Item Selection Rules in Computerized Adaptive Testing

Accuracy and Security

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

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

Abstract. The item selection rule (ISR) most commonly used in computerized adaptive testing (CAT) is to select the item with maximum Fisher information for the current trait estimation (PFI). Several alternative ISRs have been proposed. Among them, Fisher information considered in an interval (FI*I), Fisher information weighted with the likelihood function (FI*L), Kullback-Leibler information considered in an interval (KL*I) and Kullback-Leibler weighted with the likelihood function (KL*L) have shown a greater precision of trait estimation at the early stages of CAT. A new ISR is proposed, Fisher information by interval with geometric mean (FI*IG), which tries to rectify some detected problems in FI*I. We evaluate accuracy and item bank security for these six ISRs. FI*IG is the only ISR which simultaneously outperforms PFI in both variables. For the other ISRs, there seems to be a trade-off between accuracy and security, PFI being the one with worse accuracy and greater security, and the ISRs using the likelihood function the reverse.

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

One of the goals when applying a test is the accurate and efficient evaluation of the trait level (θ) of the examinees. One of the means developed for this (Lord, 1971; Owen, 1975) is computerized adaptive testing (CAT). In contrast to the case of paper and pencil tests, in CATs not all the examinees receive identical items. The item (q + 1)th presented varies according to the pattern of responses to the q items already administered. In general, a CAT can be described (Olea & Ponsoda, 2003; van der Linden & Glas, 2000a) as an iterative process of [estimation of the trait level ($\hat{\theta}$) – selection of the optimal item for $\hat{\theta}$ – response to the item], until a previously established criterion (e.g., measurement error below a certain level or application of a specific number of items) is met.

The probability of a correct response to an item is determined by the trait level of the examinee and by the parameters that characterize the item. In the three-parameter logistic model (3PLM), this probability is calculated using:

$$P(\theta) = c + \frac{1-c}{1+\mathrm{e}^{-1.7a(\theta-b)}}, \tag{1}$$

where

a is the discrimination parameter;
b is the location parameter; and
c is the pseudo-guessing parameter.

For the other two steps of the cycle, estimation of the trait level and item selection, several alternatives have been proposed. The estimation of θ is carried out from the responses to the items already administered. The majority of statistical estimation procedures seek the trait level where the probability of the pattern of responses is maximum. The Bayesian methods (Bock & Mislevy, 1982; Owen, 1975; Samejima, 1969) include suppositions about the probability distribution of the examinee population, while the maximum-likelihood method (Birnbaum, 1968) does not. For a comparison between them, Wang and Vispoel (1998) can be consulted. The likelihood function, necessary for all of them, is obtained as indicated in:

$$L(\theta,x,g) = \prod_{i=1}^{n}\left[P_i^{g_i}(\theta)\left(1-P_i^{1-g_i}(\theta)\right)\right]^{x_i}, \tag{2}$$

where

n is the item bank size;
x_i is the indicator of presentation (1)/nonpresentation (0) of items; and
g_i is the indicator of correct response (1)/noncorrect response (0) of items.

Several item selection rules (ISRs) have been also proposed. Among them, the most common consists in selecting the nonpresented item with maximum Fisher information (PFI) for $\hat{\theta}$ (Lord, 1977). For 3PLM, the Fisher information function (FIF) can be calculated with:

$$I(\theta) = \frac{2.89a^2(1-c)}{\left(c+\mathrm{e}^{1.7a(\theta-b)}\right)\left(1+\mathrm{e}^{-1.7a(\theta-b)}\right)^2}. \tag{3}$$

As the value of FIF for $\hat{\theta}$ increases, the measurement error of the estimation decreases. Asymptotically, as the number of items presented is increased, the measurement error of θ is $I_q(\theta)^{1/2}$ (Bradley & Gart, 1962). I_q is the information

DOI: 10.1027/1614-2241.5.1.7

accumulated with the q items administered. This calculation, as shown in equation (4), is the sum of the information given by each one of the presented items.

$$I_q(\theta) = \sum_{j=1}^{n} x_j I_j(\theta). \tag{4}$$

Hence, PFI seems to be the optimum rule for reducing the variance of the estimator. This ISR, however, presents some limitations in two of the goals to be optimized in CATs: the accuracy of trait level estimation and the security of the item bank. Throughout this study, we shall understand as a secure item bank that one in which examinees have low probability of knowledge, before being tested, of any of the items they will be required to answer. We shall present the limitations of PFI for each of these goals, together with the alternatives that have been proposed for the reduction of these shortcomings.

Problems Relative to Accuracy

The problems associated with accuracy have at least three sources:

1. The selection of items based on PFI (hereafter referred to simply as PFI) will be much more inadequate the larger it becomes $|\hat{\theta}_q - \theta|$. It has been established that $E\left(|\hat{\theta}_{q+1} - \theta|\right) < E\left(|\hat{\theta}_q - \theta|\right)$, in such a manner that the incidence of this problem will decrease as the number of items presented is increased (Lord, 1983).
2. It is possible to find several local maxima for the likelihood function with items calibrated according to the three-parameter model (Samejima, 1977). However, PFI only assesses the FIF for a precise value, $\hat{\theta}$. This means excluding the possible multiplicity of maxima from the selection algorithm. As was the case with the previous problem, the possible incidence of this one decreases as the number of items administered increases.
3. The Fisher information can be described as the capacity of an item to discriminate between two adjacent points of trait level. However, for an effective estimation of θ, it is better to discriminate not only between close trait levels, but also between distant levels, especially in the initial stages of test administration (Chang & Ying, 1996).

With the aim of overcoming these inconveniences, other ISRs have been proposed, whose theoretical justification we shall proceed to describe.

ISRs Alternative to PFI Focused on the Improvement of Accuracy

The different ISRs can be described as particular cases of a general rule (Veerkamp & Berger, 1997):

$$\max_{i \in B_n} \int_{\theta_{\min}}^{\theta_{\max}} V_i(\theta) W(\theta, x, g) d(\theta). \tag{5}$$

The $(q + 1)$th selected item will be that which, belonging to the set of nonadministered items (B_n), offers the maximum value for the criterion integral. W is the weighting function, which is conditional on the vector of items previously presented (x), to the correct or noncorrect response to them (g), and to the θ value. $V(\theta)$ is the valuating function. Thus, for example, in the PFI case:

$$W(\theta, x, g) = \begin{cases} 1, & \theta = \hat{\theta} \\ 0, & \theta \neq \hat{\theta} \end{cases}, \tag{6}$$

$$V(\theta) = I(\theta). \tag{7}$$

PFI evaluates FIF for a single point, $\hat{\theta}$. However, in the initial items of a CAT, it would seem appropriate to consider the information of trait levels relatively distant from $\hat{\theta}$. Thus, the alternative rules will obtain the information for an interval of values of θ, with different valuating functions (FIF and Kullback-Leibler (KL) function) and different weighting criteria (likelihood function and interval).

This gives rise to four alternative ISRs, which will be assessed in the course of this work, though they are not the only ones proposed (e.g., van der Linden, 1998, presents other ISRs that offer promising results).

ISRs Based on the FIF

These ISRs employ as a basis the same valuating function as is used with PFI. The difference is to be found in the weighting function, which allows the criterion integral to take into account more than just $I(\hat{\theta})$. Two criteria have been proposed for W. The first is the likelihood function, which would give rise to FIF by likelihood – FI*L (Veerkamp & Berger, 1997). The second proposal will be called FIF by interval – FI*I (Veerkamp & Berger, 1997). The interval is the confidence interval of θ given $I_q(\hat{\theta})$.

ISRs Based on the KL Function

The valuating function adopted for these ISRs is the KL function, which is intended to solve the third problem previously described, since it provides knowledge of the capacity for discriminating between any two trait levels.

$$\mathrm{KL}(\theta || \theta_s) = P(\theta_s) \ln\left(\frac{P(\theta_s)}{P(\theta)}\right) + (1 - P(\theta_s)) \ln\left(\frac{1 - P(\theta_s)}{1 - P(\theta)}\right), \tag{8}$$

where θ_s indicates that θ needs to be separated from θ_s.

For a more detailed description of the KL function applied in CATs, see Chang and Ying (1996) and Eggen (1999).

The weighting criteria suggested are the same as those for the ISRs based on FIF, which gives rise to two new ISRs: KL function by likelihood function – KL*L and KL function by interval – KL*I (Chang & Ying, 1996).

Of all the studies carried out to date on this topic (Barrada, Olea, & Ponsoda, 2004; Barrada, Olea, Ponsoda, & Abad, 2006; Chang & Ying, 1996; Chen & Ankenmann, 2004; Chen, Ankenmann, & Chang, 2000; Cheng & Liou, 2000; Veerkamp & Berger, 1997), only those of Barrada et al. (2004) and Chen et al. (2000) include all of these ISRs. These authors find similar results to those partially reported in previous works: (a) slight advantages of the alternative ISRs, mainly of those using the likelihood function, in comparison to PFI, which decrease as the number of items presented increases and (b) the improvements tend to be located mainly at low θ levels.

Problems Relative to Exposure Control

The second limitation of item selection by PFI concerns item exposure control. With this ISR, few items are presented to a high proportion of examinees, while a substantial part of the items bank is hardly used. In fact, the bank in use would be clearly smaller than the real bank size, and this has two disadvantages. On the one hand, it can lead to a security problem with the overexposed items, whose content may be known before their administration, thus making them a source of measurement error. On the other hand, the infra-exposed items means wasted funds, as the investment involved in their development is not recovered.

Moreover, PFI poses the problem of a high overlap rate between examinees (Way, 1998). The overlap rate is the proportion of items shared by two randomly selected examinees (Chang & Zhang, 2002). The higher its value, the more vulnerable the item bank, since an examinee has greater probability of being informed about items that will be administered to him/her (Chang, 2004). Control of exposure and overlap are related variables, since improvements in one imply improvements in the other (Chen, Ankenmann, & Spray, 2003).

ISRs Alternative to PFI Focused on Security of the Item Bank

In the last 20 years, several ISRs aimed at improving exposure control have been proposed (Barrada, Olea, Ponsoda, & Abad, 2008; Barrada, Veldkamp, & Olea, in press; Chang & Ansley, 2003; Chang & Ying, 1999; Chen & Lei, 2005; Cheng & Liou, 2003; Davey & Fan, 2000; Davey & Parshall, 1995; Li & Schafer, 2005; McBride & Martin, 1983; Revuelta & Ponsoda, 1998; Stocking & Lewis, 1998; Sympson & Hetter, 1985; van der Linden & Veldkamp, 2004), each one with its own logic and limitations. Given the goals of this study, a fuller description of them is not necessary here. For more detail, the reader can consult Georgiadou, Triantafillou, and Economides (2007).

The results found to date with the different alternative ISRs focused on improving item bank security have led some authors to assume as true the following relation: Improvements in the accuracy of measurement imply a deterioration of security, and vice versa, as explicitly stated by Chang and Ansley (2003) or Stocking and Lewis (2000). It falls to test administrators, therefore, to choose a point between these inversely related variables. At the moment there is no clear criterion for guiding this decision.

Security of the Bank and Rules Alternative to PFI Focused on Improving Accuracy

The item selection logic of the ISRs alternative to PFI allows us to suppose that they might present, simultaneously, improvements in both accuracy and security. For an item bank calibrated according to 3PLM, it is possible to find items with low exposure rate, when selecting with PFI, with suitable performance in the criterion integral when broad θ areas are considered. We can see this with an example in Figure 1 for FI*I.

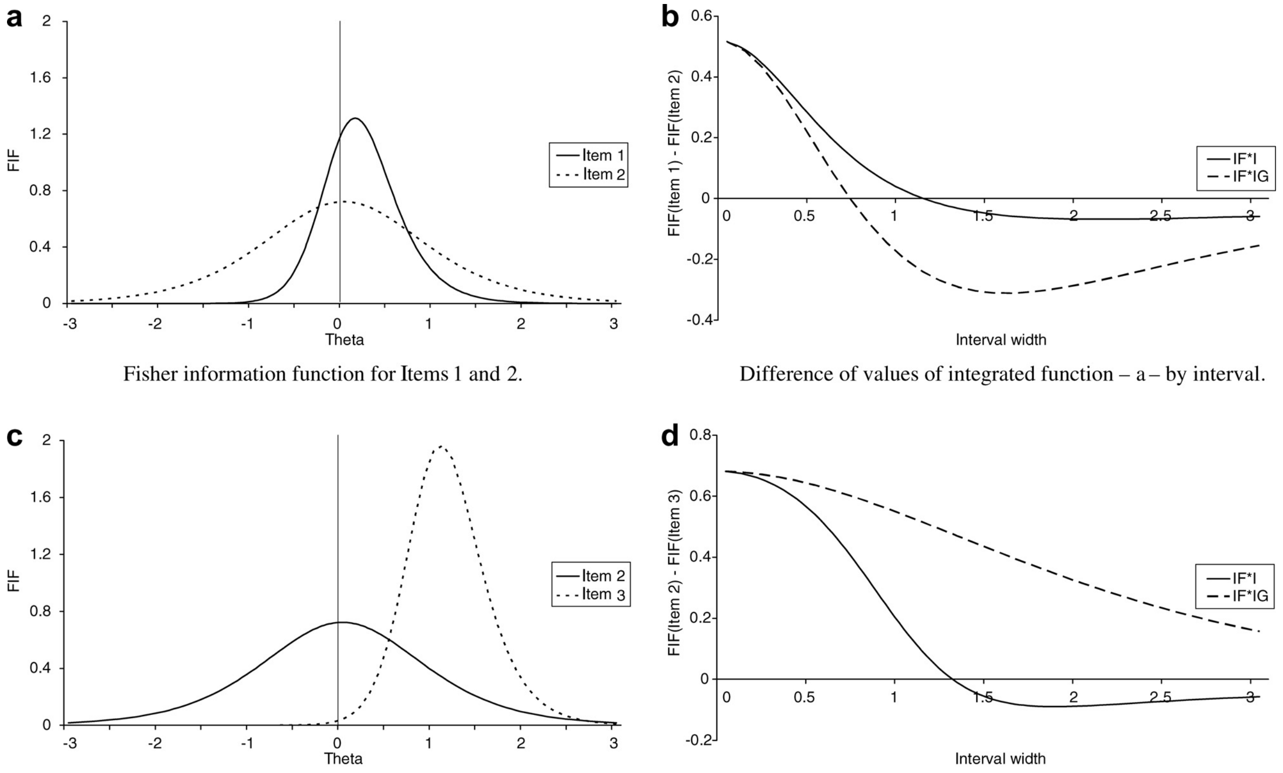
Figure 1a shows FIF for two items. Item 1 would clearly be selected for θ values around 0. However, the information function for this item is more peaked, so that, for values outside the interval (−0.3, 0.7), Item 2 would be preferable. Figure 1b shows the difference between Item 1 and 2 in the mean information for an interval focused on 0 (continuous line), in such a way that positive values imply the selection of the first item and negative values the selection of the second. For interval widths greater than 1.1, Item 2, of low value in the *a* parameter, would be preferable.

However, it is also possible to find conditions in which items of low exposure rate and maximally informative for values close to $\hat{\theta}$ cease to be selected when their performance for a wide range of trait levels is taken into account. In Figure 1c, we can see how Item 2 is that which provides most information for $\theta = 0$. This item would no longer be selected when the interval width was greater than 1.3, as Item 3, of higher value in the *a* parameter, would be chosen. This can be considered as a perversion of selection logic for FI*I: what was sought was a rule that, at the beginning of the test, selected items with appropriate performance for a wide region of θ values, not a rule that selected items with the maximum of their information distant from $\hat{\theta}$. In fact, some results (Barrada, Olea et al., 2006; Chen & Ankenmann, 2004) suggest that this could be the selection pattern with FI*L, KL*L, and FI*I.

In order to avoid FI*I risks, a modification of it was developed (Barrada, Olea et al., 2006), aimed at simultaneously improving measurement accuracy and bank security.

FI*IG

In Figure 1, it can be seen how the items with the highest maximum FIF value are also those with the greatest variance in FIF values. An ISR that incorporates, directly or indirectly, the variance of the Fisher information for the different θ points, penalizing it, would reduce or eliminate cases such as that illustrated in Panels c and d. This can be achieved through the use of the geometric mean. The geometric mean of a data set of z elements is the multiplicatory of the z values raised to the $(1/z)$th power. For data sets with the same arithmetic mean, the one with the higher geometric mean will be that with the lower variance. For instance, the sets (4, 6) and (1, 9) both have an arithmetic mean of 5, while the geometric mean of the former is 4.9 and that of the latter is 3.



Fisher information function for Items 1 and 2.

Difference of values of integrated function – a – by interval.

Fisher information function for Items 2 and 3.

Difference of values of integrated function – c – by interval.

Figure 1. Examples of information function and selected items according to the integration interval width. Parameters (*a*, *b*, and *c*): Item 1 ~ (2, 0, and 0.4); Item 2 ~ (1, 0, and 0); and Item 3 ~ (2, 1, and 0.2).

Fisher information by interval with geometric mean, FI*IG, has a similar approach to that of FI*I, and attempts to develop it and overcome some of its limitations. Putting this ISR into effect requires a slight modification in equation (5). It is no longer possible to integrate in an interval; rather, it is necessary to use the multiplicatory, as shown in equation (9). Otherwise, all the variables and functions of the equation are identical to those used for FI*I. The k values indicate the quadrature points employed:

$$\max_{i \in B_n} \prod_{j=0}^{k} V(\theta_j) W(\theta, x, g)$$

$$\theta_j = \theta_{\min} + \frac{j(\theta_{\max} - \theta_{\min})}{k}, \quad k = 80. \qquad (9)$$

Panels b and d of Figure 1 (dotted line) show the effect of applying FI*IG. In comparison to what occurred with FI*I, for no interval considered in Panel d Item 3 is selected, and the above-mentioned problem does not arise. In Panel b we see how, as FI*IG penalizes the variance, the interval necessary for selecting an item with lower a parameter is smaller, in comparison to the case of FI*I.

Goal of the Present Study

We see, with these examples, how FI*I and FI*IG: (a) take into account the global performance of the valuating function; (b) under specific conditions, select items that would not be selected with PFI; and (c) converge to PFI as the interval considered is narrowed. Similar examples can be generated for the other alternative rules.

We can expect differences in the exposure control because of the first two points, but these will be limited, because of the third point. As we have described for FI*I, with a logic easily be generalized to the other ISRs presented, there are reasons for expecting both a possible improvement and a possible deterioration of security. Previous results (Barrada, Olea et al., 2006; Chen & Ankenmann, 2004) indicate that, in fact, the security is deteriorated with the ISR based on the likelihood function and with FI*I. Until now, no study has compared all the alternative ISR with PFI at the same time. Here is the first two goals of our study: to evaluate all of them simultaneously and to identify the pattern of items selected with each ISR.

The other main goal of the study is to evaluate performance, in accuracy and security, for FI*IG, with different simulation conditions than the ones reported in Barrada, Olea et al. (2006) and comparing it with other five different ISRs. The combination of the two trends presented in Panels b and d of Figure 1, the elimination of overselection of high a-parameter items and the favoring of the selection of low a-parameter items, leads us to expect that IF*IG will present better exposure control. As FI*IG evaluates FIF not just for $\hat{\theta}$, but for the interval where it is maximally probable to find θ, improvements in accuracy are also expected.

Simulation Study

Method

A study was carried out for evaluating the accuracy and security obtained with the six ISRs described: PFI, FI*L, KL*L, FI*I, KL*I, and FI*IG.

Item Banks

As Wingersky and Lord (1984) and Chang, Qian, and Ying (2001) point out, in practice, *a* and *b* parameters are usually positively correlated. Bearing this in mind, two kinds of item bank were generated: one with noncorrelated *a* and *b* parameters and the other with r_{ab} = .5. Ten item banks were constructed for each kind, each with 500 items. The item parameters were generated randomly from the following distributions: $a \sim$ N(1.2, 0.25); $b \sim$ N(0, 1); and $c \sim$ N(0.25, 0.02).

Start

The simulation began with an initial trait level ($\hat{\theta}_0$) randomly selected from the interval (−0.5, 0.5). Before the administration of any item it is impossible to calculate the likelihood function, and the width of the interval for FI*I, KL*I, and FI*IG is infinity. To make it possible to apply, with effect from the very first item, the rules alternative to PFI, the likelihood function was calculated using two fictitious items, one correct and the other incorrect, both with the same (*a*, *b*, and *c*) parameters equal to $(0.5,\ \hat{\theta}_0,\ \text{and}\, 0)$, and the interval for the interval rules was fixed at $(\hat{\theta}_0 - 3,\ \ \hat{\theta}_0 + 3)$. After administration of the first item, the fictitious items were no longer used.

Estimation/Assignment of Trait Level

Maximum-likelihood estimation has no solution in real numbers when there is a constant response pattern, all correct or all incorrect responses. Therefore, until there was at least one correct and one incorrect response, θ 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 $(\hat{\theta} - b_{\min})/2$. In these formulas $b_{\max}$ and $b_{\min}$ refer to the highest and lowest *b* parameters, respectively, of the entire item bank. Once the constant pattern was broken, we applied maximum-likelihood estimation (Birnbaum, 1968), as indicated in:

$$\hat{\theta} = \max_{\theta \in \Theta}(L(\theta, x, g)), \tag{10}$$

$$\Theta = \{\theta : \theta = -4 + (k-1)/100\ \forall k \in N, k \leq 801\}. \tag{11}$$

In this way, unlike with other numerical approximation methods, the problem of possible multiple local maxima is avoided (Veerkamp & Berger, 1997).

Trait Level of the Simulees and Test Length

Trait level of the examinees was randomly extracted from a population N(0, 1). For each combination of simulated bank per ISR, 5,000 simulees were generated. In each simulation, up to 20 items were applied. Data for subsequent analysis were saved for every five items. Previous results have shown that, for this test length, accuracy differences between ISRs are small.

Evaluation Criteria

Five dependent variables were used for the comparison between the different ISRs: root mean square error (RMSE) relative to the measurement accuracy (equation (12)); overlap rate for evaluating item bank security (equation (13)), according to the formula developed by Chen et al. (2003); mean values of the *a* and *c* parameters administered, with the aim of analyzing the kind of item that tends to be selected by each ISR; and the correlation between the item exposure rates for each pair of ISRs, as indicative of the convergence between them:

RMSE

$$\text{RMSE} = \left(\sum_{g=1}^{r} \left(\hat{\theta}_g - \theta_g \right)^2 \Big/ r \right)^{1/2}, \tag{12}$$

where *r* is the number of replicates-examinees;

Overlap Rate

$$\hat{T} = \frac{n}{q} S_{\text{er}}^2 + \frac{q}{n}, \tag{13}$$

where

$\hat{T}$ is the large-sample approximation of the overlap rate (Chen et al., 2003);
q is the number of items administered; and
S_{er}^2 is the variance of the item exposure rates.

Chang and Zhang (2002) and Chen et al. (2003) have demonstrated that, necessarily, $q/n \leq \hat{T} \leq 1$.

Application of ISRs

The criterion function for PFI can be calculated directly with the FIF for $\hat{\theta}$, as indicated in equation (3). For the rest of the ISRs, the values and functions applied in this study for equations (5) and (9) are described in Table 1. For the ISRs based on intervals, α was set at 0.05.

Table 1. Description of the different ISRs for the selection of the qth item (Φ is the standard cumulative distribution and the other symbols correspond to the description of the text)

	FI*L	KL*L	FI*I and FI*IG	KL*I
$V(\theta_j)$	$I(\theta)$	$\mathrm{KL}\left(\theta_j \| \| \hat{\theta}\right)$	$I(\theta)$	$\mathrm{KL}\left(\theta \| \| \hat{\theta}\right)$
$W_n(x_i, g_i, \theta_j)$	$L(\theta_j, x, g)$	$L(\theta_j, x, g)$	1	1
$\theta_{\min}$	-4	-4	$\max\left(\hat{\theta} - 3, \hat{\theta} - \frac{\Phi^{-1}(.975)}{\sqrt{I(\hat{\theta})}}\right)$	$\hat{\theta} - \frac{\Phi^{-1}(.975)}{\sqrt{q-1}}$
$\theta_{\max}$	4	4	$\min\left(\hat{\theta} + 3, \hat{\theta} + \frac{\Phi^{-1}(.975)}{\sqrt{I(\hat{\theta})}}\right)$	$\hat{\theta} + \frac{\Phi^{-1}(.975)}{\sqrt{q-1}}$

The criterion integral for FI*I, developed, can be expressed as follows (Veerkamp & Berger, 1997):

$$\int_{\theta_{\min}}^{\theta_{\max}} I(\theta)d\theta = \frac{1.7a}{(1-c)}\left[c\ \ln\left(\frac{P(\theta_{\min})}{P(\theta_{\max})}\right) + P(\theta_{\max}) - P(\theta_{\min})\right]. \tag{14}$$

For the other ISRs, the criterion function cannot be solved analytically. We calculated it using quadrature points, as described in equation (15) for FI*L, KL*L, and KL*I, and in equation (9) for FI*IG:

$$\sum_{j=0}^{k} V(\theta_j)W(\theta_j, x, g)$$
$$\theta_j = \theta_{\min} + \frac{j(\theta_{\max} - \theta_{\min})}{k}, \quad k = 80. \tag{15}$$

Results

We present the results divided in five sections: (a) accuracy, (b) security, (c) relation between these two variables, (d) mean values of the a and c parameters, and (e) correlation between the item exposure rates.

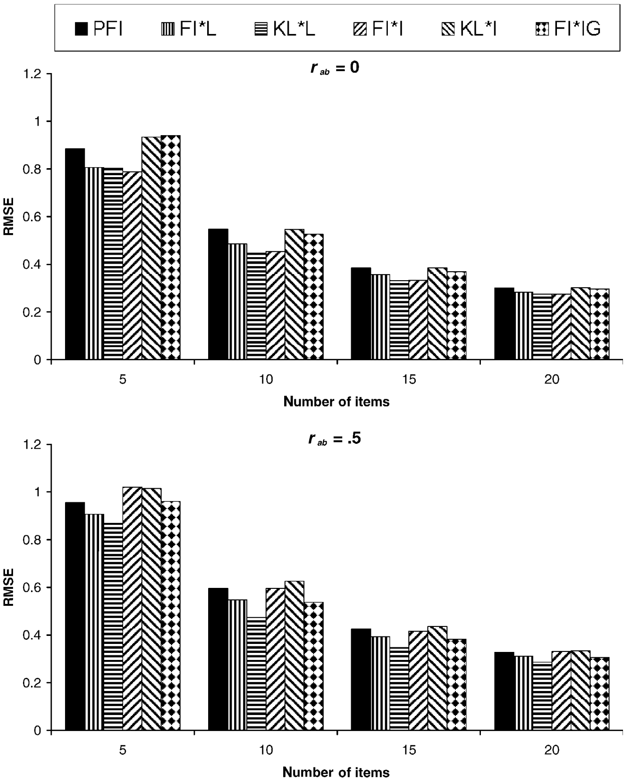


Figure 2. RMSE according to the number of items administered and the item bank kind.

Accuracy

Figure 2 shows the different RMSE values conditional on the number of items presented, for both kinds of item bank. Logically, as the number of items presented is increased, RMSE decreases. The slope of accuracy improvement is smaller the greater the number of items administered. Of the rules evaluated, the ones based on the likelihood function are those that achieve the best measurement accuracy. Among them, the one with the lowest RMSE is KL*L. The performance of FI*I and KL*I varies depending on the presence or the absence of correlation between the a and b parameters. When r_{ab} = 0, the RMSE of KL*I is very similar to that obtained with PFI, while the measurement error for FI*I is almost identical to that obtained with KL*L. However, when r_{ab} = .5, these rules have an RMSE equivalent or superior to that obtained with PFI. With more than five items administered, the RMSE obtained with FI*IG is always smaller than that of PFI. With uncorrelated a and b parameters this difference is minimal, being greater when parameters covary. The difference in RMSE between ISRs becomes smaller as the number of items presented increases, though the relative accuracy ranking remains practically constant.

Security

Overlap results for the different ISRs and item bank kinds are shown in Figure 3. Functions are almost indistinguishable for the item banks with correlated and uncorrelated parameters. Differences between ISRs are greater at the beginning of the CAT, and the order in overlap remains constant in all the evaluated conditions, except for a slight change for five items presented, when the order is reversed between KL*L and FI*L, on the one hand, and between PFI

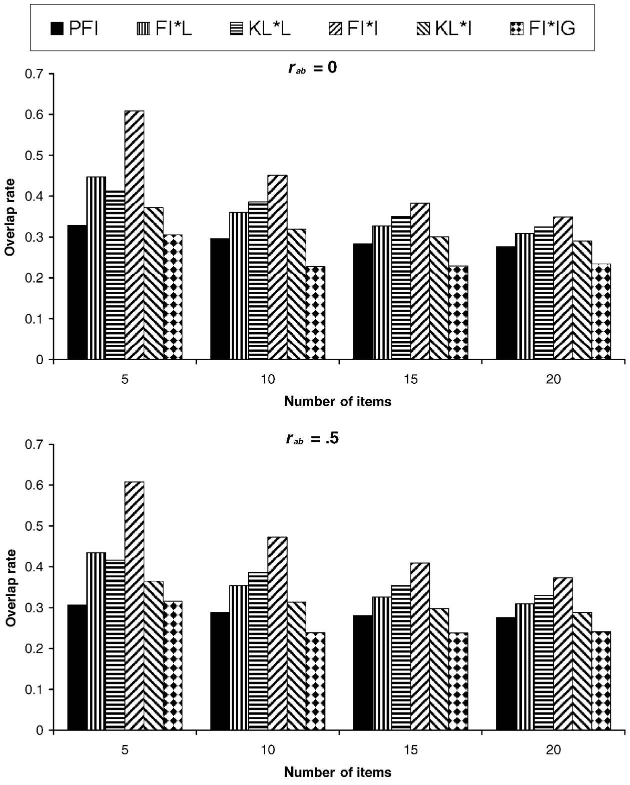


Figure 3. Overlap rate according to the number of items administered and the item bank kind.

and FI*IG, on the other. All the rules alternative to PFI show an overlap rate higher than that obtained with PFI, with the exception of IF*IG, which shows the best results in security, followed by PFI and KL*I. The rules based on the likelihood function have intermediate values, and the ISR with highest overlap is FI*I. When passing from 5 to 10 items administered, the overlap rate decreases, especially for FI*I, and subsequently remains basically constant.

Relation Between Accuracy and Security

It can be observed from our presentation of data on accuracy and security that the two variables tend to relate inversely: Greater accuracy implies less security. The data obtained seem to support a trade-off between these variables. We can find an exception for IF*IG, which achieves simultaneous improvements for the two variables, and for FI*I, in the case of banks with correlated parameters, with poor performance in security without improvements in accuracy. For a more direct study of this relationship, we show the scatter plots of RMSE and overlap in Figure 4. For correct interpretation of the plots, it must be taken into account that the scale in each one is different. It is therefore necessary to attend to the values of the axis.

We can consider an inadequately performing ISR as that for which at least one other ISR can be found offering, simultaneously, better results in accuracy and security. With this definition, three out of six of the ISRs evaluated show inadequate performance. In Figure 4, it can be seen how PFI presents better results than KL*I, for both kinds of bank and for any number of items administered. By comparison with FI*I, when the *a* and *b* parameters are uncorrelated, KL*L achieves better results for the two variables. When the parameters are correlated, these improvements are offered by PFI, FI*L, and KL*L. The ISR commonly employed in CATs, PFI, also seems to present inadequate performance: FI*IG, when more than five items are presented, has better accuracy and security levels. This result is more marked with r_{ab} = .5.

Average of the *a* and *c* Parameters of the Presented Items

Figure 5 shows these values for the banks with uncorrelated *a* and *b* parameters. For banks with r_{ab} = .5, these values are not interpretable, and are therefore omitted. All ISRs, with the exception of FI*IG, tend to administer, at the beginning of the test, items with the *a* parameter clearly above the mean of this parameter in the bank. Except for the case of FI*IG, as the number of items administered is increased and the items with higher *a* parameters have already been selected, items with lower *a* parameters are progressively administered. With FI*IG, highly discriminative items are left available for more advanced phases of the test. Because of this, the average of the *a* parameter when more than 10 items are administered is higher for FI*IG than for the rest of the ISRs. For any number of items presented, all the ISRs, discarding FI*IG, tend to select items with higher *a*-parameter values than PFI, with the exception of KL*L for five items. This trend is more marked for FI*L and for FI*I.

With regard to the *c* parameter, the studied ISRs tend to select items with *a* value in this parameter below 0.25, the mean in the item bank. As with the *a* parameter, but in reverse, as the test progresses and items with low *c* value are exhausted, the average in this parameter increases. In comparison to PFI, the other ISRs show smaller values in this variable when five items have been presented. The trend to select low *c*-parameter items is more accentuated for KL*L, for FI*I, and particularly for FI*IG. When more than 10 items have been administered, the alternative rules have exhausted the items of low *c* value, so that they present a higher average value in this variable. These combinations of average values of the *a* and *c* parameters could explain the differences in overlap rate between ISRs. FI*L, KL*L, FI*I, and KL*I tend to select, more clearly than PFI, items with high value in the *a* parameter and low value in the *c*, an odd combination of values, which raises the overlap rate of these ISRs. For these ISRs, it seems that there are no different selection strategies over the course of the test, but rather that the differences in the characteristics of the selected items are due simply to exhaustion of the available bank. Nevertheless, this is not the case for FI*IG. This ISR seeks different combinations of parameters according to the phase of the test. At the beginning, items with low *a* and *c* parameters, changing to items with high *a*-parameter values and low importance in the *c* parameter as the test advances. This variable pattern makes the overlap rate of IF*IG the smallest of the ISRs studied.

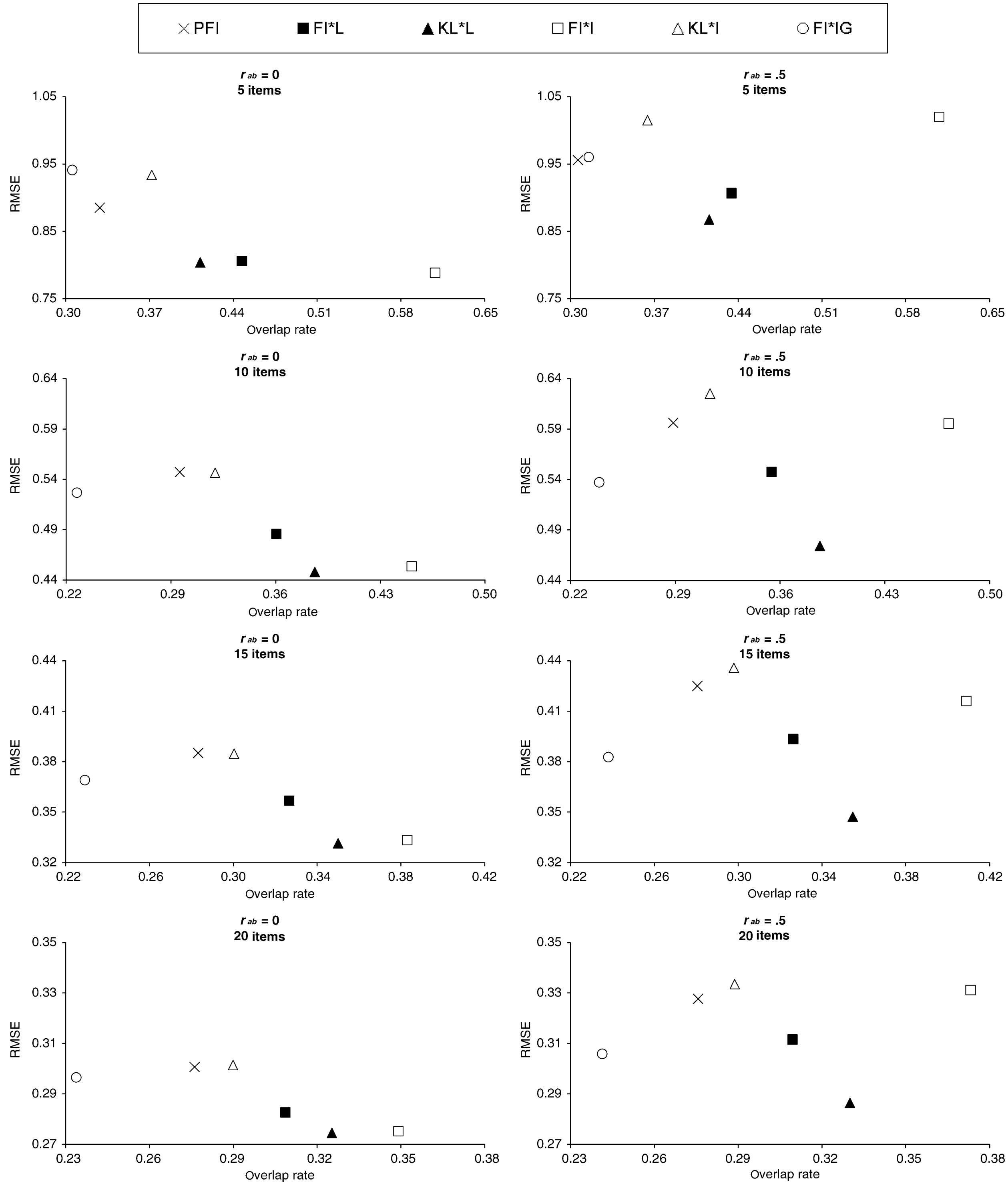


Figure 4. Scatter plots of RMSE and overlap rate according to the number of items administered and the item bank kind.

Correlation Between Item Exposure Rates for the Different ISRs

It was expected that the ISRs alternative to PFI would tend to select different items from PFI at the beginning of the test, converging with PFI as the number of items administered increased. To evaluate this we calculated, for each pair of ISRs, the correlation between the item exposure rate of the items, as shown in Table 2.

There are three particularly notable findings. First, the correlations increase as the number of items administered increases, confirming the idea of convergence between ISRs as the test progresses. Second, the coincidence in the items-selected pattern, for all the ISRs except FI*IG, is very high, even with five items administered. In Table 2 it can be seen, for instance, that the correlation between PFI and KL*I is .9 or .89 with just five items presented. Despite being ISRs with different valuating and weighting functions, the similarity in the pattern of items selected is very high (Veldkamp, 2003). Third, the ISR most differentiated from the others in its item selection patterns is FI*IG, as could be expected from what is shown in Figure 5. The correlation between exposure rates for PFI and IF*IG goes from −.03, when five items are selected, to .71, with 20 selected, markedly inferior to the correlation between the rest of the ISRs with PFI.

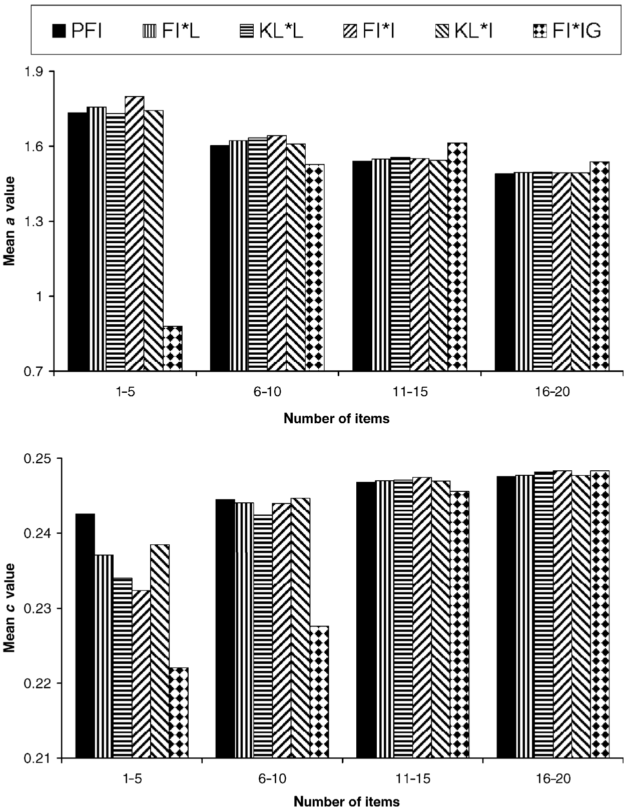


Figure 5. Average values of the *a* and *c* parameters of the administered items according to the number of items administered, for the item banks with r_{ab} = 0.

Discussion

There were two main goals in this study. On the one hand, to investigate simultaneously the accuracy and security obtained with several ISRs. On the other hand, we set out to examine the performance of a newly proposed ISR, IF*IG, designed to solve some of the possible problems with FI*I. We thus assessed the RMSE, overlap rate, pattern of selected items, and coincidence between item exposure rates.

If we do not take into account the performance of FI*IG, we have found a trade-off between accuracy and security for the different ISRs. PFI is the rule with the greatest measurement error and the smallest overlap. At the other extreme is KL*L. This result is qualified by the kind of item bank employed. The trade-off is clearer when the *a* and *b*

Table 2. Correlation between item exposure rates of the different ISRs according to the number of items administered and the item bank kind

		r_{ab} = 0					r_{ab} = .5				
		PFI	FI*L	KL*L	FI*I	KL*I	PFI	FI*L	KL*L	FI*I	KL*I
q = 5	FI*L	.85					.83				
	KL*L	.78	.87				.78	.88			
	FI*I	.73	.85	.79			.68	.79	.71		
	KL*I	.90	.90	.89	.79		.89	.90	.89	.77	
	FI*IG	−.03	−.03	−.03	−.02	−.03	−.03	−.02	−.02	−.02	−.03
q = 10	FI*L	.94					.93				
	KL*L	.90	.94				.90	.94			
	FI*I	.89	.94	.91			.86	.91	.86		
	KL*I	.96	.95	.94	.90		.96	.96	.93	.90	
	FI*IG	.31	.32	.37	.32	.32	.33	.34	.39	.30	.33
q = 15	FI*L	.96					.95				
	KL*L	.93	.96				.93	.96			
	FI*I	.93	.96	.94			.91	.94	.91		
	KL*I	.98	.97	.96	.94		.97	.97	.95	.93	
	FI*IG	.60	.60	.63	.60	.60	.60	.60	.64	.55	.59
q = 20	FI*L	.97					.97				
	KL*L	.95	.97				.95	.97			
	FI*I	.94	.97	.95			.93	.95	.93		
	KL*I	.98	.98	.97	.95		.98	.98	.97	.95	
	FI*IG	.71	.71	.73	.70	.71	.71	.71	.73	.66	.70

parameters are uncorrelated. With $r_{ab} = .5$, FI*I, in comparison to PFI, shows a high overlap rate, but without improvements in measurement accuracy. This result indicates that including the covariation between parameters, a variable not usually included, is relevant for the study of ISR performance.

In the case of FI*IG, the mentioned trade-off does not hold, since, in comparison to PFI, it achieves simultaneous improvements in RMSE and overlap. The difference in accuracy, with uncorrelated parameters, is very small, though it does hold for different test lengths. With correlated parameters, the difference in RMSE is greater. The relevance of this new ISR lies not only in the size of the improvements, but also in the fact of its breaking the trade-off assumed by many authors.

We can consider as inadequate any ISR for which there exists another rule that simultaneously improves its performance in accuracy and security. Following this criterion, we could discard the FI*I and KL*I rules. Also, and importantly, we could discard PFI, since FI*IG obtains better results for both variables.

As in the stratified methods (alpha stratified: Chang & Ying, 1999; maximum information stratified: Barrada, Mazuela, & Olea, 2006), with FI*IG the items with low *a*-parameter value are presented at the beginning of the test. This could be an additional advantage when the security of the item bank is broken. With few items administered, the effect of knowing beforehand the content and response of an item is greater the higher the discrimination of the item (Chang & Ying, 2008). Likewise, in this way, the risk of capitalization on item calibration error can be reduced (van der Linden & Glas, 2000b).

Despite having different formulations, with different valuating and weighting functions, and different values of $\theta_{\min}$ and $\theta_{\max}$, all the ISRs show high coincidence in the selected items, as in the Chen et al. (2000) and Chen and Ankenmann (2004) studies. The only ISR with a differentiated selection pattern at the beginning of the test, which converges as the number of items administered increases, is FI*IG.

The FI*IG application for which we have opted in this study, described in Table 1 and in equation (9), could be modified, so as to control the speed with which this rule converges with PFI. In this way, we could seek a continuous increase in the *a*-parameter average of the items administered, in contrast to what occurs with the current method. This could represent a possible future research line.

All the ISRs presented obtain an overlap rate above the limits usually considered as acceptable (Way, 1998). It remains for future research to examine the effects of incorporating additional methods of exposure control (Chen & Ankenmann, 2004).

Acknowledgments

This research was partly supported by two grants from the Spanish Ministry of Education and Science (projects SEJ2006-08313/PSIC and SEJ2004-05872PSIC).

References

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., & Ponsoda, V. (2004). Reglas de selección de ítems en tests adaptativos informatizados [Item selection rules in computerized adaptive testing]. *Metodología de las Ciencias del Comportamiento, Volumen Especial,* 55–61.

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2006). Estrategias de selección de ítems en un test adaptativo informatizado para la evaluación de inglés escrito [Item selection rules in a computerized adaptive test for the assessment of written English]. *Psicothema, 18*, 828–834.

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

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

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.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431–444.

Bradley, R. A., & Gart, J. J. (1962). The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika, 4*, 205–214.

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). Sage Publications.

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, H. H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73*, 441–450.

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

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.

Chen, S. Y., & Lei, P. W. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement, 29*, 204–217.

Cheng, P. E., & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. *Applied Psychological Measurement, 24*, 257–265.

Cheng, P. E., & Liou, M. (2003). Computerized adaptive testing using the nearest-neighbors criterion. *Applied Psychological Measurement, 27*, 204–216.

Davey, T., & Fan, M. (2000, April). *Specific information item selection for adaptive testing*. Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans, LA.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with adaptive testing*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249–261.

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 June 28, 2007, from http://www.jtla.org.

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. (1971). Robbins-Monro procedures for tailored testing. *Educational and Psychological Measurement, 31*, 3–31.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1*, 95–100.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and their parallel-form reliability. *Psychometrika, 48*, 233–245.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 223–236). New York: Academic Press.

Olea, J., & Ponsoda, V. (2003). *Tests adaptativos informatizados* [Computerized adaptive testing]. Madrid: UNED.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351–356.

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. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, 34*, 100.

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

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

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.

Sympson, J. B., & Hetter, R. D. (1985, Ocobter). Controlling item exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973–977). San Diego, CA.

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201–216.

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000a). *Computerized adaptive testing: Theory and practice.* Norwell, MA: Kluwer.

van der Linden, W. J., & Glas, C. A. W. (2000b). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 12*, 35–53.

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.

Veldkamp, B. P. (2003). Item selection in polytomous CAT. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 207–214). Tokyo, Japan: Springer-Verlag.

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.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109–135.

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.

Juan Ramon Barrada

Facultad de Psicologia
Universidad Autonoma de Barcelona
E-08193 Bellaterra
Spain
Tel. +34 935 813263
E-mail juanramon.barrada@uab.es