Revista Electrónica de Metodología Aplicada
2011, Vol. 16 nº 1, pp. 50-65

Development, psychometric properties and new validity evidences of the web-based computerized adaptive test of English eCat

Julio Olea[1], Francisco José Abad[1], Vicente Ponsoda[1], David Aguado[2] y Julia Díaz[2]

[1]Facultad de Psicología, Universidad Autónoma de Madrid
[2]Instituto de Ingeniería del Conocimiento

RESUMEN

Se describe la construcción del test eCat y se informa de las consecuencias que su funcionamiento ha tenido en sus propiedades psicométricas. Los resultados obtenidos en una muestra de 3224 estudiantes mostraron que el banco tiene una adecuada fiabilidad y validez convergente en relación a los autoinformes de dominio del inglés. Las simulaciones realizadas mostraron que las estimaciones carecían de sesgo y que el test ha de tener al menos 15 ítems para alcanzar una precisión razonable. A partir de las 7254 aplicaciones del test disponibles, se han vuelto a obtener las propiedades psicométricas y se han comparado con las previstas antes de la aplicación. Se aportan nuevas evidencias de validez.

Palabras clave: Test adaptativo informatizado, tests por internet, evaluación del inglés, selección de personal.

ABSTRACT

This paper describes the process of constructing the eCat test and provides information on the effects its normal operation had on its psychometric properties. Results obtained from 3224 students revealed that the 225 item pool has adequate reliability and good convergent validity with respect to self-reported English proficiency. Simulations showed that ability estimations are essentially unbiased and that a stopping criterion of over 15 items is required to achieve a reasonable level of precision. From 7254 eCat administrations the test psychometric properties have been computed again and compared with those of the preoperational eCat. Additional validity evidences are also presented.

Keywords: Computerized adaptive testing, testing by the Internet, English assessment, e-recruitment.

Contacto:
Vicente Ponsoda
e-mail: Vicente.ponsoda@uam.es
tel: 914975203 fax: 914975215

This research has been funded by the Ministerio de Ciencia e Innovacion (grants PSI2008-01685 and PSI2009-10341) and by the Chair "Psychometric models and applications" sponsored by the Instituto de Ingeniería del Conocimiento.

1.- Introduction

One of the most important advances in the theory and practice of personnel recruitment has very likely been caused by the arrival of computer and related technologies (Viswesvaran, 2003). Computerized testing via the Internet for personnel selection is being more and more popular in the last years and is one of the preferred organization practices (Bartram & Hambleton, 2006; Nye, Do, Drasgow & Fine 2008). Internet testing has important advantages, as quicker and cheaper assessment processes (Naglieri, Drasgow, Schmite, Handler, Prifitera, Margolis & Velasquez, 2004), and it makes it possible that candidates anywhere in the world can easily apply and be tested even at their own homes. However, they have also some difficulties, related mainly to the unsupervised nature of most assessments and the lack of control on the examinee behavior when responding to the test (Tippins, Beaty, Drasgow, Gibson, Pearlman, Segall & Shepherd, 2006).

Different test types, as linear tests, questionnaires and computerized adaptive tests (CAT) can be and usually are administered by the Internet. Advances in CAT have been made possible by advances in both item response theory (IRT) and in computer hardware (Ponsoda & Olea, 2003). The main idea of a CAT is to efficiently estimate the respondent ability through administering items matched to the level of proficiency demonstrated by him or her throughout the test. The respondent will not have to respond to items that very likely will pass or fail for being too easy or difficult for him or her. A CAT in general administers a reduced number of items but its precision could be even higher than of a longer linear test. The use of CATs is increasingly common in large-scale psychological and educational, certification and licensure programs, where the fundamental objective is to obtain a precise estimate of ability through the application of a reduced number of items. There are currently adaptive versions of several important knowledge and aptitude tests (such as the ASVAB, the GRE, SAT and TOEFL). The "move to CAT" is a worldwide phenomenon. There are more than 30 operational CAT programs[1] all over the world that evaluate four to six million men, women, and children each year (Fetzer, Dainis, Lambert and Meade, 2008).

CATs have the potential advantages of any computerized test (Carlson & Harvey, 2004) such as, for example, those relating to data storage and recovery, homogeneity of the testing conditions, establishment of controls to preserve the security of the test, speed in data processing, assessment through new item formats, more flexible grading procedures, preparation of automatic reports, etc. In addition to these potential advantages, CATs hold other, more specific ones: enhanced test security, reduced application time, more precise ability estimates for tests of the same length, the possibility of applying different tests to the same examinee on different occasions, and, in the long run, lower testing costs when it is necessary to assess samples of great size.

There are however specific difficulties related to CAT implementation, as it brings about the technical issue of dynamic transmission of information between the computer the examinee responds to and a server, which carries out the performance estimation process, selection of new items, their display to the examinee, data recording, and preparation of the end report. For the examinee, the time interval between their response and the display of the next item should not be perceptible. These issues become especially important when a large number of test-takers simultaneously take the adaptive test. In addition, the issues of

[1] For a list of more than 20 programs see the address http://www.psych.umn.edu/psylabs/catcentral/

preservation of the security of the item pool and anonymity of the test-takers require the establishment of specific access controls to the system.

In spite of these difficulties, web-based CATs may represent, especially for high stakes testing situations with numerous samples and several applications every year, significant progress towards the achievement of some objectives pursued by computerized assessment. Web-based CAT supplies test-takers with the most convenient and flexible conditions in which to take a test. It also provides the contracting company with advantages, such as a simplification of procedures, an enhanced company image, and an improved on-demand service.

This paper describes eCat, one of the CAT programs referred to above, web-based, developed by psychometricians and computer scientists from the Universidad Autónoma de Madrid and the Instituto de Ingeniería del Conocimiento, that assesses English proficiency. The test is operational from 2006 and it is being mostly applied in unsupervised personnel selection processes. We will describe the construction and calibration process of the item pool, the operation of the management system, the adaptive algorithms implemented, and its psychometric properties. Some validity evidences recently gathered and the results of a recalibration conducted on the responses given to the operational test will be also offered to ascertain the effects the test normal operation have on its psychometric properties.

2.- eCat Development

2.1. – Design of the item bank

The efficiency of a CAT tightly depends on the quality of the item bank from which the items have to be selected. There are a few important issues related to the item pool developing and maintenance (Wise and Kingsbury, 2000). In the eCat developing process, much attention was paid to create an appropriate item pool.

Two specialists in English Philology, in collaboration with the group of psychometricians, designed a pool containing 635 items. The psychometricians guided the item development process. They proposed a multiple choice format with 4 response options for the items, and provided guidelines for the wording of incorrect options and recommendations on the content validity of the pool and on the desirable difficulty of the items. The philologists designed a cognitive-functional model of English proficiency, that included 7 grammar categories and 46 more specific subcategories: form-related aspects (2 subcategories, 17 items in total), morphology (17 subcategories, 222 items), morphosyntax (1 subcategory, 7 items), pragmatics (2 subcategories, 20 items), lexicon (7 subcategories, 177 items), syntax (14 subcategories, 82 items), and compound categories (3 subcategories, 110 items).

For the calibration of the items, it was not possible to apply the entire bank to each examinee. Therefore, a linking design was used in order to administer different subsets of items to different samples, with linking items common to the different forms. In all, 15 forms were designed, each made up of 61 items, 20 common (the linking test) and 41 specific to each form. Both the items of the linking test and those belonging to each specific form were

chosen to be adequate samples of the difficulty of the pool and of the proportion of items in each of the 7 competency categories.

2.2.- Psychometric properties of the item bank

Five of the forms – a total of 225 items - were administered to a sample of 3224 first year students of the Pontificia Universidad Católica de Chile, who had varying levels of training in the English language. With the objective of carrying out predictive validity studies, a brief questionnaire was included with the corresponding form. It was designed to obtain information on the type of secondary school they attended (bilingual or Spanish), on the type of English language training (formal education, language academies, family, stays in English-speaking countries, etc.). A self-assessment of language proficiency in terms of reading, writing and spoken was also requested.

Several studies were carried out to ascertain the psychometric properties of the 20 items comprising the linking test and of the items of the different forms. These studies' specific aims were to a) estimate the item-test correlations, and to remove the items negatively affecting the form's reliability, b) estimate the internal consistency and mean difficulty of the five forms, to see whether equating procedures were needed, and c) verify the unidimensional assumption required by the IRT model used in the calibration. Table 1 provides the means and standard deviations, as well as the coefficient α, the average of the item-test biserial correlations, and the RMSEA index of each of the forms.

Form	No. of items	Average correct answers	Standard deviation	α	Biserial average	RMSEA
1	61	30.589	13.480	0.947	0.643	0.00702
2	61	28.443	13.037	0.944	0.615	0.00683
3	61	30.174	13.492	0.948	0.642	0.00728
4	61	31.773	14.516	0.955	0.668	0.00753
5	61	32.161	14.454	0.953	0.661	0.00750
Linking	20	9.912	4.956	0.871	0.691	0.00473

Table 1. Descriptive data, internal consistency and fit to the unidimensional solution indices for the 5 forms and the linking test

In spite of the random assignment of subjects to the 5 forms, the average number of correct responses significantly differed ($p<0.01$). It was therefore necessary to equate the metric of items and people parameters from the different forms. The chosen linking design makes this required equating possible, because this possibility was considered when the decision on the linking design was taken. The internal consistency of the different forms and of the linking test, as well as the average values obtained for the item-total biserial correlations, indicate a strong average covariation between the items of each form. There is little room left for improvement in the internal consistency indices.

In order to study the degree of unidimensionality of the different forms, factorial (exploratory and confirmatory) studies were carried out, using the programs LISREL (Jöreskog & Sörbom, 1996) and NOHARM (Fraser, 1988). Very acceptable goodness of fit indicators under the hypothesis of unidimensionality were obtained. More specifications about the results (models, estimation methods and fit tests) can be found in Olea, Abad, Ponsoda and Ximénez (2004).

The results were analyzed through the IRT three-parameter logistic model:

$$P(\theta) = c_j + (1 - c_j)\frac{e^{Da_j(\theta - b_j)}}{1 + e^{Da_j(\theta - b_j)}}$$

where θ is the examinee's trait or latent knowledge level, a_j the discrimination parameter of the item, b_j the difficulty parameter, c_j the pseudoguessing parameter, e the base of the natural logarithm, and D is a scaling constant (-1.7).

A total of 28 items were eliminated based on several psychometric criteria: a) low item-total biserial correlations, b) items in which the choice of an incorrect option correlated positively with the total score, and c) bad fit of the option response functions to the IRT model (χ^2 with $p<0.01$) accompanied by large residuals for some ability levels and/or non-monotonically increasing response functions. One or more items were removed from six content areas. All the linking items remained. Therefore, the final calibrated pool has a total of 197 items.

In order to calibrate the items of all the forms in the same metric, a concurrent calibration design was used. For the estimation, the bayesian marginal maximum likelihood estimation procedure was used, implemented in the BILOG program (Mislevy & Bock, 1990). An important decision has to be taken at this point regarding how to consider the omitted responses. There are notable cultural differences that examinees have in their tendency to omit items (Hambleton, 2005). Omission rates also depend on the instructions examinees receive. In the pretest administration, examinees were told that errors would reduce their test scores. In the calibration process, omissions can be considered as errors, missing values, or fractional correct responses. This last option was chosen, because these item parameter estimates will be ultimately applied to a CAT where no omission is possible. These differences between pretest and on-line item administration conditions make advisable the study of item parameter updating when the CAT has been applied to an appropriate number of examinees.

For the discrimination parameter *a*, most of the values were between 0.83 and 1.90 (mean (M) = 1.30; standard deviation (SD) = 0.32). For the difficulty parameter *b*, 90% of the values are between –1.26 and 2.16 (M = 0.23; SD = 1.00). For the pseudoguessing parameter *c*, the distribution was found to be centered on around 0.20 (M = 0.21; SD = 0.02) with the majority of values between 0.16 and 0.25. The only significant correlation, with a significance level of 1%, was found between parameters *a* and *c* (r = -.369), which implies that less able subjects find it less easy to respond to highly discriminating items.

One of the most common methods of studying the precision of an item pool is to obtain the information function for the different θ values. This information function indicates the maximum precision that can be attained by the item pool. For the ability range between –1 and 2, the standard error of measurement attainable if the 197 items of the pool are applied is under .2. The item pool operates better for medium to high ability levels. Clearly, ability levels below $\theta = -2.5$ cannot be estimated with precision (standard errors greater than 0.5).

A validity study of the scores on the forms was carried out through examining the fit (using AMOS 4) of a single latent variable "self-reported latent English level" in which the variables on language training and self-assessment of proficiency included in the questionnaire obtained positive loadings. This latent variable correlated .81 with the θ estimates.

Five independent variables from the self-assessment and language training questionnaire were defined, and their effects on the estimated θ levels examined with ANOVA. The main findings of these analyses illustrate that a) average ability levels increase as the levels of each of the training or self-assessment variables increase (all the multiple post-hoc comparisons through Tukey's statistical HSD were significant), and b) the values of the effect sizes are greater in the self-assessment variables of English proficiency than in the variables relating to language training (more details in Olea et al., 2004).

2.3.- Adaptive algorithm

Once the item pool was calibrated and its psychometric properties in terms of precision and validity were considered satisfactory, the next step was to create a set of programs for all the tasks a CAT requires. The programming language C++ Builder was used for this purpose. The main phases of any adaptive algorithm are presented in the following flow chart:

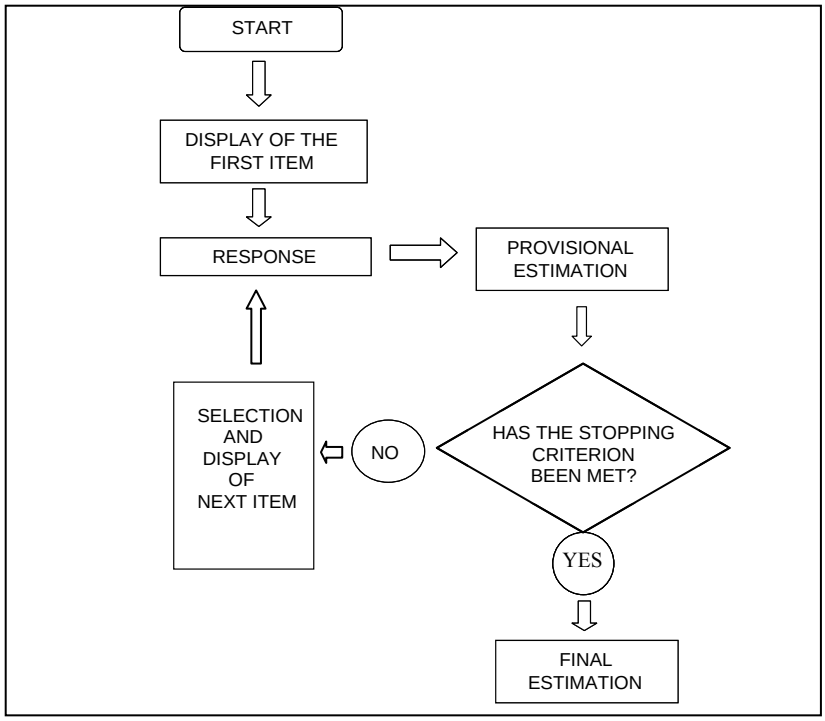

Figure 1. Flow Chart of the Adaptive Algorithm

In our case, the adaptive algorithm has the following basic characteristics:

a) *Starting procedure*. The program begins by assigning the test-taker an initial θ level between -1 and +1, obtained randomly from a normal, N(0,1), truncated distribution.

b) *Estimation of* θ. The first item to be administered is the most informative for the initially assigned ability level. As the maximum-likelihood estimation procedure in use cannot be applied when the examinee has a constant pattern of only hits or errors, the method proposed by Dodd (1990) was used and the ability estimate increased (after a hit) or decreased (after an error) in order to break this constant pattern when it occurred.

c) *Selection of item*. When a provisional θ level is estimated, the next item displayed to the test-taker is the most informative for that level.

Stopping procedure. Three different procedures are available to stop the test: the attainment of a certain precision, of a previously specified number of items, or a combination of both two. In the eCat it is important to let the user choose the stopping rule that best fits its particular aim. For example, in personnel recruitment, one may be interested in a quick screening of the candidate's English knowledge, and a fixed number of items may be appropriate for this aim. In educational assessment, one may want to know the precise English level of each student. In this situation, final score precision matters and a precision stopping rule would be preferred.

2.4.- Scores and reports

Initially, the adaptive algorithm provides the estimated θ level together with the standard error (Se) associated with that estimate. This type of information is useful in ordering examinees according to English proficiency, but it does not provide a criteria-based interpretation of their performance. In order to provide a more complete feedback to the examinees and to those in charge of the assessment process, a procedure was established to prepare computerized reports, which basically consisted of the following steps:

a) The quartile associated with each θ estimate was obtained.
b) The training on English and self-assessment characteristics of people in the same quartile was found. To this aim, a descriptive study was conducted of the responses provided by the four quarters of examinees (those below quartile 1, those between quartiles 1 and 2…) to the questionnaire on training and self-assessment of competence in English. The computerized report provides the examinee score, the decatype and percentile, and the expected training in English of those sharing the examinee's English competence.
c) One of the 4 available computerized reports is supplied to each test-taker based on his/her performance.

2.5.- Web-based application

The next step consisted in developing and testing the software necessary for administering the CAT in an on-demand format (through an Active Server Provider, ASP). The software developed can be conceived as a standard e-business application in that the implementation of the eCat is carried out on the platform hosting the application for its use on-demand or as an ASP. The development was carried out in layers: a) the database layer, which stores the information of the item pool and examinees and all actions that are carried out; b) the logical layer, where the different system algorithms reside; and c) the display layer, which is responsible for displaying the different contents to the user. These layers can be implemented on a single machine or on more than one, depending on the user demands. The database can be supported by any of the data handlers currently on the market, using either IBM DB2 UDB or Oracle in conjunction with either Windows or Linux as the machine's

operating system. The web server was Microsoft Internet Information Server. For these operating systems and database handlers, the programming languages used were (a) for display, Macromedia Flash, XSL and DHTML; (b) for data access, IBM DB2 UDB and XML;(c) for functionality, JavaScript; (d) for generating pages in the server, ASP with VBScript and JScript.

Access to eCat is obtained through Internet Explorer. The requirements of the PC used by the customer to access the application are as follows: operating system Windows 2000 or higher; web browser Microsoft Internet Explorer 5.0 or higher or Mozilla firefox 3.0 or higher; Plugin Adobe Flash Player version 5.0 or higher; Internet connection: ADSL 128 Kb (minimum recommended).

To access eCat, the examinee connects to an URL address and establishes a safe connection where the person enters a username and password. After confirming this data, eCat displays the instructions as well as four sample items which allow the examinee to get used to the application. Next, the last instructions screen appears containing a brief reminder about how to interact with the system. When the examinee clicks on the START button, the assessment begins.

First, the algorithms implemented on the server decide which item will be displayed. The system then sends it to the customer machine and displays it. The user responds by using the mouse to click on the option they consider correct. The examinee's response is sent to the server and the algorithms decide which item will be shown next, as depicted in Figure 1. When the CAT ends, a screen is displayed that thanks the examinee and indicates that the assessment is over. As described above, a report is automatically and immediately generated with the examinee results. These results are also available for the person in charge who, through a safe connection and by entering a username and password, can access the system to consult the results of all the examinees they are responsible for.

A set of tests were executed with the aim of ascertaining the system's reliability. The measures undertaken in our serves ensure a response time which is under 3.5 seconds in 90% of situations, for a load of 250 concurrent users.

2.6.- Simulation study

Before the actual implementation of the algorithm, it was considered necessary to examine specific aspects of its performance when applied to the 197 items forming the calibrated pool. First, data were collected on the item exposure rates, because the CAT requires the establishment of item exposure controls to prevent the diffusion of items in successive applications. Second, it was necessary to collect prior information on certain statistical properties of the ability estimates (bias and RMSE) for different stopping criteria.

A simulation study was designed with the following specifications: a) 10,000 subjects were simulated, extracted from a discrete normal distribution, with 17 θ values between -4 and +4, b) 4 different lengths of the CAT (15, 20, 25 and 30 items), c) 3 different exposure control criteria were established (no control, maximum rate 25% and maximum rate 40%). The *Restricted* method (Revuelta & Ponsoda, 1998) was applied to make effective these maximum rates. This method prevents the administration of an item when its exposure rate exceeds the specified maximum value. If exposure rates of the item bank are updated when the test ends,

overexposed items would decrease their exposure rates and would then be again available for future tests. With this method, no item should exceed the maximum rate and some increase in the exposure rates of items suffering underexposure is achieved (Revuelta & Ponsoda, 1998).

RMSE and bias functions were obtained under each condition. Both measures indicate whether or not trait estimates ($\hat{\theta}_i$) adequately recover the true trait levels (θ_i) of the I simulees. They are computed as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{I}\left(\hat{\theta}_i - \theta_i\right)^2}{I}}$$

$$BIAS = \frac{\sum_{i=1}^{I}\left(\hat{\theta}_i - \theta_i\right)}{I}$$

As shown in Figure 2, for a reasonable precision criterion (RMSE = 0.30), it is advisable to establish a stopping criterion of at least 15 items. RMSE values do not appreciably change across the ability values between the three exposure control conditions. The *Restricted* control exposure procedure applied was effective in preventing too high exposure rates: None item was presented in more than a 25% or 40% of the tests in the corresponding exposure control conditions, whereas a 6.1% of the items had exposure rates above 40% in the no control condition. Two items had exposure rates above 50%.

The θ estimates were mostly unbiased, even in the shortest CAT condition. Low levels of bias in the estimates were only obtained for extreme trait levels. Two reasons can be offered for this. First, the item bank lacks very informative items for these extreme ability levels, and, second, the outward bias pattern observed is characteristic of the maximum-likelihood estimation procedure applied (Kim & Nicewander, 1993). The eCat test is not ordinarily applied to extreme ability levels and these higher bias levels should then be infrequent in eCat applications.

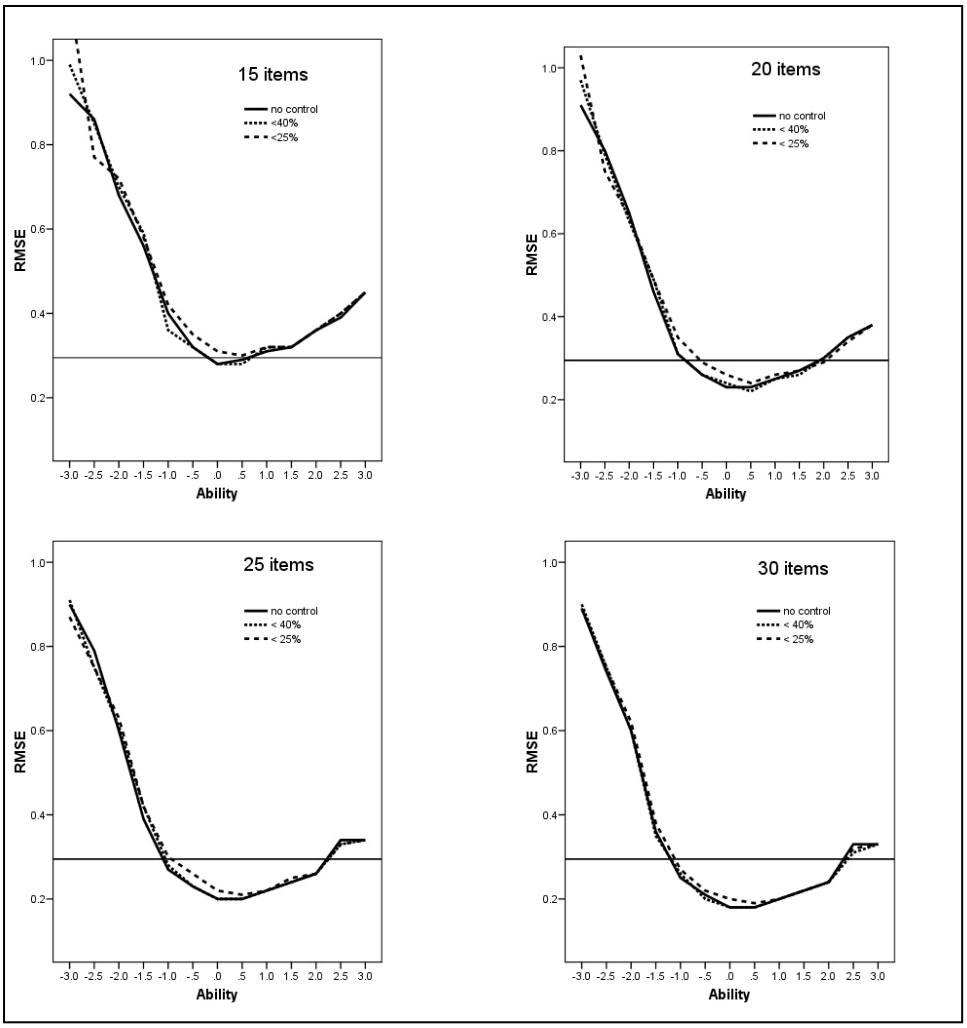

Figure 2. RMSE according to the number of items applied, ability level and item exposure control condition

From the simulation and validity results already shown it was considered that eCat scores fulfilled the basic psychometric requirements related to reliability, validity and score interpretation. Concerning security, an exposure control procedure was implemented for the item selection process based on the Restricted method (with maximum exposure rate of .25) and some additional constrains applied to the first five items (see Abad et al. 2004 for details).

3.- Operational e-Cat

The aim of this study is to check whether or not the psychometric indices of eCat have changed as a consequence of the 4-years use of the test in personnel selection contexts. To this end, the eCat application results (descriptive, reliability, item exposure rates…) will be analyzed and compared with those corresponding to the pre-operational results.

3.1.- Method

A total of 7254 eCat administrations were available for the analysis. Most of them (80%) were produced by candidates that responded to eCat in different recruitment processes of public and private organizations. A 20% of the sample administrations correspond to last

year undergraduates of a public Spanish university involved in a transversal competence program in which eCat and other assessments were conducted. The stopping criterion was set at 30 items for all the administrations.

3.2.- Results

Descriptives. The mean and standard deviation of the trait estimates are 0.67 and 0.93, meaning that eCat has been mostly applied to medium-high trait levels. The mean item response is 20.28 seconds (SD = 5.44). So, the mean administration time for eCat is below 10 minutes. Pearson's correlation between trait estimates and total test time was -0.17. This inverse, statistically significant and low correlation reveals a trend for candidates with higher trait levels to expend less time responding to the test.

Reliability. The mean of the standard error is 0.22 (SD = 0.04). So, the equivalent reliability coefficient is 0.95. A 93% of the examinees obtain a standard error below 0.30.

Validity evidences. For a small sample (n = 48), their scores in the speaking and writing TOIEC tests were also available. Pearson correlations between eCat scores and the writing, the speaking and the global TOIEC scores were 0.62, 0.49 and 0.67, all of them statistically significant at the 0.01 level.

A total of 637 students of a Spanish Language School of English responded to the eCat and the self-assessment questions of the questionnaire referred to above, regarding their reading, writing and spoken English levels. Table 2 shows the ANOVA results conducted on the ability estimates when taking as independent variables the 3 self-assessments and the school level. The means of the categories statistically differ in the four cases and, more important, increase as self-assessments and school levels do. In the four variables the lower two categories were collapsed as the students included in the lowest one were very few (less than 12).

Independent variable	Averages in estimated θ	Sig. (p<)	η^2
Reading self-assessment reading		0.000	0.258
Very low	-0.0629		
General ideas	0.2220		
Good	0.9784		
Bilingual	2.0810		
Writing self-assessment writing		0.000	0.301
Very low	-0.1284		
With effort	0.3656		
Good	1.0899		
Bilingual	2.2429		
Spoken self-assessment		0.000	0.324
Very low	-0.017		
With effort	0.4413		
Good	1.1843		
Bilingual	1.9713		
Language School level		0.000	0.372
1 and 2	-0.1733		
3	0.2832		
4	0.4248		
5	0.9956		
6	1.2369		

Table 2. Results of the ANOVAs for each of the 4 independent variables (self-assessed level in reading, writing and spoken English and level in the Language School), taking the estimated θ as the dependent variable. Averages, significant values and effect size (η^2) measures are included

Exposure rates. The observed exposure rates were compared to those expected from the previous simulation study. A total of 17 items were never applied. They are very easy items and then not appropriate for the medium-high trait level of most eCat candidates. Actual and predicted exposure rates show some noticeable differences. The rate of slight overexposed items (exposure rates above 0.15) are higher in the simulated (57.4) than in the empirical (53.4) condition. This result could be due to differences in the trait distributions. The simulated trait distribution was a standard normal distribution N(0,1), whereas the observed trait distribution mean is 0.67. The item bank has more difficult than easy items, so different exposure rates should be expected. However, a considerable percentage of items (23.9%) show exposure rates between 0.25 and 0.30 in the operational condition when this is 0 in the simulated condition. The item exposure control method in use needs a frequent update of the exposure rates. This results points out that the required update was not as frequent and it should had been. A second exposure control index is the mean percentage of

items two examinee share (overlap rate). The minimum expected overlap rate for eCat is 0.152, as this is the expected rate when items are randomly selected from the bank. The observed overlap rate found is 0.222.

Item exposure rates were correlated with item parameters. As expected, a positive significant correlation was found with the *a* parameter ($r_{ER,a} = 0.636$; $p < .01$) and a negative with the *c* parameter ($r_{ER,c} = -0.345$; $p < 0.01$), as items with high *a* and low *c* values are in general more informative. Exposure rates also significantly correlate with *b* values ($r_{ER,b} = 0.356$; $p < .01$), due to the medium-high level of the examinees: Item more frequently administered are the most difficult items because the adaptive algorithm selects items with a difficult parameter similar to the examinee trait level.

4. – Discussion

This report details the work conducted in the construction of eCat, an adaptive test administered via the Internet to estimate proficiency in English as a second language. The results show that eCat provides a reliable, valid, and informative measure. Each examinee receives a normative and a criterion referenced interpretation of his/her test score.

In terms of reliability, the standard error of measurement of the pool is under 0.2 for a wide range of the ability level continuum (between -1 and 2). Ability estimates are unbiased even in the conditions of shortest test length (15 items), but it is necessary to establish a stopping criterion of over 15 items when the usual precision criterion (RMSE = 0.30) is demanded.

Validity evidences regarding eCat were collected. A detailed procedure was applied for the item bank development (validity evidence based on content). Unidimensionality checks were also conducted (evidence based on the internal structure of the test). Analyses were also conducted to relate eCat scores to training in English (evidence based on relations to other variables). In fact, eCat ability estimates correlate 0.81 with self-reported English knowledge provided by the examinees. It was also found that groups differing in self-assessment and training in English also differed in eCat scores.

The test eCat has been in use for some years and we have also analyzed its performance in actual recruitment processes. The standard error of a 93% of examines is below 0.30 (reliability coefficient above 0.91). So, eCat provides precise measures. We have also shown that eCat scores correlate with other English measures (writing and spoken TOIEC tests) and again with the operational tests data a clear relationship is found between eCat scores and self-assessment and Language school levels. A 23.9 of items have exposure rates slightly above the allowed (0.25) rate, indicating that the updating demanded by the exposure control method in use should be done more frequently.

The results presented show certain limitations which must be born in mind. First, the precision study clearly shows that the pool does not allow a precise estimation of ability levels below -2.5. Although the percentage of people with ability levels below -2.5 is very small (0.6%), adding to eCat items appropriate for these low levels would be advisable. Moreover, only 225 items of the original 653 items have been calibrated. New pilot applications must be conducted to increase the set of calibrated items. An item bank increase would also make it possible the addition of content control procedures to the item selection

algorithm, providing additional content validity support for the eCat. Finally, more validity evidence studies should be conducted, to learn more on how eCat scores relate to other linguistic and non-linguistic performance criteria.

5.- References

Bartram, D. & Hambleton, R. K. (2006). *Computer-based testing and the internet issues and advances*. Chichester, West Sussex: Wiley.

Carlson, J.F. & Harvey, V.S. (2004). Using computer-related technology for assessment activities: ethical and professional practice issues for school psychologists. *Computers in Human Behavior*, *20*, 5, 645-659.

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

Fetzer, M., Dainis, A. Lambert, S. & Meade, A. (2008). *Computer adaptive testing (CAT) in an employment context*. White paper. Previsor.

Fraser, C. (1988). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. NSW: University of New England.

Hambleton, R.K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R.K. Hambleton, P.F. Merenda & Ch.D. Spielberger (Eds.). *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). NJ: LEA.

Jöreskog, K. G. & Sörborn, D. (1996). *LISREL8: User's Reference Guide*. Chicago, IL: Scientific Software International.

Kim, J.K. & Nicewander, A. (1993). Abilty estimation for conventional tests. *Psychometrika, 58*(4), 587-599.

Mislevy, R.J., & Bock, R.D. (1990). *BILOG: Item analysis and test scoring with binary logistic models* [computer program]. Chicago: Scientific Software, Inc.

Naglieri, J.A., Drasgow, F., Schmite, M., Handler, L., Prifitera, A., Margolis, A. & Velasquez, R. (2004) Psychological Testing on the Internet: New problems, old issues. *American Psychologist,* 59, 150–162.

Nye ,C. D., Do, B., Drasgow, F. & Fine, S. (2008). Two-Step Testing in Employee Selection: Is score inflation a problem? *International Journal of Selection and Assessment*, 16, 2, 112–120.

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

Ponsoda, V. & Olea, J. (2003). Adaptive and tailored testing. Including IRT and non-IRT Application. In R. Fernández-Ballesteros (Ed.). *Encyclopedia of Psychological Assessment* (pp. 9-13). London: Sage.

Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in CAT. *Journal of Educational Measurement, 35*, 4, 311-327.

Tippins, N.T., Beaty, J., Drasgow, F., Gibson, W.M., Pearlman, K., Segall, D.O. & Shepherd, W. (2006) Unproctored Internet Testing in Employment Settings. *Personnel Psychology*, 59, 189–225.

Viswesvaran, C. (2003). Introduction to Special Issue: Role of technology in shaping the future of staffing and assessment. *International Journal of Selection and Assessment*, 11, 107–112.

Wise, S.L. & Kingsbury, G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica*, *21*, 135-155.