Psicothema 2006. Vol. 18, nº 1, pp. 156-159
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

SOFTWARE, INSTRUMENTACIÓN Y METODOLOGÍA

Maximum information stratification method for controlling item exposure in computerized adaptive testing

Juan Ramón Barrada, Paloma Mazuela and Julio Olea
Universidad Autónoma de Madrid

The proposal for increasing the security in Computerized Adaptive Tests that has received most attention in recent years is the a-stratified method (AS - Chang and Ying, 1999): at the beginning of the test only items with low discrimination parameters (*a*) can be administered, with the values of the *a* parameters increasing as the test goes on. With this method, distribution of the exposure rates of the items is less skewed, while efficiency is maintained in trait-level estimation. The pseudo-guessing parameter (*c*), present in the three-parameter logistic model, is considered irrelevant, and is not used in the AS method. The Maximum Information Stratified (MIS) model incorporates the *c* parameter in the stratification of the bank and in the item-selection rule, improving accuracy by comparison with the AS, for item banks with *a* and *b* parameters correlated and uncorrelated. For both kinds of banks, the blocking *b* methods (Chang, Qian and Ying, 2001) improve the security of the item bank.

Método de estratificación por máxima información para el control de la exposición en tests adaptativos informatizados. La propuesta para aumentar la seguridad en los tests adaptativos informatizados que ha recibido más atención en los últimos años ha sido el método *a*-estratificado (AE - Chang y Ying, 1999): en los momentos iniciales del test sólo pueden administrarse ítems con bajos parámetros de discriminación (*a*), incrementándose los valores del parámetro *a* admisibles según avanza el test. Con este método la distribución de las tasas de exposición de los ítems es más equilibrada, manteniendo una adecuada precisión en la medida. El parámetro de pseudoadivinación (*c*), presente en el modelo logístico de tres parámetros, se supone irrelevante y no se incorpora en el AE. El método de Estratificación por Máxima Información (EMI) incorpora el parámetro *c* a la estratificación del banco y a la regla de selección de ítems, mejorando la precisión en comparación con AE, tanto para bancos donde los parámetros *a* y *b* correlacionan como para bancos donde no. Para ambos tipos de bancos, los métodos de bloqueo de *b* (Chang, Qian y Ying, 2001) mejoran la seguridad del banco.

One of the most significant advances of the last decade in psychometric practice (Hambleton, 2004) has been the more generalized application of Computerized Adaptive Tests (CATs). In these tests, the next item to be presented to an examinee is selected from an item bank according to performance on the items answered previously. In this way, we obtain quicker and/or more reliable measures of the trait levels of examinees than with conventional paper-and-pencil tests.

CATs can be described as an iterative process of [estimation of the trait level ($\hat{\theta}$) - selection of the next item]. A standard approach to item selection has been to select the item with the maximum Fisher information as the next item (Lord, 1977). In doing so, certain items tend to be used more often than others, while some are never presented, making item exposure rates quite uneven. This has resulted in two main problems, the first economic, given the money spent on developing the unused items, and the second security-related, because of the risk of item-sharing among the often-used items.

Various alternative item selection rules have been proposed to remedy this situation, some dealing with underexposure (progressive method – Revuelta and Ponsoda, 1998; *a*-stratified method – Chang, Qian and Ying, 2001; Chang and Van der Linden, 2003; Chang and Ying, 1999) and others dealing with overexposure (restricted method – Revuelta and Ponsoda, 1998; Sympson-Hetter method – Sympson and Hetter, 1985; van der Linden, 2003). That which has probably aroused most interest in

Fecha recepción: 25-11-04 • Fecha aceptación: 7-6-05
Correspondencia: Juan Ramón Barrada
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
E-mail: juanra.barrada@uam.es

the last five years is the a-stratified (AS) method. It is applied as follows:

1. Prior to administration of the test to any examinee (fixed-length test of L items), we proceed to the stratification of the item bank of m items:
   a. The number of strata (s), the number of items belonging to each stratum (ni_s) and the number of items to be administered for each stratum (na_s) are defined in such a way that $\sum_{i=1}^{s} ni_i = m$ and $\sum_{i=1}^{s} na_i = L$.
   b. The items in the bank are arranged in increasing order according to their value in the item discrimination (a) parameter.
   c. The first ni_1 items belong to the first stratum; the next ni_2 items according to the value in the a parameter belong to stratum 2; and so on, until the final ni_s belonging to stratum s.

2. For administration of the test to an examinee, item selection is carried out as follows:
   a. The first na_1 can only be selected from stratum 1, the next na_2 belongs to stratum 2… and the na_s belong to stratum s.
   b. The selected item is that which minimizes the difference, in absolute value, between $\hat{\theta}$ and the difficulty (b) parameter of the item.

With the AS method, at the beginning of the test, the items used are those never usually employed with the maximum Fisher Information rule; items with high a values are left for the final part of the test, when the differences between $\hat{\theta}$ and θ are assumed to be small, so that these items are more appropriate. Chang and Ying (1999) showed how this method greatly balanced item usage within the pool, while maintaining accuracy in trait estimation.

However, the AS method assumes that the distribution of the b values among strata will be basically the same, and this does not hold when a and b are correlated. In practice, a and b parameter estimates are often positively correlated (Wingersky and Lord, 1984). In order to deal with this, Chang, Qian and Ying (2001) developed the AS method with b blocking (AS-B). The basic idea is to force each stratum to have a balanced distribution of b values. The strata are created as follows (assuming that ni_i and na_i are constant in all the strata):

1. Divide the item bank into m/n_s blocks, in such a way that the first block contains items with the lowest b values and the (m/n_s)th block contains items with the highest b values.
2. Arrange the items within each block according to their increasing a value.
3. Combine all the first items of each block to form the first stratum, the second ones to form the second stratum… and so on, until the sth items are combined to form the s stratum. The selection rule applied in the AS method with b blocking is the same as that used in the AS method: select the item that minimizes $|\hat{\theta} - b|$.

Chang, Qian and Ying (2001) showed that the AS-B method outperformed the AS method in precision and exposure control when an item pool with correlated a and b item parameter estimates was used.

As can be seen from the description of the two methods, they take into account just two parameters: a and b. However, the three-parameter logistic model (3PLM) has one more parameter, the pseudo-guessing parameter (c), not used by AS and AS-B for either the stratification or the selection of items. As far as we know, there has been no attempt to incorporate the c parameter into the stratification in CATs. In fact, Chang and Ying (1999) considered it as basically irrelevant.

When the c parameter is taken into account, two principles present in the 2PLM no longer hold (Hambleton and Swaminathan, 1985). First, in the 2PLM the ranks of each item of the bank according to their a parameters and according to their maximum in the Fisher information function ($I(\theta)_{max}$) are the same. This is not true in the 3PLM. Second, the maximum of the item Fisher information function (θ_{max}) is no longer attained in b, as is the case in the 2PLM. These two differences can cause AS and AS-B to perform below their optimum when the 3PLM is employed.

Two simple modifications are introduced into the methods for incorporating the c parameter. Instead of using the a parameter for stratifying the item bank, the proposal is to substitute it by the maximum attained by an item in the Fisher information function $I(\theta)_{max}$. This value is given in Equation 1.

$$I(\theta)_{max} = \frac{1.7^2 a^2}{8(1-c^2)}\left[1 - 20c - 8c^2 + (1+8c)^{3/2}\right] \quad (1)$$

$I(\theta)_{max}$ increases as the discrimination parameter increases, and decreases as the c parameter approaches 1.

Secondly, we will substitute the b value in the selection rule of items in AS and AS-B and in the stratifying process in AS-B by θ_{max}. The θ value where θ_{max} is attained is given in Equation 2.

$$\theta_{max} = b + \frac{1n\left[1 + (1+8c)^{1/2}\right] - 1n(2)}{1.7a} \quad (2)$$

θ_{max} will always be shifted to the right by comparison with b. The difference between θ_{max} and b is related positively to the value in the c parameter and negatively related to the value in the a parameter.

Because of these two differences from the AS method, we shall call our alternative method Maximum Information Stratified (MIS). Keeping the same logic as in the AS methods, two item selection rules are proposed: one without blocking θ_{max} (MIS-NOB) and the other with blocking (MIS-B).

Because MIS uses the available information of the item parameters in a more exhaustive way, an improvement in the accuracy achieved with it, compared to the AS method, is expected. The size of this expected effect was investigated through simulation studies.

Method

Item banks: two kinds of item banks were randomly generated. In the first of them, there was no correlation between a and b parameters. In the second, the correlation between a and b values was 0.5. Twenty item banks of 250 items were generated, ten of each kind. The distributions for the parameters were: $a \sim N(1.2, 0.25)$; $b \sim N(0, 1)$; $c \sim N(0.25, 0.02)$.

Trait level of the simulees, test length and starting rule: the trait level of the simulees was randomly generated for a population N(0, 1). For each one of the twenty item banks, 5000 simulees were sampled. The test length was fixed at 30 items. The starting $\hat{\theta}$ was chosen at random from the interval (-0.5, 0.5).

Stratifying of the banks: the bank was divided into five strata, with 50 items in each. Six items of each stratum were administered to each examinee.

Estimation/assignment of trait level: maximum-likelihood estimation has no solution in the 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, 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 $(\hat{\theta} - b_{min})/2$. Since the constant pattern was broken, we applied maximum-likelihood estimation (Birnbaum, 1968).

Performance measures: two dependent variables were used for the comparison between methods: RMSE for the accuracy and χ^2 to measure the skewness of the exposure rate of the items.

RMSE

$$RMSE = \left(\frac{\sum_{i=1}^{r} \left(\hat{\theta}_i - \theta_i \right)^2}{r} \right)^{1/2} \quad (3)$$

where *r* is the number of simulees.

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(er_i - L/m \right)^2}{L/m} \quad (4)$$

where er_i is the observed exposure rate of the *i*th item.

χ^2 measures the discrepancy between the observed and ideal item exposure rates and quantifies the efficiency of item bank usage (Chang and Ying, 1999).

Results

Table 1 summarizes the simulation results. We shall present them according to the different independent manipulations we introduced.

Table 1
RMSE and χ^2 for AS and MIS with and without blocking

		RMSE		χ^2	
		AS	**MIS**	**AS**	**MIS**
r_{ab} = 0.0	NO-B	0.289	0.274	8.787	8.900
	B	0.290	0.274	4.767	4.913
r_{ab} = 0.5	NO-B	0.302	0.282	18.035	17.950
	B	0.291	0.279	4.658	4.895

Effects of blocking: as expected, blocking (*b* or θ_{max}) for stratifying the item bank when r_{ab} was equal to 0.0 had no effect in the observed RMSE by comparison with the NO-B condition. When the *a* and *b* parameters correlated, the methods that employed strata-generated blocking outperformed the ones that did not.

The B conditions, by comparison with the NO-B conditions, always presented lower χ^2 values. In accordance with Chang, Qian and Ying (2001), these were the expected results when *a* and *b* correlated. About 74% of the skewness in the B methods was reduced relative to the AS method ($1-\chi_{BB}^2 / \chi_{NO\text{-}BB}^2$) with r_{ab} equal to 0.5.

The unexpected result was that B also improved the exposure control when *a* and *b* parameters were uncorrelated. In fact, under this condition B reduced approximately 45% of the skewness of the distribution of exposure rates when no B was applied. After some consideration of this surprising result, a possible explanation was proposed. Let us imagine three items assigned to the same strata with *b* values of (1, 1.1, 1.2). The interval of $\hat{\theta}$ that would lead to the selection of the second item is quite narrow, just when $\hat{\theta} \in [1.05, 1.15]$. When we stratify the items blocking *b*, these three items would be assigned to different strata, and the variance of the interval width that leads to selection of each item is expected to be reduced.

Figure 1 illustrates the distribution of item exposure rates for the four different methods and for the two kinds of banks. As can be seen there, the distribution in the unblocked conditions is more skewed, with more underexposed and overexposed items, by comparison with the blocked methods.

Effects of taking into account c parameter: increasing the information employed for the stratifying and selection of items, with the incorporation of the *c* parameter for both processes, improved accuracy of the estimation of trait level. For all the

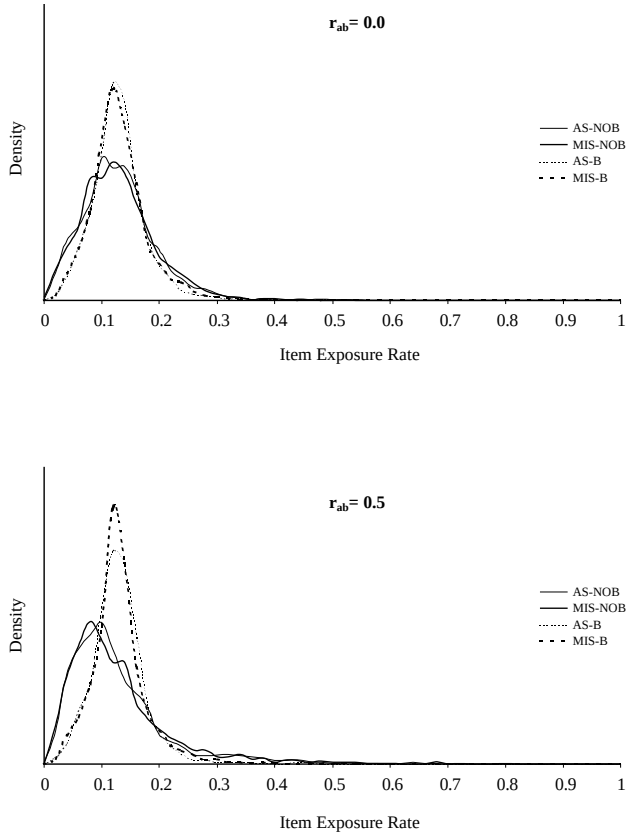

Figure 1. *Distribution of the item exposure rates for a and b parameters uncorrelated and correlated by method*

evaluated conditions, $RMSE_{MIS}$ was lower than $RMSE_{AS}$. Overall, MIS reduced the RMSE of AS by 5%.

Incorporating the *c* parameter into the methods slightly reduced item exposure control for three of the four conditions. Solely when *a* and *b* correlated and no blocking was applied, was χ_{MIE}^2 smaller than χ_{AE}^2. The exposure control that can be achieved with this stratifying approach is conditioned by the extent to which the distribution of the *b* parameters or θ_{max} values is similar to the distribution of the $\hat{\theta}$ (Chang and Ying, 1999; Cheng and Liou, 2003). As can be derived from (2), θ_{max} is always greater than *b*, so that, for an item bank with *b* parameters following the standard normal distribution, the distribution of θ_{max} will not be distributed N(0, 1). In the item banks employed in the simulations, the distribution of θ_{max} was N(0.16, 1). As in this study the real trait levels were generated from a standard normal distribution, this discrepancy between distributions can be assumed as the reason for the slightly greater χ^2 with MIS than with AS.

As observed in Figure 1, differences between the AS and MIS methods are quite negligible in the distribution of their item exposure rates.

Discussion

The purpose of this study was to check whether, as Chang and Yi (1999) noted, and has been assumed since then, incorporating the *c* parameter into the stratifying approach in CATs is irrelevant. In order to check this, we changed the way the item bank was stratified, taking into account not the *a* parameters, but $I(\theta)_{max}$, and we changed the item selection rule, choosing not the item with the *b* parameter closest to $\hat{\theta}$ but that with θ_{max} closest to $\hat{\theta}$. As can be seen in the simulation results, using all the available information of the item bank with the MIS method improved the accuracy of the trait estimations when compared with the AS method. Although MIS, in general, slightly decreased the extent of the exposure control achieved with AS, both of these methods, when a blocking strategy was applied for stratifying, attained a performance very close to perfect. Another relevant finding is the importance of blocking for stratifying the item bank. Chang *et al* (2001) showed that doing so is important when the *a* and *b* parameters of the item bank are correlated. Our study suggests that blocking is also useful when there is no correlation. Based on all these results, our recommendation is to use MIS-B whenever a stratifying methodology is chosen for the exposure control in CATs.

The AS method has been developed in recent years to incorporate content control in CATs (Leung, Chang and Hau, 2003; van der Linden and Chang, 2003), the use of linear programming for stratifying the pool (Chang and Van der Linden, 2003) or the imposition of a maximum exposure rate (Leung, Chang and Hau, 2002; Parshall, Harmes and Kromrey, 2000), or to adapt the method to variable-length CATs (Wen, Chang and Hau, 2000). The MIS can easily incorporate all these improvements of the original AS method. Furthermore, all the open issues in relation to the AS method are also relevant to the MIS method: the optimal number of strata, the minimum acceptable a values, which characteristics of the item pool would make it unsuitable for the stratifying methods, and so on.

Acknowledgements

This research has been supported in part by a DGES-MEC grant (project BSO2002-01485).

References

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

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

Chang, H.-H. and van der Linden, W.J. (2003). Optimal stratification of item pools in alpha-stratified computerized adaptive testing. *Applied Psychological Measurement, 27*, 262-274.

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

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

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. (2004). Theory, methods and practices in testing for the 21st century. *Psicothema, 16*, 696-701.

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

Hau, K.-T. and Chang, H.-H. (2001). Item selection in computerized adaptive testing: should more discriminating items be used first? *Journal of Educational Measurement, 38*, 249-266

Leung, C.-K., Chang, H.-H. and 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.

Leung, C.-K., Chang, H.-H. and Hau, K.-T. (2003). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement, 63*, 257-270.

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

Parshall, C., Harmes, J.C. and Kromrey, J.D .(2000). Item exposure control in computer-adaptive testing: the use of freezing to augment stratification. *Florida Journal of Educational Research, 40*, 28-52.

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

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

Sympson, J.B. and 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: 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. and Chang, H.-H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement, 27*, 107-120.

Wen, J.-B., Chang, H.H. and Hau, K.-T. (2000). *Adaptation of the a-stratified method in variable length computerized adaptive testing.* Paper presented at the American Educational Research Association Annual Meeting.

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