Psicothema
2001. Vol. 13, nº 1, pp. 152-158

ISSN 0214 - 9915 CODEN PSOTEG

SOFTWARE, INSTRUMENTACIÓN Y METODOLOGÍA

Analysis of the optimum number alternatives from the Item Response Theory

Francisco J. Abad, Julio Olea and Vicente Ponsoda
Universidad Autónoma de Madrid

The optimum number of alternatives in multiple choice items is studied. The new procedure applied is based on Item Response Theory and only needs a single sample. Concretely, a 5-alternative 221 items English Vocabulary test was administered to 452 people. Their answers to the 1/2/3 worst alternatives were randomly reassigned to generate their hypothetical answers to items of 4/3/2 alternatives, respectively. Changes in item parameters, test information function and ability estimation were analysed. The worst results were provided by 2-alternative items. As the results in the 3- and 4-alternative items do hardly differ from those obtained in the original 5-alternative case, and in the 3-alternative case the role of partial knowledge is less important, 3-alternative items were considered the best choice. The effects of number of alternatives on ability estimation was also discussed.

Análisis del número óptimo de opciones desde la teoría de la respuesta al ítem. En este estudio se analiza el número óptimo de alternativas en tests de elección-múltiple desde la Teoría de la Respuesta al Ítem (TRI). Se estableció un nuevo procedimiento que permite el estudio de esta cuestión desde la TRI con una sola muestra. Concretamente se aplicó un Test de Vocabulario Inglés de 221 ítems de 5 alternativas a una muestra de 452 personas. Las respuestas de las 1/2/3 peores alternativas fueron reasignadas aleatoriamente para construir las respuestas a los ítems de 4, 3 y 2 opciones, respectivamente. Se analizaron los cambios en los parámetros *a*, *b* y *c* de los ítems, en la información del test y en la estimación de la habilidad. Se obtuvieron los peores resultados con los ítems de 2 alternativas. La reducción a 3 o 4 alternativas no tuvo efectos prácticos relevantes. Se recomiendan los ítems de 3 alternativas por ser menos susceptibles a la presencia de conocimiento parcial. Se comentan los efectos del número de opciones para diferentes niveles de habilidad.

There is a broad discussion about the optimum number of options in multiple-choice item formats, typical in achievement tests (eg. Haladyna & Downing, 1989). This format introduces an error in the estimation of ability which is inversely related to the quantity of alternatives. Although there are procedures to correct the effect of guessing, there is none effective for all kinds of conditions of application and/or contents (Ben-Simon, Budescu & Nevo, 1997); This is one of the reasons why a great and reasonable number of alternatives (4 or 5) is traditionally advised to minimize the presence of guessing. On the other hand, there are some advantages in using items with small number of alternatives (Haladyna & Downing, 1993): a) Constructing many incorrect functional options is difficult and expensive; b) The printing and application costs are lower; c) More items can be managed in a given time, permitting the increase of the content validity and test reliability. Hence, it has been considered important to what point we can reduce the number of options without affecting the psychometric quality of the test.

From the classical test theory (CTT), many researches have been done both from a theory (eg. Grier, 1975; Lord, 1980) as well as empirical perspectives (eg. Cizek, Robinson & O'Day, 1998; Crehan, Haladyna and Brewer, 1993; Delgado & Prieto, 1998; Haladyna & Downing, 1993; Owen & Froman, 1987). Generally, the same test is applied to several samples, modifying the number of the options of their items. Starting from an original test of 5 or 4 already applied alternatives, tests with fewer number of options are generated, and each test is applied to a new sample. Successively, options with a low choice proportion and, or a non-decreasing trace line are eliminated (eg. Haladyna & Downing, 1993).

Correspondencia: Francisco J. Abad
Facultad de Psicología
Universidad Autónoma de Madrid
28049 Madrid (Spain)
E-mail: fjose.abad@uam.es

Generally, the changes in the indexes of difficulty *p,* in the items discrimination indexes (item-test biserial correlations) and in the test reliability (indexes a or r_{xx}) are analysed. Occasionally, these and other parameters (eg.; the number of inefficient options) are analysed according to the level of ability.

The most consistent result of this type of study is that the reduction of the number of alternatives has little psychometric effect, with the number of options being reducible to 2 or even 3 (Downing, 1992). Specifically, the reduction of the number of alternatives: *a)* slightly reduces the difficulty, *b)* does not affect in a systematic, practical and, or significant way, neither the item discrimination nor the reliability; *c)* can reduce the time of application (Owen & Froman, 1987), though it is not clear if a systematic relation exists between the time to solve a test and the number of options (Budesco & Nevo, 1985); *d)* affects subjects with different levels of ability in different ways: Green, Sax & Michael (1982) and Weber (1978) compared the test reliability of 3, 4 and 5 alternatives tests, without obtaining significant differences for the high-ability group, but for the low ability group. Furthermore, Trevisan, Sax & Michael (1991) proved that in the high and average ability groups, there was a greater number of unused distractors.

So, the most common conclusion of these studies is that, it is difficult to draw up more than 3 effective options. Furthermore, the "optimum number" of options may not be independent of the subjects' levels of ability. The number and quality of wrong options is irrelevant to the subjects who know the answer. On the other hand, higher ability subjects have a greater probability of discarding the wrong options (i.e.: partial knowledge; see Hutchinson, 1997; Waller, 1989). The effect of reducing the number of alternatives should be greater for subjects with a lower ability level. However, some contradictions exist about this topic (eg: Trevisan, Sax & Michael, 1991).

Few studies have been carried out from the IRT, even when they constitute the ideal setting for the analysis of this problem, since they give more appropriate indicators to compare the test accuracy according to the ability level being measured (information function), the estimates of the parameters being independent of the sample variations. By the means of simulation using the parameters of an empirically calibrated bank, Lord (1980; page 110) manipulated the guessing parameter without changing the other parameters of the logistic model. He found that when the number of alternatives is reduced (and the number of items is proportionally increased) in the high-ability subjects, the longer the test, the higher the level of information obtained. For lower-ability test levels, shorter tests (and with more options) turned out to be more informative. In the same line, Levine & Drasgow (1983) analysed the answer patterns of subjects in items of 5 alternatives; only 1 or 2 options were chosen for the high-ability subjects, while lower-ability subjects used more options.

Although the results pointed in the expected direction, in the Lord's (1977;1980) study, where the effect of the number of options of the test is analysed according to the information of a test, the law of proportionality is assumed. Furthermore it is a simulation study. On the other hand, none of the previous studies analysed the direction of ability changes when items with different number of options are used. The main reason for not applying the IRT to this problem is probably the cost of applying each form of test with different number of options to big samples so that the parameters could be accurately estimated. In this context of research, the following goals are set out: a) show the effectiveness of a new strategy to reduce the number of options which permit the application of the IRT and which results plausible from a psychological perspective; b) study the variations produced in the parameters (*p,* biserial correlation, α) considered from the CTT, and which should not be different from those found in previous studies carried out from this approach; c) study the variations produced in the parameters from the IRT (*a,b,c).* The fact that only *c* or more parameters are affected informs of the absence or presence of partial knowledge (Lord, 1980); d) analyse the effects on the accuracy for the different levels of ability. It is expected that the reduction of choices will affect the information function more in the low-ability subjects.

Method

Materials and Samples

The English vocabulary test (Olea, Ponsoda, Revuelta & Belchí, 1996; Ponsoda, Wise, Olea & Revuelta, 1997) was applied in this study. This test, which was calibrated for this research with the BILOG program is made up of 221 items with 5 alternatives, the mean of 1.005 for *a* (standard deviation, 0.374), 0.000 for *b* (standard deviation, 1.573) and 0.215 for *c* (standard deviation of 0.069). Standardization was done on the theta values. The correlations between the estimations of the parameters *a, b* and *c* were not significant (they ranged between -.071 and -.013). This test has already been used as an items bank in several studies on adaptive testing and computerized self-adapted testing (Ponsoda, Olea, Rodriguez & Revuelta, 1999; Ponsoda, Wise, Olea & Revuelta, 1997). In these studies, the information on the psychometric properties can be extended. The English vocabulary test was used in a sample of 452 subjects of heterogeneous levels of ability (secondary school students, undergraduates and university teachers).

Procedure

From the empirical answers of the subjects, the rest of the conditions were established through the following procedure: a) The choice of the least functional alternatives (1, 2 or 3). The least functional criterion was the least choice proportion. In this study, we consider frequency of response as the single indicator of the *plausibility* of each option. As pointed out by one reviewer, information about discriminating power might also be considered as an additional functional criterion; b) The detection of the subjects that have chosen these least functional alternatives; c) Random equiprobable re-allocation of the subjects that have chosen the least functional alternatives to the remaining alternatives (including the correct one). The subjects that did not choose the least functional alternatives maintained their chosen alternatives; d) Re-calibration of each data set resulting from the re-allocation. The re-allocation of the least functional alternatives produced the condition of 4–alternative items. The re-allocation of the 2 least functional alternatives produced the condition of 3-alternative items. The re-allocation of the 3 least functional alternatives produced the condition of 2-alternatives items.

Data Analysis

Each of the new conditions were calibrated (reducing 1,2 or 3 least functional alternatives) with the ITEMAN program (ASC, 1988), to obtain the CTT parameters, and BILOG (Mislevy & Bock, 1989) to obtain the parameters from the IRT. For the joint

IRT estimation of parameters, the Marginal Maximum Likelihood Bayesian (MMAP) was used, to avoid the Heywood cases in the item parameter estimation. For the ability estimation, the Maximum Likelihood procedure was used. The treatment of the omitted answers in the estimation consisted of allocating them a probability of success equal in reciprocation to the number of alternatives. After carrying out the estimation in each test, those items that fitted the model in all the tests were selected, with the aim that the compared items would always be the same. The χ^2 ($p>0.05$) statistics and the parameter estimation standard error *a* (lesser than 0.5), *b* (lesser than 0.5) and *c* (lesser than 0.1), were used as the selection criteria. In this way it was assured that the differences found did not result from the estimation problems of specific items. 113 items met these criteria. Finally, the metric scale of ability, estimated with these 113 items, was equated in the 3 new tests to the metric in the original test of 5 alternatives; the mean and standard deviation method (Kolen & Brennan, 1995) was used. In calculating the constants (bias and intercept) those subjects estimated with an elevated standard error ($S_\theta > 0.5$) were not considered. The final sample for this calculation was 398 subjects. The corresponding transformations were later applied to *a* and *b*.

Each of the analysis from the CTT was repeated twice, one considering the items bank, and the other limited to the 113 items fitted to the 3-parameter logistic model. For each of the conditions, we obtained: *a)* parameter descriptive statistics from the CTT (indexes *p*, biserial correlations) and the correlations of these parameters in the new conditions with those of the original condition; *b)* parameter descriptive statistics from the IRT (*a*,*b* and *c*) already re-scaled , and the correlations of these parameters in the new conditions with those of the original condition; *c)* test coefficients a; *d)* ANOVA contrasts and post hoc comparisons (Bonferroni) among the different conditions, using as dependent variables the indexes *p* (applying the logit transformation ln[p/(1-p)]) and the item-test biserial correlations (transformed into Z Fisher); *e)* ANOVA contrasts and post hoc comparisons (Bonferroni) among the conditions, taking as an additional independent variable the original item difficulty (with three equiprobable levels) and taking as dependent variables the parameters *a, b* and *c*; *f)* the test information function; *g)* the average differences of the θs estimated in the new conditions as regards the θs in the original condition; for this analysis, only the 398 subjects that had been estimated with a higher precision in the original condition were considered.

Results

Classical Test Theory Parameters (CTT)

In Table 1 we can see the descriptive statistics (total-item biserial correlations, *p* indexes, α coefficients) when 1, 2 or 3 options are eliminated. No significant differences were observed in the discrimination indexes ($F_{3.880}$=0.688, p=0.559 for the total bank; $F_{3.448}$= 0.930, p=0.426 for the fitted items). Logically, the changes in the α coefficients go with the changes in the items discrimination and they are not very relevant from a practical viewpoint. The correlation between the discrimination indexes of the 3 new conditions with the original condition were: 0.992, 0.973 and 0.526 (0.991, 0.953 and 0.390 for the fitted items as the 1, 2 or 3 least functional altematives were being eliminated. This points out that, though there were no changes in the mean discrimination level, the order of the items as regards its discrimination level differs in a certain way in the condition in which the 3 least functional alternatives were eliminated.

Table 1
Means for the Classical Test Theory parameters and the Item Response Theory; the data for the CTT is shown for the total items Bank (N=221) and for the items adapted to the IRT model (N=113)

		Original	-1 option	-2 option	-3 option
TOTAL DATA SET	Biserial	.34	.33	.33	.34
	Difficulty(p)	.60	.62	.64	.72
	α	.97	.96	.96	.97
FITTED ITEMS	Biserial	.38	.37	.37	.39
	Difficulty(p)	.55	.57	.59	.69
	α	.95	.94	.94	.95
FITTED ITEMS (EASY ITEMS)	*a*	1.01	1.01	1.02	.75
	b	-.73	-.71	-.66	-.98
	c	.23	.27	.34	.45
FITTED ITEMS (AVERAGE DIFFICULTY ITEMS)	*a*	1.06	1.11	1.10	.92
	b	.35	.41	.33	-.06
	c	.24	.29	.32	.37
FITTED ITEMS (DIFFICULTY ITEMS)	*a*	.95	.96	.89	.96
	b	1.48	1.50	1.36	.57
	c	.20	.23	.25	.28
FITTED ITEMS (TOTAL)	*a*	1.01	1.03	1.01	.88
	b	.37	.40	.34	-.16
	c	.22	.27	.30	.37

The difficulty indexes *p* increased significantly (the items become easier), ($F_{3.880}$=9.685, p<0.000; $F_{3.448}$= 13.871, p<0.000) when the 3 least functional alternatives were eliminated (all the post hoc comparisons with this condition turned out significant with p<0.000). The correlations between the *p*'s of the 3 conditions with the original condition were 0.999, 0.998 and 0.989 (0.999, 0.998 and 0.980 for the fitted items). This would mean that even if there were variations in the difficulty, the order of the items as regards *p* is maintained.

Guessing Parameter

As was expected, the guessing parameter increased significantly when the number of alternatives were reduced (see Table 1), from 0.22 to 0.37 ($F_{3.440}$=71.916; p<0.000; all the post hoc comparisons, p < 0.01). In the original condition and the condition where the least functional alternative is eliminated, the mean of *c* is slightly higher than the expected probability for very low levels of ability (0.2 and 0.25, respectively); in the conditions where 2 or 3 alternatives are eliminated the mean of *c* is lower than expected (0.33 and 0.5). Figure 1a shows the item frequency distribution in the *c* parameter. The distribution shifts to the right as the alternatives are eliminated. On the other hand, when the 3 least functional alternatives are eliminated, the distribution becomes flatter (with a higher range for *c*), which indicates that the increase in size of *c* is not homogeneous for all the items. In fact, the correlation between the *c*'s of the original condition and those of the condition where the 3 least alternatives functional are eliminated is 0.630; this correlation is considerably inferior to those obtained with the condition where only 1 or 2 alternatives are eliminated (0.974 and 0.900, respectively).

In table 1, we can see that the increase in c observed when the 3 least functional alternatives are removed is smallest for items that are difficult in the original condition (Interaction effect, $F_{6.440}$=8.021; p<0.000). That is to say, when the item has its location point essentially in the highest abilities, the elimination of alternatives has a lesser effect on *c*.

The Difficulty Parameter

As shown in table 1, the reduction of alternatives produces a significant decrease in *b* ($F_{3.440}$= 50.496, p<0.000) with the elimination of the 3 least functional alternatives (all the comparisons are post hoc, p<0.000), but not when 1 or 2 alternatives are eliminated (p>0.05). While the difficulty in the first conditions oscillates between 0.34 and 0.40, the mean difficulty is below 0 (-0.16) in the conditions where 3 alternatives are eliminated. This tendency can also be observed in fig.1b, where the frequency distribution of parameter *b* for each condition is shown. While the distribution hardly changes after eliminating 1 or 2 alternatives, the elimination of 3 alternatives produces an increase in the number of items with *b* between –1.75 and –2.25, and a decrease of items with *b* higher than 0.50. When the 3 less functional alternatives are eliminated, the decreases of most practical importance are again obtained in the most difficult original items (Interaction effect, $F_{6.440}$= 6.189; p<0.000). Despite all these differences, the order of the difficulties seems to remain stable: the correlations between the parameter *b* of each newly reduced condition are 0.998, 0.989 and 0.915.

The Discrimination Parameter

The reduction of alternatives also produces a significant decrease in *a* ($F_{3.440}$=8.807, p<0.000; see table 1) in the elimination of the 3 less functional alternatives (all the comparisons are post hoc, p<0.000) but not when 1 or 2 alternatives are eliminated (p>0.05). The elimination of 3 alternatives produces a decrease in the mean of *a* from a value slightly above 1, to a considerably lower value (0.88). The frequency distributions of *a* can be see in fig. 1c. When the 2 least functional alternatives are eliminated, although there are no differences in the mean, the distribution of *a* is more peaked. However, the clearest change is the shift of the distribution towards lower values when 3 alternatives are eliminated. Once again, this shift is not homogeneous because the discrimination numbers between 0.2 and 0.4 do not increase, but decrease. In fact, the correlations between the *a* of each new condition with the original condition are 0.972, 0.872 and 0.448, showing that the least stable order with the elimination of alternatives is given in parameter *a*, especially with the elimination of 3 alternatives.

In Table 1 we can see that the decrease in *a*, when the 3 less functional alternatives are eliminated, is bigger for the easiest items (Interaction effect, $F_{6.440}$=3.658; p<0.000).

The Information Level and Differences in the estimation of θ

Figure 2A shows the test information in each of the 4 conditions. The elimination of an alternative shows that the test preci-

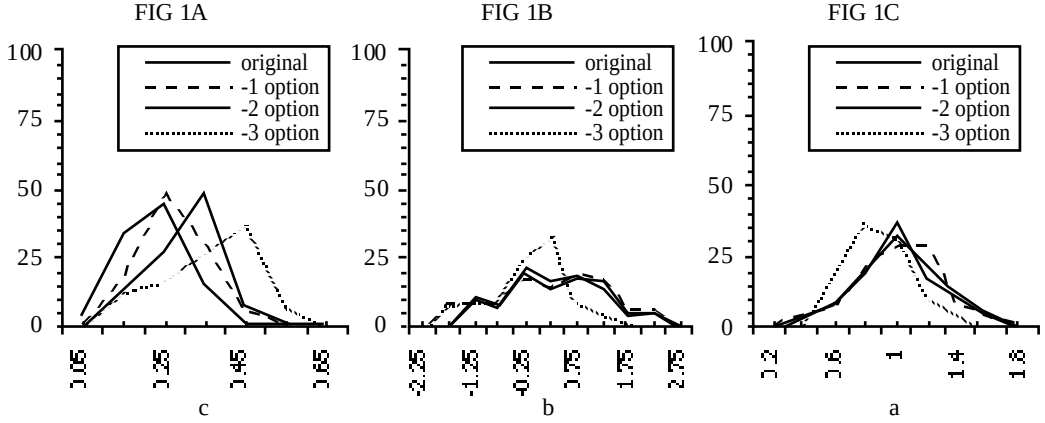

Figure 1. *Frequency distribution for the IRT parameters for the 113 adapted items*

sion is maintained for the high-ability subjects, while the information function for the low-ability subjects is slightly lesser. When the 2 least functional alternatives are eliminated, the information reduces considerably for all the levels of ability, while the elimination of 3 alternatives seems to accentuate that decrease in the higher-ability level subjects. These results indicate that when the changes in parameter *c* are small, (reduction of an alternative) they mainly affect the low-ability subjects, while when they are bigger, they seem to affect all the ranges of ability. Finally, the biggest decrease in the information function (especially for the high-ability subjects) when 3 alternatives are removed is probably not only related to the increase in *c* but also to the decrease in *a* and in *b*.

Figure 2B shows the average differences in the estimation of θ between the original and the 3 new conditions (considering only the 113 fitted items). For the high-ability subjects and compared to their original estimates, it is shown that the reduction of alternatives produces a progressively lesser estimation. For the low-ability subjects, the effect is the reverse. These effects are manifested especially in the condition where 3 alternatives are eliminated.

Discussion

It must be pointed out in the first place, that our results on the CTT parameters are coherent with those of other studies. For the difficulty level *p*, the results are consistent; the elimination of alternatives increases the *p* of the items. While some authors (Crehan, Haladyna & Brewer, 1993) found that the reduction from 4 to 3 alternatives produces a significant increase in *p*, other authors did not consider this tendency important with the reduction from 5 to 4 alternatives (Cizek & O'Day, 1994), nor with the reduction from 5 to 3 alternatives (Owen & Froman, 1987). These results are not contradictory because they indicate that the elimination of alternatives has greater effects in accordance with its quality. In fact, Cizeck Robinson & O'Day (1998) found that some changes do exist for some items. Nevertheless, these and other studies coincide that the difference of difficulty is of little practical importance.

For the lower discrimination level, no significant difference was found either, for the reduction from 5 to 4 alternatives, like in other studies (Cizek & O'Day, 1994; Crehan, Haladyna & Brewer 1993; Owen & Froman, 1987). Cizek, Robinson and O'Day (1998) found that the reduction from 5 to 4 alternatives produced significant decreases in the discrimination in some items, while they produced significant increases in others. Trevisan, Sax and Michael (1991) did not find any significant differences in the reliability with the reduction to 4 or 3 alternatives, either. The fact that the results are consistent with those of other previous studies constitutes a certain validation of the random re-allocation method proposed in this study.

As regards the reduction to 2 alternatives, the results from the classical theory were less consistent. Though the discrimination and reliability levels are maintained at their initial values, the elimination of the 3 least functional alternatives gives place to items with a significantly higher *p* index and a somewhat different discrimination order. So, some of the psychometric properties at the item level may have been affected (eg: the discrimination order).

The conclusions are similar from the IRT, with the results being more illustrative than what occurs at items level. The *c* parameter varies systematically in the expected direction with the reduction of each least functional alternative, and it very well seems to reflect the quality of the incorrect alternatives. For example, when the 3 least functional alternatives are eliminated, the mean value of *c* obtains a value (.37) lower than expected for a 2-options item (.50) while removing one alternative we obtain a *c* mean value of .27 larger than expected for a 4-options item (.25). Finally, the increases in parameter *c* are not homogeneous, being lesser for the difficult items in the original bank. It is also in these items where the greatest difficulty decrease is found. The upper-middle ability subjects chose the third least functional alternative in a higher degree than the low-ability subjects; hence it could happen that while *c* has little changes (the correct response probability of the low-ability subjects is unchanged), *b* changes quicker (because there are more upper-middle ability subjects that are giving correct answers to the item).

As regards parameter *a*, it stays relatively stable when 1 or 2 alternatives are eliminated. However, the elimination of the 3- least functional alternatives produces some changes. These changes can be considered more important than those which occurred for *b* and *c*, as what the item is measuring may be changing (differences in the orders of *a*) and it would be more difficult to identify (lesser mean of *a*). On the other hand, the changes in the order of *a* are

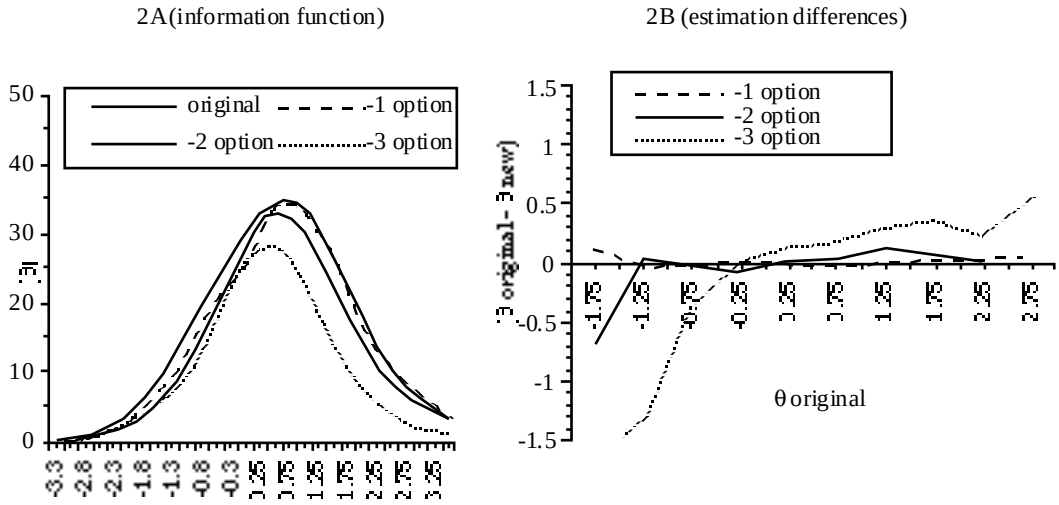

Figure 2. *Effects on ability (accuracy and estimation differences) for each condition*

consistent with those that occurred from the CTT for the biserial correlations.

Finally, the effects in the precision for the different levels of ability indicate, contrary to the results of Lord (1980), that the elimination of options does not affect only the low-ability subjects. While the elimination of an alternative fits the results of Lord, the elimination of 2 and especially 3 alternatives affects the subjects of all the ranges of ability.

Conclusions

So, as regards the 5-alternative items: a) The elimination of 1 alternative does not seem to have any practical importance on the difficulty or discrimination of the items, nor on the information test function or ability estimation; b)The elimination of 2 altematives seems to produce important differences in comparison to the original estimation, especially for the high-ability subjects. This effect however, is probably due to the elimination of the partial knowledge advantage, and it can be considered a point in favour of the use of 3 options, because a purer measurement of ability would be achieved; considering that the items bank psychometric properties are not especially affected and that the advantages mentioned in the introduction is verified to be the best option; c)The elimination of 3 alternatives produced the worst results. The information properties of this bank are quite inadvisable from the IRT perspective, with an important decrease of the difficulty and discrimination of the items.

Among the principal limitations we observed in our research, we have: a) the conclusions are based only on the items in which a precision level and a fit to the most adequate IRT has been achieved. That is, we did not analyse the properties of each condition, but the type of changes given at items level; eg., many of the items not considered in the 5-alternatives set had good fit and were not considered, nevertheless. This items selection was not made at random as shown by the fact that the mean difficulty of the 113 items bank increased in contrast to the mean difficulty of the original test of the 221 items. This problem could be avoided with bigger and more heterogeneous subject samples; b) As pointed out by one reviewer, the size of the sample is at the limit of what is acceptable, and this makes recommendable to carry out a replica of the study with bigger samples; however, although small samples provide more unstable parameters, this problem was minimized using the parameter standard error as one fit criteria; c) The random re-allocation method itself constitutes a new procedure, and it needs additional evidence to verify its effectiveness, starting from empirical studies. It could happen that the elimination of an alternative would provoke different proportions of altemative choices different from those obtained from the supposition of our procedure. However, the fact that the results obtained are similar to those of other previous studies, constitutes a point in our favour; d) this study has been carried out with an English Test Vocabulary; it would be interesting to do researches with another type of tests (e.g., spatial rotation) in which the alternatives are probably much easier to construct, and are more discriminative; e) Finally, only the influence of one item factor (difficulty) on parameter change was studied. It might also be interesting to explore the influence of other item characteristics (e.g., option elimination criteria, item discrimination,...) on parameter change.

Finally, this study shows that with different number of options, we could obtain different estimations of ability for the subjects. This points out an important advantage of the polytomous models which consider the information given by the wrong altematives. Research on a similar topic, the optimal number of categories on Likert scales, (e.g.: Hernández-Baeza, Muñiz y García-Cueto, 2000) shows the usefulness of applying polytomous models (i.e.: graded response model).

Referencias

Assesment Systems Corporation (1988). *User's manual for the microCAT testing system version 3.* St.Paul, MN: Author.

Ben-Simon, A., Budescu, D.V. & Nevo, B. (1997). A comparative study of partial knowledge in multiple-choice tests. *Applied Psychological Measurement* , 21,65-88.

Budescu, D.V. & Nevo, B. (1985). Optimal number of options: An investigation of the assumption of proportionality. *Journal of Educational Measurement,* 22, 183-196.

Cizek, G.J. & O'Day, D. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement,* 54, 861-872.

Cizek, G.J., Robinson K.L. & O'Day, D. (1998). Nonfunctioning options: a closer look. *Educational and Psychological Measurement, 58, 605-611.*

Crehan, K. D., Haladyna, T, M. & Brewer B. W. (1993). Use of an inclusive option and the optimal number of options for the multiple-choice items. *Educational and Psychological Measurement.* 53, 241-247.

Delgado, A. & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment*, 14 (3), 197-201.

Downing, S. M. (1992). True-false and alternative-choice formats: a review of research. *Educational and Psychological Measurement.* 29, 565-570.

Green, K., Sax, G. & Michael, W.B. (1982). Validity and reliability of tests having different numbers of options for students of differing levels of ability. *Educational and Psychological Measurement, 42, 239-245.*

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement,* 12, 109-113.

Haladyna, T. M. & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education,* 2, 37-50.

Haladyna, T.M. & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement,* 53, 999-1010.

Hernández Baeza, A., Muñiz , J. y García Cueto, E. (2000). Comportamiento del modelo de respuesta graduada en función del número de categorías de la escala. *Psicothema,* 12, (2), 288-291.

Hutchinson, T. P. (1997). Mismatch models for formats that permit information to be shown. En Van del Linden, W. J. & Hambleton, R. K., *Handbook of Modern Item Response Theory.* New York: Springer.

Kolen, M. J. & Brennan, R. L. (1995). *Test Equating. Method and Practices.* New York: Springer.

Levine, M. V. & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement,* 43, 675-685.

Lord, R. M. (1977). Optimal number of choices per item – a conparison of four approaches. *Journal of Educational Measurement.*14, 33-38.

Lord, R. M. (1980). *Applications of Item Response Theory to practical test problems.* Hillsdale, NJ: Lawrence Erlbaum.

Mislevy, R. J. & Bock, R. D. (1989). *BILOG: Item Analysis and Test Scoring with Binary Logistic Models.* Chicago: Scientific Software.

Olea, J., Ponsoda, V., Revuelta, J. & Belchí, J. (1996). Propiedades psicométricas de un test adaptativo informatizado de vocabulario inglés. *Estudios de Psicología.* 55, 61-73.

Owen, W. V. & Froman, R. D. (1987). What's wrong with three option multiple choice items? *Educational and Psychological Measurement.* 47, 513-522.

Ponsoda, V., Wise, S.L., Olea, J. & Revuelta, J. (1997). An investigation of Self-Adapted Testing in a Spanish High School Population. *Educational and Psychological Measurement.* 57, 210-221.

Ponsoda, V., Olea, J., Rodriguez, M.S. & Revuelta, J. (1999). The effects of test difficulty manipulation in computerized adaptive testing and self-adapted testing. *Applied Measurement in Education*, 12 (2), 167-184.

Trevisan, M. S., Sax, G. & Michael W. B. (1991). The Effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement.* 51, 829-837.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology,* 1, 386-391.

Waller, M. I. (1989). Modelling guessing behaviour: a comparison of two IRT models. *Applied Psychological Measurement.* 13 (3) 233-243.

Weber, M. B. (1978). *The effect of choice format on internal consistency.* Paper presented at the National Council on Measurement in Education Annual Meeting, Toronto, Canada.

Aceptado el 11 de julio de 2000