Psicológica (2000) 21, 157-173.

Psychometric and psychological effects of review on computerized fixed and adaptive tests

J. Olea*, J. Revuelta, M.C. Ximénez and F J. Abad

Autonoma University of Madrid

Two computerized versions of an English vocabulary test for Spanish speakers (an adaptive and a fixed one) were applied in a Spanish sample of first-year psychology undergraduate students. The effects of test type (computer-adaptive vs. computerized-fixed) and review condition (allowed vs. not allowed) on several psychological variables were examined. Within-subject variables were measured both before and after review to study the effects of review on the psychological and psychometric variables for the review condition in both tests. Two major results were obtained after review: a) a significant increase of correct responses and estimated ability, and b) a decrease of the state-anxiety level. The differences were not significant for measurement error. Interaction effects (test type by moment) were not significant. These and other results concerning the assessment conditions established in this and previous papers are discussed. Finally, the implications that the results may have to establish review conditions in computerized adaptive tests are commented.

Key words: computerized adaptive tests, fixed-item tests, item review

The opportunity to review and therefore to change the answers that a subject initially gives to an achievement test is something usually done in conventional paper-and-pencil tests. However, opportunities for item review and answer change are far less common in computerized tests. Both in fixed-item tests (FITs) and in computerized adaptive tests (CATs) the majority of the examinees manifest a clear preference for item review because they perceive that the test is more fair and consider it a disadvantage if review is disallowed (Vispoel, 1993). In assessment contexts where the tests scores have important consequences for subjects, it is assumed that the opportunity to review increases the comfort of the examinees and contributes to control their emotional state towards the stressful testing situation. For certain subjects, this can increase their performance level in the test: "highly tests anxious examinees might engage in item revision for a

* This research was supported in part by the research project of the D.G.E.S. PB97-0049. The authors wish to thank Eulogio Real for his help in the data gathering process and for his comments.

variety of affective reasons (e.g. to gain a perception of control over a stressful testing situation)" (Wise, 1995, p. 4). There is a general agreement for allowing review in conventional paper-and-pencil tests. However, the inclusion of review in CATs is not so clear because it may have negative effects (e.g. increasing error measurement and obtaining ability inflated scores) that will be discussed below.

An analysis of previous studies about the effects of item review on paper-and-pencil tests (for example the meta-analysis by Waddle and Blankenship, 1995; or the papers by Vispoel, 1998; and Wise, 1995) indicates that: a) only a small percentage of answers are changed (between 3 and 5%); b) about 85% of the examinees change some of their initial answers; c) 68% of subjects who change answers improve their scores, 15% deteriorate them and 17% remain the same; and d) 57% of modified answers represent changes from wrong to right and this contributes to increasing the test score.

Review is usually not allowed in computerized FITs. This is perceived as frustrating by the examinees (Vispoel, Wang, de la Torre Bleiler & Dings, 1992) and may question its equivalence with the paper-and-pencil FIT versions (Bunderson, Inouye & Olsen, 1989; Sykes & Ito, 1997). One of the few studies aimed specifically to study the effects of review on computerized FITs (Vispoel, in press) found: a) a similar number of changed answers than in paper-and-pencil FITs (4%); b) a lower percentage of subjects modifying answers (45%); c) among them, more than the 50% increased their performance in the test; d) significant inverse linear relations between anxiety with performance and estimation precision; e) non-significant effects of review (allowed vs. disallowed) and its interaction with anxiety as related to performance and estimation precision; f) significant effects over testing time; g) various patterns of answer change in high and low ability subjects: high ability subjects showed less changes but increased their ability level more after review (more changes from wrong to right and less from right to wrong) as compared with low ability subjects; h) a significant positive linear relation between anxiety and positive attitude to the review condition.

The possibility of reviewing and changing answers may have more important consequences in CATs. There is a strong evidence about the efficiency of CATs as compared to FITs regarding of the psychometric properties of ability estimates (Lord, 1980a). CATs make more precise estimations with the same number of items or, in other words, reach the same precision level as FITs but with less items. This makes it possible to apply a greater number of tests in the same time period and decrease the costs of evaluation. CATs are specially used in psychological and educational testing contexts with large samples. For example, the Graduate

Record Examination (GRE) uses an adaptive version. Other examples are the Differential Aptitude Tests, the Armed Services Vocational Aptitude Battery, the COMPASS Placement Tests, the Computerized Adaptive Reporting and Testing and the CATs to assess knowledge and skills in different undergraduate exams, certifications, personal selection processes or student admissions (see Drasgow & Olson-Buchanan, 1999). Among the operative CATs there is only one which incorporates item review. It is the one developed with certification aims by the American Association of Clinical Pathologists (Bergstrom & Lunz, 1999). In order to avoid the use of illegitimate strategies to make score gains (e.g. to answer the items deliberately wrong in the first application and to answer them correctly after review) this CAT monitors the rate of correct responses after each item. When the rate is small, the test administers the most informative items for the pass point and not for the provisional ability estimate.

There are a variety of reasons for denying item review in CATs: a) It would increase the testing time and therefore suppose higher costs; this may represent a serious disadvantage from an operational point of view. b) It would reduce the precision in estimation because a CAT without review successively presents the more informative items for each examinee. c) It could produce illegitimate score gains because some subjects could use this option to answer the items correctly without knowing the correct answer (Wise, 1995).

The absence of item review in CATs is not well perceived by examinees. They feel a loss of control over the test and perceive that the test is unfair and increases their anxiety level (Stocking, 1997). The examinees argue legitimate reasons to include review (Wise, Freeman, Finney, Enders & Severance, 1997). Additionally this does not help the CATs, which are more efficient, to be socially more accepted. Also, the adaptive algorithm in itself makes the subjects answer correctly not much more than 50% of the items. This breaks with the idea of "the more you guess right, the better you perform". This can increase the anxiety level and affect the performance in the test. Both characteristics in the CATs have motivated different research trends which intend to establish the most appropriate testing conditions; which means the most similar to the conventional FITs without losing efficiency. First of all, adaptive variants have been proposed (e.g. the self-adapted-tests, SATs, which allow the subject to choose the item difficulty level, Wise, Ponsoda & Olea, in press). Secondly, the easy-CATs have been proposed (Lunz & Bergstrom, 1994). They permit a greater correct responses rate. Other easy-variants both in CATs and SATs have been proposed in Ponsoda, Olea, Rodriguez and Revuelta (1999). Finally, a more recent research trend is aimed to study the effects of review in CATs.

Previous studies: design and main results

This last research trend starts with an article by Wise, Barnes, Harvey and Plake (1989) where no significant differences were found among the group of subjects with and the one without review. Other empirical studies have studied the effects, mainly the psychometric ones, of item review in CATs (Lunz & Bergstrom, 1994; Lunz, Bergstrom & Wright, 1992; Stone & Lunz, 1994; Vispoel, 1998; in press; Vispoel, et al., 1992). A second set of studies both with real and simulated data has focused on the study of different answer strategies (related with 'test wiseness') to illegitimately increase the ability level in a CAT with review (Gershon & Bergstrom, 1995; Stocking, 1997; Vispoel, Rocklin, Wang & Bleiler, 1999; Wise et al., 1997). If we focus on the first group of studies and consider only the item review conditions in CATs (there is a set of studies which includes review conditions for SATs and a last set whose aim is to study omissions or answer deferring rather than review), the most typical design consists of establishing one condition of non-review and another of review. Within the review condition the effects of the differences before-after review on the ability level and the precision (information or measurement error) is studied. Additionally, the rate of changed answers in each of the three possible directions (wrong-right, wrong-wrong, right-wrong) and the percentage of subjects who modify their ability level (also in the three directions: increased, decreased or same performance) is studied. Some studies consider as independent variables the ability levels and the anxiety level and include the testing time as dependent variable (e.g. Vispoel, 1998). Other studies focus on the effect of review in the reliability of the pass/fail classification (Lunz & Bergstrom, 1994; Lunz, et al., 1992; Stone & Lunz, 1994). Only one of the studies establish the review and non-review conditions in a computerized FIT (see Vispoel, in press). Only one of the studies consider simultaneously a CAT and a FIT (see Vispoel, et al., 1992) but it does not include the review conditions for both tests. Some authors recognize that a limitation of their studies was that they were made in low-stakes conditions (Lunz, et al., 1994; Vispoel, 1998).

Among the main results of these studies we want to emphasize the following: a) about 60% of examinees changed at least one answer; b) only a small percentage (between 2 and 5%) of answers were changed; c) among the changed answers, about 50% of them are changed from wrong to right; d) among the subjects who change answers, between 42 and 52% improve their ability level in the test. Only a small percentage of subjects (between 10 and 15%) reduces it; e) slight losses of precision (ratios of variance error before and after review greater than 0.97) but never significant ones; f) correlations between estimated ability before and after review greater than

0.98 and mean differences between 0.02 and 0.07; g) negative and significant correlation between anxiety level and ability. Absence of significant effect for anxiety x review interaction; h) significant effect in testing time (review increases time between 37 and 61%); i) when ability is considered as independent variable, it is found that at the greatest levels of ability least answers are changed, the changes are mostly from wrong to right and least from right to wrong. Therefore, the highest levels of ability are the ones which most take advantage of review.

As can be seen, the majority of the previous studies are oriented to study the psychometric effects of review. Concerning these studies the main innovations of our study are: a) to establish a between-subject design to consider as independent variables: the computerized test type (CAT vs. FIT) and the review (allowed vs. not allowed) and study their main and interaction effects in the psychometric and psychological variables; b) to establish a mixed design to study the psychometric and psychological effects due to the revision in both tests. Among the psychological variables included as dependent variables are the state-anxiety level and the degree of comfort experienced by the examinees during the test.

With the consideration of state-anxiety level measures, it is expected to obtain a decrease in anxiety (posttest-pretest negative differences), a greater comfort in the review conditions and a decrease in the within-subject anxiety after review. Concerning the psychometric characteristics, it is expected that after review occurs: a) an increase in number of correct responses and mean level of estimated ability; b) a decrease on estimation precision; and c) more than the half of answer changes from wrong to right.

METHOD

Participants. 184 first year undergraduate students (143 females and 41 males) of Psychology from two Spanish universities participated in the study (Autónoma University of Madrid and University of Santiago). Ages of the subjects ranged between 17 and 19 years.

Design. In order to study the effects of allowing and not allowing review a 2 x 2 between-subjects design was defined. The first independent variable is the test type (CAT vs. FIT) and the second the review condition (review vs. non-review). Both the main and interaction effects of the two variables were studied in the dependent variables posttest-pretest anxiety and comfort.

To study the effects of review a 2 x 2 mixed design was defined. The between-subjects variable is the test type (CAT vs. FIT) and the within-subjects variable is the review moment (before vs. after review). Both the main and interaction effects of the two variables were studied in the dependent variables estimated ability, measurement error, correct responses, testing time and anxiety.

Given that a FIT of the same size as a CAT will provide less precise ability estimations, the subjects who answer only the FIT conditions will also be given a CAT to obtain more reliable estimations of their ability level.

Instruments. Both the computerized CAT and the FIT include items to assess the English vocabulary level for the Spanish speaking population. Each item consists of an English word and five Spanish alternatives of translation. Of the five alternatives, subjects choose the correct one with the cursor arrow from the keyboard and press backspace when they are sure of their response. They have 15 seconds to answer each item. The remaining time available to answer can be seen at the right top corner of the screen. During the 15 seconds the subject has to select and confirm a response alternative. In doing so the test administers the subsequent item. If the subject has not confirmed a response after the 15 second the test takes the last alernative selected by the subject. If no alternative has been selected after the 15 seconds the test considers that the response is incorrect and continues with the subsequent item.

The item pool for the CAT consists of 221 items calibrated with the three parameters logistic model (more details about its construction, calibration, assumptions verification and parameter distributions may be found in Olea, Ponsoda, Revuelta & Belchí, 1996; Ponsoda, Wise, Olea & Revuelta, 1997).

The CAT algorithm selects items according to the maximum information principle (Ponsoda, Olea & Revuelta, 1994). After each answer, the algorithm estimates the provisional ability level of the subject by conditional maximum likelihood. The entry point of the CAT is a random ability level between –0.40 and 0.40. The test includes a correct answer and a wrong one in an easy item (b = -4) and a difficult one (b = 4) to obtain finite maximum likelihood estimates when all the actual responses are correct or incorrect. To avoid extreme ability estimates the algorithm implements the solution proposed by Revuelta and Ponsoda (1997). The procedure to control item exposure is ‘the progressive method’ (see Revuelta & Ponsoda, 1998) which consists of increasingly weighing the item information as the test progresses. The CAT stopped when the examinee answered 20 items.

The computerized FIT consisted of 20 items ordered by difficulty. Items were selected from the pool with the criteria that they were optimal for the ability distribution in the psychology students population (normal (0.57; 0.92); see Olea, et al., 1996). Figure 1 includes the information function for both the CAT and the FIT.

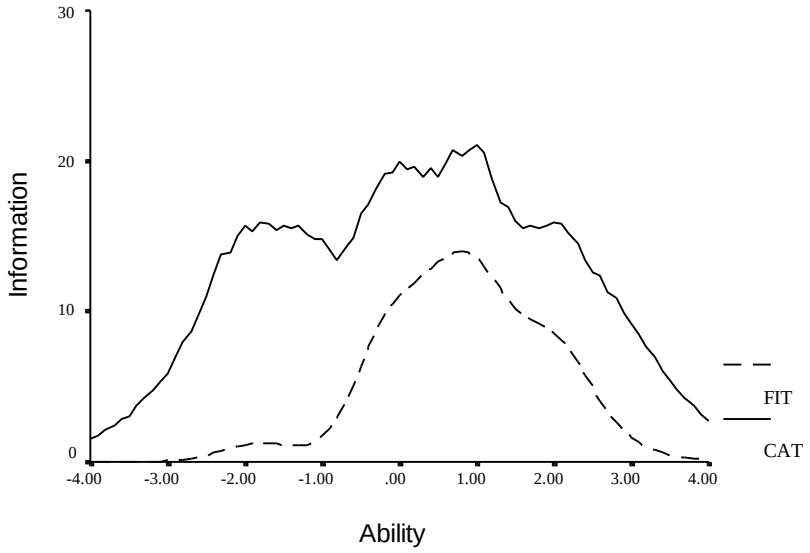

Figure 1. Information function of the adaptive and fixed test

The state-anxiety of subjects was measured with the Spanish version of the State-Anxiety Scale, SAS (Spielberger, Gorsuch & Lushene, 1970). The 20 items of the scale were divided in two equivalent parts; one to be applied before the vocabulary tests and the other as a measure of state-anxiety posttest. The equivalence and factor validity of both parts were studied in Ponsoda et al. (1999).

The comfort towards the tests was measured with two computerized items with five ordered response categories. Item 1 refers to the calmness level during the test administration and item 2 to the degree to which the subjects perceive that the test reflects their true English vocabulary level.

Procedure. As in other computerized tests studies where state-anxiety is measured, a main concern in planning the research was to establish the most realistic testing conditions. The students should perceive that it was a real testing situation (not an experiment) and that their performance could have some consequences. In order to reach medium-stakes conditions: a) the

sample consisted of the first year undergraduate psychology students; b) data gathering took place in their first three weeks of class; c) a teacher of Psychometric methods persuaded all of them to take the English test; d) they were told about the importance of knowing English to understand the advances that occur in the Psychology field; e) they were told that the results would be displayed on the notice board in class; f) they were informed that the head of the Psychology department was the promoter of the English assessment. Each student who agreed to participate registered on a certain day and hour to answer the test. After all data were gathered, results were anonymously published. They included their identity number, their percentile and a brief explanation.

The testing sessions took place in two computer rooms, one with 30 positions and the other one with 20. Students were randomly assigned to each of the four experimental conditions. Then, all conditions appeared in each session a variable number of times depending on the number of subjects in each session. Once the students were seated in front of the computer, a researcher gave few instructions about the procedure and informed the subjects that they would receive punctual instructions through the computer.

The testing session for each examinee was as it follows: 1) to write down their name and identity number. 2) General instructions, 3 examples and 10-item SAS pretest. 3) Instructions for the English vocabulary test: the subjects in the non-review condition were told that once the answer was given it could not be modified. The subjects in the review condition were told that in the end of the test would be allowed to review and modify their answers. 4) Test administration: 4 examples and 20 English vocabulary items. It was established a time-limit of 15 seconds for each item. 5) The examinees in the review condition were asked if they wanted to review their answers. For the ones who decided to review, the same items and their initial answers were presented and they were allowed to modify each of them in a 15 seconds time-limit. 6) SAS posttest and comfort items. Additionally, the subjects assigned to the FIT answered a 20-item CAT based on the original item pool but not including the FIT items.

Data analysis. The effects of test type, review condition and test type x review on post-pretest anxiety and comfort of the between-subject design were evaluated using two-way analyses of variance for fixed effects factors. Concerning the mixed design, the effects of the between-subjects factor (test type), the within-subjects factor (moment: before and after review) and the interaction (test x moment) on ability estimates, measurement error, correct responses, testing time and anxiety were evaluated using two-way analyses of variance with repeated measures in one factor and fixed effects in the

other factor. Both analyses were made with the general linear model SPSS subroutine. Some correlations between variables (e.g. ability with and without review) were studied with the Pearson correlation coefficient. The relations between ability level (dichotomized by the median) and score gains (positive, negative or null changes) after review were evaluated using the chi-square test.

RESULTS

Before data analysis two subjects were eliminated (the standard error of estimation was 2.47 for the first subject and the estimation algorithm did not converge for the second subject). Hence, the final sample consisted of 182 subjects.

Allowed review versus not allowed review

Significant correlations between the estimated ability level and the anxiety measures (pretest, posttest or post-pretest) were not found either in the total sample or in any of the four conditions.

Table 1 shows the analyses of variance results for the between subjects model. The examinees allowed to review their answers show a decrease in their anxiety level mean after the test (post-pretest mean =-0.91). However, the ones who are not allowed to review show an increase in their anxiety level (post-pretest mean = 0.51). Results were not significant for the main effect of test type and for the interaction.

Concerning the comfort with the testing situation, subjects who answer the FIT show more calmness in the posttest (item 1 mean = 3,24) than the ones who answer the CAT (item 1 mean = 2,97). Results about item 2 were not significant. None of the interaction effects was significant.

Table 1. ANOVA results for post-pretest anxiety and comfort by test type and review condition in the total sample

	CAT		FIT		Statistical test		
	Review (*n* = 45)	No-Review (*n* = 46)	Review (*n* = 47)	No-Review (*n* = 44)	Test	Review	Test x Review
Post-pretest anxiety	-1,64 (4,02)	0,35 (4,52)	-0,21 (3,62)	0,68 (2,88)	*ns*	*p* = 0,012	*ns*
Comfort							
ITEM 1	3,02 (0,99)	2,91 (0,94)	3,36 (0,82)	3,11 (0,72)	*p* = 0,039	*ns*	*ns*
Item 2	2,07 (0,86)	2,07 (0,83)	2,04 (0,78)	2,09 (0,83)	*ns*	*ns*	*ns*

Note: **The CAT and FIT cells represent the means and standard deviations (in parentheses) of variables. The last three columns contain the results of significance tests of the ANOVA F-ratios (*ns*: *p* > 0,05). Item 1 for comfort refers to the calmness level during the test administration, and item 2 to the degree to which the subjects perceive that the test reflects their true English vocabulary level.**

Table 2. ANOVA results for estimated ability, measurement error, correct responses, time and anxiety by test type and moment in the review condition sample

	CAT (*n* = 45)		FIT (*n* = 47)		Statistical test		
	Before	After	Before	After	Test	Moment	Test x Moment
Estimated ability	0,53 (0,69)	0,64 (0,71)	0,86 (0,72)	0,97 (0,76)	*p* = 0,027	*p* < 0,0001	*ns*
Measurement error	0,24 (0,02)	0,25 (0,04)	0,32 (0,06)	0,31 (0,05)	*p* < 0,0001	*ns*	*ns*
Correct responses	13,02 (1,48)	13,76 (1,94)	11,49 (3,14)	11,96 (3,40)	*p* = 0,002	*p* < 0,0001	*ns*
Testing time	195,22 (63)	295,11 (120)	191,65 (61)	295,40 (119)	*ns*	*p* < 0,0001	*ns*
Anxiety	20,42 (5,07)	18,78 (3,84)	18,81 (4,53)	18,60 (3,07)	*ns*	*p* = 0,002	*ns*

Note: The CAT and FIT cells represent the means and standard deviations (in parentheses) of variables. The last three columns contain the results of significance tests of the ANOVA F-ratios (*ns*: *p* > 0,05).

Answer changing behavior within the review condition

Of the 92 subjects who were allowed to review, 75 (a 81.52% of the total) chose to do it. In the CAT, 80% chose to review and among them 66.7% improved their ability estimates, 25% decreased them and 8.3% remained the same than before review. In the FIT, 83% chose to review and among them 59% improved their ability estimates, 20.5% decreased them and 20.5% remained the same as before review. The chi-square statistic for ability level (above and below the median) and change level (positive, negative and null) was not significant.

In terms of changed answers, in general 13.5% of the answers were changed. In the CAT, 12.6% answers were changed and among them 42% were from wrong to wrong, 43% from wrong to right and 15% from right to wrong. In the FIT, 14.4% answers were changed and among them 54% were from wrong to wrong, 31% from wrong to right and 15% from right to wrong.

Review effects

The Pearson correlation between ability levels before and after review was 0,951 in the CAT and 0,941 in the FIT.

After review, the correlations between estimated ability and standard error of estimation were significant (0.349 in the CAT and 0.685 in the FIT). This indicates that the lower-levels of ability in this sample are best measured than the higher-levels.

Table 2 shows the analyses of variance results for the mixed model with the review condition sub-sample.

Several main effects of test type were significant. The mean of ability estimates in the FIT (0.9158) is significantly greater than its corresponding in the CAT (0.5831). The mean number of correct responses is greater inthe CAT (13.39) than in the FIT (11.72). The standard error of estimation is greater in the FIT (0.3146) than in the CAT (0.2457).

The first of these results may appear strange. Among the possible explanations are the following: 1) the true ability levelof the subjects who answered the FIT was greater than the one of the subjects who answered the CAT, 2) the maximum-likelihood algorithm implemented might overestimate the ability levels for the FIT group, and 3) the maximum likelihood estimation method overestimates the true ability level of the subjects. Although it is not possible to clarify the supposed bias because the true ability levels of the subjects are unknown, some additional verifications were made to clear up the reasons for these differences. First, to estimate the

ability parameters for these subjects (using the BILOG program) not considering the answers to the two fictitious items. The differences between means estimated after review with this program (0.894 vs. 0.975) were not significant ($p > 0.01$). The differences in standard error of estimation were null (0.31 in both). Thus, it seems that the problem is not in the estimation algorithm used. Second, because the subjects who answered a FIT also answered to a CAT (let us call them CAT_{FIT}), means in ability and standard error of estimation among groups were compared. The differences between means in the CAT and CAT_{FIT} estimates were not significant. Therefore, it may be that the maximum likelihood method overestimates the true ability level of the subjects in the FIT.

Except for the standard error of estimation, all effects of moment (before and after review) were significant. After the review, the mean in estimated ability is significantly increased (0.8058 after vs. 0.6932 before), the number of correct responses is significantly increased (12.86 after vs. 12.26 before), the testing time is significantly incremented (295.26 after vs. 193.44 before), and the level of anxiety decreases significantly (18.69 after vs. 19.61 before). None of the interaction effects was significant.

DISCUSSION

The main objectives of the present study were: a) to study several psychological effects of allowing vs. non-allowing answer review in a computerized adaptive and fixed-item test; b) to study the psychological and psychometric effects of review in both tests types.

Concerning the first objective, the group which was allowed to review showed a significant decrease in the state-anxiety. On the average, the subjects who were not allowed to review increased their anxiety level after the test. These results confirm some of the reasons to include review in real testing situations (Stocking, 1997). Subjects exhibited more calmness level in the FIT than in the CAT. This result is not consistent with the anxiety results, where these differences do not appear. We consider that the anxiety results showed be more credited because the SAS is a longer and more reliable test and because the means considered in the SAS (post-pretest) are more appropriate to control the random unbalances that may occur between the different conditions in the psychological variables. Also, the subjects who are allowed to review do not consider that the test is more appropriate to reflect their true ability level.

Concerning to the second objective, the two major results are: a) after review there is a significant decrease of state-anxiety in both tests types; b)

review increases significantly the number of correct responses and the estimated ability level without loosing precision in the estimation. The first of these results is novel since no previous studies obtain post-anxiety measures. The state-anxiety level decreasing confirms that the experiment was carried out in a minimum level of stake conditions. If future studies replicate this result, this would represent a reason to include review in operative CATs.

Concerning the effects obtained in the psychometric dependent variables, none of the previous studies (Lunz & Bergstrom, 1994; Lunz, et al., 1992; Stone & Lunz, 1994; Vispoel, 1998; in press) reached significance for the increased estimated ability after review. There are different reasons for the result found in our study. It is evident that the sample of Spanish undergraduate psychology students changed more items (81.52% of examinees review their answer and 13.5% of answers are changed) if we compare them with the data of previous studies (no more than the 60% of examinees reviewing answers and 5% of answers changed). After review, more than 60% of subjects increase their ability level and this also represents a greater percentage as compared with previous studies. Some of the testing conditions planned in this study may be on the basis of these results. First of all, the present work establish a time-limit of 15 seconds for each item. This is different to previous studies where there are no time-limits for each item. Secondly, it is the first time that a study of this characteristics is done with a Spanish sample of examinees, who are less familiar with answering performance tests. Third, the decreases in state-anxiety confirm that the planned testing situation is of medium-stake conditions for examinees and this can stimulate the revision. If this last reason were true and the significant increase in mean estimated ability was obtained in other studies with high-stake conditions, this would be essential for the authors to establish the best conditions for examinees in large scale testing situations. The greater answer changes and score gains make the correlation between ability estimates before and after review (0.95 for both tests) to be slightly smaller as compared to previous studies.

Similar as in previous research we have assumed that the item parameters do not change under the revision mode (Vispoel, 1998; in press; Stocking, 1997; Stone & Lunz, 1994). However, the revision conditions may invalidate the items of subject parameters. The item pool was calibrated in the without revision mode, but since revision makes that the subjects double the time to answer to a single item, this may imply a decrease in the item difficulty. A second explanation could be that the revision condition changes the definition of theta because the properties of the items differ from the non-revision condition. Further research would be necessary in order to

clarify if the revision modify the items parameters, the subjects abilities, or both.

Another topic to discuss is the legitimacy of the score gains obtained by examinees after review. Among the possible reasons for a subject to obtain illegitimate score gains are three: a) one item may serve as a clue for others; b) the examinee may intentionally answer items incorrectly in the test application and answer them correctly during the review (this is called the Wainer strategy); c) to detect the incorrect items in the first application from the difficulty inferred from the subsequent items because the subject realizes that after a wrong answer the test presents an easier item (this is called the Kingsbury strategy). To which degree may our examinees benefit fromone of these strategies?. It is very difficult that the first reason has occurred because of the established controls for the items design (Olea, et al., 1996) and the verification of the assumptions of the model (Ponsoda, et al., 1997). Also, since the options of items were ramdomly established from the words of an English dictionary this minimizes the possibility of items dependence. It is also difficult that the Wainer strategy has occurred because in Spain there is not an operative CAT and the first year psychology undergraduates do not know how an adaptive test works. Concerning the Kingsbury strategy, we do not have data about its incidence but some studies (Wise, et al., 1997) have found how difficult it is for examinees to calibrate items difficulty, detect their wrong answers in the first items of the test and take advantage of this kind of strategy. Moreover, it would be necessary to know how the CAT algorithm works to benefit from this strategy.

Some studies (Vispoel, 1998; in press) suggest that high ability subjects benefit more from review in the sense that they obtain greater changes from wrong to right and less from right to wrong than low ability subjects. Our study does not replicate this result. It is possible that the specific characteristics of the sample used (with a medium-high ability level) are not the most appropriate to confirm this result but it could be confirmed by using more heterogeneous samples in English vocabulary.

The differences obtained between both tests types in mean estimated ability (significantly greater in the FIT) represented an unexpected result. These differences were not significant when the CAT measures were considered in the FIT condition. These results may indicate that the maximum likelihood estimations of ability in FITs with specific properties (e.g. high difficulty for the sample) could overestimate the examinees level of ability. It would be interesting to carry out simulation studies to confirm this possible bias and manipulate variables such as the test size, its difficulty, or the statistical method of parameter estimation.

Finally, the testing time is significantly increased (in a 52%) with review. This aspect should be assessed by the authors of large scale CAT and FIT real applications.

To summarize, and only within the conditions and sample used in the present study, it can be concluded that the review condition contributes to decrease state-anxiety of examinees and to increase the estimated ability level. The study was conducted with a sample where it is unlikely that illegitimate strategies have been used. Not allowing review both in CATs and in FITs would contribute to a greater level of discomfort and to underestimate the ability levels of the majority of the subjects. Under these conditions, there are no arguments to advise against the inclusion of item review.

RESUMEN

Efectos psicométricos y psicológicos de la revisión de respuestas en tests fijos y adaptativos informatizados. Se aplicaron dos versiones informatizadas de un test de vocabulario inglés para hispanohablantes (uno fijo y otro adaptativo) a una muestra de estudiantes españoles de primer curso de Psicología. Se estudiaron los efectos del tipo de test (fijo versus adaptativo) y de la condición de revisión (permitida versus no permitida) sobre diversas variables psicológicas. Se analizaron los efectos de la revisión en ambos tests (diferencias antes-después) en una serie de variables psicológicas y psicométricas. Después de la revisión, dos fueron los resultados más destacables: a) un incremento significativo del número de aciertos y de la habilidad media estimada, y b) un descenso significativo del nivel de ansiedad estado de los evaluandos. No se obtuvieron diferencias significativas en precisión. Tampoco resultó significativo el efecto de la interacción entre el tipo de test y el momento (antes versus después de la revisión). Se discuten estos resultados y otros relativos a las condiciones de evaluación establecidas en el presente trabajo y en otros realizados previamente. Finalmente, se comentan las implicaciones que los resultados pueden tener para permitir la revisión en la aplicación real de tests adaptativos informatizados.

Palabras clave: Tests adaptativos informatizados, tests fijos, revisión de respuestas

REFERENCES

Bergstrom, B. & Lunz, M. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Buchanan (Eds.). *Innovations in computerized assessment*. Mahwah, NJ: LEA.

Bunderson, C.V., Inouye, D.K. & Olsen, J.B. (1989). The four generations of computerized educational measurement. In R.L. Linn (Ed.). *Educational measurement* (3rd. Edition). New York: McMillan.

Drasgow, F. & Olson Buchanan, J.B. (Eds.) (1999). *Innovations in computerized assessment*. Mahwah, NJ: LEA.

Gershon, R.C. & Bergstrom, B. (1995). *Does cheating on CAT pay: Not*. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA.

Lord, F. (1980a). *Applications of IRT to practical testing problems*. Hillsdale, NJ: LEA.

Lunz, M.A. & Bergstrom, B.A. (1994). An empirical study of computerized adaptive test administration conditions. *Journal of Educational Measurement*, 31, 251-263.

Lunz, M.A., Bergstrom, B.A. & Wright, B.D. (1992). The effect of review on student ability and test efficiency for computerized adaptive tests. *Applied Psychological Measurement*, 16, 33-40.

Olea, J., Ponsoda, V., Revuelta, J. & Belchí, J. (1996). Propiedades psicométricas de un test adaptativo de vocabulario inglés [Psychometric properties of a CAT for the measurement of english vocabulary]. *Estudios de Psicología*, 55, 61-73.

Ponsoda, V., Olea, J. & Revuelta, J. (1994). ADTEST: A computer adaptive test based on the maximum information principle. *Educational and Psychological Measurement*, 54, 3, 680-686.

Ponsoda, V., Olea, J., Rodríguez, M. S. & Revuelta, J. (1999). The effects of test difficulty manipulation in computerized adaptive testing and self-adapted testing. *Applied Measurement in Education*, 12, 167-184.

Ponsoda, V., Wise, S., Olea, J. & Revuelta, J. (1997). An investigation of self adapted testing in a Spanish high school population. *Educational and Psychological Measurement*, 57, 210-221.

Revuelta, J. & Ponsoda, V. (1997). Una solución a la estimación inicial en los tests adaptativos informatizados. [A solution to initial estimation in CATs]. *Revista Electrónica de Metodología Aplicada*, 2, 1-6.

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

Rocklin, T.R. & O'Donnell, A.M. (1987). Self-adapted testing: A performance improving variant of computerized adaptive testing. *Journal of Educational Psychology*, l79, 315-319.

Spielberger, C.D.; Gorsuch, R.L. & Lushene, R.E. (1970). *Manual for the State-trait anxiety inventory*. Palo Alto: Consulting Psychologists Press. Spanish adaptation by TEA ediciones S.A. (1988). 3th edition.

Stocking, M.L. (1997). Revising item responses in computerized adaptive tests: a comparison of three models. *Applied Psychological Measurement*, 21, 129-142.

Stone, G.E. & Lunz, M.E. (1994). The effect of review on the psychometric characteristics of computerized adaptive tests. *Applied Measurement in Education*, 7, 211, 222.

Sykes, R.C. & Ito, K. (1997). The effects of computer administration on scores and item parameter estimates of an IRT-based licensure examination. *Applied Psychological Measurement*, 21, 51-63.

Vispoel, W.P. (in press). Reviewing and changing answers on computerized fixed-item vocabulary tests. *Educational and Psychological Measurement*.

Vispoel, W.P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement*, 35, 328-345.

Vispoel, W.P. (1993). Computerized adaptive and fixed item versions of the ITED vocabulary subtest. *Educational and Psychological Measurement*, 53, 779-788.

Vispoel, W.P., Rocklin, T.R., Wang, T. & Bleiler, T. (1999). Can examinees use a review option to positively bias their scores on a computerized adaptive test? *Journal of Educational Measurement*, 36, 2, 141-157.

Vispoel, W.P., Wang, T., de la Torre, R., Bleiler, T. & Dings, J. (1992). *How review options, administration mode and anxiety influence scores on computerized vocabulary tests*. Paper presented at the Meeting of the National Council on Measurement in Education, San Francisco (ERIC Document Reproduction service, No. TM018547).

Waddell, D.L. & Blankenship, J.C. (1995). Answer changing: a meta-analysis of the prevalence and patterns. *Journal of Continuing Education and Nursing*, 25, 155-158.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, 12, 15-20.

Wise, S.L. (1999). Tests autoadaptados informatizados: fundamentos, resultados de investigación e implicaciones para la aplicación práctica. [Self-Adapted Testing: Fundamentals, research results and implications for applied contexts]. In J. Olea, V. Ponsoda and G. Prieto, *Tests informatizados: fundamentos y aplicaciones*. [Computerized Testing: Fundamentals and applications] Madrid, Pirámide.

Wise, S.L. (1995). *Item review and answer changing in computerized adaptive tests*. Paper presented at the Third European Conference on Psychological Assessment. Trier, Germany.

Wise, S.L., Barnes, L.B., Harvey, A.L. & Plake, B.S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. *Applied Measurement in Education*, 2, 235-241.

Wise, S.L., Freeman, S.A., Finney, S.J., Enders, C.K. & Severance, D.D. (1997). *The accuracy of examinee judgements of relative item difficulty: Implications for computerized adaptive testing*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.