Psicothema 2009. Vol. 21, nº 2, pp. 326-330
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Evaluation of five guidelines for option development in multiple-choice item-writing

Rafael J. Martínez, Rafael Moreno, Irene Martín and M. Eva Trigo
Universidad de Sevilla

This paper evaluates certain guidelines for writing multiple-choice test items. The analysis of the responses of 5013 subjects to 630 items from 21 university classroom achievement tests suggests that an option should not differ in terms of heterogeneous content because such error has a slight but harmful effect on item discrimination. This also occurs with the «None of the above» option when it is the correct one. In contrast, results do not show the supposedly negative effects of a different-length option, the use of specific determiners, or the use of the «All of the above» option, which not only decreases difficulty but also improves discrimination when it is the correct option.

Evaluación de cinco directrices para la construcción de opciones en ítems de elección múltiple. El artículo evalúa algunas directrices para la construcción de ítems de elección múltiple. El análisis de las respuestas de 5.013 sujetos a 630 ítems de 21 exámenes de una materia universitaria sugieren que ninguna de las opciones debería diferenciarse del resto en términos de contenido, porque tal error tiene un ligero pero perjudicial efecto sobre la discriminación del ítem, algo que también sucede con la opción «Ninguna de las anteriores» cuando es la correcta. Por el contrario, los resultados no muestran los supuestos efectos negativos de la longitud diferencial de una opción, del uso de determinantes, o de la opción «Todas las anteriores», la cual no disminuye la dificultad, sino que mejora la discriminación cuando es la opción correcta.

Multiple-choice items are used extensively in classroom achievement tests. This explains the importance attributed to understanding how to write these items correctly. Multiple-choice test items may be presented in a question or completion format, with negative, positive, simple or complex sentences, with language that is more or less suited to the subjects being tested, with a set number of options, presented vertically or horizontally, and with different lengths or grammatical characteristics. The adequacy of these aspects was considered by different guidelines (Haladyna, 2004; Haladyna & Downing, 1989a; Haladyna, Downing, & Rodríguez, 2002; Marrelli, 1995; Martínez, Moreno, & Muñiz, 2005; Moreno, Martínez, & Muñiz, 2004, 2006; Osterlind, 1998; Roid & Haladyna, 1982). However, the possible importance of these aspects for the quality of test items is an assumption based more on common sense than on the findings of well-founded research. Apart from studies on the number of options, the amount of empirical studies concerning item-writing guidelines is clearly insufficient as has been stated by Haladyna and Downing (1989b), Haladyna et al. (2002).

The present study deals with five of the guidelines on options development included in the above-mentioned studies. These guidelines are: i) *Keep all options homogeneous in content*; ii) *Avoid specific determiners*; iii) *Keep the length of options about equal*; iv) *Avoid «All of the above» option*; and v) *Avoid «None of the above» option*. This study examines a) test-maker non-compliance with these guidelines and b) the influence this may have on item quality, using a broad set of items taken from university classroom achievement tests that were written without these guidelines being considered.

The study's first objective was to evaluate the frequency of errors made by test-makers who did not consider these guidelines, and their distribution between correct and incorrect options. Secondly, as in other similar research (e.g., Sireci, Wiley, & Keller, 1998), where the sample of items allows us to do so, this study evaluates whether the statistical indexes of difficulty and discrimination vary depending on compliance or non-compliance with the stated guidelines. In addition, it adds two other aspects. On the one hand, it studies whether these item indexes vary depending on whether non-compliance occurs in a correct or incorrect option. On the other, it analyzes whether non-compliance with one of these guidelines in the correct option affects equally the proportion of subjects that choose it in the high and low competence groups.

Method

Participants

The sample of participants in this study came from a population of first-year undergraduate students in Psychology at the

Fecha recepción: 16-6-08 • Fecha aceptación: 6-11-08
Correspondencia: Rafael Moreno
Facultad de Psicología
Universidad de Sevilla
41018 Sevilla (Spain)
E-mail: rmoreno@us.es

University of Seville during several academic years. A study into the profile of these students conducted in the 2007-08 academic year (by the Coordinating Group of the Tutorial Action Plan, University of Seville, 2008) showed that 85.2% of this population were women, with the majority having completed a baccalaureate in Humanities and Social Sciences (70.8%) performed in the state education system (72.2%), with mean entrance marks of 7 out of 10 (M= 7.00; *SD*= 0.69). All chose Psychology as their first option, and during the first term 80.9% were full-time students, while the remainder combined studies with a job. 52.8% came from the province of Seville, 32.4% from other provinces in Andalusia, and 14.8% from the rest of Spain, with a very small number of overseas students (<0.1%).

The sample consisted of the 5013 students who opted voluntarily to take the exam in the subject entitled *Methodological Foundations in Psychology;* this number represented 50.1% of all first-year Psychology students. The sample consisted of 3998 women and 1015 men with a mean age of 20.7 years (*SD*= 4.54), whose age ranged from 18 to 53 years.

These students took classroom achievement tests between February 1995 and February 2002, and the average number of students per test was 238.71 (SD= 146.37). The course syllabus and teachers remained unchanged during this period, which provided stable conditions for this study.

Instruments

A database was built with all 630 items included in 21 classroom achievement tests. Each test has 30 question items about two texts summarizing research or psychological interventions. The average internal consistency of the test scores was .60 (95% confidence interval, CI_{lower}= .57; CI_{upper} = .67) that must be considered in the context of tests composed of item-bundles. Each item consists of a sentence that is completed by four explicit options, which are arranged vertically. Every item has six response options, of which only one is correct. Of the six options, four appear explicitly in each item, while the other two —«None of the above» (NOTA) and «All of the above» (AOTA) referring to the explicit options— appear at the top of each page of the test and do not appear in each item. All of the items therefore were in non-compliance with the guidelines on both options.

The guidelines on the «None of the above» and «All of the above» options were clear enough in the reference taxonomies. However, the content of the above mentioned guidelines i, ii, iii was not clear enough to categorize database items. As a result, a system of categories was obtained of errors or ways in which non-compliance with these guidelines occurs.

These categories of errors allow us to determine whether, in each item analyzed, there is an explicit option that is the only one to have, or the only one to lack, one or more of the following characteristics:

- Differential *Content* of an option compared to the rest because a different terminology is used that changes appearance. This difference usually takes the form of nouns, adjectives and non-copulating verbs. This category specifies the guideline on options of homogeneous content.
- Differential *Determiner* of an option compared to the rest in terms of adverbs that modify, limit or condition the content of an option. This involves a specification of the guideline referring to terms such as *Always*, *Never*, *Completely*, *Absolutely*, referred to as specific determiners by Haladyna et al. (2002), enlarged here to include others such as *Exclusively*, *Only*, *Solely* and *Necessarily*.
- Differential *Length* of one option compared to the rest, consisting of a surplus or deficit of four or more words that is visually noticeable because it extends beyond the rest of the options. This is a specification detailing the generic guideline on length.

Procedure

The 630 items from the database were categorized with the aforementioned categories of non-compliance with guidelines for option development. In addition, it was stated whether the correct option for each item was «None of the above», «All of the above» or one of the explicit options.

Statistical indexes of difficulty and discrimination were estimated for each item in the database, giving the *p*-value and point-biserial correlation respectively. To empirically evaluate the impact of non-compliance with item-writing guidelines, all the applied items were used, regardless of whether or not their statistical indexes were suitable. An estimation was also made of the proportions of correct alternatives chosen in the extreme groups of the least and the most competent subjects, those below 27% and those above 73% of the overall total of correct choices, respectively.

Results

In the database as a whole the difficulty index presents a mean of .51 (*SD*= .22), and the mean of the discrimination index was .18 (*SD*= .11). According to Kolmogorov-Smirnov's test, the distributions of these two indexes are normal (*Z*= 1.3 y *Z*= 0.81, respectively for *p*-value y r_{pb}). The proportion of correct responses in the most competent group of subjects presents a mean of .66 (*SD*= .23), while the mean in the least competent group is .36 (*SD*= .22). In neither case is the distribution of these two variables normal, according to the Kolmogorov-Smirnov's test (*Z*= 2.07 y *Z*= 2.27, respectively), because, as was to be expected, they are biased to the right and the left respectively.

All 630 items fail to comply with the recommendations about avoiding the use of the NOTA and AOTA options, and 121 contain one or more explicit options with differential errors in one or more of categories i, ii or iii. In the majority of these items (*n*= 112) only one error is committed, with two occurring in the remaining items (*n*= 9).

Regardless of whether they appear in the same item or not, the total frequency of differential errors is 130. These errors are not distributed equally among the different categories, with non-homogeneous Content being the most frequent (65.4%), while the presences of a Determiner (16.9%) and different Length (17.7%) have a similar, lower frequency. The differential errors analyzed occur more frequently in the fourth option, which is the last of the explicit options (56.2%). In addition, in absolute terms these errors of Content, Determiners or Length, occur mainly in incorrect options rather than in correct ones. However, bearing in mind the proportion of three explicit distracting alternatives to one correct one, the occurrence of errors is fairly similar in both cases (90/3 vs. 40/1, respectively).

Table 1
Comparisons of item difficulty and discrimination in terms of compliance with guidelines

	Compliance			Non-compliance				
Guideline	***n***	***M***	**(*SD*)**	***n***	***M***	**(*SD*)**	*z*	η^2
				Difficulty				
Content	545	.51	(.22)	85	.54	(.25)	-1.18	.002
Determiners	608	.51	(.22)	22	.52	(.25)	-0.05	.000
Lengths	607	.51	(.22)	23	.51	(.22)	-0.02	.000
				Discrimination				
Content	545	.18	(.11)	85	.14	(.11)	-3.53*	.019
Determiners	608	.18	(.11)	22	.19	(.13)	-0.84	.001
Lengths	607	.18	(.11)	23	.15	(.15)	-0.73	.002

* $p<.05$ asymptotic two-tailed probability Mann-Whitney test

Table 2
Item difficulty and discrimination by guideline non-compliance in correct vs. incorrect options

	Incorrect option			Correct option				
Non-compliance	***n***	***M***	**(*SD*)**	***n***	***M***	**(*SD*)**	*z*	η^2
				Difficulty				
Content	56	.51	(.26)	29	.59	(.22)	-1.37	.023
Determiners	18	.54	(.25)	4	.42	(.26)	-0.60	.033
Lengths	16	.52	(.23)	7	.46	(.20)	-0.60	.017
«All of the above»	591	.51	(.22)	39	.48	(.22)	-0.92	.001
«None of the above»	544	.53	(.21)	86	.37	(.21)	-6.13*	.060
				Discrimination				
Content	56	.14	(.12)	29	.13	(.09)	-0.08	.001
Determiners	18	.20	(.13)	4	.12	(.16)	-1.02	.062
Lengths	16	.14	(.17)	7	.17	(.08)	-0.33	.011
«All of the above»	591	.17	(.11)	39	.23	(.09)	-3.43*	.017
«None of the above»	544	.18	(.11)	86	.15	(.13)	-1.95	.008

* $p<.05$ asymptotic two-tailed probability Mann-Whitney test

When comparing the statistical indexes of items that present the different types of errors, categorized as non-compliance with guidelines, and those items that comply with those guidelines (Table 1) there is no statistically significant difference in item difficulty. However errors resulting from differential Content produce a drop in the point-biserial correlation, which is statistically significant. In any event, bearing in mind the effect size shown by the η^2 index, these contrasts would only explain approximately 2% of the differences found in the items' statistical indexes.

As for the comparison between items that present non-compliance with guidelines in the correct option and those that present it in an incorrect one, the only statistically significant differences appear over the presence of the «All of the above» or «None of the above» alternatives as the correct option (Table 2). When the «None of the above» alternative is the correct one, lower p-values and r_{pb} are observed. These differences are only statistically significant for *p*-value, with an effect size of 6% over item difficulty. However, when the «All of the above» alternative is the correct one, the point-biserial correlation increases statistically significant, with a small effect size of almost 2% over item discrimination.

Table 3
Mean proportions of correct option choice in terms of compliance of guidelines and subjects competence

	Subjects competence			
	High		Low	
Correct option	***M***	**(*SD*)**	***M***	**(*SD*)**
Compliance of guidelines	.68	(.21)	.38	(.21)
Non-compliance of content	.71	(.19)	.46	(.24)
Non-compliance of determiners	.54	(.30)	.27	(.15)
Non-compliance of length	.63	(.21)	.32	(.18)
«All of the above»	.64	(.23)	.30	(.21)
«None of the above»	.51	(.26)	.24	(.18)

Table 3 presents the mean proportions of choice of the correct option in the extreme groups of most and least competent subjects, when the option in question contains at least one of the differential errors considered. When «None of the above» is the correct option there is a drop in the choice of the correct alternative. This is

statistically significant according to the Mann-Whitney test, in both the most competent (z= -5.78; p<.001; η^2= .083) and in the least competent group (z= -5.80; p<.001; η^2= .059). When «All of the above» is the correct option, the proportion choosing the correct response drops in comparison with items that do not present any error, but this is only statistically significant in relation to the proportion of less competent subjects that choose it (z= -2.51; p= .01; η^2= .012).

Discussion

The descriptive analysis highlights the relatively important frequency of errors of differential Content, something that merits special precaution when writing test items. It should also be noted that errors seem to be committed as much in correct as incorrect options, for which reason care must be taken in the elaboration of all options. This analysis also provides an answer to a question not proposed initially on the number of options. It seems advisable to avoid a fourth option because this is the one in which most errors tend to be committed. It is possible that this affects subject matter where it is hard to think of a fourth option and this may lead to format errors. Whatever the reason, our preference for three options ties in with the general opinion in existing studies (e.g., Abad, Olea, & Ponsoda, 2001; Bruno & Dirkzwager, 1995; Delgado & Prieto, 1998; Haladyna et al., 2002; Rogers & Harley, 1999).

As far as the relations studied are concerned, there are data that support some of the proposed guidelines. It seems that options should be built with homogeneous Content, as the presence of this kind of error leads to a drop in discrimination. However, this recommendation is supported by very small effect sizes, something already mentioned in the literature for this and other item errors (for example, Freedle & Kostin, 1999; Knowles & Welch, 1992) It is possible that repeated small effects in single items may build up to affect the overall properties of the test as a whole, something that could focus interest in specific future research.

Regarding the «None of the above» option, our results show that it produces lower p_values, in other words greater difficulty, when it is the correct option. Therefore, this option would unnecessarily increase item difficulty (e.g., Crehan & Haladyna, 1991; Crehan, Haladyna, & Brewer, 1993). This would allow us to support the assumption of the supposedly negative effects of that option, but not necessarily in the sense of rewarding examinees with serious knowledge deficiencies (Dochy, Moerkerke, De Corte, & Segers, 2001; Gross, 1994), as this was subjected to revision by Knowles and Welch (1992) on the basis of a meta-analytic review of previous research.

Together with this supportive finding, results were obtained that seem to temper other guidelines or even contradict them. As for the «All of the above» option, if we look at the p-value, our results fail to demonstrate a general increase in difficulty that is supposed to accompany this option, because it only presents higher difficulty for the least competent subjects. Furthermore, the items that have it as the correct option present better r_{pb}, which contradicts bibliographical data (Dudycha & Carpenter, 1973; Mueller, 1975). If on the basis of these results one opts for the use of this option, one would in any event need to bear in mind the bibliography's warning that this option may only mean that more than one option is correct, not that all the previous options are correct.

Finally, neither do the data found support the precaution of avoiding differential length or determiner of an option, at least with the specification given in this study. Our findings concerning option length do not follow the general tendency observed in the review by Haladyna and Downing (1989b), which stated that the longer an option the easier it becomes. Neither is the use of determiners associated to changes in difficulty or item discrimination, although our findings provide the only empirical information about determiners in multiple-choice items that has been presented to date.

On a methodological level, the results from our study illustrate that for the evaluation of multiple-choice item-writing guidelines it may be useful to consider sets of items that have not undergone previous analysis. Furthermore, they show that it is insufficient to consider the non-compliance of guidelines in overall terms and their effects on difficulty and discrimination indexes. Empirical evaluations into the validity of some of these guidelines seem to be conditioned by the correct or incorrect option in which a differential error is introduced, and also whether these errors affect subjects' responses depending on their demonstrated level of competence.

These results are limited to university-level students subjects, who are able to focus on item content and ignore any problems of wording. In any event, the ex post facto nature of the research performed —a methodology present in the bibliography (e.g., Freedle & Kostin, 1999; Hansen & Dexter, 1997; Jozefowicz et al., 2002; Sireci et al., 1998)— may be producing problems of control of variables that were not taken into account. It might therefore be advisable to conduct experimental replicas with manipulation of errors in the items being studied. One line of future research ought to establish whether the findings are upheld depending on whether the option is differential because it is the only one that does, or the only one that does not, present an error of non-compliance with guidelines. Considering the guidelines existing in the bibliography on test item construction, the results and conclusions from the evaluation performed here demonstrate the need for further evaluation to establish the effectiveness of those guidelines that lack sufficient empirical basis.

In any event, the present study —in conjunction with findings in related literature— has allowed us to make certain practical observations that could be used by professionals when constructing items and tests.

Great care must be taken when writing each item, given the high rate at which the recommendations in the literature are ignored. Effort should be made to construct options that are homogeneous in content and appearance, ensuring that none stand out over others. Although such a factor may have little effect on one specific item, the effect may build-up over the whole test if it is repeated in all items.

As for the debate surrounding the optimum number of options, it must be emphasized that three appears to be an appropriate number. A higher number increases the risk of constructing options that are differentiated from the rest, because not all subjects contents allow several plausible options.

In practice, when faced with a lack of relevant contents, the options «None of the above» and «All of the above» are often used. On this point it is worth remembering the following: they can both unnecessarily increase the difficulty of an item, particularly when it is the correct option. Our results show that this would occur in the «All of the above» option, especially in less competent subjects.

If this option is used in items with more than three options it should be remembered that the subject translates it as «More than one option is correct», instead of all of them being correct, for which reason s/he does not have to check all the options.

In general, it seems reasonable to keep to the guidelines about item construction, including those for which the empirical data are inconclusive. Even though they may be contradictory or insufficiently clear, they should be taken as warnings of the problems that may arise through non-compliance with specific guidelines. These effects will probably vary depending on specific situations that are beginning to become clearer. For example, this study found that certain results vary depending on whether the subject is more or less competent, or the options with errors are correct or incorrect.

References

Abad, F.J., Olea, J., & Ponsoda, V. (2001). Analysis of the optimum number alternatives from the item response Theory. *Psicothema, 13*(1), 152-158.

Bruno, J.E., & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55*(6), 959-966.

Crehan, K.D., & Haladyna, T.M. (1991). The validity of two item-writing rules. *Journal of Experimental Education, 59*(2), 183-192.

Crehan, K.D., Haladyna, T.M., & Brewer, B.W. (1993). Use of an inclusive option and the optimal number of options for 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.

Dochy, F., Moerkerke, G., De Corte, E., & Segers, M. (2001). The assessment of quantitative problem-solving skills with «none of the above» items. *European Journal of Psychology of Education, 16*(2), 163-177.

Dudycha, A.L., & Carpenter, J.B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology, 58*, 116-121.

Freedle, R., & Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? *Language Testing, 16*(1), 2-32.

Gross, L.J. (1994). Logical versus empirical guidelines for writing test items: The case of «none of the above». *Evaluation and the Health Professions, 17*(1), 123-126.

Haladyna, T.M. (2004). *Developing and validating multiple-choice test items* (3rd). Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T.M., & Downing, S.M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*(1), 37-50.

Haladyna, T.M., & Downing, S.M. (1989b). The validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 1*(1), 51-78.

Haladyna, T.M., Downing, S.M., & Rodríguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Hansen, J.D., & Dexter, L. (1997). Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. *Journal of Education for Business, 73*(2), 94-97.

Jozefowicz, R.F., Koeppen, B.M., Case, S., Galbraith, R., Swanson, D., & Glew, R.H. (2002). The quality of in-house medical school examinations. *Academic Medicine, 77*(2), 156-161.

Knowles, S.L., & Welch, C.A. (1992). A meta-analytic review of item discrimination and difficulty in multiple-choice items using «none-of-the-above». *Educational and Psychological Measurement, 52*(3), 571-577.

Marrelli, A.F. (1995). Writing multiple-choice test items. *Performance and Instruction, 34*, 24-29.

Martínez, R., Moreno, R., & Muñiz, J. (2005). Construcción de ítems. En J. Muñiz, A.M. Hidalgo, E. García-Cueto, R. Martínez y R. Moreno: *Análisis de los ítems* (pp. 9-52). Madrid: La Muralla.

Moreno, R., Martínez, R.J., & Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema*, *16*, 490-497.

Moreno, R., Martínez, R., & Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology*, *2,* 65-72

Mueller, D.J. (1975). An assessment of the effectiveness of complex alternatives in multiple choice achievement test items. *Educational and Psychological Measurement, 35*, 135-141.

Osterlind, S.J. (1998). *Constructing test items: multiple choice, constructed-response, performance and other formats* (2nd). Boston: Kluwer.

Rogers, W.T., & Harley, D. (1999). An empirical comparison of three- and four-choice items and tests: Susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement, 59*(2), 234-247.

Roid, G.H., & Haladyna, T.M. (1982). *Toward a technology for test-item writing.* New York: Academic Press.

Sireci, S.G., Wiley, A., & Keller, L.A. (1998). *An empirical evaluation of selected multiple-choice item writing guidelines.* Paper presented at the annual meeting of the Northeastern Educational Research Association.

University of Seville, Faculty of Psychology, Coordinating Group of the Tutorial Action Plan (2008). *Perfil de ingreso de los estudiantes de Psicología de la Universidad de Sevilla* [Profile of first-year undergraduate students in Psychology at the University of Seville]. Unpublished manuscript. University of Seville.