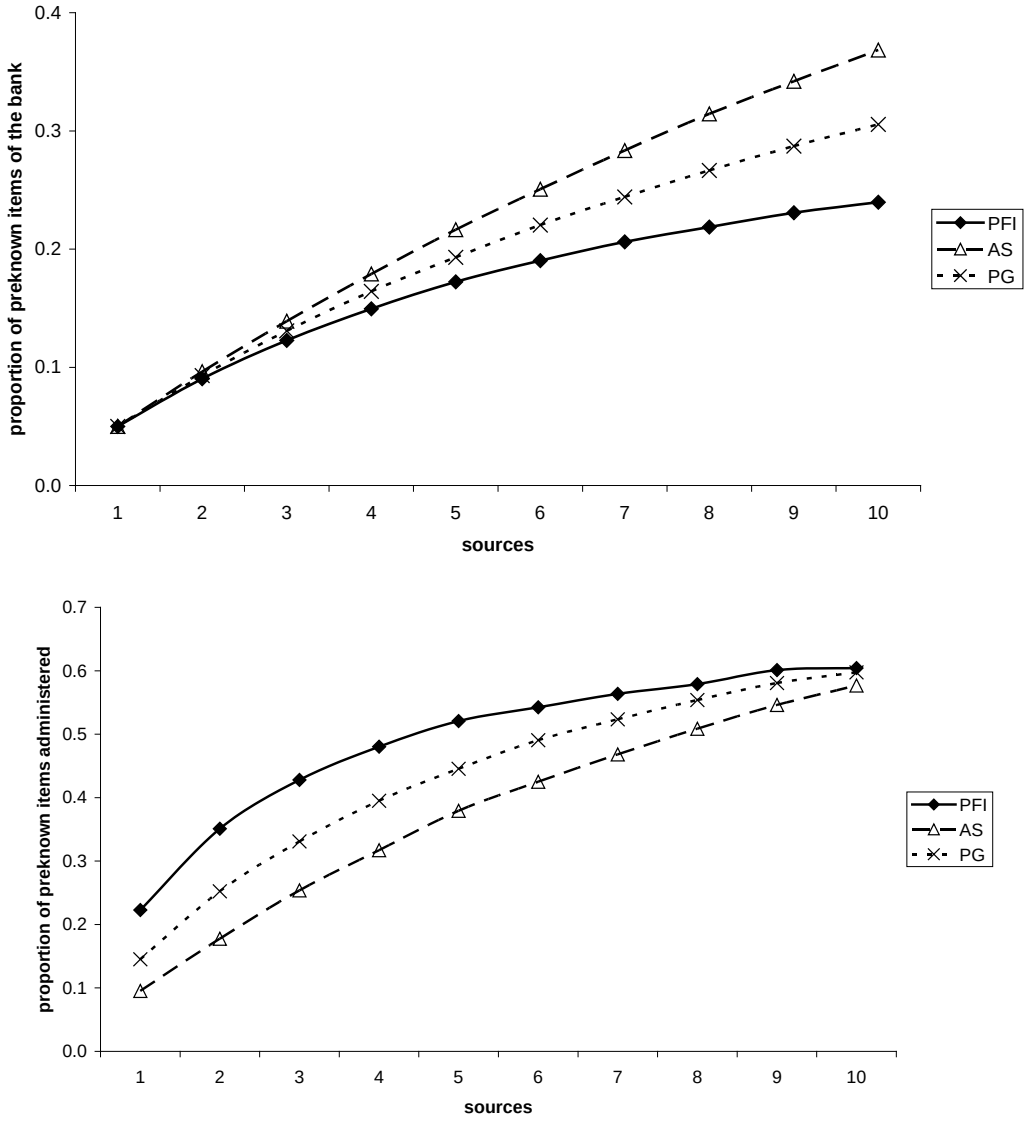
$$Bias = \sum_{g=1}^{v} \left(\hat{\theta}_g - \theta_g \right) / v \qquad (10)$$

Results

The upper panel of Figure 3 shows the proportion of pre-known items in the bank according

to the number of sources. The higher the number of sources, the higher is this proportion. With a high overlap rate, as evaluated in Study 1, the sources offer redundant information. Because of this, the higher the overlap rate the lower the proportion of pre-known items in the bank. Thus, the item selection method leading to higher knowledge of the bank was the AS method.

Figure 3. Proportion of Pre-Known Items in the Item Bank and Proportion of Pre-Kwon Items Administered According to the Number of Sources and Item Selection Method



The proportion of pre-known items administered in the test can be seen in the lower panel of Figure 3. Again, the higher the number of sources, the higher this proportion is. The higher the overlap rate, the higher the proportion of items in the test for which there is pre-knowledge. The method for which the highest proportion of pre-known items was administered is the PFI method; the one with the lower proportion was the AS method. The results of both this figure and the previous figure follow what could be expected from the results of Study 1.

Figure 4 shows the bias (overestimation) and RMSE when item disclosure is present. The selection with worse resistance to item disclosure was PFI. Contrary to the hypotheses derived from the overlap rate and item exposure rates in Study 1, the next was the AS method. The method that was less affected by bank disclosure was the PG method. The presence of each new source produced an increment in bias and RMSE that was higher for the AS method than for the PG method. In any case, the presence of just one source led to bias and RMSE values that could be considered unacceptable.

Discussion

Two results from this study deserve special attention. First, the method with the lower overlap rate, the AS method, was not the method with the lower impact of item disclosure. Second, the method with the lower proportion of pre-known items in the test, the AS method again, was not the method with the greater resistance to bank disclosure. As we hypothesized above, the PG method was the one offering greater security when the item bank was disclosed.

The AS method has two problems in terms of resisting item pre-knowledge: (1) there are not many different paths for the estimation of high θs, and (2) it is easier than with the PG method to incorporate one of these paths from the beginning of the test. In this way, we can explain why with the AS method we find a higher bias and RMSE when item disclosure is simulated.

It should be noted that we have not chosen the worse conditions when simulating the AS method. Randomly assigning the initial θ level in the interval (–0.5, 0.5), as we have done, reduces the probability of item pre-knowledge of the first administered items for the AS and PFI methods, while it does not affect this probability for the PG method, as it starts with random selection. The common practice of starting the CAT with a fixed θ level would deteriorate the performance of the PFI and AS methods, while not affecting the PG method.

That higher impact that is attained with fewer items with pre-knowledge is probably due to the fact that the items with pre-knowledge are mainly situated at the beginning of the test. When this happens, θ estimation is highly shifted to the positive extreme, and many items are required to reduce the overestimated θ level and many items are administered providing low information at the real θ level. When the pre-known items are presented at later stages of the test, as with the PG method, the likelihood-function is more peaked, so the overestimation due to pre-known items is smaller.

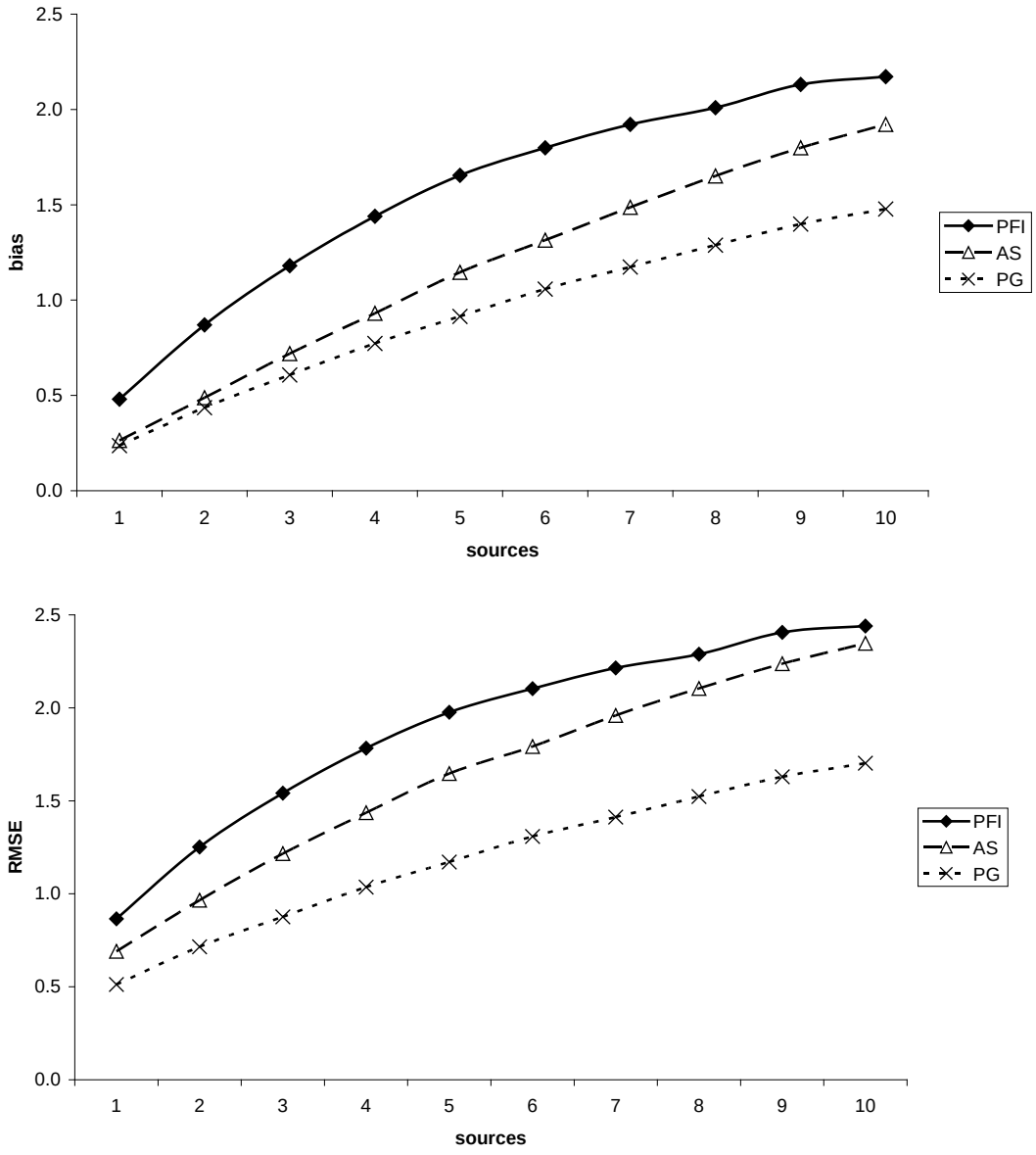
In Study 2, we have evaluated the effects of item disclosure from random sources. Another approach, taken in Study 3, is to evaluate the impact of different trait levels of the sources and recipients.

Study 3: Effect of Disclosure According to the Trait Level of the Sources and the Recipients

Method

The method of this study was equivalent to that employed in the previous studies, except as follows: the source and recipient θ levels were manipulated with 9 levels, from –2 to 2 in increments of 0.5; each recipient had just one source; and each of the 81 conditions (9 levels of source × 9 levels of recipient) was simulated 1,000 times. As dependent variables we used: (1) the proportion of items of the test that were pre-known, and (2) the bias.

Figure 4. Bias and RMSE According to the Number of Sources and Item Selection Method



Results

In Table 3, the proportion of pre-known items according to θ levels of sources and recipients is shown. This table should be compared with Table 2, where the overlap rate between θ levels in the condition of no bank disclosure was shown. While Table 2 was symmetric, Table 3 is not: a high-level source giving information to a low-level recipient does not have the same effect as the opposite case. In Table 2, the maximum overlap rate was on the diagonal, that is, between examinees of the same θ level. In Table 3, the pattern of results depends on the item selection method. For the PFI method, the maximum proportion of pre-known items was achieved with a source with a θ level 0.5 or 1 points above the recipient. For the AS method, except for low θ

recipients, the maximum pre-known items were offered by high-level sources. For the PG method, the maximum proportion was given by sources with equal or slightly greater (+0.5) θs than the recipients.

Table 3. Proportion of Pre-Known Items According to θ Levels of the Source and the Recipient and the Item Selection Method. Bold-Faced Values Correspond to the θ Levels of the Source Where the Proportion of Pre-Known Items Was Greater With the Recipient (Maximum by Row)

Method	Recipient θ Level	Source θ Level								
		–2	–1.5	–1	–0.5	0	0.5	1	1.5	2
PFI	–2.0	.35	**.38**	**.38**	.28	.22	.17	.14	.13	.13
	–1.5	.26	.33	**.39**	.32	.23	.18	.16	.14	.13
	–1.0	.15	.24	.33	**.37**	.29	.20	.17	.16	.16
	–0.5	.08	.13	.22	.31	**.35**	.26	.21	.21	.20
	0.0	.06	.07	.09	.19	.30	**.35**	.32	.31	.29
	0.5	.05	.05	.05	.09	.19	.33	.40	**.41**	.39
	1.0	.03	.03	.04	.07	.11	.22	.37	.48	**.51**
	1.5	.03	.03	.03	.04	.06	.12	.27	.46	**.58**
	2.0	.02	.02	.03	.04	.05	.08	.16	.34	**.55**
AS	–2.0	**.27**	.23	.16	.07	.05	.04	.05	.07	.11
	–1.5	.17	**.19**	.17	.09	.05	.05	.05	.08	.11
	–1.0	.08	.11	.13	.11	.07	.06	.06	.10	**.14**
	–0.5	.04	.05	.08	.10	.10	.08	.09	.12	**.17**
	0.0	.02	.03	.05	.07	.11	.11	.12	.17	**.23**
	0.5	.02	.02	.03	.04	.07	.13	.17	.23	**.34**
	1.0	.02	.02	.02	.03	.06	.11	.20	.32	**.47**
	1.5	.01	.02	.02	.03	.04	.07	.15	.33	**.60**
	2.0	.02	.02	.02	.03	.04	.06	.14	.36	**.63**
PG	–2.0	.31	**.33**	.27	.13	.06	.04	.03	.03	.02
	–1.5	.23	.29	**.31**	.20	.09	.04	.03	.03	.02
	–1.0	.13	.22	**.29**	.27	.15	.07	.04	.03	.03
	–0.5	.06	.09	.18	**.25**	**.25**	.16	.07	.05	.03
	0.0	.03	.04	.08	.16	**.26**	.25	.16	.10	.07
	0.5	.02	.02	.04	.07	.15	.27	**.29**	.21	.14
	1.0	.02	.02	.02	.03	.08	.18	.30	**.34**	.30
	1.5	.02	.02	.02	.02	.03	.09	.20	.36	**.42**
	2.0	.02	.02	.02	.02	.03	.06	.12	.28	**.42**

In Table 4 we present the bias (overestimation) according to θ levels of sources and recipients. The pattern of results is markedly different between the PFI and AS methods, on the one hand, and the PG method on the other. For both the PFI and AS methods (with some exceptions in the PFI for low level recipients), the higher the θ of the sources, the higher the overestimation. For the PG method, the higher overestimation comes from gaining information from a source slightly above (+0.5 or +1) the θ level of the recipient.

Table 4. Bias According to θ Levels of the Source and the Recipient and the Item Selection Method. Bold Faced Values Correspond to the θ Levels of the Source That Produced the Greater Gain in the Estimated θ Levels of the Recipient (Maximum by Row). Cells Where the Bias Was Equal to or Over .5 Are Shaded in Gray

Method	Recipient θ Level	Source θ Level								
		–2	–1.5	–1	–0.5	0	0.5	1	1.5	2
PFI	–2.0	.76	.96	**1.15**	1.08	1.05	.99	.89	.89	.86
	–1.5	.39	.58	.84	**.89**	.82	.83	.88	.84	.80
	–1.0	.17	.31	.48	.72	.73	.71	.76	**.83**	.81
	–0.5	.10	.13	.22	.43	.65	.67	.75	**.88**	.86
	0.0	.09	.09	.11	.23	.43	.67	.84	1.03	**1.06**
	0.5	.07	.07	.06	.10	.20	.47	.77	1.02	**1.19**
	1.0	.04	.04	.06	.07	.12	.24	.52	.91	**1.23**
	1.5	.04	.03	.03	.05	.07	.10	.28	.63	**1.07**
	2.0	.02	.04	.04	.04	.05	.09	.13	.33	**.79**
AS	–2.0	.61	.61	.50	.30	.28	.31	.41	.49	**.74**
	–1.5	.30	.36	.42	.32	.26	.31	.38	.58	**.66**
	–1.0	.13	.18	.28	.29	.24	.30	.35	.55	**.74**
	–0.5	.10	.11	.14	.20	.27	.32	.41	.55	**.77**
	0.0	.06	.07	.12	.13	.22	.30	.41	.64	**.92**
	0.5	.06	.07	.09	.09	.14	.27	.41	.67	**1.14**
	1.0	.04	.06	.06	.06	.11	.16	.37	.72	**1.26**
	1.5	.03	.03	.05	.04	.04	.09	.20	.54	**1.30**
	2.0	.03	.03	.03	.05	.04	.07	.14	.46	**1.05**
PG	–2.0	.63	**.78**	.74	.40	.22	.17	.12	.10	.07
	–1.5	.31	.47	**.63**	.48	.28	.14	.12	.11	.06
	–1.0	.12	.23	.38	**.49**	.34	.20	.13	.11	.11
	–0.5	.04	.08	.17	.31	**.41**	.35	.21	.16	.12
	0.0	.04	.05	.06	.15	.32	**.43**	.37	.30	.21
	0.5	.04	.03	.05	.06	.13	.34	**.51**	.50	.41
	1.0	.03	.03	.04	.03	.07	.17	.41	.65	**.70**
	1.5	.02	.02	.03	.04	.04	.08	.20	.52	**.81**
	2.0	.05	.03	.03	.03	.04	.06	.10	.31	**.67**

Discussion

If you are an examinee seeking to inflate your score by means of item pre-knowledge, which θ level source should you look for? According to studies in which no item disclosure was simulated, that examinee should try to find an examinee with a similar θ level. When item disclosure is simulated, the answer changes and depends on the item selection method implemented in the CAT. Both PFI and AS offered the same answer (with slight exceptions): in general, look for a source with a high θ level. Or, in other words, one source fits all the recipients. The source θ level leading to higher benefit is not the one with the higher overlap rate

when there is no bank disclosure. If the CAT uses the PG method, examinees trying to boost their estimated θ level should try to find sources with slightly higher θ levels. Again, overlap rate, as shown in Study 1, would lead to incorrect predictions.

In Studies 2 and 3,(1) all the examinees had item pre-knowledge, (2) the sources could give perfect information about the items they received, (3) the recipients could remember perfectly all the items the sources shared with them, and (4) item pre-knowledge was equal to a probability of a correct response equal to 1. Although all of these conditions were useful for capturing how item disclosure works, in the next study we present a more realistic simulation. Our goal was to check whether, when changing the four points noted above, the pattern of results still holds.

Study 4: Effect of Disclosure According to Examinee Position in the Item Bank Life

Method

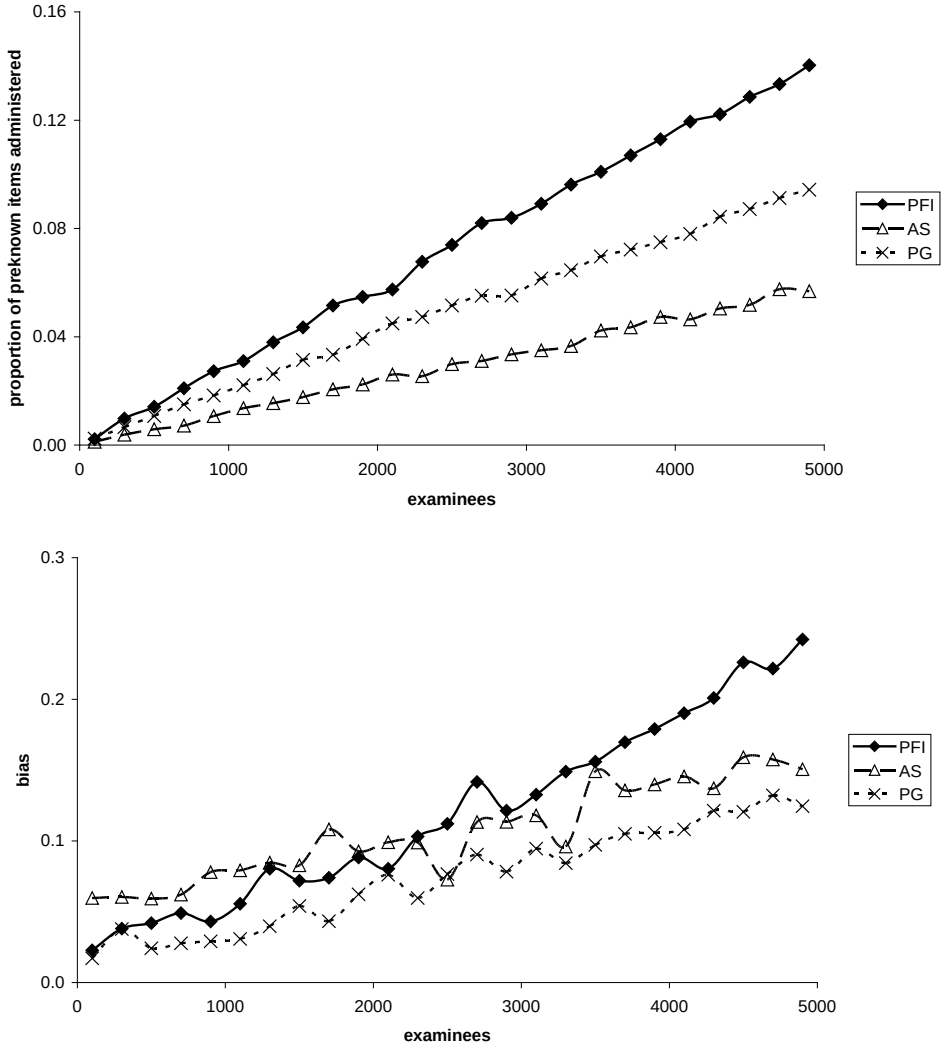
In this study, the longer the item bank has been in use, the higher the probability of an examinee knowing one or several sources. We set the probability of an examinee knowing each previously examined person as equal to 0.001. For the *(h+1)*th examinee, for each of the *h* previous examinees, a random number was extracted from a uniform distribution (0, 1). Only if the number was lower than 0.001, did that examinee become a source. The probability of the source giving information about each single item he/she received was equal to 0.15. Whenever he shared the item, the probability of a correct response to that item was fixed at 1 for the recipient. The probability of the source sharing the content of each item can also be viewed, in this context, as the probability of the recipient remembering it. Clearly, the probabilities chosen are arbitrary, but serve to show the effect of bank disclosure under different conditions from those simulated previously. Unreported simulations with different values led to equivalent patterns of results. As dependent variables, we present the number of pre-known items in the test, the bias, and the RMSE. To improve the clarity of the figures, we show the results averaged for each 200 examinees.

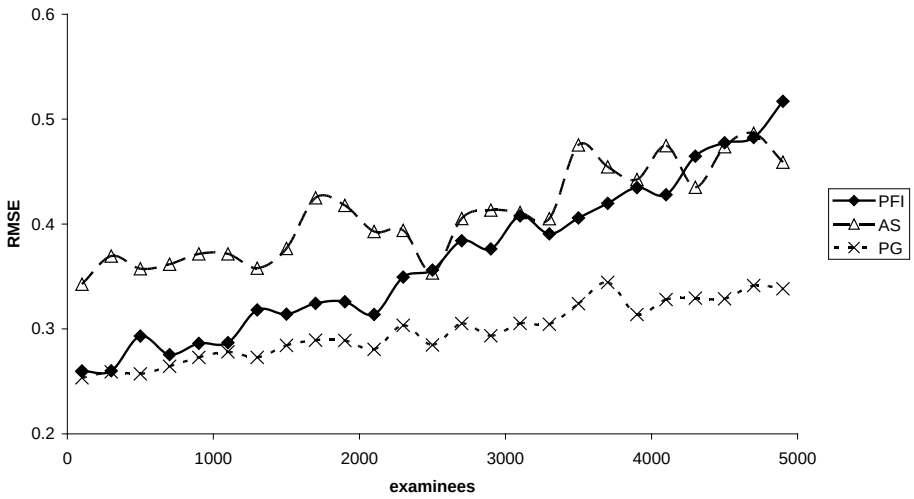
Results

As expected due to the simulation procedure, the higher the examinee position in the life time of the item bank, the higher the number of pre-known items in the test. This can be seen in the upper panel of Figure 5. Consistent with the results reported in Figure 3, the item selection method that led to the higher number of items administered with prior information was the PFI method. The method where fewer pre-known items were administered was the AS method. The PG method was between these two extremes.

The middle and lower panels of Figure 5 show how these pre-known items affected the accuracy of θ estimation. The later the position of the examinee, the higher both the bias and the RMSE, as the expected number of sources increases. Consistent with Table 1, at the beginning of the item bank life the PFI method had a lower measurement error than the AS method and a bias and RMSE equivalent to those obtained with the PG method. But, as the PFI method is the item selection method where the measurement error increased more rapidly, when 2,400 examinees or more had been tested, its bias was higher than with the AS method and for the last 1,000 examinees its RMSE was also higher. For AS and PG, the speed with which they increment the measurement error appeared equivalent. As the PG method started with higher accuracy, this method showed throughout the bank life lower bias and RMSE than the AS method.

Figure 5. Proportion of Pre-Known Items Administered, Bias and RMSE According to Examinee Position and Item Selection Method





Discussion

In this study, we have changed the way in which item disclosure was simulated. In this case, the sources could also have an inflated θ level. Neither the source nor the recipient perfectly memorized the content of the item. In these different conditions, again, the item selection method with lower resistance to bank disclosure was the PFI method. The order of the different item selection methods in terms of accuracy depended on the length of the item bank life. At the beginning, the worst method was AS; after some examinees, it was the PFI method. Importantly, for any examinee position, the method with the lowest measurement error was the PG method. As in Study 2, we found that administering a lower number of pre-known items, as the AS method does, did not lead to a lower effect of overestimation of θ.

CONCLUSIONS

The purpose of these four studies was to evaluate the validity of two variables that are frequently used when evaluating test security in CATs: distribution of the item exposure rates and overlap rate. It has been commonly assumed that the more balanced the distribution and the lower the overlap rate, the higher would be the resistance to item disclosure. We have argued that this idea could be incorrect. First, we have shown the results of what could be considered a typical study (Study 1): in the condition of no disclosure, the conclusion should be that the AS method is the safer one. In the following three studies, we have shown that this conclusion does not hold. The AS method, when compared with the PG method: (1) was more affected by the presence of sources, and any new source increased the bias and RMSE more (Study 2); (2) has a "golden source", a source with a high θ level that will inflate the θ estimation of all the recipients the most (Study 3), and this "golden source" is not the source with higher overlap as measured in Study 1; and (3) is never, when considering the item bank life under conditions of bank disclosure, the item selection method to be preferred in terms of accuracy (Study 4). Taking all these facts into account, we can conclude that the usual variables reported in the extensive literature on test security and item exposure control in CATs should be considered with caution.

We foundnd another interesting result: a higher number of pre-known items does not lead directly to a higher effect on the accuracy of estimation. The effect of bank disclosure cannot be tested by the overlap rate, the distribution of item exposure rates, the percentage of the item bank

that is pre-known, or by the percentage of the items administered during the test that are pre-known. It seems that the only way to detect the safer item selection methods is by simulation of item disclosure.

What seems clear is that one way of improving test security is to increase randomness in item selection at the beginning of the test, as the PG method does. Thus, we reduce the overlap rate when the test starts and increase the number of possible combinations of items leading to an estimation of high θ levels. Several other options that could improve the resistance to bank disclosure could also be considered. For instance, the testing agency could construct an item bank with a higher mean or standard deviation of the *b* parameter distribution, so more items could be available at the high extreme and, thus, probably reducing the overlap rate at the high levels. Another option would be to use methods for the restriction of a maximum rate conditional on θ levels (Stocking & Lewis, 2000; van der Linden & Veldkamp, 2007), reducing the r^{max} value especially for high θ levels.

In these studies, we have not used any method to try to detect the examinees with item pre-knowledge by means of their pattern of responses (Bradlow, Weiss, & Cho, 1998; McLeod & Lewis, 1999; McLeod et al., 2003; Nering, 1997; van Krimpen-Stoop & Meijer, 2001) or by means of their response times (van der Linden & van Krimpen-Stoop, 2003; van der Linden & Guo, 2008). Clearly, these lines of research are useful but, perhaps, problematic in practice. What should a testing agency do with an examinee who probably has item pre-knowledge? As the evidence is only probabilistic, it is hard to believe that any examinee could be failed or be asked to retake the exam with such evidence. Segall (2004) presents an interesting idea: instead of making a decision at the end of the test, it could be possible to adapt the items to be presented, switching to infrequently exposed items when pre-knowledge is suspected.

Our approach is different. Instead of detecting recipients, we try to identify the item selection method that would produce the lowest benefit for recipients. The lower the benefit, the lower the probability of an examinee spending time trying to find a source.

The algorithm we have employed adapts, after each item administered, the next item to be presented. Another option is to reduce the number of times of selection and selecting predefined packages of items, as done with multistage testing (Luecht & Nungester, 1998). With this option, and considering how item packages are currently built (e. g., Belov & Armstrong, 2008; Breithaupt & Hare, 2007), the probability of item pre-knowledge at the beginning of the test is necessarily greater than the probability with the PG method. It remains for future research to study the differential effect of item disclosure for CAT within multistage testing.

References

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness to the Fisher information for improving item exposure control in CATs. *British Journal of Mathematical and Statistical Psychology, 61*, 493-513.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (In press). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement.*

Barrada, J. R., Veldkamp, B. P., & Olea, J. (2009). Multiple maximum exposure rates in computerized adaptive testing. *Applied Psychological Measurement, 33*, 58-73.

Belov, D.I., & Armstrong, R. D. (2008). A Monte Carlo approach to the design, assembly, and evaluation of multistage adaptive tests. *Applied Psychological Measurement, 32*, 119-137.

Birnbaum, A. (1968). Some latent ability models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.) *Statistical theories of mental test scores* (pp. 392-479). Reading, MA: Addison-Wesley.

Bradlow, E. T., Weiss, R. E., & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. *Journal of the American Statistical Association, 93*, 910–919.

Breithaupt, K., & Hare, D. R. (2007). Automated simultaneous assembly of multistage testlets for a high-stakes licensing examination. *Educational and Psychological Measurement, 67*, 5-20.

Chang, H. H.. (2004). Understanding computerized adaptive testing – From Robbins-Monro to Lord and beyond. In David Kaplan (Ed.) *The SAGE handbook of quantitative methodology for the social sciences* (pp. 117-133). Thousand Oaks, CA: Sage Publications.

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

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

Chen, S. Y., 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.

Dodd, B. G. (1990) The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement, 14,* 355-366.

Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8). Retrieved February 17, 2009, from http://escholarship.bc.edu/jtla/vol5/8/.

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.

Li, Y. H., & Schafer, W. D. (2005). Increasing the homogeneity of CAT's item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. *Journal of Educational Measurement, 42,* 245-269.

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

Luecht, R. M., & Nungester R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35,* 229-249.

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23,* 147-160.

McLeod, L. D., Lewis, C, & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement, 27,* 121-137.

Mills, G. N., & Steffen, M. (2000). The GRE computer adaptive test: Operation issues. In W. J. van der Linden and C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75-100). Boston: Kluwer Academic Press.

Nering M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21,* 115-127.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 311-327.

Rulison, K. L., & Loken. E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33,* 83-101.

Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics, 27,* 163–179.

Segall, D. O. (2004). A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 29,* 439-460.

Stocking, M. L., & Lewis, C. L. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice* (pp. 163-182). Dordrecht, The Netherlands: Kluwer Academic.

Stocking, M. L., Ward, W. C., & Potenza, M. T. (1998). Simulating the use of disclosed items in computerized adaptive testing. *Journal of Educational Measurement, 35,* 48-68.

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. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 28,* 249-265.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika, 73,* 365-384.

van der Linden, W.J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika, 68,* 251-265.

van der Linden, W. J., & Veldkamp, B. P. (2007).Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics 32,* 398-418.

van Krimpen-Stoop, E. M. L. A., & Meijer R. R. (2001). CUSUM-based person-fit statistics for adaptive testing. *Journal of Educational and Behavioral Statistics, 26,* 199-217.

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

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