The Spanish Journal of Psychology
2011, Vol. 14 No. 1, 500-508
doi:10.5209/rev_SJOP.2011.v14.n1.45

ISSN 1138-7416

Varying the Valuating Function and the Presentable Bank in Computerized Adaptive Testing

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

[1]Universidad Autónoma de Barcelona (Spain)
[2]Universidad Autónoma de Madrid (Spain)

In computerized adaptive testing, the most commonly used valuating function is the Fisher information function. When the goal is to keep item bank security at a maximum, the valuating function that seems most convenient is the matching criterion, valuating the distance between the estimated trait level and the point where the maximum of the information function is located. Recently, it has been proposed not to keep the same valuating function constant for all the items in the test. In this study we expand the idea of combining the matching criterion with the Fisher information function. We also manipulate the number of strata into which the bank is divided. We find that the manipulation of the number of items administered with each function makes it possible to move from the pole of high accuracy and low security to the opposite pole. It is possible to greatly improve item bank security with much fewer losses in accuracy by selecting several items with the matching criterion. In general, it seems more appropriate not to stratify the bank.
Keywords: computerized adaptive testing, item selection rule, item bank security, overlap rate.

En los tests adaptativos informatizados, la función de valoración más comúnmente empleada es la función de información de Fisher. Cuando el objetivo es mantener al máximo la seguridad del banco de ítems, la función de valoración que parece más adecuada es el criterio de proximidad, con el que se valora la distancia entre el nivel de rasgo estimado y el punto donde es máxima la información proporcionada por un ítem. Recientemente, se ha propuesto no mantener la misma regla de valoración constante a lo largo de todo el test. En este estudio, expandimos la idea de combinar el criterio de proximidad con la función de información de Fisher. También manipulamos el número de estratos en los que se divide el banco. Encontramos que la manipulación del número de ítems administrados con cada función hace posible moverse desde el extremo de alta precisión y baja seguridad hasta el extremo opuesto. La selección de varios ítems con el criterio de proximidad hace posible mejorar en gran medida la seguridad del banco con pérdidas escasas en precisión. En general, parece más adecuado no estratificar el banco.
Palabras clave: tests adaptativos informatizados, regla de selección de ítems, seguridad del banco de ítems, tasa de solapamiento.

This research was supported by a grant from the Spanish Ministerio de Ciencia e Innovación (project number PSI2009-10341).

Correspondence concerning this article should be addressed to Juan Ramón Barrada. Facultad de Psicología. Universidad Autónoma de Barcelona. 08193 Bellaterra. Barcelona. (Spain). Phone: +34-935813263. E-mail: juanramon.barrada@uab.es

The lower costs and higher calculation speed of computers have popularised computerized adaptive testing (CAT) as a technique for evaluating educational or psychological contents (van der Linden & Glas, 2010). A CAT allows, when compared with a paper and pencil test, faster and/or more accurate estimation of the examinees' trait level.

The item selection process when a CAT is applied seeks to maximize, at least, two objectives. The first is measurement accuracy. The satisfaction of this objective is commonly measured, in simulation studies, with the root mean squared error (RMSE):

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

where r is the number of examinees, θ_g is the (real) trait level of the g-th examinee and θ_g is the estimated trait level for that examinee.

The second objective to maximize is the item bank security. A CAT allows for greater flexibility in test scheduling: two examinees can be evaluated at different moments, with a totally or partially identical item bank. If the first examinee informs the second about the items he received, the second could get correct responses not due to his trait level, but because of the leakage of the bank's content, which would lead to the over-estimation of his trait (H. H. Chang, 2004). The greater the proportion of items that is presented to both examinees, the greater this risk. Overlap rate, defined as the mean proportion of items shared by two examinees (H. H. Chang & Zhang, 2002; Chen, Ankenman & Spray, 2003), is one of the most commonly employed variables for evaluating item bank security. The following equation is the one used to calculate this (Chen et al., 2003):

$$\hat{T} = \frac{n}{Q} S_{er}^2 + \frac{Q}{n}, \quad (2)$$

where $\hat{T}$ is the large-sample approximation of the overlap rate, n is the item bank size, Q is the test length and S_{er}^2 is the variance of the exposure rates of the items.

Several authors have declared that an improvement in one of these objectives means a reduction in the satisfaction of the other (S. W. Chang & Ansley, 2003; Stocking & Lewis, 2000).

Most item selection rules proposed until now (H. H. Chang & Ying, 1996; H. H. Chang & Ying, 1999; van der Linden, 1998; Veerkamp & Berger, 1997) can be described by means of two different steps. The first consists of the definition of the subset of items in the bank that can be selected for a given examinee and for a given item position in the test. The second step seeks the item belonging to the presentable subset that optimizes a certain valuating function. We now provide a more detailed description of these steps.

Definition of the presentable bank

B_q is the subset of items belonging to the item bank that can be presented to the examinee in the q-th position in the test. We first consider the whole item bank to compose B_q. After this definition, restrictions are applied that reduce the size of B_q to below the item bank size. A restriction applied in all of the item selection rules is to remove from B_q, for an examinee, all the items they have already been administered to him.

In addition to this general restriction, two other non-excluding methods for defining B_q have been applied in CATs:

- *B_q composition variable according to examinee*: A way to limit over-exposure is for the items which, if applied with no additional restriction would have an exposure rate above a threshold considered to be the maximum tolerable rate (r^{max}), not to belong to B_q for all the examinees, but just for a proportion of them. This proportion will be more limited the greater the over-exposure of the item. This reasoning has been applied in trying to fix the exposure rates of the items below r^{max} in the Sympson-Hetter method (Sympson & Hetter, 1985), the restricted method (Revuelta & Ponsoda, 1998) and the item-eligibility method (van der Linden & Veldkamp, 2004). Among them, the most commonly employed is the Sympson-Hetter method (van der Linden, 2003), although the item-eligibility method is the one that seems to be preferable (Barrada, Abad, & Veldkamp, 2009).
- *B_q composition variable according to the item's position in the test:* To guarantee that the information supplied by the items increases as the number of administered items rises, H. H. Chang and Ying (1999) proposed dividing the bank into S strata, in such a way that the n/S items of the lowest a parameter belong to the first stratum, the next n/S items are in the second stratum, etc. For the first Q/S items to be administered, B_q will be composed only of the items forming the first stratum; for the second Q/S items to be presented, B_q will be composed of the items belonging to the second stratum; etc. In operative banks, it is common for the a and b parameters of the items to be positively correlated (Wingersky & Lord, 1984). If the bank is stratified by only taking into account the a parameters, the distribution of the b parameters will be moving to the right from stratum to stratum. For these cases, H. H. Chang, Qian, and Ying (2001) proposed stratifying

the bank blocking *b*. To do this, the items are ordered according to the value of their *b* parameter. The first *S* items will then be be distributed between the *S* strata according to their *a* parameter value. This operation is repeated with the next *S* items and so on, until the whole bank has been divided. When the items in the bank have been calibrated using the three-parameter model, there is not a perfect correlation between the *a* parameter of the items and the maximum in the Fisher information function ($I(\theta)_{max}$). The value of $I(\theta)_{max}$ for item *i* can be calculated using the following equation (Hambleton & Swaminathan, 1985):

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

In these cases, the θ value where the maximum information is reached is not equal to the *b* parameter, as it is in the case of the items calibrated according to the one and two-parameter models, but a value moved towards the right in relation with *b*, this movement being greater the lower the value of the *a* parameter and the greater the value of the *c* parameter. We will call θ^{max} the value θ where the maximum of the information function is reached. For item *i*, it is calculated according to Equation 4 (Hambleton & Swaminathan, 1985):

$$\theta_i^{\max} = b_i + \frac{\ln\left[1+\left(1+8c_i\right)^{1/2}\right]-\ln(2)}{1.7a_i}. \quad (4)$$

Barrada et al. (2006) have shown the convenience of stratifying, instead of employing the *a* and *b* parameters, using θ^{max} and $I(\theta)_{max}$. In this way, accuracy is improved, while we increase the security, in comparison with the *a*-stratified method without blocking (H. H. Chang & Ying, 1999), also with banks without correlation between the *a* and *b* parameters.

Combinations of these two methods for restricting B_q are possible. For instance, Leung, Chang and Hau (2002) have proposed defining B_q by stratifying the bank and applying the Sympson-Hetter method. More recently, Barrada, Veldkamp, and Olea (2009) have proposed using the item-elegibility method while increasing the size of B_q as the test goes on, as the value of r^{max} also increases.

Valuating function

Some of the developed valuating functions have focused on maximizing the measurement accuracy in CATs. Others have tried to offer a less skewed item exposure distribution, without losing accuracy. Among the former, the most commonly used valuating function is the Fisher information function for the estimated trait level (Lord, 1980). Being V_i the value of the *i*-th item, this valuating function would be:

$$V_i = I_i(\hat{\theta}) = \frac{2.89\, a_i^2\left(1-c_i\right)}{\left(c_i+e^{1.7a_i(\hat{\theta}-b_i)}\right)\left(1+e^{-1.7a_i(\hat{\theta}-b_i)}\right)^2}. \quad (5)$$

When this valuating function (called hereafter FI) is employed, the selected item *j* is the one which, belonging to B_q, maximizes Equation 5:

$$j = \arg\max_{i\in B_q} I_i(\hat{\theta}). \quad (6)$$

With this valuating function, and when the item bank is not stratified, the item exposure rates correlate positively and strongly with their *a* parameters (Li & Schaffer, 2005). This implies a high variance in the item exposure rates, with some of them highly over-exposed and others never presented to any examinee. Following Equation 2, this means a high overlap rate.

Some of the valuating functions that make it possible to greatly reduce the risks to the item bank security are the matching criteria (Li & Schafer, 2005). As commented previously, in the one and two-parameter models, θ_{max} meets the *b* parameter of the item. Because of this, when the item selected is the one with the minimum distance between the estimated trait level and the *b* parameter of the item, it is guaranteed, for these two IRT models, that the information capitalized is maximized. Also, as the *a* parameter is irrelevant in this valuating function, there is no correlation between the exposure rates and the discrimination parameters. The valuating function, which we will call B-MC, would be:

$$V_i = \left|b_i - \hat{\theta}\right|. \quad (7)$$

With this function, the item would be selected as follows:

$$j = \arg\min_{i\in B_q}\left|\hat{\theta} - b_i\right|. \quad (8)$$

With items calibrated according to the three-parameter model it is more appropriate to replace the *b* value in Equation 7 and 8 for the $I(\theta)_{max}$ value, which leads to the reduction of RMSE (Barrada et al., 2006). This function,

which we will call TM-MC (because of the theta-maximum matching criterion), would follow the next equation:

$$j = \arg\min_{i \in B_q} \left| \hat{\theta} - \theta_i^{\max} \right| . \quad (9)$$

New approaches to item selection rules

Recently, it has been proposed not to hold the valuating function of the items constant throughout the entire length of the CAT (Leung, Chang, & Hau, 2005; Li & Schafer, 2005). At the beginning of the test, as the trait level estimation is unstable and not very accurate, applying FI could not be the best strategy. Because of this, Leung et al. (2005) have proposed dividing the item bank into two strata. During the first half of the test, B-MC would be applied; FI would be the used in the second part.

Given the variety of valuating functions and methods for defining B_q, the combinatory that could be established is very wide. We have opted, like Leung et al. (2005) to study the combination of two of the valuating functions that are most commonly used and situated near the extremes that can be found. On the one hand, FI is one of the most accurate options, with the inconvenience of a high overlap rate. On the other hand there are the matching-criteria, which are the opposite pole: high security with losses in RMSE.

For a test length of Q items, accepted that during the test the valuating function will be changed no more than once and that, if changed, it will be from MC to FI and not the reverse, it is possible to define (Q+1) different patterns to define the valuating function according to the item position in the test, from the selection of zero items with FI to the selection of the Q items based on FI. Leung et al. (2005) have not studied all the possible points, but just three of them, the extremes (Q items with FI or Q items with MC) and the central point (a half and a half).

Our goal is to study each possible combination of FI and MC. We consider that the definition of the number of items selected according to each valuating function will make it possible to situate any CAT with flexibility between the limits of accuracy and security that define the two functions. We also wanted to evaluate the effect of the stratification of the bank, checking whether this is an appropriate strategy for defining B_q. With these objectives in mind, we developed a simulation study with two item banks were employed, each one with its own test length. This was done to check if the pattern of results hold under different conditions.

Simulation study

Method

Item banks: The first item bank was composed with items from 9 ACT Mathematics test forms (ACT, 1997). The item parameters of 520 items were available in the documentation of the ICL software (Hanson, 2002). The mean, standard deviation, maximum and minimum for the *a*, *b* and *c* parameters were (1.01, 0.33, 2.35, 0.33), (0.06, 1.12, 2.82, -3.66) and (0.17, 0.08, 0.50, 0.03), respectively. The correlation of the *a* and *b* parameters was .50.

As second item bank, we used eCAT, an item bank for evaluating the knowledge of English grammar (Olea, Abad, Ponsoda, & Ximénez, 2004). The bank had 197 items. Although the parameters of the bank have been recently updated (Abad, Olea, Aguado, Ponsoda, & Barrada, 2010), we used the original ones. The mean, standard deviation, maximum and minimum for the *a*, *b* and *c* parameters were (1.30, 0.32, 2.20, 0.43), (0.23, 1, 3.42, -2.71) and (0.21, 0.03, 0.29, 0.11), respectively. The correlation of the *a* and *b* parameters was .17.

Test lengths: For the ACT bank, the test length was fixed at 30 items, as was done in another study using the same item bank (Ban, Hanson, Wang, Yi, & Harris, 2001). For eCAT, the test length used was 20 items. Although, in practice, the test length of eCAT depends on the needs of the companies which contract its use, this is long enough for the common goals (Olea et al., 2004).

Valuating functions and stratification of the bank: All the possible combinations of FI and MC were simulated (31 for ACT bank and 21 for eCAT). The matching-criterion function used was TM-MC. The item bank was divided into one, two and five strata. The items presented from each stratum were equal to test length divided by S (number of strata). For the ACT item bank, each stratum was composed of the same number of items. For eCAT, when S was two, 99 items belonged to the first stratum and 98 to the second; when S was set to five, 39 items corresponded for each of the even strata and 40 to the odd strata. The bank was stratified according to $I(\theta)_{max}$ while blocking in accordance with θ_{max}. In other words, the method used both for stratification and selection was the one proposed by Barrada et al. (2006). We have given the name L05 to Leung et al.'s proposal (2005) of using two strata and applying as a valuating function in the first half of the test the FI and MC in the second half. L05 is not a different condition incorporated in our studies, but a condition included in our exhaustive manipulation of the number of items valuated with each function.

Trait level of the simulees and starting rule: 500,000 simulees were sampled. Its real trait level was randomly extracted from a distribution $N(0, 1)$. For each condition [2 (number of item banks) * 31 or 21 (number of items selected with each valuating function) * 3 (number of strata)], the same 500,000 simulees were employed. The starting $\hat{\theta}$ was chosen at random from the interval (-0.5, 0.5).

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. In order to avoid this, 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 $(\hat{\theta} - b_{min})/2$, where b_{max} and b_{min} correspond to the maximum and minimum b parameter values in the bank. After the constant pattern was broken 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: Two dependent variables were used for the comparison between conditions: RMSE and overlap rate, calculated with the Equation 1 and 2.

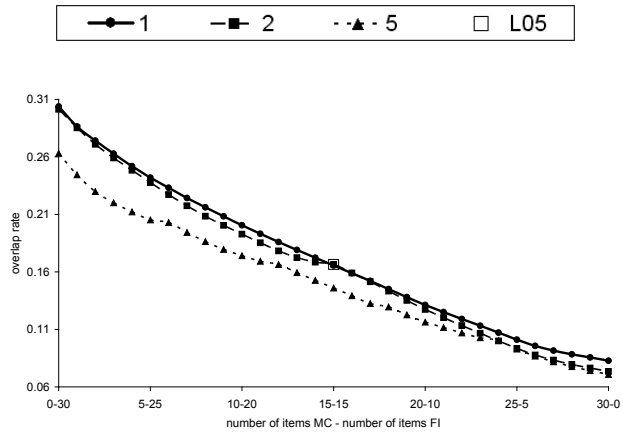
Results

In Figure 1 overlap rates for both item banks according to the number of items selected with MC and FI and to the number of strata are shown. As we move from left to right in the axis of abscissas, the number of items selected with MC is increased and, therefore, the number of items selected with FI is reduced. Given the item bank size and the test length, the minimum overlap rate possible (Q/n) was equal to .06 for the ACT bank and .10 for eCAT.

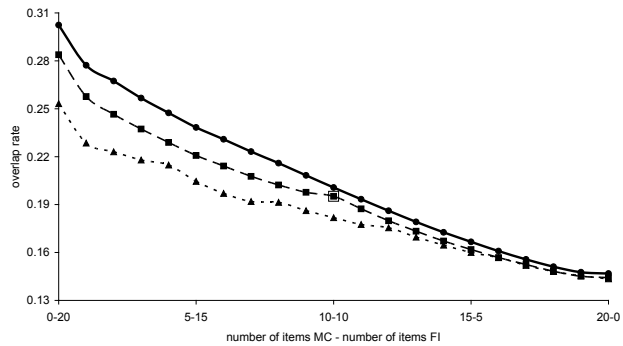
For both item banks, the pattern of results was basically the same. As expected, as we increased the number of items selected according to MC, overlap rate was reduced. When FI was the only valuating function employed, the overlap rates obtained were over the values considered acceptable (Way, 1998). When MC was the valuating function applied throughout the test, overlap was located near the minimum possible value. With L05, the overlap was in the middle of the two extremes, slightly nearer the condition of only using MC.

The differences in the overlap according to the number of strata are clear when the items were selected completely or basically with FI. In these cases, a greater number of strata implied a lower overlap. These differences were reduced as the number of items selected by means of MC was increased.

In Figure 2, the same conditions as in Figure 1 are shown, but with RMSE as the dependent variable. Again, the results for eCAT mimic the results obtained with the ACT bank, with the difference that the latter achieves a higher accuracy, as more items are administered. As expected, reductions in the number of items selected with FI implied increments in RMSE. The greater the number of strata, the lower the difference between a selection solely based on FI and the selection solely based on MC. A possible explanation for this is that, with a high number of strata, it is guaranteed that all the examinees receive some highly informative items. When just one stratum was used and MC was the valuating function employed, one examinee could, through chance, receive only poorly informative items.



(a) ACT item bank.



(b) eCAT.

Figure 1. Overlap rate according to the number of items selected with MC and FI and to the number of strata.

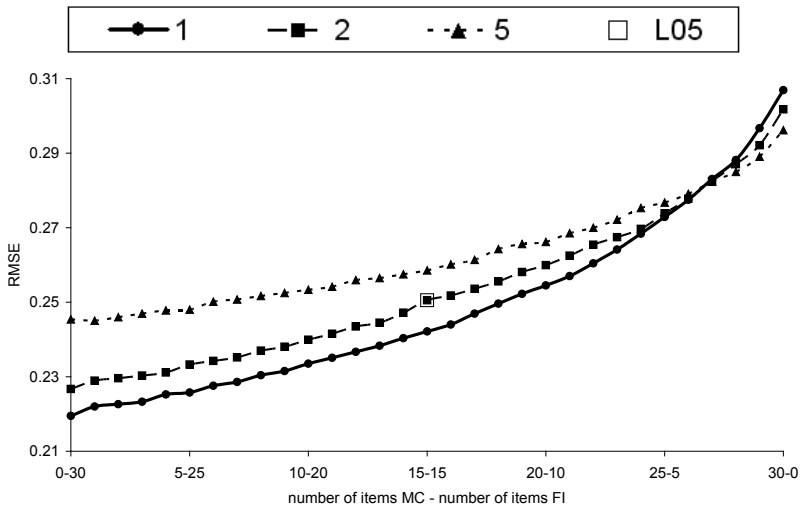
When all or most of the items were selected with FI, a greater number of strata meant lower accuracy. The restriction of not always having the best items available for selection worsens the results. On the contrary, when most of the items were selected by means of MC, a greater number of strata implied better accuracy. The L05 condition offers slightly worse RMSE than the condition of using just FI.

The information, as presented in Figures 1 and 2, allows us to evaluate the impact of the different variables manipulated. However, this way of presenting the information makes decision making complex. Our goal, in a practical setting, is to decide, for a given RMSE or overlap rate value, which of the different alternatives is the one that offers a lower value in the other variable. For this, Figure 3 was built, where the scatter plot of both dependent variables is shown for the three different numbers of strata simulated. The marks (squares, triangles and circles) on the left correspond to the selection completely based on FI. The ones on the right, are selected solely according to MC.Keeping the trade-off between accuracy and security commented by some authors and what we have seen in

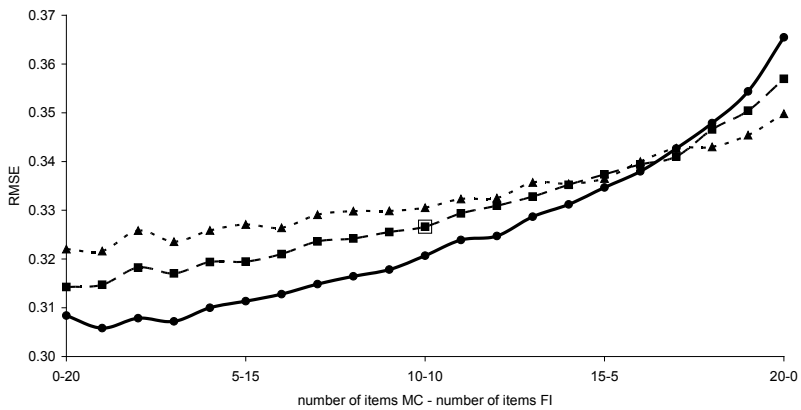
Figures 1 and 2, reductions in the overlap rate meant increases in RMSE. However, the relation between overlap and RMSE was not lineal, but seems more logarithmic (with a logarithm base lower than 1). This means that, with respect to the selection completely based on FI, major improvements in the overlap rate had small effects in RMSE. In the opposite case, when attending to the results when the whole selection is done with MC, small increments in the overlap rate imply high reductions in RMSE. So, it is possible, in comparison with a selection solely based on FI, to substantially improve the item bank security with effects in the accuracy that can be considered negligible. We can also see that it is a better option not to stratify the bank, unless we are interested in overlap rate values near the minimum possible. According to this figure, and although the differences are minimal, it seems that L05 could be a method that is never preferred, because there are some other conditions that can achieve the same overlap rate with lower RMSE.

Discussion and conclusions

Our goal was to study the effect of the variation of the item valuating functions and the use of changing B_q composition throughout a CAT. To do this, we have chosen MC and FI as valuating functions, following Leung et al. (2005). The first one represents the option of strict control of item bank security with important losses in measurement accuracy. The valuating function FI means a very accurate and very risky option for item bank security. To define the content of B_q we have used the stratifying strategy, as suggested by H. H. Chang and Ying (1999) or Hau and Chang (2001), among others.



(a) ACT item bank.



(b) eCAT.

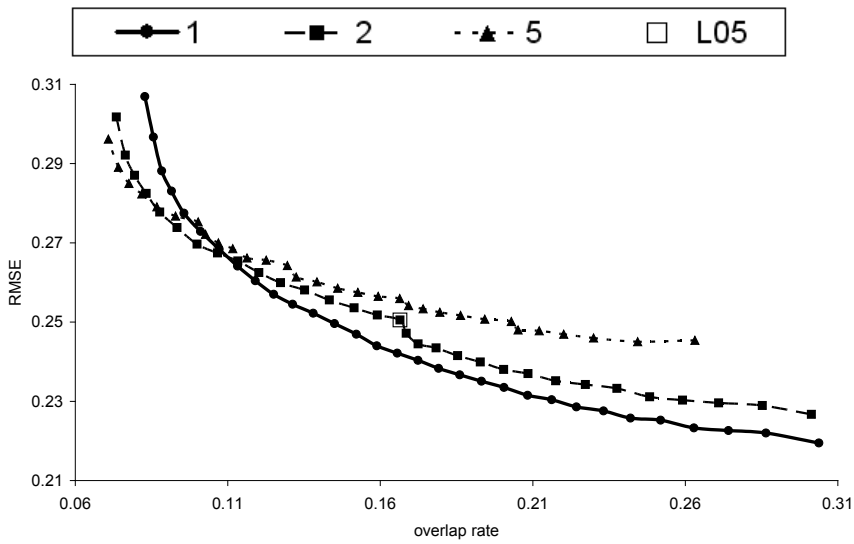
Figure 2. RMSE according to the number of items selected with MC and FI and to the number of strata.

Our results were consistent among the two item banks and test lengths used. It is possible, by defining the number of items to be selected according to each valuating function, to obtain the security or accuracy desired for our CAT between the extremes of FI and MC. If accuracy is a priority over security, we should choose FI as the basic function in our test. Otherwise, we should increase the number of items selected by means of MC. We consider our proposal to be more flexible than the one suggested by Leung et al. (2005).

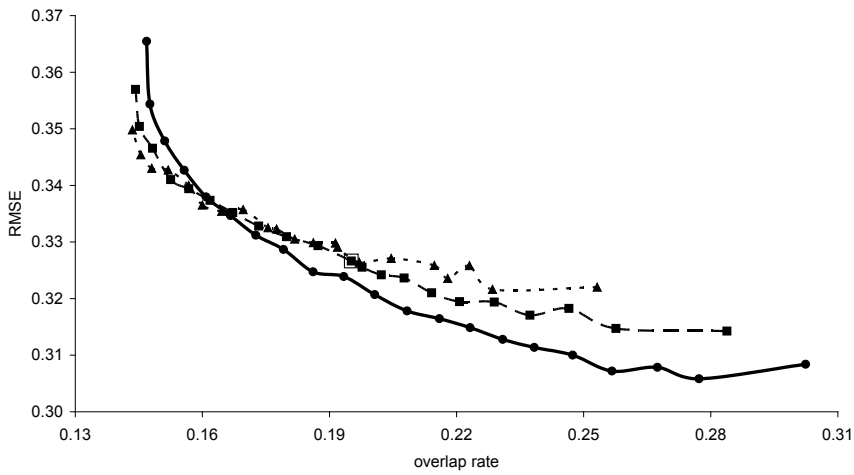
Our results also indicate that it is possible to markedly improve bank security with very low losses in accuracy. Taking this into consideration, we consider that the selection based solely on FI should only be used in exceptional cases where no risk at all is expected for the integrity of the item bank. It should be noted that, for these cases, item selection rules more accurate than FI are available (Barrada, Olea, Ponsoda, & Abad, 2010).

As opposed to the proposals to stratify the item bank, our data suggests that, unless we are interested in an overlap rate near the minimum possible, it is more convenient not to stratify the bank.

The presentation of data, as has been done in Figure 3, facilitates decision making, as it simultaneously includes the information of both criteria variables, overlap rate and RMSE. In this way, it is possible to choose the most convenient selection strategy for a specific CAT use.



(a) ACT item bank.



(b) eCAT.

Figure 3. Overlap rate and RMSE according to the number of items selected with MC and FI and to the number of strata.

References

Abad, F. J., Olea, J., Aguado, D., Ponsoda, V., & Barrada, J. R. (2010). Deterioro de parámetros de los ítems en tests adaptativos informatizados: Estudio con eCAT [Item parameter drift in computerized adaptive testing: Study with eCAT]. *Psicothema, 22*, 340-347.

ACT, Inc. (1997). *ACT assessment technical manual*. Iowa City, IA: Author.

Ban, J., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item-calibration/scaling methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 191-212. doi:10.1111/j.1745-3984.2001.tb01123.x

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., Ponsoda, V., & Abad, F. J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement, 34*, 438-452. doi:10.1177/0146621610370152

Barrada, J. R., Veldkamp, B. P., & Olea, J. (2009). Multiple maximum exposure rates in computerized adaptive testing. *Applied Psychological Measurement, 33*, 58-73. doi:10.1177/0146621608315329

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., Qian, J., & Ying, Z. (2001). a-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement, 25*, 333-341. doi:10.1177/01466210122032181

Chang, H. H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229. doi:10.1177/014662169602000303

Chang, H. H., & Ying, Z. (1999). a-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211-222. doi:10.1177/01466219922031338

Chang, H. H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67*, 387-398. doi:10.1007/BF02294991

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. doi:10.1111/j.1745-3984.2003.tb01097.x

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.doi:10.1111/j.1745-3984.2003.tb01100.x

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. doi:10.1177/014662169001400403

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

Hanson, B. A. (2002). *IRT Command Language.* Computer software manual. Retrieved from: http://www.b-a-h.com/software/irt/icl/icl_manual.pdf

Hau, K. T., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38*, 249-266. doi:10.1111/j.1745-3984.2001.tb01126.x

Leung, C. K., Chang, H. H., & Hau, K. T. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement, 26*, 376-392. doi:10.1177/014662102237795

Leung, C. K., Chang, H. H., & Hau, K. T. (2005). Computerized adaptive testing: a mixture item selection approach for constrained situations. *British Journal of Mathematical and Statistical Psychology, 58*, 239-257. doi:10.1348/000711005X62945

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. doi:10.1111/j.1745-3984.2005.00013.x

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

Olea, J., Abad, F. J., Ponsoda, V., & Ximénez, M. C. (2004). Un test adaptativo informatizado para evaluar el conocimiento de ingles escrito: Diseño y comprobaciones psicométricas [A computerized adaptive test for the assessment of written English: Design and psychometric properties]. *Psicothema, 16*, 519-525.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-327. doi:10.1111/j.1745-3984.1998.tb00541.x

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: Kluwer Academic.

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

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201-216. doi:10.1007/BF02294775

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. doi:10.3102/10769986028003249

van der Linden, W. J., & Glas, C. A. W. (Eds.) (2010). *Elements of adaptive testing*. New York, NY: Springer.

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational & Behavioral Statistics, 29*, 273-291. doi:10.3102/10769986029003273

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new ítem selection criteria for adaptive testing. *Journal of Educational & Behavioral Statistics, 22*, 203-226. doi:10.3102/10769986022002203

Way, W. D. (1998). Protecting the integrity of computerized testing ítem pools. *Educational Measurement: Issues and Practice, 17*, 17-27. doi:10.1111/j.1745-3992.1998.tb00632.x

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. doi:10.1177/014662168400800312

Received October 23, 2009
Revision received April 13, 2010
Accepted June 14, 2010