New guidelines for developing multiple-choice items

Rafael Moreno, Rafael J. Martínez, and José Muñiz[2]

University of Seville, University of Oviedo[2]

Spain

Correspondence address:
José Muñiz
Faculty of Psychology
University of Oviedo
Plaza Feijoo, s/n
33003 Oviedo (Spain)

E-mail: jmuniz@uniovi.es

Abstract

The rigorous construction of items constitutes a field of great current interest for psychometric researchers and practitioners. In previous studies we had reviewed and analyzed the existing guidelines for the construction of multiple-choice items. From this review emerged a new proposal for guidelines that is now, in the present work, subjected to empirical assessment. This assessment was carried out by users of the guidelines and by experts in item construction. The results endorse the proposal for the new guidelines presented, confirming the advantages in relation to their simplicity and efficiency, as well as permitting identification of the difficulties involved in drawing up and organizing some of the guidelines. Taking into account these results, we propose a new, refined set of guidelines that constitutes a useful, simple and structured instrument for the construction of multiple-choice items.

Key words: Guidelines, Multiple-choice items, Psychometric tests, Item construction, Performance

New guidelines for the construction of multiple-choice items

The multiple-choice format is one of those most commonly used today in psychological and educational tests. Constructing items of this type is more demanding than for other formats, since it requires, in addition to writing the stem, working out the different response options. The principal reference for this task, especially for tests measuring aptitudes or performance, is constituted by the sets of guidelines aimed at promoting systematic construction. Proposals such as those of Haladyna (2004), Hoepfl (1994), Marrelli (1995), Martínez, Moreno and Muñiz (2005) and Osterlind (1998) are good examples of such sets of guidelines. Two others are also especially noteworthy, namely that of Haladyna and Downing (1989a), which synthesizes over 40 taxonomies from the previous 54 years, and the subsequent update by Haladyna, Downing and Rodríguez (2002), with 36 guidelines.

However, these proposals present certain difficulties, such as overlapping and duplication in their content, imprecise language or an excessive number of guidelines. After identifying such problems, Moreno, Martínez and Muñiz (2004) drew up a new and more efficient set with just 12 guidelines. Explicit in this set is the basic principle that guidelines should help to increase the validity of the instrument in construction, this being understood in terms of congruence with the purpose of the assessment. Furthermore, this new set excluded guidelines referring to non-central aspects, and reorganized and reduced in number those with relevant content, incorporating some as particular cases of other, and eliminating redundancies. Two of the guidelines relate to the *content* to be assessed: 1) Should be a sample representing the content featured in a specification table, avoiding trivial items, 2) The representativeness should guide how simple or complex, specific or abstract, memoristic or reasoning that the item should be, as well as the best way of expressing it. Another three guidelines refer to the *expression of content* in the item: 3) The main point should be expressed in the statement. Each option should agree grammatically with the statement, 4) The syntax or grammatical structure must be correct. Avoid items that are excessively short or long, ambiguous or confusing, and use negative expressions with care, 5) The semantics should match the content and the subjects being assessed. The rest of the guidelines refer to the *construction of the options*: 6) There should only be one correct option, accompanied by plausible distractors. 7) The correct option should be spread around in different places, 8) Three is the preferred number of options, 9) Options should be presented vertically, 10) The set of options for each item should appear to be structured, 11) Options should be autonomous, without overlapping or referring to others. For that reason, the options “All of the above” or “None of the above” should be avoided, 12) No option should stand out from the rest in either content or appearance.

Having constructed the set of guidelines, it seemed appropriate to submit it to empirical assessment in order to identify and measure its potential and its limitations. This would permit the introduction of improvements to the set of guidelines, which would in turn make it easier to construct multiple-choice items, as well as assisting research – as yet insufficient – on the empirical foundation of guidelines (Haladyna & Downing, 1989b; Haladyna, Downing & Rodríguez, 2002; Millman & Greene, 1989; Roid & Haladyna, 1982). For a more beneficial evaluation, it was considered appropriate to take into account different perspectives, associated with people with different knowledge and interests in relation to the construction of multiple-choice items. Two important groups of assessors can be identified, that of professionals in psychological and educational measurement, *experts* in (or at least closely familiar with) item construction, and that of teachers of different subjects who, without being professionals in measurement, need to construct items, thus making them potential *users* of guidelines.

The main objective of the present work is to draw up a new and improved version of the guidelines for developing multiple-choice items. In order to achieve this main goal, two other objectives were previously investigated: a) the opinions of two group of assessors on the clarity and utility and other aspects of the twelve guidelines submitted for evaluation, and b) the similarities and differences between the two groups' assessments.

Method

Assessors

The group of experts or professionals in measurement was made up from three public lists of e-mail addresses: participants in the AERA 2004 Meeting, members of the International Tests Commission (ITC), and members of the Spanish Association of Behaviour Sciences Methodology (AEMCCO). Of a total of 159 experts, 29 (18.24%) returned their assessment of the guidelines. As regards the group of potential guidelines users, this was made up of teachers in a range of subjects from the University of Seville; all of them were familiar with the guidelines submitted for assessment, as they had voluntarily attended the course on *Construction and analysis of multiple-choice items* imparted by two of the authors of the present work and based in part on these guidelines. From a total of 98 teachers, 51 (52.04%) returned their assessment. Of these, 32.6% were from the area of Science and Technology, 14.3% from Health Sciences and 53.1% from Social and Human Sciences. Of the total eventual participants, 73.5% reported having experience in the use of multiple-choice items.

Instrument

We used a questionnaire (http://www.personal.us.es/rmoreno/cuesei.htm), with common questions for the two groups of assessors and some specific ones for the users' group. There were a total of 37 questions common to the two groups: 24 Likert-type questions with five assessment options, with one question on the clarity and another on the utility of each of the twelve guidelines; another 9 with the same format enquiring about the following aspects of the set of guidelines: Utility, Conceptual foundation, Exhaustiveness, Simplicity of phrasing, Efficiency or cost-benefit ratio, Overlap avoidance, Coherence, Respondent's preparedness to use them, and General assessment; and 4 open questions: one offering the possibility of explaining each closed response, two requesting respondents to indicate positive and negative aspects not covered by previous questions, and a fourth for any further comments the respondent wished to make. In order to aid sample description, the user group was also asked about their area of knowledge and their experience in the construction of multiple-choice items for exams in the subjects they taught.

In order to help participants make informed responses, they were referred to a document (see in http://www.personal.us.es/rmoreno/resdiri.htm) containing a summary that explained the basis of the guidelines proposed and the reasons why they were considered an improvement on their predecessors. This document and the questionnaire were written in Spanish and English so as to make them accessible to all the assessors.

Procedure

The request for participation was made by means of an individual e-mail presenting the objectives of the assessment and indicating the address of the questionnaire which could be filled out anonymously and sent via Internet. The e-mail also included a link to the document explaining the basis of the proposed guidelines, but only for the experts group, as the users had become familiar with the set of guidelines through the course they attended, and because of which they were selected as assessors. Approximately 30 days after the initial request, a second request was sent to those who had not yet replied, with a note of thanks to those who had already done so.

Results

Reliability of the ratings given in the questionnaire, measured by Cronbach's alpha, was .917 for the total of 33 closed questions, and also for the 9 referring to the set of guidelines; for the 24 questions on the utility and clarity of the different guidelines, reliability was .868. Assessments of the set of guidelines on the 1 to 5 scale varied in the experts group (see Table 1), from a minimum of 3.22 referring to Exhaustiveness, to a maximum of 4.35 for Avoidance of contradictions, with a mean of 3.87 (*SD* = 0.35). Rated with a score of less than 4 were the aspects of Foundation, Exhaustiveness, Clarity and simplicity, Efficiency, Respondent's preparedness to use them and Soundness of the set, and with a score of more than 4, Utility, Avoidance of overlap and Avoidance of contradictions. These ratings were completed with the comments obtained through the open questions referring to the set of guidelines. In these, the experts mention as positive aspects the parsimony and synthesis achieved with the guidelines, their consistency with published work on the topic, their utility and their contribution to improving the quality of the items to be constructed. As aspects to be modified they suggest: expressing more clearly and in more detail some of the guidelines, so as to avoid ambiguities and lack of clarity (in terms such as "specification table", "representativeness" or "structured set of options"); relativizing some of the guidelines that seemed too restrictive (6, 8 and 11); and revising the number and organization of the guidelines, dividing up the content of some (such as 2 and 4), grouping together that of others (9, 10 and 12, in that they all refer to formal aspects), and adding some that were lacking, in relation to aspects such as the need to revise the items written, the need to ensure impartiality in the response options, and the need to include the appropriate number of items in the test.

INSERT TABLE 1

As regards the user group they rate all the aspects higher than 4, even giving more than 4.5 to the aspects Utility, Avoidance of contradictions and Respondent's preparedness to use the guidelines, the mean rating being 4.35 (*SD* = 0.21). As positive aspects they mention the simplicity and brevity of the set of guidelines, as well as its utility for a regulated construction of multiple-choice items. As negative aspects they refer to the restrictions and demands involved in following the set of guidelines and the complex language employed in some of them. Furthermore, the users' rating is higher in all aspects than that of the experts, with statistically significant and medium-sized differences in the case of Foundation, Efficiency, Respondent's preparedness to use them and Soundness of the set, and large-sized differences in Exhaustiveness (see Table 1).

In the assessments of each of the 12 guidelines obtained from the first 24 questions, with regard to clarity (see Table 2), the experts' mean is 4.17 (*SD* = 0.58), those referring to guidelines 2, 5 and 10 being lower than 4 and those referring to guidelines 6, 9, 11 and 12 being higher than 4.5. For the users' group, the mean of the assessments is 4.38 (*SD* = 0.34), being lower than 4 for guidelines 2 and 10 and higher than 4.5 for numbers 3, 7, 8, 9, 11 and 12. The comments made, by both experts and users, refer to syntactic and semantic difficulties in 1, 2, 5 and 10. Furthermore, the users' ratings are higher than those of the experts, except in the cases of guidelines 6, 11 and 12, in which the opposite occurs. Differences between the two groups considered with R^2 are small, except in the assessments of guidelines 1, 2 and 5, where they are moderate. The differences in ratings of guidelines 1, 2 and 6 are statistically significant.

INSERT TABLE 2

As regards the utility of each guideline (see Table 2), in the experts' group the mean of the assessments is 4.00 (*SD* = 0.63), being below this value those referring to guidelines 2, 8, 9 and 10, and over 4.5 those for guidelines 4, 6 and 12. In the users' group, the mean of the assessments is 4.32 (*SD* = 0.25), with that for guideline 10 being below 4 and those for guidelines 1, 3 and 6 being higher than 4.5. The comments made highlight the following aspects, especially in the experts group: four or more options may also be appropriate, and not only three, as suggested in guideline 8; the option "None of the above" may be useful and clear; sometimes it is appropriate ask for the most correct option, and not just the only correct one, as suggested in guideline 6; finally, some users fail to see the sense of guideline 10. The experts' assessments are lower for all the guidelines, except numbers 4, 5 and 12. Size of the differences is small in all cases, except those of 8 and 9, where the sizes are large and moderate, respectively; these are also the only two cases where the differences are statistically significant.

Proposal for new guidelines

On the basis of adequate reliability of the ratings obtained with the questionnaire, there is an interesting body of data on the guidelines assessed. Both groups of assessors stress the utility of the set, which results from its parsimony and synthesis of other proposals, and rate as adequate the avoidance of overlap and contradictions between the guidelines. The users' group, moreover, indicate high preparedness to use the set of guidelines. The two groups agree on the need to rewrite some guidelines that are ambiguous and unclear, especially numbers 1, 2, 5 and 10.

With regard to the remaining aspects, the two groups' assessments differ. The less favourable rating by the experts' group in almost all cases is probably due to their greater knowledge of the topic, which permits them to appreciate more details than the users. These differences in assessment may also be due to a procedural factor, insofar as the two groups responded to the same questionnaire that provided a summarized version of the guidelines, but had different levels of information on them: the users could take into account the detailed information received in the course on each one of the guidelines, including the full text from Moreno et al. (2004); on the other hand, the experts did not have this information (unless they had read the text in question on their own initiative – highly unlikely in the case of the non-Spanish speakers). Nevertheless, it is perhaps because of this, which may be seen as a problem, that more beneficial comments for the assessment were made.

In this regard, it is appropriate to consider three other suggestions. First of all, that some guidelines (6, 8, 9 and 11) that are too restrictive should be relativized. Indeed, on attempting to offer a parsimonious set, we probably over-simplified the content of the guidelines indicated, especially in the summarized version, losing shades of meaning and leaving implicit some aspects that were found to be lacking in the assessment. Moreover, the assessors advise revision of the number and organization of the guidelines, dividing up the content of some (specifically, 2 and 4) and considering the possible grouping of others (9, 10 and 12, all of which refer to formal aspects). Finally, the experts mention a lack of exhaustiveness of the set of guidelines. As stated in the introduction to the present work, and also in the document provided to the experts, we are confident that the set proposed incorporates in full the relevant content of the guidelines of reference from Haladyna, Downing and Rodríguez (2002), as well as observing the AERA, APA, NCME (1999) Standards, as recommended by some of the experts. Nevertheless, it is true that some content included as particular cases of certain guidelines is not made sufficiently explicit. This reflects the importance of making explicit all the relevant content of each guideline even in the summarized version in the form of a table, as well as reviewing the reference works in search of content that may not yet have been included.

We shall continue by presenting a new version of the guidelines, which incorporates almost all of the assessments and suggestions obtained. The new guidelines continue to be derived from the principle of validity or fit of the items and tests to the objectives of the assessment to be made. They have been grouped in three sections, the third of which is subdivided. The first includes aspects related to foundations, prior to construction itself; the second presents general criteria of construction for each item and the test they make up; and the third section constitutes a guide focusing on response options, the differentiating element of the multiple-choice format.

A. On foundations

1. In order to improve the validity of the test, the objective and domain of the assessment should be defined as detailed as possible.

In addition to deciding whether the intention is to describe a construct, identify the subjects with respect to a feature or aptitude, or place them within a group, differentiating them from one another (Crocker & Algina, 1986), it is necessary to specify components and indicators of the domain to be assessed. Failure to do this increases the likelihood of obtaining items that are easy to construct but irrelevant to the objective set. Such specification is facilitated by procedures such as review of the literature on the topic of interest, surveys of experts and, where necessary, observation of situations relevant to the topic.

2. *It is necessary to specify the context in which the items are to be used, which includes the population to which they are oriented and the circumstances in which they will be applied.*

It is important to take into account characteristics that may limit or distort comprehension of the item. Significant aspects tend to be age, educational level, mother tongue, physical or mental limitations and possible special features of subjects, as well as the language used in the domain and context assessed; also to be considered is the possibility of deciding the location and conditions of the assessment, whether it will be individual or collective, and the resources that will be made available to the subjects (Aguerri, Galibert, Zanelli, & Attorresi, 2005; Elosua & Lopez, 2005; Hidalgo, Gómez-Benito, Padilla, 2005; Tomás-Sábato & Gómez-Benito, 2005). Failure to consider these specifications will increase the likelihood of inappropriate language, content or format of the items.

B. On the expression of the domain and context in each item and test

3. The objective, domain and context of interest should be the determining criteria in construction. Each item should cover a significant unit of this referent and form with the others a relevant test.

What constitutes a significant unit is given by the domain and context of interest, with no universal rule beyond such referent. Sometimes it will be a specific and simple aspect, such as remembering the date of a historical event; in other cases it will be one that implies the solution of complex problems. In any case, the domain and context of interest should be studied in their entirety if their size permits it. If they are excessively large, representative samples should be chosen using the standard procedures, starting out from definitions of the domain and context of reference in terms of their individual components, such as each unit of knowledge in a school subject, or in terms of groupings such as thematic units. Consequently, the interpretation of the results should take into account the degree of representativeness obtained. Therefore, it must be avoided that the difficulties for constructing a given item lead to the construction of another with characteristics other than those required by the domain of reference. This would occur, for example, on constructing a memory-based item due to failure to overcome the difficulties of constructing an item for assessing a particular reasoning process.

4. *Each item should clearly show the intended content. Both the syntax and the semantics should fit with those of the domain and context of reference, without the addition of unnecessary difficulties.*

Unless the item is constructed to assess the ability to understand complex expressions, the item should be presented in as clear a way as possible, without involving unnecessary and irrelevant difficulties. The norms of the code employed should be respected, be it verbal, graphic, formal numerical or any other; thus, for example, if axes of coordinates or algebraic expressions are used, each element must be in the place and with the meaning that corresponds to it. With verbal codes, it is preferable to use affirmative or clearly interrogative expressions rather than negative ones, which tend to be more difficult to understand. Moreover, the precise meaning of technical terms employed should be respected, and special attention paid to the polysemy of many terms in the everyday language used in item construction. It should also be borne in mind that the clarity of meaning of an item depends on the specification of circumstances that frame the chosen content; for example, a general knowledge item that asks about the meaning of the term “root” should specify whether the question refers to the mathematics, linguistics or agricultural field. In brief, it is important to avoid items that are confusing or ambiguous, too wordy or too succinct.

Even so, the criteria of clarity and simplicity cannot be defined in a universal way, but rather in relation to the domain and contexts of reference. An item that is confusing or inappropriate in one context may not be so in another. A correctly-written legal text may be confusing for the lay subject; numeric-formal language may be relevant in that domain, but may not help clarity in populations unfamiliar with such expressions. Bear in mind that it is often taken for granted that the referent for the items is everyday language, when this is not always the case, so that it is not the only type of language that should dictate the rules of correctness and clarity.

5. Once the items have been constructed, it has to be made sure that they fit the domain and context of reference, especially as regards their number and their distribution in the test.

The total number of items in the test and for each portion of the domain should be that which, while not being excessive, gives a reasonable degree of trust in the representativeness of the test, one of the features of the multiple-choice format that should not be taken advantage of. As regards the order of the items, they can be grouped by type of content, or it may be preferred to mix them; given that both approaches can be defended, a decision on this should be made according to the objective of each assessment. In any case, the different items should be as independent of one another as possible, even when it is intended to focus on certain content; in this case, items relating to similar content should have different appearances.

C. On response options

C.1. Aspects that should facilitate the expression of the domain of interest

6. Each option should be the shortest possible continuation or response to the stem.

Moving the bulk of the content to the different options would mean including in each option an excessive quantity of information, sometimes with repetition, which would make it more difficult to understand what was being asked.

7. *Construction tends to be more efficient when there is just one correct option, otherwise, the criteria involved should be clarified.*

If the items contain different numbers of correct responses there may arise difficulties whose relevance would have to be evaluated, especially in the case of more insecure subjects, if it is not specified how many correct answers have to be identified in each item. Furthermore, if it is considered as “correct” that which is most correct among several that are only partially correct, it must be ensured that the scale on which such a maximum degree of correctness lies is made sufficiently explicit.

8. Spatial disposition of the options should aid perception of the item’s content.

In general, constructors should avoid practices that hinder perception of the items, such as using small print or leaving too little space between items or between lines. Furthermore, the options tend to be more clearly identified when presented vertically, though horizontal layout may be more appropriate when gradations are requested, since many metric scales are constructed in this way, and subjects are thus more accustomed.

9. The content of each option should be independent of the rest. Caution should therefore be exercised in using the options “All of the above” and “None of the above”.

The different options should not overlap or refer to one another. If this recommendation is not observed, there is a risk of introducing unnecessary problems for choosing an option or rejecting others. In order to maintain the independence mentioned above, caution should be exercised in using the options “All of the above” and “None of the above”. Nevertheless, if it is decided to use them, it is important to bear in mind the following: the former appears to introduce an additional difficulty (Dudycha & Carpenter, 1973; Mueller, 1975), especially for subjects with low levels of knowledge (Martínez, Moreno, Martín, Trigo & López, 2004), probably because it requires them to know that at least two of the above are correct. For its part, the option "None of the above" has a general difficulty effect (Dochy, Moerkerke, De Corte & Segers, 2001; Haladyna, Downing & Rodríguez, 2002), at least when it is constructed as the correct option (Martínez et al., 2004), probably because it involves negative language and logic, referring to what things *are not*, an indirect and normally more complicated form than referring to them in positive terms.

10. The options for each item should appear in order, and not require being put in order as a prior task.

If the options are presented out of order, the subjects are obliged to take on a task of prior organization different from the intended task, distracting them from the main objective of the item and affecting its validity, as occurs with the item below, corrected in the version that follows it. In general, if the options are qualitatively different, they should be organized according to some criterion of their content or appearance, and presented in order where applicable, as in the case of quantities or dates.

C.2. Aspects that should prevent undue induction of an incorrect response.

11. The options should be plausible for the subject that does not know the correct response, permitting those that do know it to identify it and reject the others.

Plausibility of the distractor options tends to be obtained by two compatible routes. One is empirical, and consists in utilizing common errors committed by subjects in the assessed domain. Another, conceptual, utilizes content close to that of the correct answer that is credible for subjects without knowledge of it.

12. Clues to the correctness or incorrectness of one or more options should be avoided. Do not use terms that may provide (undesirable) information to supplement that given in the stem.

The sources of such clues are varied, though nearly all of them involve concordance in syntactic or semantic aspects between the initial statement and the options. In terms of content it would be an error to use an option that is clearly exclusive due to its difference or incoherence, such as the option "*China*" used together with others referring to African countries in a question on Africa. It would also be inappropriate to give a clue to the correct option by including only one that fits in with the statement. Also to be avoided are modifiers, normally adverbs or adverbial phrases, that rule out or highlight certain options. Terms or expressions such as *sometimes, it may be, usually* or *generally* tend to be perceived by subjects as associated with true content, whilst others, such as *always, never, all* or *only* are associated with false content. Even so, this is mainly the case with regard to domains and contexts of everyday language and content, and it may indeed be relevant to use such modifiers in other, more specific contexts. Thus, they may be appropriate in items on urgent medical attention, where the subject should know never or always to do something in certain circumstances, since in the opposite case the risk of the patient dying is very high.

13. It is important to avoid characteristics which, without constituting clear indications of the correctness or otherwise of an option, set it apart from the rest and give rise to a suspicion in the subject that this difference may be significant.

This is the case, for example, when one of the options is much longer (or shorter) than the rest, or is clearly different in appearance or content.

14. The number of options to be included should permit the plausibility of all the options for the subject who does not know the correct one. Three is usually adequate, though if the domain so permits, a higher number may also be permissible.

It is important to bear in mind the different criteria relevant to this decision. From the point of view of probability, it would be desirable to increase as much as possible the number of options so as to reduce the possibility of random correct answers. However, with more options construction becomes more complicated, on increasing the possibility that the topic to be assessed does not permit construction of all the plausible options, thus making more likely the incorporation of options that are easily rejectable, even for subjects who do not know the answer. Weighing up these criteria, the literature seems to incline towards three options for the majority of disciplines (Abad, Olea & Ponsoda, 2001; Bruno & Dirkzwager, 1995; Delgado & Prieto, 1998; Haladyna, Downing & Rodriguez; 2002; Rodriguez, 2005; Rogers & Harley, 1999), though if the domain in question allows it, higher numbers are also admissible.

15. *Care should be taken that the set of items itself does not include any type of key or clue leading to the correct responses. Therefore, it is advisable to revise the entire test according to the guidelines*.

Such ‘keys’ may be incorporated inadvertently. Among other aspects, care should be taken that the position of the correct option in the different items does not give respondents clues, and that the content of the options in some items does not provide information that can assist the response to others.

In conclusion, the assessment carried out has made possible a new version of the guidelines (see summary in Table 3) which, while maintaining the efficiency achieved in the previous version, has been corrected in its principal defects, such as ambiguity, grouping of diverse content and lack of flexibility in certain cases; moreover, some content that was not made explicit previously, or not sufficiently so, has now been made explicit. For all of these reasons, we have every reason to trust in the utility of the guidelines now offered.

INSERT TABLE 3

References

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

Aguerri, M. E., Galibert, M. S., Zanelli, M. L., & Attorresi, H. F. (2005). Detección errónea del funcionamiento diferencial del item. Una comparación de métodos. *Psicothema, 17*, 350-355.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

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.

Crocker, L. & Algina, G. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston.

Delgado, A. R. & Prieto, G. (1998). Further evidence favouring 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.

Elosua, P. & Lopez, A. (2005). Clases latentes y funcionamiento diferencial del item. *Psicothema, 17*, 516-521.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items.* (2nd ed.). Hillsdale, NJ: LEA.

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 test item. *Applied Measurement in Education, 1* (1), 51-78.

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

Hidalgo, M. D., Gómez-Benito, J., & Padilla, J. L. (2005). Regresión logística: alternativas de análisis en la detección del funcionamiento diferencial del ítem. *Psicothema, 17*, 509-515.

Hoepfl, M. C. (1994). Developing and evaluating multiple choice tests. *Technology Teacher, 53*(7), 25-26.

Marrelli, A. F. (1995). Writing multiple-choice test items. *Performance and Instruction, 34* (8), 24-29.

Martínez, R., Moreno, R., Martín, I., Trigo, E., & López, J. (2004). *Evaluation of Multiple-choice Item-Writing Guidelines.* Paper presented at the VII European Conference on Psychological Assessment. Málaga.

Martínez, R., Moreno, R, & Muñiz, J. (2005) Construcción de ítems. In J. Muñiz, A. A. M. Fidalgo, E. García-Cueto, R. Martínez & R. Moreno. *Análisis de ítems*. Madrid: La Muralla.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Lindd (ed.), *Educational Measurement* (3rd ed.) (pp. 335-366). New York: Macmillan.

Moreno, R., Martínez, R. J. & Muñiz, J. (2004) Directrices para la construcción de ítems de elección múltiple. *Psicothema, 16* (3), 490-497.

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 ed.). Boston: Kluwer Academic Publishers.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24(2),* 3-13.

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). *A technology for test-item writing*. New York: Academic Press.

Tomas-Sábato, J & Gómez-Benito, J. (2005). Construction and validation of the death anxiety inventory (DAI). *European Journal of Psychological Assessment, 21*, 108-114.

Note. The authors would like to express their sincere thanks to all those who took the time to reply to the questionnaire that made this study possible. Funds for this research were provided by a grant from the Spanish Ministry of Science and Culture, BSO2001-3653.

Table 1. Opinions of Experts and Users on Different Aspects of the Set of Guidelines

	Experts		Users		Difference		
	Mean	(*SD*)	Mean	(*SD*)	*F*	d.f.	R^2
Useful set	4.32	0.86	4.68	0.68	3.57	1, 46.1	.05
Well founded	3.80	1.08	4.37	0.77	5.61*	1, 36.5	.09
Exhaustive	3.22	1.12	4.17	0.79	15.47*	1, 40.1	.20
Clear and simple	3.82	0.90	4.04	0.75	1.17	1, 48.1	.01
Efficient	3.69	1.04	4.19	0.82	5.32*	1, 76.0	.07
No overlap	4.17	1.02	4.21	0.85	0.03	1, 78.0	.00
No contradictions	4.35	0.91	4.52	0.67	0.91	1, 78.0	.01
Preparedness to use it	3.76	1.24	4.52	0.78	8.08*	1, 35.5	.12
Soundness of the set	3.74	1.09	4.49	0.75	12.57*	1, 77.0	.14

* $p < .05$ asymptotic two-tailed probability Snedecor or Welch *F* test

Table 2. Opinions of Experts and Users on the Clarity and the Utility of the Different Guidelines

Clarity

Guidelines	Experts		Users		Differences		
	Mean	(*SD*)	Mean	(*SD*)	*F*	d.f.	R^2
1	4.00	0.96	4.47	0.73	6.05*	1, 79.0	.07
2	2.82	1.36	3.71	0.88	9.78*	1, 42.3	.13
3	4.34	0.72	4.55	0.85	1.17	1, 79.0	.02
4	4.27	0.99	4.49	0.83	1.05	1, 79.0	.01
5	3.89	1.23	4.39	0.89	3.59	1, 45.0	.06
6	4.82	0.38	4.49	0.83	6.08*	1, 75.5	.05
7	4.27	0.92	4.55	0.94	1.57	1, 79.0	.02
8	4.44	1.15	4.54	0.90	0.15	1, 78.0	.00
9	4.51	0.82	4.57	0.82	0.08	1, 75.0	.00
10	3.34	1.39	3.61	1.22	0.78	1, 77.0	.01
11	4.69	0.47	4.55	0.85	0.66	1, 79.0	.00
12	4.66	0.55	4.64	0.79	0.01	1, 79.0	.00

Utility

	Experts		Users		Differences		
Guidelines	Mean	(*SD*)	Mean	(*SD*)	F	d.f.	R^2
1	4.48	0.87	4.74	0.68	2.20	1, 79	.03
2	3.89	0.97	4.10	0.90	0.92	1, 74	.01
3	4.27	0.70	4.51	0.73	1.94	1, 79	.02
4	4.55	0.73	4.49	0.73	0.13	1, 79	.00
5	4.22	0.93	4.17	0.97	0.04	1, 77	.00
6	4.55	0.68	4.56	0.90	0.01	1, 79	.00
7	4.20	0.90	4.33	1.03	0.30	1, 79	.00
8	2.41	1.45	4.06	1.11	27.78*	1, 77	.29
9	3.27	1.33	4.23	0.91	11.60*	1, 74	.16
10	3.64	1.36	3.85	1.14	0.50	1, 74	.00
11	4.10	1.01	4.39	0.96	1.60	1, 79	.02
12	4.51	0.57	4.49	0.78	0.03	1, 79	.00

* $p < .05$ asymptotic two-tailed probability Snedecor or Welch *F* test

Table 3. New guidelines for the construction of multiple-choice items

A. On foundations

1. In order to improve the validity of the test, the objective and domain of the assessment should be defined as detailed as possible.

2. It is necessary to specify the context in which the items are to be used, which includes the population to which they are oriented and the circumstances in which they will be applied.

B. On the expression of the domain and context in each item and test

3. The objective, domain and context of interest should be the determining criteria in construction. Each item should cover a significant unit of this referent and form with the others a relevant test.

4. Each item should clearly show the intended content. Both the syntax and the semantics should fit with those of the domain and context of reference, without the addition of unnecessary difficulties.

5. Once the items have been constructed, it has to be made sure that they fit the domain and context of reference, especially as regards their number and their distribution in the test.

C. On response options

C.1. Aspects that should facilitate the expression of the domain of interest and not add unnecessary difficulties

6. Each option should be the shortest possible continuation or response to the stem

7. Construction tends to be more efficient when there is just one correct option, otherwise, the criteria involved should be clarified.

8. Spatial disposition of the options should aid perception of the item's content.

9. The content of each option should be independent of the rest. Caution should therefore be exercised in using the options "All of the above" and "None of the above".

10. The options for each item should appear in order, and not require being put in order as a prior task.

C.2. Aspects that should prevent undue induction of an incorrect response

11. The options should be plausible for the subject that does not know the correct response, permitting those that do know it to identify it and reject the others.

12. Clues to the correctness or incorrectness of one or more options should be avoided. Avoid terms that may provide (undesirable) information to supplement that provided in the stem.

13. It is important to avoid characteristics which, without constituting clear indications of the correctness or otherwise of an option, set it apart from the rest and give rise to a suspicion in the subject that this difference may be significant.

14. The number of options to be included should permit the plausibility of all the options for the subject who does not know the correct one. Three is usually adequate, though if the domain so permits, a higher number may also be permissible.

15. Care should be taken that the set of items itself does not include any type of key or clue leading to the correct responses. Therefore, it is advisable to revise the entire test according to the guidelines.