Applied Psychological Measurement
34(6) 438–452

DOI: 10.1177/0146621610370152
http://apm.sagepub.com

A Method for the Comparison of Item Selection Rules in Computerized Adaptive Testing

Juan Ramón Barrada[1], Julio Olea[2], Vicente Ponsoda[2], and Francisco José Abad[2]

Abstract

In a typical study comparing the relative efficiency of two item selection rules in computerized adaptive testing, the common result is that they simultaneously differ in accuracy and security, making it difficult to reach a conclusion on which is the more appropriate rule. This study proposes a strategy to conduct a global comparison of two or more selection rules. A plot showing the performance of each selection rule for several maximum exposure rates is obtained and the whole plot is compared with other rule plots. The strategy was applied in a simulation study with fixed-length CATs for the comparison of six item selection rules: the point Fisher information, Fisher information weighted by likelihood, Kullback-Leibler weighted by likelihood, maximum information stratification with blocking, progressive and proportional methods. Our results show that there is no optimal rule for any overlap value or root mean square error (RMSE). The fact that a rule, for a given level of overlap, has lower RMSE than another does not imply that this pattern holds for another overlap rate. A fair comparison of the rules requires extensive manipulation of the maximum exposure rates. The best methods were the Kullback-Leibler weighted by likelihood, the proportional method, and the maximum information stratification method with blocking.

Keywords

computerized adaptive testing, item exposure control, test security, item selection

There are at least two objectives to be maximized in a computerized adaptive test (CAT): the first is measurement accuracy; the second, item bank security. An item bank is considered more secure if the probability is low that an examinee is aware of the item content before a test is taken.

To evaluate the level of measurement accuracy, the root mean square error (RMSE) is the most commonly used variable, computed according to Equation 1:

[1]Universidad Autónoma de Barcelona, Spain
[2]Universidad Autónoma de Madrid, Spain

Corresponding Author:
Juan Ramón Barrada, Facultad de Psicología, Universidad Autónoma de Barcelona, Barcelona 08193, Spain
Email: juanramon.barrada@uab.es

$$\mathrm{RMSE} = \left(\sum_{g=1}^{r} \left(\hat{\theta}^g - \theta^g \right)^2 \Big/ r \right)^{1/2}, \quad (1)$$

where r is the number of examinees, θ^g is the trait level of the gth examinee and $\hat{\theta}^g$ is the estimated trait level for that examinee.

Improvement in item security is related to measurement accuracy. If an examinee receives an item that is known beforehand, a correct response may be expected. As the probability of a correct response no longer depends on the examinee's trait level and on item parameters, test validity is compromised. The relevance of test security will vary between CATs. In some applications, such as personality measurement (e.g., Reise & Henson, 2000) or patient-reported outcomes (e.g., Cella, Gershon, Lai, & Choi, 2007), examinees are probably motivated to get an assessment that is as accurate as possible, so security becomes a minor issue. In the case of high-stakes CATs, however, some examinees would try to inflate their score if they could (H. H. Chang, 2004; Davey & Nering, 2002).

The overlap rate, which has been used as an index of item bank security, is defined as the proportion of items that are shared, on average, by two randomly selected examinees (Way, 1998). The higher this is, the less secure the bank. Chen, Ankenmann, and Spray (2003) have shown that the overlap rate is linearly related to the variance of the item exposure rates [$S^2_{P(A)}$]. When this variance approaches zero, the overlap rate approaches its minimum, Q/n, where Q is the test length and n is the item bank size. Throughout the paper, the focus will be on fixed-length CATs, so Q is constant for all the examinees. The overlap rate (T) can be calculated with Equation 2 (Chen et al., 2003):

$$T = \frac{n}{Q} S^2_{P(A)} + \frac{Q}{n}. \quad (2)$$

A wide range of item selection rules has been proposed. Some of them are focused, mainly, on increasing accuracy, whereas others seek to reduce security risks. In general, a trade-off between accuracy and security has been found: increases in one variable mean reductions in the other. It is assumed that this tradeoff holds both within and between rules (S. W. Chang & Ansley, 2003; Stocking & Lewis, 2000). For any given rule, it is understood that any manipulation that is effective in reducing the overlap rate (as, for instance, restrictions on the maximum exposure rate—see below), will yield increases in RMSE. For any two different rules, it is expected that the rule with the lower RMSE will necessarily be the one with greater overlap rate.

An example may help to better explain this. Let us imagine that it is desired to compare X and Y item selection rules. For Rule X, the RMSE is .25 and the overlap rate is .18. For Rule Y, the resulting RMSE is .27 and the overlap rate is .12. This would be an example of the abovementioned tradeoff. Results of this kind are usually described with sentences like this: ''The Y rule is the most convenient, as it allows security improvements with a small impact in accuracy.'' From our point of view, in conditions where there is a tradeoff between accuracy and security, it is not possible to draw any conclusion. The X rule could be modified by incorporating, for instance, restrictions on the maximum exposure rate, which may yield an overlap rate equal to that obtained by Y rule. In that case, it is not known if the RMSE found with rule X would be greater, lower, or equal to the RMSE of rule Y. Any conclusion drawn from studies where no item selection rule dominates over the other (it is better in terms of one variable and equal or better in terms of the other) should be considered with caution.

Our aim is to present a method that allows a better comparison between different item selection rules. The article is structured as follows. First, a comparison procedure is shown that allows

Table 1. Relation Between the Indicator Variable v and the Corresponding r^{max}, Overlap Rate, and RMSE

v	$r^{\max}$	Overlap rate	RMSE
1	$r_1 = Q/n$	T_1	RMSE_1
2	r_2	T_2	RMSE_2
. . .	. . .	. . .	. . .
V	$r_V = 1$	T_V	RMSE_V

Note: RMSE = root mean square error.

us to establish, for a given level of accuracy or security, which selection rule is to be preferred. Second, a method for restricting the maximum exposure rate is presented. Third, some of the rules that have been proposed until now are described, and fourth, the proposed method is illustrated with a simulation study where six item selection rules are compared.

A Method for the Comparison of Item Selection Rules

As described above, two item selection rules should share the same value in one of the variables of interest (accuracy or security) for a fair comparison of their efficiency. When this happens, a safer conclusion on their relative quality can be drawn by comparing their performance on the other variable. The probability that two selection rules show, without any additional manipulation, the same value in one of the variables is, of course, very small. This probability is even smaller if the number of rules to be compared is more than two and if all of them should share the same value for one variable.

The use of methods that restrict the maximum exposure rate of the items (r^{max}) has been the most common solution to this problem, as they are effective in reducing the overlap rate of the rule with lower item bank security. An $r^{\max}$ value that can provide a similar overlap rate for both selection rules is obtained and then the measurement accuracy obtained with each of the rules is compared. H. H. Chang and Ying (1999) provide one example of this approach.

This method presents at least two limitations. First, as the $r^{\max}$ values used for the comparison are tentatively established, the overlap rates for all the rules usually show some differences. Second, although for a given common overlap rate the RMSE obtained with one selection rule could be lower than the RMSE of another rule, it should not be taken for granted that this pattern of results will hold for any other security level. It is possible that the item selection rule to be preferred varies according to the security control desired.

An alternative strategy for comparing as many selection rules as needed for the whole range of possible values of accuracy and security can be achieved by manipulating $r^{\max}$. Our proposal is to manipulate the $r^{\max}$ for each item selection rule, in V different values ranging from $r_1^{\max} = Q/n$, which is the minimum possible value for $r^{\max}$, to $r_V^{\max} = 1$, which is equivalent to not applying any restriction on the maximum exposure rate. This idea mimics the method used by Barrada, Olea, and Abad (2008) for the comparison between rotating items banks and the restriction on maximum exposure rates in a master bank.

With this strategy, tables of results are obtained with a structure as shown in Table 1: the RMSE and overlap rate for V different conditions are found, starting with the maximum item exposure control and finishing with no item exposure control. There is one independent variable ($r^{\max}$) and two dependent variables (RMSE and overlap). With these data, it is possible to obtain the curves that relate $r^{\max}$ with RMSE and $r^{\max}$ with the overlap rate. Also, with this information it is possible to plot the graph that relates the overlap rate to the RMSE.

As many tables as item selection rules are generated. In this way, a curve can be plotted for each item selection rule. Thus, it is possible to make the desired comparisons: holding RMSE (or overlap) constant, it is possible to check the differences in overlap (or RMSE). For example, imagine that the X axis represents the overlap rate and the Y axis corresponds to the RMSE. If the curve of an item selection rule is always below the curve for another item selection rule, the former should be preferred, because for any value of one variable it offers better results in the other variable. If two curves cross at some point, this means that the optimal selection rule depends on the degree of security (or accuracy) that is desired, and that no rule is superior to the other over the whole range of possible values.

Following Barrada, Olea, and Abad (2008), the different $r^{\max}$ values are defined by means of Equation 3:

$$r_v^{\max} = \frac{Q}{n} + \frac{\left(1 - Q/n\right) \sum_{f=1}^{v} (f-1)^2}{\sum_{f=1}^{V} (f-1)^2}, \tag{3}$$

where f is a dummy variable used only for calculations and v is used for defining the position in the V different $r^{\max}$ values ($r_1^{\max}$ the minimum and $r_V^{\max}$ the maximum).

This equation leads to unequally spaced values of $r_v^{\max}$, a characteristic that is desirable, given the usual form of the curves relating overlap rate to RMSE (Barrada, Olea, & Abad, 2008). An important aspect of this comparison method relies on the control of maximum exposure rates.

Restriction of Maximum Exposure Rate

The most common approach to improving item bank security is to reduce $r^{\max}$ (van der Linden, 2003). The methods aimed at restricting maximum exposure rate eliminate the problem of item overexposure and reduce the overlap rate, although this is typically accompanied by an increase in the RMSE. Each method establishes the probability that an item will be eligible, $P(E_i)$, for administration. This probability will be lower for items with higher exposure rates (when no restriction on $r^{\max}$ is applied) than for underexposed items. The methods differ in how the $P(E_i)$ values are computed and in their range. When the Sympson-Hetter method (Sympson & Hetter, 1985), the restricted method (Revuelta & Ponsoda, 1998), and the item-eligibility method (van der Linden & Veldkamp, 2004) are compared, the proposal put forward by van der Linden and Veldkamp is the one that seems to be preferable (Barrada, Abad, & Veldkamp, 2009).

In the item-eligibility method, the calculation of the $P(E_i)$ parameters for the $(m+1)$th examinee is made according to Equation 4:

$$P^{(m+1)}(E_i) = \begin{cases} 1 & \text{if} \quad P^{(1..m)}(A_i)/P^{(m)}(E_i) \leq r^{\max} \\ r^{\max} P^{(m)}(E_i)/P^{(1..m)}(A_i) & \text{if} \quad P^{(1..m)}(A_i)/P^{(m)}(E_i) > r^{\max} \end{cases}, \tag{4}$$

where $P^{(1..m)}(A_i)$ is the probability of the administration (exposure rate) of the ith item computed from the responses from the first to the mth examinee. Further details about this method can be found in van der Linden and Veldkamp (2004, 2007).

A general framework has been presented that allows us to compare different item selection rules. The next step is to present some of these rules, that will be compared by means of the proposed method in a simulation study.

Item Selection Rules

Our choice of item selection rules is based on their relevance to our questions, as well as the amount of research available on their performance. The descriptions will not be exhaustive (more information can be obtained from the references).

Point Fisher Information (PFI)

The item selection rule most commonly used in CATs for the selection of the qth item, q being the indicator of the item's position on the test, is the selection of the item with maximum Fisher information for the estimated trait level (PFI rule; Lord, 1980). The selected item j is given by

$$j = \arg\max_{i \in B_q} I_i(\hat{\theta}) \tag{5}$$

where $I_i(\hat{\theta})$ is the Fisher information of item i for the estimated trait level and B_q is the subset of items belonging to the item bank that can be presented to the examinee in the qth position in the test. If no restriction is active, B_q consists of those items not presented to that examinee in the $q-1$ previous items. If the item-eligibility method is used, B_q is the intersection between the nonpresented items and those items marked as eligible.

Fisher Information Weighted by Likelihood (FI-L)

The PFI rule has several limitations. On one hand, it does not take into account the values of the information function for trait levels different from the estimated trait level. On the other, the likelihood function, $L(\theta)$, is merely used to locate its maximum, its shape playing no role at all, which can vary from being mainly flat, as at the beginning of the test, to more peaked, as the test goes on. In addition, it does not take into account the possibility of various local maxima in the likelihood function (Samejima, 1977). Veerkamp and Berger (1997) proposed a more exhaustive use of both functions with the item selection rule called Fisher information weighted by likelihood (FI-L), described in Equation 6:

$$j = \arg\max_{i \in B_q} \int_{-\infty}^{\infty} I_i(\theta) L(\theta) d(\theta). \tag{6}$$

The entire trait-level range affects the FI-L rule. This allows for greater accuracy of this rule when compared with PFI (Veerkamp & Berger, 1997), especially for low trait levels (Chen, Ankenmann, & Chang, 2000), although this is achieved with an increment in the overlap rate (Chen & Ankenmann, 2004).

Kullback-Leibler Function Weighted by Likelihood (KL-L)

The Kullback-Leibler (KL) information function evaluates the item discrimination capacity between any possible pairs of trait levels. This means that KL is a global information measure (H. H. Chang & Ying, 1996). H. H. Chang and Ying proposed to weight the KL measure with the posterior trait level distribution. As maximum-likelihood estimation is used in this study, likelihood is used to weight KL:

$$j = \arg\max_{i \in B_q} \int_{-\infty}^{\infty} KL_i\left(\theta||\hat{\theta}\right) L(\theta) d(\theta), \tag{7}$$

where $KL_i(\theta||\hat{\theta})$ is calculated as follows:

$$KL_i\left(\theta||\hat{\theta}\right) = P_i(\hat{\theta}) \ln\left[\frac{P_i(\hat{\theta})}{P_i(\theta)}\right] + \left[1 - P_i(\hat{\theta})\right] \ln\left[\frac{1 - P_i(\hat{\theta})}{1 - P_i(\theta)}\right]. \tag{8}$$

When compared with the PFI, KL-L offers a lower RMSE (Chen et al., 2000), although with a greater overlap rate (Chen & Ankenmann, 2004). When compared with the FI-L, KL-L reduces the RMSE and increases the overlap rate (Barrada, Olea, Ponsoda, & Abad, 2009).

Maximum Information Stratification Method With Blocking (MIS-B)

The logic of stratified methods, first proposed by H. H. Chang and Ying (1999), is to administer low-informative items at the beginning of the test and to increase the administration of more highly informative items as the test goes on. In all the stratified methods, those items belonging to B_q are determined according to their position in the test length. The formulation proposed by Barrada, Mazuela, and Olea (2006) is followed, as it outperforms the original in both security and accuracy.

In the three-parameter logistic model, the maximum Fisher information of item i ($I_i^{\max}$) is equal to (Hambleton & Swaminathan, 1985)

$$I_i^{\max} = \frac{1.7^2 a_i^2}{8(1 - c_i^2)}\left[1 - 20c_i - 8c_i^2 + (1 + 8c_i)^{3/2}\right], \tag{9}$$

where a_i is the discrimination parameter and c_i is the pseudoguessing parameter for item i.

The trait level at which this maximum information is achieved (θ_i^{max}) can be calculated according to Equation 10 (Hambleton & Swaminathan, 1985):

$$\theta_i^{\max} = b_i + \frac{\ln\left[1 + (1 + 8c_i)^{1/2}\right] - \ln(2)}{1.7a_i}, \tag{10}$$

where b_i is the location parameter of item i.

In the MIS-B method, prior to any item being administered, the item bank is stratified. First, the n items of the bank are placed in increasing order according to their θ_i^{max} values. The first S items (S being the number of strata into which the item bank will be divided) are rearranged, placing them in an ascending order according to their $I_i^{\max}$ value. The first item of this item set S will be assigned to the first stratum, the second to the second, and the Sth item to the Sth stratum. This process is repeated for the n/S blocks of size S that can be obtained.

In a CAT of length Q, during the first Q/S items of the test administration, the B_q item set will be formed by the n/S items of the first stratum, the n/S items of the second stratum will compose B_q for the next Q/S items of the test, and so on. As the test goes on, the mean $I_i^{\max}$ of the items belonging to B_q increases, leaving the items with high a parameters and low c parameters for use at the end of the test. Stratifying by blocking θ_i^{max} makes the distribution of θ_i^{max} as similar as possible between strata. Once the B_q set is defined for each item, the selection will be made according to Equation 11:

$$j = \arg\min_{i \in B_q} \left| \hat{\theta} - \theta_i^{\max} \right|. \tag{11}$$

The stratified methods, compared with PFI, improve the security of the item bank, leading to an overlap rate near the minimum possible overlap rate (e.g., H. H. Chang, Qian, & Ying, 2001; Cheng, Chang, & Yi, 2007) while decreasing accuracy (H. H. Chang & Ying, 1999).

Progressive Method (PG)

Revuelta and Ponsoda (1998) proposed the progressive method (PG). This method selects the item for which the sum of a random component and the Fisher information is highest. At the beginning of the test, when the trait estimation error is high, the weight of the random component is most important. The weight of the Fisher information (W_q) increases as the number of items administered increases. The PG method can be described as follows:

$$j = \arg\max_{i \in B_q} \left[(1 - W_q) R_i + W_q I_i(\hat{\theta}) \right], \tag{12}$$

where the weight W_q is the contribution of item information to 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, Olea, Ponsoda, and Abad (2008) have proposed the following equation to relate W_q to q:

$$W_q = \begin{cases} 0 & if \quad q = 1 \\ \dfrac{\sum_{f=1}^{q} (f-1)^t}{\sum_{f=1}^{Q} (f-1)^t} & if \quad q \neq 1 \end{cases}. \tag{13}$$

The t parameter marks the speed at which the weight of the random component is reduced and, thus, the speed at which the importance of item information increases. This parameter defines the improvement in bank security and the reduction in accuracy, in comparison with PFI. With a t equal to 1, marked improvements in security are obtained with hardly any impact on accuracy.

Proportional Method (PP)

All the methods presented so far are deterministic, in the sense that the item to be selected maximizes (or minimizes, in the case of the stratified methods) the selection function. Segall (2004) has proposed a stochastic method, where the selection function value is not used to order the items but is used to select the first item and to calculate the probability of selecting each item. Hence, no item would have a probability of being administered equal to 0. In this method, which will be called proportional (PP; Barrada, Olea, Ponsoda, & Abad, 2008; Segall, 2004), the probability of selecting the items is given by Equation 14:

$$P(S_i) = \frac{z_i I_i(\hat{\theta})^{H_q}}{\sum_{i=1}^{n} z_i I_i(\hat{\theta})^{H_q}}, \tag{14}$$

where z_i indicates whether the item belongs (1) or not (0) to B_q. Once the probabilities of each item being selected are obtained, a cumulative distribution of probabilities is formed. Then,

a random number drawn from the uniform interval (0, 1) is used to identify the item to be selected.

When H_q is 0, selection is random. The higher the H_q, the higher is the probability of selecting the item with maximum Fisher information, making the selection by the PP and PFI methods more similar.

Barrada, Olea, Ponsoda, and Abad (2008) have proposed defining H_q according to Equation 15, which as can be seen, clearly resembles the equation that defines the values of W_q for the PG method:

$$H_q = \begin{cases} 0 & \text{if} \quad q = 1 \\ \dfrac{Q\sum_{f=1}^{q}(f-1)^s}{\sum_{f=1}^{Q}(f-1)^s} & \text{if} \quad q \neq 1 \end{cases}. \tag{15}$$

According to this function, the test starts with random selection, and for common CAT lengths, the selection of the last item will be similar to selection with PFI. The s parameter has the same role as the t parameter in the PG rule: It defines the speed with which the method reduces random selection of items.

The six rules described are, in our view, an acceptable sample of the available selection rules for CAT: PFI is the current standard in item selection, whereas FI-L and KL-L illustrate the pole of maximum accuracy with minimum security; MIS-B is at the opposite extreme, having a low overlap rate with accompanying increases in RMSE; finally, PG and PP are methods that obtain satisfactory results in accuracy, comparable to those obtained with PFI, with improvements in security. More importantly, these six item selection rules will allow us to show the use of our proposed comparison procedure.

Simulation Study

Method

Wingersky and Lord (1984) and H. H. Chang et al. (2001) have pointed out that in practice, the a and b parameters of the items are usually correlated. The performance of different item selection rules can vary depending on whether the item banks used have correlated parameters or not (Barrada, Olea, et al., 2009). Thus, two kinds of item banks were generated, one with uncorrelated a and b parameters and the other with correlated parameters ($r_{ab} = .5$). A total of 10 banks of 500 items each were obtained for the correlated and uncorrelated item bank types. The parameter distributions were: $a \sim N(1.2, .25)$, $b \sim N(0, 1)$ and $c \sim N(.25, .02)$. For each item bank, 5,000 examinees were generated randomly, with trait levels extracted from a distribution $N(0, 1)$. Two different test lengths, 20 and 40 items, were used. The initial trait level, $\hat{\theta}_0$, was selected randomly from the uniform interval (–.5, .5). Dodd's (1990) procedure was applied for the trait-level estimation until each examinee obtained correct and incorrect responses: 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 $(\hat{\theta} - b_{\min})/2$. Once the constant pattern was broken or the test was finished, maximum-likelihood estimation was applied, with the restriction that $\hat{\theta}$ had to be in the interval [–4, 4].

The likelihood function cannot be computed when no item has been administered. Because of this, for the selection of only the first item with the FI-L and KL-L rules, two fictitious items were added to the response vector to obtain the likelihood function, one correct and one incorrect, both with the same parameters: $a = .5$, $b = \hat{\theta}_0$ and $c = 0$ (Barrada, Olea, et al., 2009).

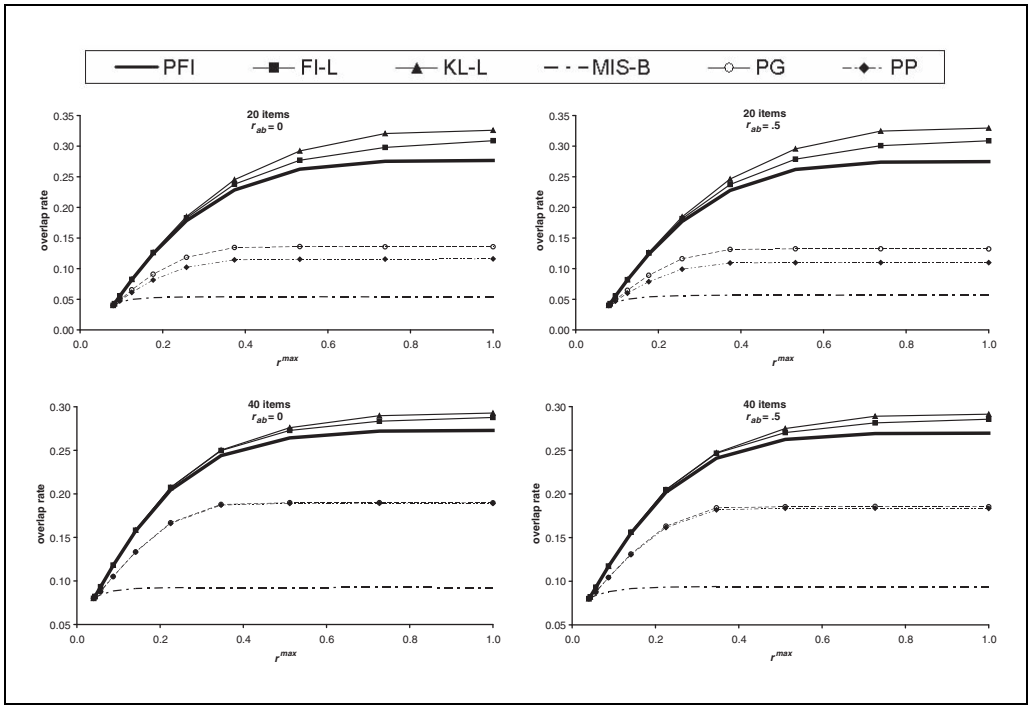

Figure 1. Relation between r^{max} and overlap rate

In the MIS-B rule, the item bank was stratified into five strata, all of them having the same number of items. The number of items extracted from each stratum was held constant. In the PG method, the *t* parameter was set equal to 1. The same value was given for the *s* parameter in the PP method.

For the restriction of the maximum exposure rates, the item-eligibility method was used (van der Linden & Veldkamp, 2004).

The variables for comparing the different methods were the RMSE and the overlap rate, as defined in Equations 1 and 2. The strategy of simulating several different r^{max} values for each item selection rule was applied. Barrada, Olea, and Abad (2008) showed that 10 values for r^{max} are enough to a plot the desired graphs correctly, so *V* (Equation 3) was fixed at 10.

Results

The relation between the overlap rate and r^{max} can be seen in Figure 1. Results for each dot in the plot were based on 50,000 examinees (10 banks $\times$ 5,000 simulees). The results maintain the same pattern independently of the number of items administered or the correlation between parameters. When no restriction on maximum rate is applied (i.e., r^{max} equal to 1), the expected results were found: The KL-L rule produced the highest overlap rates, followed by FI-L and, after these, PFI. Higher security is achieved with the PG and PP rules. When the test length is 20 items, PP is more secure than PG; however, with 40-item tests both rules offer the same overlap rate. The rule that offers the highest security level is MIS-B, as its overlap rate, even when r^{max} is 1, is very close to the minimum possible value.

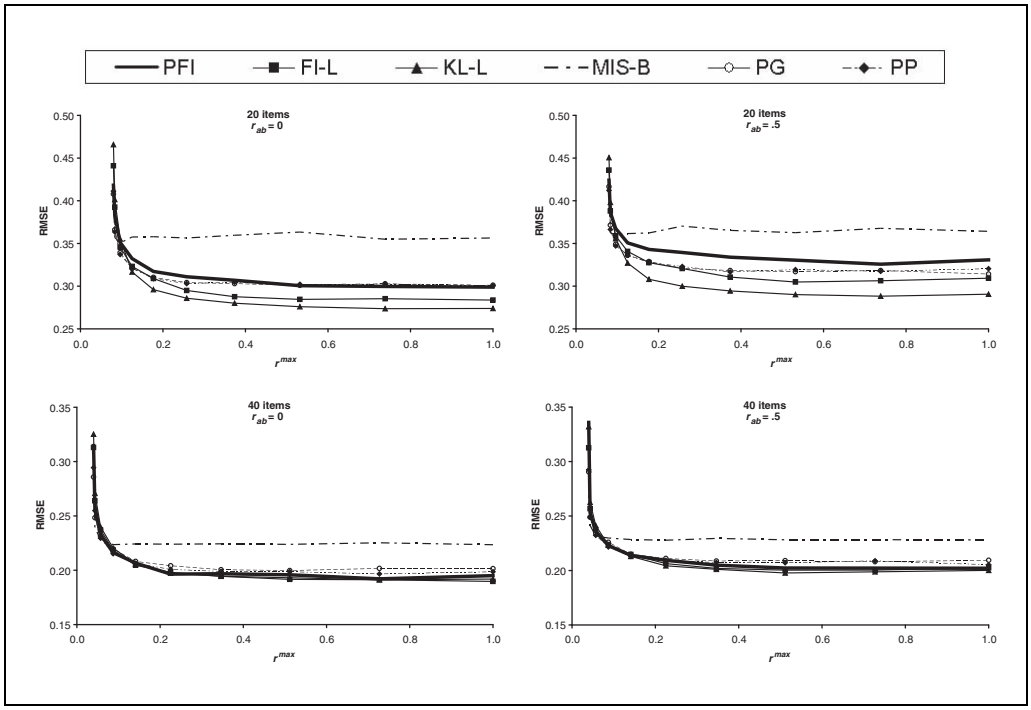

Figure 2. Relation between r^{max} and root mean square error (RMSE)

The effect of restricting r^{max} is not the same for all the selection rules. Although for the rules with greater overlap, when r^{max} is 1, small changes in r^{max} reduce overlap; for the rules with better exposure control, a greater reduction in r^{max} is needed to improve security. In other words, the rules with greater overexposure problems are more sensitive to changes in r^{max}.

The effects of reducing r^{max} on the overlap rate are more evident with low r^{max}: The effect of changing r^{max} from 1 to .9 is smaller than the effect of changing from .2 to .1. In any case, when the minimum possible r^{max} is imposed, minimum overlap is obtained.

Figure 2 shows the relation between r^{max} and RMSE. As expected, increasing test length improves accuracy. The correlation among parameters increases RMSE, as in these banks the information available around the average trait levels is lower than in banks with uncorrelated parameters. Increasing test length reduces the differences in accuracy between rules.

When there is no restriction on r^{max}, the selection rule that offers greater accuracy is KL-L, followed by FI-L, although in 40-item tests the difference is negligible. The rule with greater measurement error is MIS-B. For a length of 20 items and correlated parameters, PFI obtains higher RMSE than PG and PP, whereas for the rest of the conditions, PFI is more accurate.

The different selection rules allow important restrictions on r^{max} without translating this into increments in RMSE. Fixing, for instance, r^{max} to .3 does not imply any noticeable losses in accuracy. As r^{max} approaches its minimum possible value, the speed with which RMSE increases is accelerated.

Figure 3 depicts the relation between the overlap rate and RMSE. These plots are, in our view, the most relevant for deciding which selection rule to choose for a CAT. As expected, there is a within-rule trade-off between accuracy and security. However, important improvements in security can be achieved with negligible reductions in accuracy.

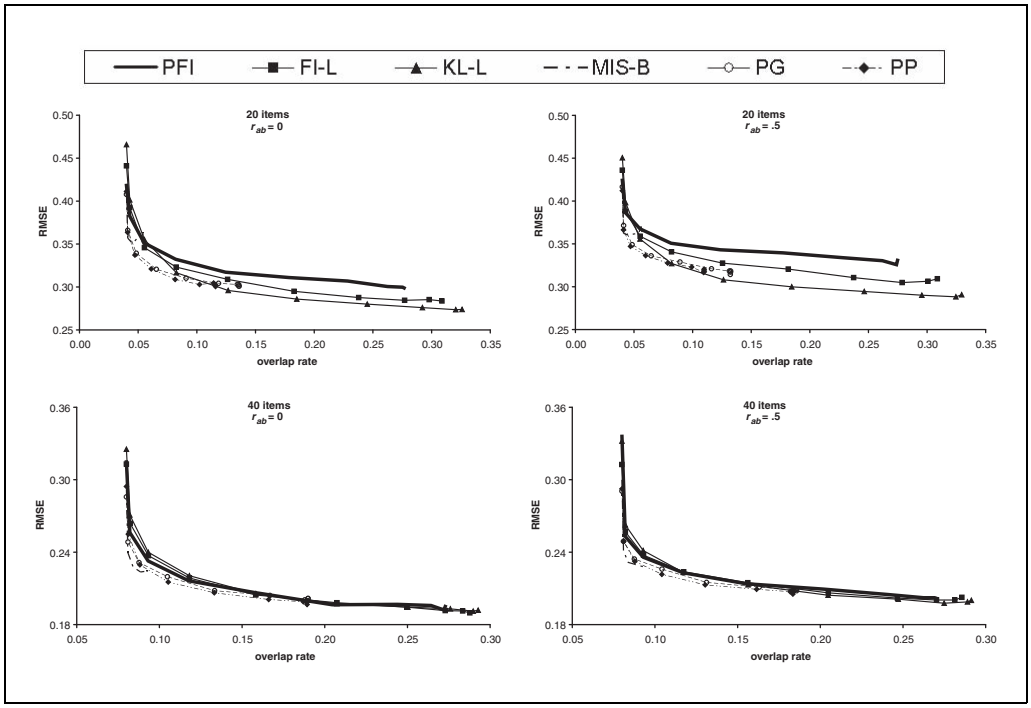

Figure 3. Relation between overlap rate and root mean square error (RMSE)

The main comparison is between rules. The rule to be selected will depend on the required accuracy level and on the acceptable security risk. Let us consider first the results for a test length of 20 items. If our CAT tolerates overlap rates greater than .11, in the case of uncorrelated banks, or more than .08 when r_{ab} is equal to .5, the rule to be used is KL-L, as for these overlap rates it provides the highest accuracy levels. If a more strict control of bank security is needed, the most convenient rule is PP. Looking at the same data from the perspective of accuracy, if RMSE values between .27 (minimum possible value) and .30 are desired, for the condition of uncorrelated banks, or between .28 and .32 in the correlated case, the rule to use is KL-L. If a reduction in accuracy in order to increase security seems appropriate, the best alternative is PP. None of the other four rules tested would be selected for any level of security or accuracy.

The differences between selection rules when the test length is 40 items are much smaller. When maximizing accuracy is the main objective or when high overlap rates can be tolerated, the most convenient rules are KL-L, FI-L, and PFI. When a greater exposure control is desired and only a small increment in measurement error can be tolerated, the rule to be used is PP. If we want the highest possible item bank security, given these conditions of test length and bank size, the rule to choose is MIS-B.

Discussion and Conclusions

Several item selection rules are available for CAT. Some of them, like PFI, FI-L, and KL-L, focus on measurement accuracy, whereas others, like MIS-B, PG, and PP, are focused on item bank security. The comparison of the results provided by different selection rules is not an easy task, as rules offering better accuracy are usually lower on security indicators. The

proposed strategy enables an improved comparison of item selection rules, as rules can be compared in one indicator (accuracy or security) while holding the other constant. Two main features of the proposed strategy are, first, that it can be easily applied to more than two selection rules and, second, that it compares the rule's global performance rather than just its efficiency for a particular pair of accuracy–security values.

The strategy was applied for the comparison of six selection rules and provided these main results:

1. The item selection rule most commonly used, PFI, is never the best alternative. At most, its use can be recommended for a test of 40 items with uncorrelated parameters when poor exposure control can be tolerated or test security is a minor issue, although in this case its performance is equivalent to that provided by KL-L and FI-L. A possible reason to prefer PFI in these conditions could be its lower computational complexity, although it is considered that, with modern computers, differences will be negligible in terms of CPU time.
2. The FI-L rule seems, also, to be outperformed almost continuously by KL-L. The PP rule is always a slightly better alternative than PG. So it seems that three of the six rules (PFI, FI-L, and PG) could be discarded.
3. There is no optimal rule for any value of overlap or RMSE. The fact that a rule, for a given level of overlap, has a lower RMSE than another does not imply that this pattern has to hold for another overlap rate. So a fair comparison of the rules requires an extensive manipulation of the maximum exposure rates, as does the proposed strategy, in order to obtain more than one pair of accuracy–security indicators and to enable the comparison of the global efficiency curve of one selection rule against the others. Studies lacking such extensive manipulation should be considered with caution.
4. The point at which PP becomes preferable to KL-L depends on the kind of item bank used and on the test length. The MIS-B rule is a viable option only for a test length of 40 items.

Some limitations of the proposed method should be noted. Although overlap rate has been used as the only variable for assessing test security, several other variables have been proposed, namely χ^2 (H. H. Chang & Ying, 1999) and the number of items unused from the bank and maximum exposure rate. In plots such as the one shown in Figure 3, the overlap rate is held constant, but the other indicators of security could differ between rules; thus security conditions that are, in fact, unequal could be taken as equal. For performing the comparisons, a single measure of security is needed. In this way, readable plots can be drawn. This study has offered an imperfect but interpretable solution. Also, it is believed that the other variables for measuring test security are redundant or more limited.

1. χ^2 is as follows (H. H. Chang & Ying, 1999):

$$\chi^2 = \frac{\sum_{i=1}^{n} \left[P(A_i) - Q/n\right]^2}{Q/n}. \tag{16}$$

It is a measure of departure from uniform usage of items. With some substitutions, it can be shown that χ^2 is equal to $(nT - Q)$. For item banks of the same size and tests of the same length, χ^2 is a linear transformation of T, so the ordering of rules will not change.

2. The overlap rate takes into account the whole distribution of item exposure rates to produce a single number, which is easy to interpret. Maximum exposure rate and number of items unused are restricted to just one of the extremes of this distribution.

Because of this, it is believed that the overlap rate is the measure to be used in these conditions, when only one value is desired. Along this line, Yi, Zhang, and Chang (2008) have shown that holding constant the maximum exposure rate, the item selection rule with a lower overlap rate, the alpha-stratified method, could better resist an environment of item bank disclosure when compared with PFI, thus indicating that the overlap rate is a valid measure of test security.

The results of this study provide general guidance on choosing an item selection rule when fixed-length CATs are used. The generalization of this design to the case of variable-length CATs would not be direct. The problem is that it is unclear how to interpret the overlap rates for variable-length CATs. Is terms of security, is an overlap rate of .25 the same on a test with mean length of 16 as on a test with mean length of 24? With variable-length CATs, it is unclear that sharing the value on overlap can be interpreted as sharing item bank security. As long as this point is not solved, the proposed method should probably be restricted to use with fixed-length CATs.

The simulation conditions do not exhaust all the relevant variables. For instance, different item bank sizes or other parameter distributions have not been considered. So for the final decision on an optimal rule for a specific item bank and goals, our advice is to conduct an ad hoc simulation study including, at least, the KL-L, PP, and MIS-B rules. Having decided beforehand the desired level of accuracy or security, a plot such as that shown in Figure 3 would help decide which selection rule better fits our needs. Plots like those shown in Figures 1 and 2 would help establish the $r^{\max}$ value needed to obtain the target values.

Declaration of Conflicting Interests

The authors declared no conflicts of interests with respect to the authorship and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research and/or authorship of this article:

Two grants from the Spanish Ministry of Science and Innovation (project numbers PSI2009-10341 and PSI2008-01685).

References

Barrada, J. R., Abad, F. J., & Veldkamp, B. P. (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema, 21*, 313-320.

Barrada, J. R., Mazuela, P., & Olea, J. (2006). Maximum information stratification method for controlling item exposure in computerized adaptive testing. *Psicothema, 18*, 156-159.

Barrada, J. R., Olea, J., & Abad, F. J. (2008). Rotating item banks versus restriction of maximum exposure rates in computerized adaptive testing. *The Spanish Journal of Psychology, 11*, 618-625.

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. (2009). Item selection rules in computerized adaptive testing: Accuracy and security. *Methodology, 5*, 7-17.

Cella, D., Gershon, R., Lai, J., & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research, 16*, 133-141.

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

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

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

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

Chang, S. W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40*, 71-103.

Chen, S. Y., & Ankenmann, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement, 41*, 149-174.

Chen, S. Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*, 241-255.

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.

Cheng, Y., Chang, H. H., & Yi, Q. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement, 31*, 467-482.

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.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Hingham, MA: Kluwer.

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

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO-PI–R. *Assessment, 7*, 347-364.

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

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 233-247.

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, Netherlands: Kluwer Academic.

Sympson, J. B., & Hetter, R. D. (1985). 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., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics, 29*, 273-291.

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.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203-226.

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

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement, 8*, 347-364.

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.