Psicothema 2011. Vol. 23, nº 4, pp. 802-807
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

eCAT-Listening: Design and psychometric properties of a computerized adaptive test on English Listening

Julio Olea[1], Francisco José Abad[1], Vicente Ponsoda[1], Juan Ramón Barrada[2] and David Aguado[1]
[1] Universidad Autónoma de Madrid and [2] Universidad Autónoma de Barcelona

In this study, eCAT-Listening, a new computerized adaptive test for the evaluation of English Listening, is described. Item bank development, anchor design for data collection, and the study of the psychometric properties of the item bank and the adaptive test are described. The calibration sample comprised 1.576 participants. Good psychometric guarantees: the bank is unidimensional, the items are satisfactorily fitted to the 3-parameter logistic model, and an accurate estimation of the trait level is obtained. As validity evidence, a high correlation was obtained between the estimated trait level and a latent factor made up of the diverse criteria selected. The analysis of the trait level estimation by means of a simulation led us to fix the test length at 20 items, with a maximum exposure rate of .40.

eCat-Listening: diseño y propiedades psicométricas de un test adaptativo informatizado de comprensión auditiva de la lengua inglesa. En este trabajo se describe eCAT-Listening, un nuevo test adaptativo informatizado para la medición del nivel de comprensión auditiva del inglés. Se describe la elaboración del banco de ítems, el diseño de anclaje para la recogida de datos y el estudio de las propiedades psicométricas del banco de ítems y del test adaptativo. La muestra de calibración fue de 1.576 personas. Se obtienen unas buenas garantías psicométricas: el banco es unidimensional, los ítems se ajustan satisfactoriamente al modelo logístico de 3 parámetros y se consigue una estimación precisa de los diferentes niveles de rasgo. Como prueba de validez, se obtuvo una alta correlación entre el rasgo estimado y un factor latente de nivel de inglés compuesto por las diferentes puntuaciones criterio utilizadas en el estudio. El análisis de la estimación del nivel de rasgo mediante simulación nos lleva a fijar la longitud del test adaptivo en 20 ítems, con una tasa máxima de exposición de 0,40.

Computerized adaptive testing (CAT) is an assessment method in which, in comparison with a fixed form test, items are administered according to the examinee's trait level (Olea & Ponsoda, 2003). Among the main advantages of CATs we find: (a) improvement in test security, (b) reduction in testing time, and (c) improvement in the accuracy with the same number of items as a fixed test. CATs have been made possible due to the evolution of the psychometric theory with the Item Response Theory (IRT) models, and to the progress in computer technology, which has allowed the implementation and integration in the test of algorithms for item selection and examinees' trait level estimation.

The use of CATs in psychological and educational assessment is widely spread in countries like the United States or the Netherlands, where some important testing programs are being applied adaptively. Wainer (2000) posited an exponential growth in the number of CATs administered, and his predictions seem to be fulfilled. However, although there has been an expansion of CATs in Spain (i.e., López-Cuadrado, Pérez, Vadillo, & Gutiérrez, 2010; Olea, Abad, Ponsoda, & Ximénez, 2004; Rebollo, García-Cueto, Zardaín, Cuervo, Martínez, Alonso, Ferrer, & Muñiz, 2009; Rubio & Santacreu, 2004), we are still far from the level of other countries.

The assessment of English knowledge is a topic in which several CATs have been developed, with adaptive versions of the Test of English as Foreign Language (TOEFL) and the Business Language Testing Service (BULATS). However, the most commonly applied English test in the organizational context, the Test of English for International Communication (TOEIC), has no adaptive format.

To cover this lack, a CAT of English Grammar was initially developed (eCAT-Grammar; Olea et al., 2004) and updated (Abad, Olea, Ponsoda, Aguado, & Barrada, 2010). However, despite satisfactory validity evidence in terms of internal structure and relation with other variables, it can be argued that the measurement of the English level lacks of content validity if Listening skills are not evaluated. Our current purpose is to present the development and psychometric properties of a new CAT, called eCAT-Listening, designed for the assessment of the English level with orally administered items.

Method

Participants

A sample of 1.576 people (n_1= 592, n_2= 605, n_3= 379 for each subtest) was selected, mainly participants in selection processes

Fecha recepción: 2-12-10 • Fecha aceptación: 16-3-11
Correspondencia: Julio Olea
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
e-mail: julio.olea@uam.es

where English assessment was required. An important part of the sample (n_1= 190, n_2= 267, n_3= 187) comprised students from the *Escuela Oficial the Idiomas* (EOI; Official School of Languages).

Measures

Development of the item bank. Two experts in English philology, with the collaboration of three experts in psychometrics, developed an initial item bank of 227 items. The English experts followed a functional theoretical framework, from which they proposed verbal contents about daily situations. Taking into account the criteria established by Common European Framework of Reference for Languages (CEFR; Council of Europe, 1999), items for 6 difficulty levels were written. Items varied according to the processes required to understand them (i.e., to obtain specific information, to grasp the global idea or to infer the speakers' intentions). The English experts redacted the items, assigning to each one an estimation of its difficulty, and made suggestions about the recording (i.e., dialogue rhythm, kinds of voices, sex of the characters…). Item content was reviewed by two native English speakers, who assigned (independently of the philologists) difficulty levels to the items. The correlation between the difficulty level estimated by the philologists (one level for each item, agreed by both philologists) and the mean level estimated by the native English speakers was .663.

Each item had a brief introduction (i.e., «Listen to this short dialogue»), followed by an audio with the item content (an interactional dialogue, a transactional dialogue or a monologue). After playing the audio, a written question was presented about what had been listened to with three response options, only one of them correct. The recording process of the items was carried out in a professional studio by native British or North American actors. The items of the two first difficulty levels were read slower, whereas the other items were read at speakers' usual speed.

Development of the subtests. In the item bank application, for its subsequent calibration, an anchoring design was established in which the predicted item difficulty was considered. For this first version of eCAT-Listening, three subtests were elaborated, each one with 42 items: 12 as the anchor test (common for all the subtests) and 30 specific items for each subtest. All items were chosen to properly represent the 6 difficulty levels (2 items per level in the anchor test and 5 for the specific part). The items with higher inter-judge agreement in the assignment of difficulty were selected. In this initial bank of 102 items, the correlation between the difficulty level established by the philologists and the native English speakers was .864.

Criteria measures. With the goal of obtaining data about the validity of the scores, eCAT-Grammar (Olea et al., 2004) and a self-report questionnaire about English knowledge and studies were also applied. In this questionnaire, the participants informed about: (a) the type of school where they had attended their middle-studies (bilingual-English or others), (b) their perceived mastery in English (reading, writing and conversation), and (c) their training in English (primary and secondary education, academies, family, stays in Anglo-Saxon countries and others). Finally, the EOI students informed about the level to which they were assigned according to the CEFR (Basic 1 or 2, Intermediate 1 or 2, Advanced 1 or 2) and their educational level (no studies, primary studies, secondary studies, university studies).

Data analysis

For the study of the unidimensionality assumption, a confirmatory factor analysis was performed in each subtest with M*plus* 5 (Muthén & Muthén, 2006). We analyzed the tetrachoric correlations with the RWLS method, recommended for dichotomous items. Model fit was evaluated with the indexes CFI, TLI, RMSEA and SRMR.

Items were calibrated according to the 3-parameter logistic model (normal metric). To calibrate the items of the different subtests in the same metric, concurrent calibration was used (Hanson & Béguin, 2002), so the responses of the non-applied items are considered missing values. Parameters were estimated with the Bayesian marginal maximum-likelihood procedure, as implemented in MULTILOG 7.0 (Thissen, Chen, & Bock, 2003). The following prior distributions were assumed: (a) for the ability, a standard normal distribution; (b) for the *a* parameters, N(1, 0.588), which corresponds to N(1, 1) in the logistic metric; (c) for the *b* parameters, N(0, 2); and (d) for the logit of *c*, N(-0.69, 0.5), which corresponds approximately to a mean of .33 for the *c* parameter.

Several approaches were used to evaluate item fit. Firstly, the χ^2/*df ratios* were calculated with the program MODFIT (Stark, 2001). These ratios are taken as heuristics to make decisions about the size of the discrepancies between the expected and observed frequencies for the possible response patterns to an item, to pairs of items or to triplets. Ratio values lower than 3 are usually considered indicators of a good fit (Drasgow, Levine, Tsien, Williams. & Mead, 1995). This approach is especially sensitive to the detection of local dependence between item pairs or triplets. Secondly, the empirical and expected item characteristic curves (ICCs) were obtained with the program MODFIT. Finally, the GOODFIT program (Orlando & Thissen, 2003) was used to study the statistical significance of the differences between the observed and expected probabilities of correct response as function of the test score. Thus, we analyzed whether the theoretical probability, which in our case follows the 3-parameter logistic model, is flexible enough to model the empirical ICC.

Various statistical (ANOVAs, *t*-tests and Pearson correlations) tests were performed to establish the relation between the results with eCAT-Grammar and the scores in the questionnaire of English training and the estimated trait level in Listening: as dependent variables, we used the trait level of Listening for each examinee estimated from their responses to the corresponding subtest; as independent variables, each one of the items of the questionnaire and the eCAT-Grammar estimation.

The predictive capability of eCAT-Listening, compared with eCAT-Grammar, was also analyzed for each criterion variable separately. In this case, we applied two statistical models: linear and probit regression. In probit regression, the categorical condition of the criterion variables is considered: the independent variables (Listening and Grammar) predict the probability of belonging to each ordered response category of the criterion variable.

We also tested the predictive value of the eCAT-Listening and Grammar scores for a latent variable of self-reported English, constructed by the categorical variables of reading, writing, conversation, years of stay, English at home and EOI level. The parameters were computed with M*plus*, using RWLS estimation procedure.

To study the psychometric properties of the CAT, mainly to define the test length, a simulation study was performed with 50,000 examinees extracted from a standard normal distribution.

The bank was composed of the final 95 items, with their calibrated parameters. The implemented adaptive algorithm is described in detail in Olea et al., (2004).

As independent variables, we considered the test length (15, 20, 25 and 30 items) and the maximum exposure rate allowed for an item (two levels: .25 and .40), according to Barrada, Abad and Veldkamp (2009). The test lengths of 25 and 30 could not be combined with the restriction of maximum exposure rate equal to .25, as these lengths are 26 and 32% of the full bank. As dependent variables, we considered RMSE, bias, the proportion of examinees whose estimated standard error was lower than 0.3 or 0.4 (p_SE_0.3 and p_SE_04), the correlation between the real and estimated trait level ($r_{\theta\hat{\theta}}$), the overlap rate (T; the main proportion of items shared by two examinees) and the proportion of infra-exposed items (p_infra; items administered less than 1%).

Results

Psychometric analysis and dimensionality. The mean time for responding to each item (the audio part not included) was 13 seconds (*SD*= 4.76). In all three subtests, all of them with 42 items, the mean number of correct responses ranged between 27.37 and 28.89 (*SD* range: 7.16-7.68). The differences in the mean number of correct responses in the subtests were statistically significant ($F_{2,1573}$= 5.599, *p*= .004), which indicates the need to equate the metric of the items and subjects parameters. Item difficulty varied between .26 and .98 (Mean= .69, *SD*= .17) and the item-test correlation between .14 and .80 (Mean= .51. *SD*= .14). The alpha coefficients for the three subtests and the anchor test were, respectively, .889, .869, .862 and .638. Three items with item-test correlation below .1 (non-significant values with a confidence level of 95%) were eliminated, thus increasing the alpha coefficients of the three subtests to .893, .873 and .865.

The results of factor analysis are shown in Table 1. The unidimensional solution shows a good fit (CFI and TLI >.95, RMSEA <.05, SRMR <.09).

The presence of a predominant first factor is clear in all three cases, as the percentage of explained variance only increased minimally with the extraction of a second factor. Additionally, if two factors are extracted, both are highly correlated (Subtest 1: .653, Subtest 2: .524, Subtest 3: .608). Lastly, when inspecting the results of the unidimensional solution, all the loadings were significant (*p*<.05), ranging between .15 and .84, and there was no high modification index (over 3.84) in any of the subtests.

Table 1
Fit indexes (CFI, TLI, RMSEA y SRMR) and percentage of explained variance (EV) for the models with 1 and 2 factors

	CFI	TLI	RMSEA	SRMR	% EV
Subtest 1					
1 factor	.982	.988	.021	.071	32.7
2 factors	.991	.994	.015	.062	36.3
Subtest 2					
1 factor	.962	.972	.028	.086	28.7
2 factors	.985	.989	.017	.073	33.4
Subtest 3					
1 factor	.947	.960	.031	.090	25.7
2 factors	.973	.980	.022	.079	30.2

Fit and parameter estimation. The results obtained with MODFIT for the three subtests showed that, for all the items, pairs and triplets of items tested, the ratio was lower than 3. The χ^2 analysis with GOODFIT showed that: (a) for 87% of items, the model fit the data (*p*>.05); (b) the *p* value of 12 items was between .01 and .05; (c) four items had a *p* value between .001 and .01; and (d) the *p* value of one item was lower than .001. Despite the overall good fit, some items had extreme parameters: three items were too easy (*b*-parameter lower than -4) and one showed an exceptionally high *c*-parameter (.59). These items were deleted.

The final calibrated item bank is comprises 95 items. The mean, standard deviation, 10th percentile and 90th percentile for *a*, *b* and *c* parameters were, respectively, {1.09, 0.38, 0.47 and 1.75}, {-0.31, 0.95, -2.17 and 0.95} and {.28, .07, .14 and .39}. The difficulty of the items is, in general, medium-low, and the *c* parameter reflects the quality of the incorrect options, as it is below 1/3 (3 is the number of alternatives). Significant correlations (*p*<.01) were found between *a* and *b* parameters (r_{ab}= .29) and *b* and *c* parameters (r_{bc}= -.38), which implies that the most difficult items tend to be more discriminative. The negative r_{bc} can be explained as a lack of information (few people of low English level) for the easy items to estimate the *c* parameter, so their estimated value is dominated by the prior distribution, whose mean (.33) is higher than the mean *c* value of the item bank.

Reliability. The Fisher information and standard error for the full bank according to the trait level is shown in Figure 1. For trait levels between -1.3 and 1.7, the standard error if the entire item bank was administered is equal to or lower than .3, which is approximately equivalent to a reliability coefficient of .91. The item bank is more accurate for medium-high trait levels. Ability levels below -2.2 or higher than 2.4 cannot be accurately estimated (standard errors over .5)

Validity. Descriptive statistics of estimated trait level according to the criterion measures are shown in Table 2. To analyze the predictive validity (relation between Listening and criterion variables), a summary of the statistically significant results is shown in Table 3.

The estimated trait level in Listening differs according to: (a) training (the number of years in academies, years in EOI correlated in the EOI sample at .18, .24 and .14 respectively, with all *p*s<.001); (b) presence of English at home (t_{1366}= -6.11, *p*<.001), Mean= 1.01 and .06, SD= .69 and .90, respectively for presence/absence,

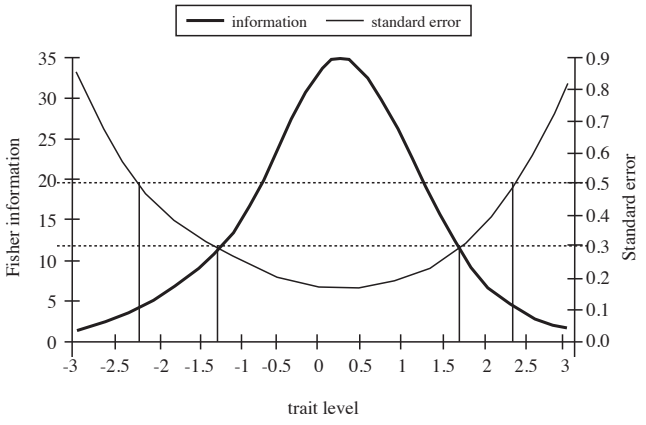

Figure 1. *Standard error according to the trait level for the full item bank (95 items)*

(c) level achieved in the EOI ($F_{4,635}$= 89.584, p<.001), where the mean estimations were -0.84 (Basic 1 and 2 collapsed), -0.30 (Intermediate 1), -0.18 (Intermediate 2), 0.35 (Advanced 1) and 0.52 (Advanced 2); (d) educational level ($F_{2,639}$= 24,097, p<.001), the mean estimations being equal to -0.66 (no studies or primary studies), -0.32 (secondary education) and -0.04 (university). The estimated trait level in Listening did not differ according to the linguistic nature of the attended school ($F_{3,1362}$= 1,211, p= .304) or to whether or not the student had an EOI degree (t_{1366}= -0.450, p= .653). The results in relation to the EOI level are noteworthy, because examinees' level in these centres is determined according to the criteria established by the European Council (CEFR recommendations).

In Table 4 is shown the proportion of variance that is explained by the estimated scores in Listening and eCAT-Grammar, both separately and combined. Listening scores explained approximately 45% (50% with probit regression) of the variance in the criterion in self-perceived conversational skill. The same scores explain approximately 34% (40% with probit regression) of the self-reported ability in reading and writing. Comparatively, eCAT-Grammar scores had a higher predictive power in all the cases. However, when we include both predictors in the probit regression (dependent variable «conversation»), the variance explained by eCAT-Grammar alone (50%) increased by 4%.

For the rest of the criterion variables, the predictive value of the two tests is very similar. The specific patterns depend on the regression model. With linear regression, eCAT-Grammar is the best predictive test for all the criteria, although the differences in terms of explained variance are always lower than 4%. In this linear model, Listening scores explained 35% of the variance in the EOI level, 14% of years of stay and 2.7% of presence of English at home. The proportions of explained variance when eCAT-Grammar is incorporated increase to 42, 18 and 3.2%, respectively. When considering the probit regression, Listening scores are more predictive of the years of stay and presence of English at home, although the differences are lower than 5%. The scores in the subtest of eCAT-Listening explained 39% of the EOI level, 35% of years of stay and 27% of presence of English at home. When including eCAT-Grammar, these proportions increase to 46, 38 and 28%, respectively.

Figure 2 includes a structural model relating Listening and Grammar estimates with a latent variable of self-reported English level (criterion measures). Fit indexes were acceptable (RMSEA= .070, TLI= .984, CFI= .982).

Before testing the structural model, we found that the proportion of explained variance just including eCAT-Listening was 40%,

Table 2
Descriptive statistics of the estimated trait level according to variables related to training in English

	Mean	SD	Frequency	%
School				
Public	0.05	0.94	763	55.9
Private non-bilingual	0.15	0.80	430	31.5
Private bilingual English	0.17	1.06	89	6.5
Private bilingual others	0.06	0.86	84	6.1
Presence of English at home				
No	0.06	0.90	1334	97.5
Yes	1.01	0.69	34	2.5
Holding of a degree				
No	0.09	0.91	1356	99.1
Yes	0.20	0.52	12	0.9
EOI level[a]				
Basic 1 or 2	-0.84	0.65	183	28.7
Intermediate 1	-0.30	0.66	190	29.7
Intermediate 2	-0.18	0.64	92	14.4
Advanced 1	0.35	0.58	109	17.0
Advanced 2	0.52	0.55	66	10.3
Education level[b]				
None or Primary studies	-0.66	0.72	83	12.9
Secondary studies	-0.32	0.76	268	41.7
University	-0.04	0.77	291	45.3

Notes: [a] Only 5 people in Basic 1. Multiple comparisons (Bonferroni). All statistically significant with exception of comparisons 2-3 and 4-5 (p>.05); [b] only 3 people with no studies

Table 3
Relation between the estimated trait level in Listening and the criterion variables

Criterion variables	ANOVAs, correlations or t-tests
School	$F_{3,1362}$= 1.211, p= .304
Reading	r= .586, p<.001
Writing	r= .586, p<.001
Conversation	r= .663, p<.001
Years in academy	r= .182, p<.001
Degree	t_{1366}= -0.450, p= .653
Years in EOI	r= .143, p<.001
Presence of English at home	t_{1366}= -6.108, p<.001
Years of stay[a]	$F_{3,1364}$= 80.070, p<.001 // r= .239, p<.001
Level in EOI[a]	$F_{4,635}$= 89.584, p<.001 // r= .588, p<.001
Educational level[a]	$F_{2,639}$= 24.097, p<.001 // r= .261, p<.001
eCAT-Grammar	r= .822, p<.001

Note: [a] Analysis considering the criterion variables as categorical (ANOVAs) and continuous (correlations)

Table 4
Proportion of explained variance for the probit and linear regression models with eCAT-Listening and eCAT-Grammar as predictors (separately and combined)

	Probit regression			Linear regression		
	Listening	Grammar	Both	Listening	Grammar	Both
Reading[a]	.405	.468	.479	.336	.400	.412
Writing[a]	.386	.488	.492	.342	.432	.439
Conversation[a]	.482	.499	.538	.434	.453	.489
Years of stay[a]	.353	.339	.380	.141	.175	.179
Presence of English at home[a]	.273	.230	.277	.027	.031	.032
EOI Level[b]	.391	.416	.462	.347	.387	.420

Notes: [a] n= 1291, eCAT-Grammar M= 0.42, SD= 0.64; [b] n= 635, eCAT-Grammar M= 0.24, SD= 0.57

whereas it was 48% when only considering eCAT-Grammar. When both predictors are considered simultaneously, the percentage reaches 51%. The test with higher predictive power is eCAT-Grammar. So, due to the high correlation between Listening and eCAT-Grammar (.754 in this sample), the increment in explained variance when adding Listening to eCAT-Grammar scores is small, about 3%.

Psychometric properties of the CAT. The main results of the simulation study are included in Table 5 and Figure 3. Table 5 shows that when the test is short (15 items) or the maximum exposure rate is low (.25), although $r_{\theta\theta'}$ is high (over .9), high RMSEs are found (equal to or greater than .45) and a bias equal to or greater than .04. If the maximum exposure rate is fixed at .40, greater increments in accuracy are obtained when increasing the number of items from 15 to 20. Therefore, it seems that the best solution is to set a CAT of 20 items combined with a maximum exposure rate of .40. Thereby, testing time will be minimized without too much loss of accuracy (compared with a 30-item test length). Eighty-five percent of the examinees will have a standard error lower than .4. (If all the examinees had a standard error of .4, the reliability coefficients would be .86). In average, examinees will share about 6 out of 20 items (32%).

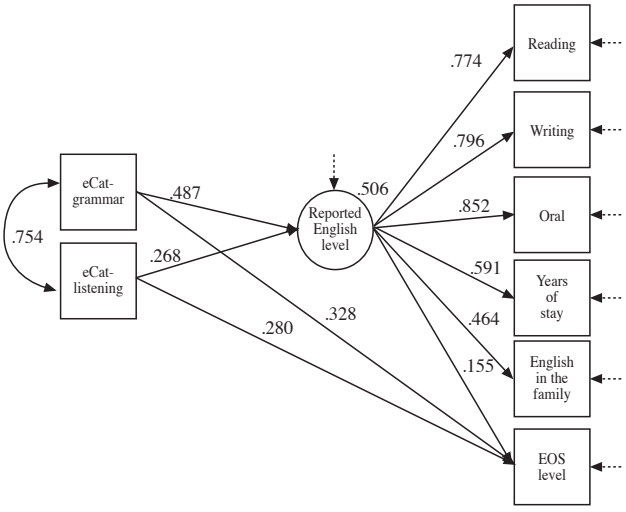

Figure 2. *Model relating eCAT-Listening and eCAT-Grammar to a latent variable of self-reported English level. Standardized parameters. For the latent variable, the proportion of variance explained by the predictor variables is shown in italics*

Table 5

Bias, RMSE, proportion of examinees with a standard error lower than 0.3 and 0.4 (p_SE_0.3 and p_SE_0.4), correlation between the real and estimated trait level ($r_{\theta\theta'}$), overlap rate (T) and proportion of infra-exposed items (p_infra) according to test length and maximum exposure rate (r^{max})

Length	r^{max}	Bias	RMSE	p_SE_0.3	p_SE_0.4	$r_{\theta\theta'}$	T	p_infra
15	.25	.05	.49	.11	.64	.91	.22	.11
20	.25	.04	.45	.29	.70	.92	.23	0
15	.40	.05	.46	.20	.74	.92	.31	.24
20	.40	.03	.40	.51	.85	.94	.32	.11
25	.40	.03	.37	.61	.89	.94	.34	0
30	.40	.02	.35	.61	.90	.95	.36	0

In Figure 3, the results as function of decile trait level are shown. The same conclusions can be reached. In the upper panel, bias is very high for examinees of the higher decile. These results could be expected, as the item of maximum difficulty had a *b* parameter of 1.16. This is no problem from an applied point of view, as the bias is positive and, therefore, these examinees would remain in the same decile. In the other two panels, the examinees whose real trait levels are between the fourth and the eighth deciles are estimated with high accuracy.

Conclusions

We presented the development of a new CAT for the assessment of English Listening. Bank construction, anchoring design, fit and

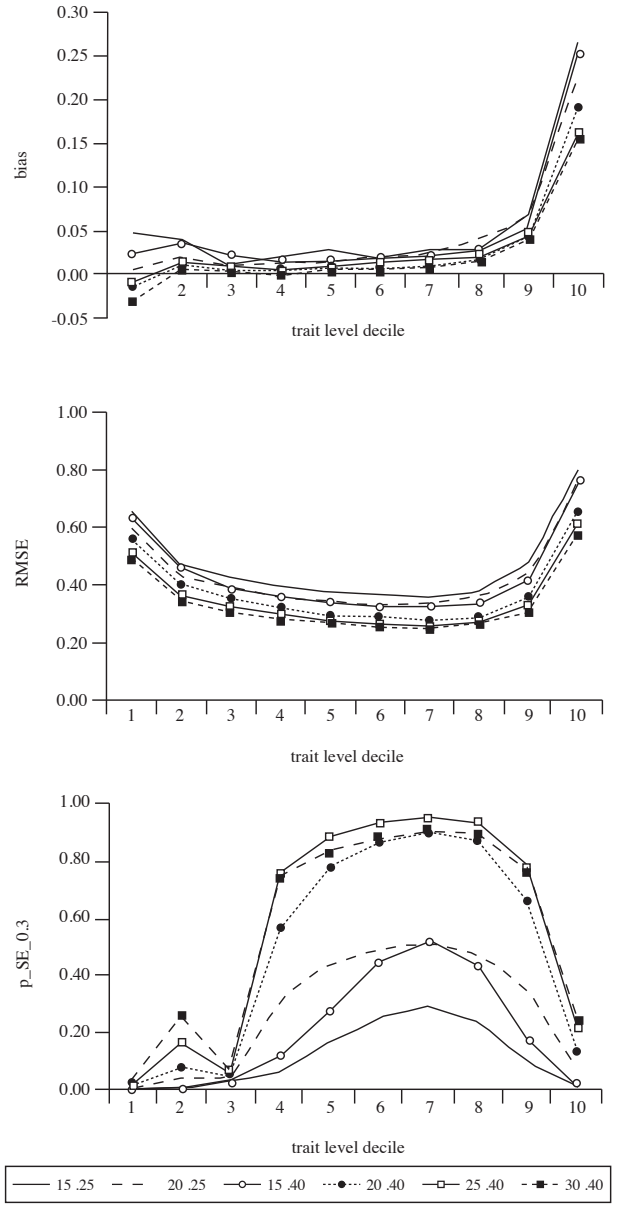

Figure 3. *Bias (upper panel), RMSE (middle panel) and proportion of examinees with a standard error lower than 0.30 (lower panel) as a function of the real trait level decile. In the legend, the first number is test length and the second number is maximum exposure rate*

bank calibration and psychometric properties of the bank and the CAT were explained in detail. Results show that eCAT-Listening has satisfactory psychometric properties with regard to the dimensionality of the bank, the item fit to the 3-parameter logistic model, the validity of the scores, and accuracy and efficiency of the CAT. Specifically, in relation to the dimensionality of the bank, a dominant factor underlies examinees' responses to the items. In addition, a solution with a single factor yielded a good fit in a confirmatory factor analysis. Also in relation to validity, the estimated scores are related to diverse indicators of the participants' English linguistic ability. Considering the different self-reported variables as indicators of a latent variable of Self-informed knowledge of English, eCAT-Listening explains between 37 and 40% of the variance of this latent variable. The relation found between the scores in eCAT-Listening and the examinees' EOI level (r= .59, p<.001) is especially relevant, as the EOI classifies their students according to international criteria (CFER). In addition, the incremental validity of eCAT-Listening over eCAT-Grammar has been shown. As previous studies had revealed (Olea et al., 2004), the predictive value of eCAT-Grammar is very high and, given the high correlation between both tests, the increment in explained variance when Listening is included in a hierarchical model is small. Lastly, the adaptive administration, analyzed by means of a simulation study, has proved to be accurate and efficient with a test length of 20 items and a maximum exposure rate of .40.

From a practitioner's point of view, the satisfactory psychometric properties of eCAT-Listening combined with the method of administration (web-delivered) allows the test to be used in educational and organizational contexts as a first approach to examinees' English competence. This first approach is highly predictive of the examinees' level according to CEFR, with the added value that it can be obtained in a short testing time.

Acknowledgements

This research was supported by two grants from the Spanish Ministry of Science and Innovation (project numbers PSI2009-10341 and PSI2008-01685) and by the UAM-IIC Chair «*Psychometric Models and Applications*».

References

Abad, F.J., Olea, J., Aguado, D., Ponsoda, V., & Barrada, J.R. (2010). Deterioro de parámetros de los ítems en test adaptativos informatizados: estudio con eCat. *Psicothema, 22*, 340-347.

Barrada, J.R., Abad, F.J., & Veldkamp, B.P. (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. *Psicothema, 21*, 318-325.

Council of Europe (1999). *Relating language examinations to the common european framework of reference for languages: Learning, teaching, assessment (CEFR)*. Language Policy Division, Strasbourg.

Drasgow F., Levine M.V., Tsien S., & Williams B. (1995). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement, 19*, 143-165.

Hanson, B.A., & Beguin, A.A. (2002). Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*, 3-24.

López-Cuadrado, J., Pérez, T.A., Vadillo, J.A., & Gutiérrez, J. (2010). Calibration of an item bank for the assessment of Basque language knowledge. *Computers and Education, 55*, 1044-1055.

Muthén, B., & Muthén, L.K. (2006). *Mplus*. Los Angeles: Muthén & Muthén.

Olea, J., Abad, F.J., Ponsoda, V., & 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.

Olea, J., & Ponsoda, V. (2003): *Tests adaptativos informatizados*. Madrid: UNED.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of *S-X2*: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289-298.

Rebollo, P., García-Cueto, E., Zardaín, P.C., Cuervo, J., Martínez, I., Alonso, J., Ferrer, M., & Muñiz, J. (2009). Desarrollo del CAT-Health, primer test adaptativo informatizado para la evaluación de la calidad de vida relacionada con la salud en España. *Medicina Clínica, 133*, 241-251.

Rubio, V., & Santacreu, J. (2003). *TRASI. Test adaptativo informatizado para la evaluación del razonamiento secuencial y la inducción como factores de la habilidad intelectual general*. Madrid: TEA Ediciones.

Stark, S. (2001). *MODFIT: A computer program for model-data fit*. Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.

Thissen, D., Chen, W.H., & Bock, R.D. (2003). *Multilog (version 7) [Computer sotware]*. Lincolnwood, IL: Scientific Software International.

Wainer, H. (2000). CATs: Whither and whence. *Psicológica, 21,* 121-133.