New Guidelines for Developing Multiple-Choice Items

Article *in* Methodology European Journal of Research Methods for the Behavioral and Social Sciences · January 2006

DOI: 10.1027/1614-2241.2.2.65

3 authors:

Rafael Moreno Rodríguez
Universidad de Sevilla
57 PUBLICATIONS **392** CITATIONS
SEE PROFILE

Rafael Martínez
Universidad de Sevilla
58 PUBLICATIONS **367** CITATIONS
SEE PROFILE

José Muñiz
Nebrija Universidad
333 PUBLICATIONS **8,702** CITATIONS
SEE PROFILE

New Guidelines for Developing Multiple-Choice Items

Rafael Moreno,[1] Rafael J. Martínez, [1] and José Muñiz[2]

[1]University of Seville, Spain
[2]University of Oviedo, Spain

Abstract. The rigorous construction of items constitutes a field of great current interest for psychometric researchers and practitioners. In previous studies we have 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.

Keywords: guidelines, multiple-choice items, psychometric tests, item construction, performance

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 Rodriguez (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 others and eliminating redundancies. Two of the guidelines relate to the *content* to be assessed: (1) The content should be a sample representing the content featured in a specification table, avoiding trivial items; and (2) the representativeness should guide how simple or complex, specific or abstract, and memorization- or reasoning-based 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, and each option should agree grammatically with the statement; (4) the syntax or grammatical structure must be correct, items should not be ambiguous, confusing, or excessively short or long, and negative expressions should be used with care; and (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 distracters; (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 this reason, the options "All of the above" or "None of the above" should be avoided); and (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 et al., 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.

DOI 10.1027/1614-1881.2.2.65

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 12 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 American Educational Research Association (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% were from health sciences, and 53.1% were from social and human sciences; 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: First, 24 Likert-type questions with five assessment options, with one question on the clarity and another on the utility of each of the 12 guidelines. Second, another 9 with the same format inquiring 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 last, 4 open questions, the first offering the possibility of explaining each closed response, the next two requesting respondents to indicate positive and negative aspects not covered by previous questions, and the last for any further comments the respondents wished to make. In order to aid sample description, the users 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 (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 the 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, which was the reason 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

The 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, the 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 the guidelines, and soundness, and rated with a score of more than 4 were 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 mentioned 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 suggested 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*, and *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.

Table 1. Opinions of experts and users on different aspects of the set of guidelines

	Experts		Users		Difference		
	Mean	*SD*	Mean	*SD*	*F*	*df*	R^2
Utility	4.32	0.86	4.68	0.68	3.57	1, 46.1	.05
Conceptual foundation	3.80	1.08	4.37	0.77	5.61*	1, 36.5	.09
Exhaustiveness	3.22	1.12	4.17	0.79	15.47*	1, 40.1	.20
Clarity and simplicity	3.82	0.90	4.04	0.75	1.17	1, 48.1	.01
Efficiency	3.69	1.04	4.19	0.82	5.32*	1, 76.0	.07
Avoidance of overlap	4.17	1.02	4.21	0.85	0.03	1, 78.0	.00
Avoidance of contradictions	4.35	0.91	4.52	0.67	0.91	1, 78.0	.01
Preparedness to use guidelines	3.76	1.24	4.52	0.78	8.08*	1, 35.5	.12
Soundness	3.74	1.09	4.49	0.75	12.57*	1, 77.0	.14

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

As regards the users group, they rate all the aspects higher than 4, even giving more than 4.5 to the aspects of 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 the guidelines, and soundness, 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 ($SD = 0.58$), with 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 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.

As regards the utility of each guideline (see Table 2), in the experts group the mean of the assessments is 4.00 ($SD = 0.63$), those referring to Guidelines 2, 8, 9, and 10 being below this value and those for Guidelines 4, 6, and 12 being above 4.5. 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; and finally, some users fail to see the sense of Guideline 10. The experts' assessments are lower for all the guidelines, except 4, 5, and 12. The 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, indicates 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 1, 2, 5, and 10.

With regard to the remaining aspects, the two groups' assessments differ. The less favorable 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, the assessors said that some guidelines (6, 8, 9, and 11) are too restrictive should be relativized. Indeed, on attempting to offer a parsimonious set, we probably oversimplified the content of the guidelines indicated,

Table 2. Opinions of experts and users on the clarity and the utility of the different guidelines

	Experts		Users		Differences		
Guidelines	Mean	*SD*	Mean	*SD*	*F*	*df*	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

	Experts		Users		Differences		
Guidelines	Mean	*SD*	Mean	*SD*	*F*	*df*	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

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

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 advised 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 mentioned 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 et al. (2002), as well as observing the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (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 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 in as much detail 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 are 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 only 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, but this is not always the case, and this 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. As regards the order of the items, they can be grouped by type of content, or it may be preferable 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 they are 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 and Not Add Unnecessary Difficulties

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 to it.

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 et al., 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. 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.

The plausibility of the distracter options tends to be obtained by two compatible routes. One is empirical, and consists of utilizing common errors committed by subjects in the assessed domain. Another is conceptual, and 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, while 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 that, without constituting clear indications of the correctness or incorrectness 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, 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 toward three options for the majority of disciplines (Abad, Olea, & Ponsoda, 2001; Bruno & Dirkzwager, 1995; Delgado & Prieto, 1998; Haladyna et al., 2002; Rodriguez, 2005; Rogers &

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 in as much detail 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. Do not use terms that may provide (undesirable) information to supplement that given in the stem.
13. It is important to avoid characteristics that, without constituting clear indications of the correctness or incorrectness 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.

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) that, 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.

Acknowledgments

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 two grants from the Spanish Ministry of Science and Culture, IB05-027 (FICYT) and SEJ2005-08924/PSIC.

References

Abad, F. J., Olea, J., & Ponsoda, V. (2001). Analysis of the optimum number of alternatives from the item response theory. *Psicothema, 13,* 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. [Erroneous detection of differential item functioning. A comparison of methods.] *Psicothema, 17,* 350–355.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psycholog-*

ical 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 & Psychological Measurement, 55,* 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,* 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,* 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. [Latent classes and differential item functioning.] *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,* 37–50.

Haladyna, T. M., & Downing, S. M. (1989b). The validity of a taxonomy of multiple-choice test item. *Applied Measurement in Education, 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,* 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. [Logistic regression: Alternatives of analysis in the detection of differential item functioning.] *Psicothema, 17,* 509–515.

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

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

Martínez, R., Moreno, R., Martín, I., Trigo, E., & López, J. (April 2004). Evaluation of multiple-choice item-writing guidelines. *Paper presented at the VII European Conference on Psychological Assessment*, Málaga, Spain.

Martínez, R., Moreno, R, & Muñiz, J. (2005). Construcción de ítems. [Construction of items.] In J. Muñiz, A. A. M. Fidalgo, E. García-Cueto, R. Martínez, & R. Moreno, *Análisis de ítems* (pp. 9–52). Madrid, Spain: 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. [Guidelines for the construction of multiple-choice items.] *Psicothema, 16,* 490–497.

Mueller, D. J. (1975). An assessment of the effectiveness of complex alternatives in multiple choice achievement test items. *Educational & 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 & Practice, 24,* 3–13.

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

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

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

José Muñiz

Faculty of Psychology
University of Oviedo
Plaza Feijoo, s/n
33003 Oviedo
Spain
E-mail jmuniz@uniovi.es

Table 1: Summary of MCQ Types: MTF, OCA and SBA

Features	Multiple True False (TF)	One Correct Answer (OCA)	Single Best Answer (SBA)
Example	Regarding hip muscles A. Gluteus maximus is involved in flexion of the hip (T) B. Bicep femoris is located at lateral side of the hip (T) C. Hamstrings is inserted over the knee joint (T) D. Iliopsoas is originated from lesser trochanter of the femur (F) E. Obturator internus is innervated by L3 and L4 (F)	The function of muscle X is primarily to flex and abduct the hip. Based on the above statement, which is muscle X? **A. Gluteus maximus**** B. Bicep femoris C. Hamstrings D. Iliopsoas E. Obturator internus	A 65-year-old man has difficulty rising from a seated position and straightening his trunk, but he has no difficulty flexing his leg. Which of the following muscles is most likely to have been injured? **A. Gluteus maximus**** B. Gluteus minimus C. Hamstrings D. Iliopsoas E. Obturator internus
Stem	No scenario, simple instruction for students to perform the task	A short statement for students to perform the task	A scenario with sufficient clues for students to perform the task
Number of options	5	3 to 5	Optimum 3
Option visualization	 True and False	 True and False	 Least correct and most correct (depend on the scenario)
Dimension of option	The options are different from each other	The options can differ on a single dimension	The options must differ on a single dimension
Cover test (can we still get answer if we cover the stem)	Not applicable	Negative	Negative
Learning taxonomy	Recall (C1), Understand (C2)	Recall (C1), Understand (C2)	Application (C3), Analysis (C4)
Ease of construction *(in comparison between the three types of MCQ)*	Easy	Medium	Challenging

Features	TF	OCA	OBA
The common guidelines	**A. On Foundation** 1) In order to improve the validity of the test, the **objective** and **domain** of the assessment should be **defined** in as much detail 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 ad 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 (i.e. arrangement of words) and the semantics (i.e. the meaning of words) should fit with those of the domain 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** <u>C1. Aspects That Should Facilitate the Expression of the Domain of Interest and Not Add Unnecessary Difficulties</u> 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 (i.e. layout) 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 options "All of the above" and "None of the above." 10) The options for each item should appear in order, where applicable, according to some criterion of their content or appearance. <u>C2.Aspects That Should Prevent Undue Induction of an Incorrect Response</u> 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. Do not use terms that may provide (undesirable) information to supplement that given in the stem. 13) It is important to avoid characteristics that, without constituting clear indications of the correctness or incorrectness 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. 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 review the entire test according to the guidelines.		

Features	TF	OCA	OBA
Specific guidelines	1. Typically short stem 2. Responses start with • capitals if complete sentences e.g Regarding, Concerning etc • small letters if completing the stem 3. Tests broad/ integrated aspects (responses independent of each other) 4. Authors' names – no apostrophe e.g Down syndrome, Fallot Tetralogy 5. Punctuations: no full stops or comma following stem and at the end of responses 6. Double facts in one item to be avoided if possible 7. Look out for particular pattern e.g., first response always true	Not available, however it can follow the general guideline of MCQ and OBA.	1. Stems are usually longer with scenario 2. Stem ends usually with a question e.g., which is the best/most appropriate diagnosis/ treatment/course of action 3. Question can be answered even without the options provided. 4. Identify the focus of question. 5. Avoid rare conditions and esoteric (very specialized) question that have little clinical application. 6. Questions should test application of knowledge to pose medical decisions. 7. Avoid 'recall questions' that assess candidate knowledge of definition knowledge of definitions or isolated facts. 8. Provide detailed scenarios – mixture positive and negative findings candidate need to sort through patient info, synthesis important findings & conclude the scenario. 9. Options should be short and no additional data should be provided in the options 10.Write distractors that are plausible and the same relative length as the correct answer • All distractors - should be homogenous (same category with correct answer) • Avoid using absolutes (always, never), vague terms (usually, frequently), negatively phased (except, not) • Clues to the correctness or incorrectness of one or more options should be avoided. 11. For each question, explain why the keyed answer is most correct, why each distractors less correct. (give references for supporting evidence if available) 12. **Three option is usually adequate**, though if the domain so permits, a higher number may also be permissible.

Features	TF	OCA	OBA
Negative scoring	Negative scoring** • minus ½ mark and carry forward for MD for each wrong response • minus 1 mark and not carried forward for each wrong response ***justification must be valid, for examples penalizing candidates for unforgiveable errors related to:* • *a critical or essential step(s) in the resolution of a problem,* • *a step(s) in which examinees are most likely to make errors in the resolution of the problem, or* • *a difficult or challenging aspect in the identification and management of the problem in practice.*		
Item-writing strategies for facilitate test security	Content concerns • Use novel material to test higher level concepts. Paraphrase textbook language to avoid testing for simple recall. • Avoid overly specific and overly general content. • Application of knowledge/skills Writing the choices • Develop as many effective choices as you can. • Make sure that only one of these choices is the right answer. • Simple completion item • Complex problem-solving • Short answer or fill in the blank Online/computer-based assessment • Randomly order response choices		

References:

Case, S. and D. Swanson (2003). Constructing Written Test Questions for the Basic and Clinical Sciences. Philadelphia, US, National Board of Medical Examiners.

Clay B. (2001). A Short Guide to Writing Effective Test Questions. Kansas Curriculum Center, US.

Downing SM, Yudkowsky R. Assessment in Health Professions Education. New York, NY: Routledge; 2009.

Featherstone, C., and Hurst, Y. (2014), adapted from Case, S. M and Swanson, D. B. (2001) Constructing Written Test Questions for the Basic Clinical Sciences, NBME: Philadelphia.

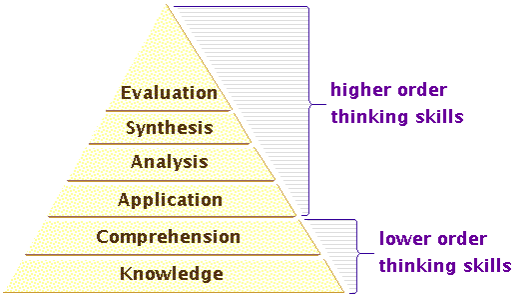
Lane S, Raymond MR, Haladyna TM. (2016). Handbook of Test Development. Routledge.

Moreno R, Martinez RJ, Muniz J. (2006). New guidelines for developing multiple-choice items. Methodology; 2 (2): 65-72.

Tan, L and J. McAleer (2008). "The introduction of single best answer questions as a test of knowledge in the final examination for the fellowship of the Royal College of Radiologists in Clinical Oncology." Clinical Oncology 20(8): 571-576.

EXTENDED MATCHING QUESTION (EMQ)

- Emerged in 1993 largely by the work of Case and Swanson
- Refers to any matching format with more than 5 options
- Extends the traditional selected-response formats.
- Can be thought of as MCQs turned upside down. See example below.
- Ideal to test higher-order cognitive knowledge. (As per medicine this relates to investigations, diagnosis, management)



Anatomy of EMQ

- A well-constructed Extended-Matching set includes four components:
    1. A theme: General topic addressed by a set of items.
    2. An option list: Response choice that apply to the item (Commonly six to 25)
    3. A lead-in statement: Specify the "task" for examinees and indicate relationship between stems and options.
    4. At least two item stems: Questions to be answered by examinees.

Theme

Theme: **Fatigue**

Options:

A. Acute leukemia
B. Anemia of chronic disease
C. Congestive heart failure
D. Depression
E. Epstein-Barr virus infection
F. Folate deficiency
G. Glucose 6-phosphate dehydrogenase deficiency
H. Hereditary spherocytosis
I. Hypothyroidism
J. Iron deficiency
K. Lyme disease
L. Microangiopathic hemolytic anemia
M. Miliary tuberculosis
N. Vitamin B_{12} (cyanocobalamin) deficiency

Options

Lead-in: For each patient with fatigue, select the most likely diagnosis.

Lead-in statement

Stems:

Stems

1. A 19-year-old woman has had fatigue, fever, and sore throat for the past week. She has a temperature of 38.3 C (101 F), cervical lymphadenopathy, and splenomegaly. Initial laboratory studies show a leukocyte count of 5000/mm³ (80% lymphocytes, with many lymphocytes exhibiting atypical features). Serum aspartate aminotransferase (AST, GOT) activity is 200 U/L. Serum bilirubin concentration and serum alkaline phosphatase activity are within normal limits.

Ans: E

2. A 15-year-old girl has a two-week history of fatigue and back pain. She has widespread bruising, pallor, and tenderness over the vertebrae and both femurs. Complete blood count shows hemoglobin concentration of 7.0 g/dL, leukocyte count of 2000/mm^3, and platelet count of 15,000/mm^3.

Ans: A

- The number for options is limited by the constraint of answer sheet design. Therefore options number must be tailored to this.
- There are two types of EMQs.
    - R-type (as above. Only 1 best answer required)
    - Pick N format (may be similar to either the extended-matching format; the primary difference is that the examinee is told to pick 2, 3, 4, or even 5 of the options listed. The format was developed to replace negative items or items with double options)

Sample Pick N Set

A. Calcium
B. Fluoride
C. Folic acid
D. Iron
E. Vitamin A
F. Vitamin B_1 (thiamine)
G. Vitamin B_6
H. Vitamin B_{12} (cyanocobalamin)
I. Vitamin C
J. Vitamin D
K. Vitamin E

For each child, select the appropriate vitamin or mineral supplements.

1. A 1-month-old infant is brought to the physician for a well-child examination. He has been exclusively breast-fed, and examination shows normal findings. (SELECT 2 SUPPLEMENTS). **Ans: B, J**
2. A 6-year-old girl has cystic fibrosis. She has been taking no medications. (SELECT 3 SUPPLEMENTS). **Ans: E, J, K**

<u>Known Advantage of EMQ</u>

- The format of themes aid the organisation of the examination, and the use of blueprinting is a natural aid to the process of writing EMQs
- As questions are written in themes or general topic it allows the teacher to write many questions for that theme and then share these questions out randomly to create more than one examination paper
- Good questions provide a structure designed to assess application of knowledge rather than purely recall of isolated facts
- The approach to writing these questions is systematic, which is very important when several people are contributing questions to one exam
- The extended list of options allows the inclusion of all relevant options, and reduces the opportunity for students to 'guess' the correct answer as in MCQs
- EMQs were found to be more discriminating than two and five option versions of the same questions resulting in a greater spread of scores, and reliability was higher as a consequence of this.
- Once the item writers master the basic, they will find EMQ is easier to create since several items are written on a common theme at the same time. In fact, the same option set can be used on future tests.

Table 1. Relative characteristics of 4 types of examinations.*

	Essay	Short Answer	Multiple-Choice	Extended-Matching
Application of Knowledge	Excellent	Good	Poor	Good, can be improved with justification
Assessment	Excellent	Good	Poor	Poor to good if justification is required
Coverage of Topic	Poor	Good	Excellent	Excellent
Reliability of Score	Poor to Fair	Good	Excellent	Excellent
Ease of Scoring	Poor	Moderate	Excellent	Excellent
Preparation time	Minimal to Moderate	Moderate	Large, if properly done	Moderate
Total Costs	Large	Moderate	Low**	Low**
Cheating (Sneak-a-Peek)	Most Difficult	Difficult	Easy	Easy unless justification is required

* Characteristics of examinations vary depending on the construction and context of the questions. A well-constructed multiple choice question might better assess cognitive skills than a poorly-constructed essay question. The relative characteristics here assume well-constructed questions of all types.

** Particularly with large numbers of examinees.

Overview of the Steps for Writing Extended-Matching Items

1. ***Identify the theme for the set.***
   The theme can be a chief complaint (eg, chest pain, fatigue), a disposition situation (eg, admission/discharge from the emergency department), a drug class (eg, antihypertensive agents, antibiotics).
2. ***Write the lead-in for the set***
   (eg, For each patient described below, select the most likely diagnosis). The lead-in indicates the relationship between the stems and options, clarifying the question posed for examinees. It is an essential component of an Extended-Matching set.
3. ***Prepare the list of options.***
   The list of options should be single words or very short phrases. List the options in alphabetical order unless there is a logical order.
4. ***Write the items.***
   The items within a set should be similar in structure. Most often, patient vignettes are appropriate.
5. ***Review the items.***
   Check to make sure that there is only a single “best” answer for each question. Also make sure that there are at least four reasonable distractors for each item. As a final check, it is recommended that you ask a colleague to review the items (without the correct answer indicated). If the colleague has difficulty determining the correct answer, modify the option list or the item to eliminate the ambiguity.

Overview of the Steps to Write Type-R EMQ in a Group

1. Define the content domain of the exam.
2. Train a group of faculty members to serve as item writers. Training should include a brief discussion of the purpose of the exam, some sample items, and the procedures to be followed during item writing.
3. Divide the group into pairs to write items. Each pair is assigned to write on 2-4 chief complaints; they generate (or modify) a list of diagnoses for each assigned complaint and write one or more patient descriptions for the diagnoses they included in their option list. Expect 20 to 60 item stems from each pair (10 to 20 per complaint). Use of computers will save considerable time in the long run.

4. Stress the following guidelines for writing stems.
5. Merge the pairs into a larger group to review the items. One approach is to have the author read the item aloud; others attempt to provide the correct answer. The group reviews the option list and modifies the item or the option list to eliminate any ambiguity. Other approaches are outlined above.
6. Type, edit, and subject the items to external review. Items should be reviewed without the correct answer indicated after they are in their final form.
7. Construct the test. Select a sample of items from each complaint; save the remaining items for subsequent exams. Items can be converted into one-best-answer items by adding a lead-in and the best five (or more) options from the option list.

Constructing Tips:

1. Time allotted per item depends on the length of the stem, skills to be measured and known relationship between response time and those skills. If more stems are included, the later stems requires less time.
2. Individual options should be phrased short enough to be scanned quickly.
3. Option lists should be listed in alphabetical order.
4. Pictorial material can be used to replace option list as well.
5. In contrast to option list, item stems can be quite long. It should dictate the thoroughness of questions/patient descriptions.
6. Item stems should challenge the examinees to identify key information.
7. In reviewing the items, check to make sure that there is only a single "best" answer (for type-R) for each question. Also make sure that there are at least four reasonable distractors for each item.

Please Avoid:

1. Sets without specific lead in, such as *"Match each with the best options".* They generally pose ambiguous tasks for examinees and they might answer incorrectly.
2. Setting too many EMQ on the same test. This might impose oversampling on certain topic and under sampling of other topics – a threat to the test validity!

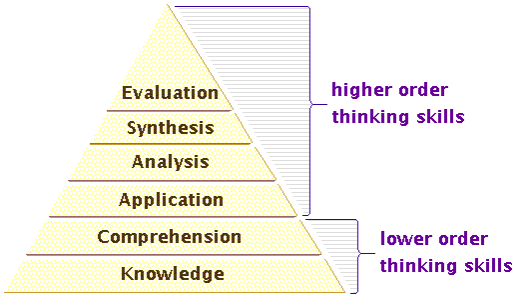
References:

Case SM, Swanson DB: Extended matching items: a practical alternative to free response questions. Teaching and Learning in Medicine 1993, 5:107-15

Downing SM, Yudkowsky R. Assessment in Health Professions Education. New York, NY: Routledge; 2009.

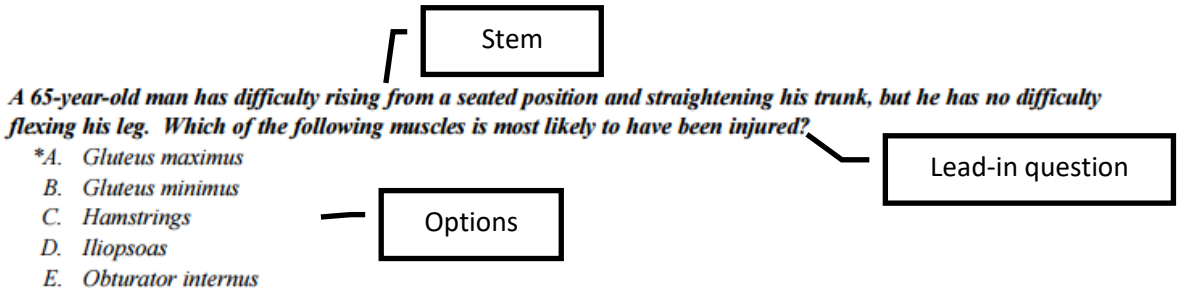
SINGLE BEST ANSWER (SBA)

- The most widely used multiple-choice-item format
- Require application of knowledge, allowing assessment of both an examinee's information base plus ability to use that information.
- Note the question difference
    - Recall: What area is supplied with blood by the posterior inferior cerebellar artery?
    - Application: A 62-year-old man develops left-sided limb ataxia, Horner's syndrome, nystagmus, and loss of appreciation of facial pain and temperature sensations. What artery is most likely to be occluded?

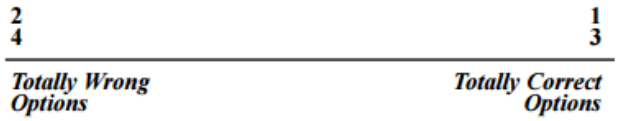


Anatomy of SBA

- Stem
- Lead-in question
- Followed by a series of choices (Typically one correct answer and four distractors)



❖ *Note that in True/False, the options are absolutely true or false with no ambiguity*



- *But in SBA, the incorrect answers are not completely wrong, they are less correct than the "keyed answer."*

Note that the incorrect options are not totally wrong. The options can be diagramed as follows:

D	C	A	E	B
Least Correct				*Most Correct*

Overview of the Steps in Writing SBA

1. **Identify an area of the blueprint which is in need** of more questions or the area of the blueprint which you have been allocated.
2. **Identify the topic area and the level of thinking** you want to test, and write a question around this which mimics tasks that successful candidates must be able to undertake at the next stage of training**. Ideally, questions should be pitched at the level of integration/interpretation** (questions which require "putting the pieces together") and problem solving (questions which require "clinical judgement"), not simple recall (questions which can be answered with a Google search).
    a. Example recall questions, to avoid:
        i. What is the name/definition of this procedure?
        ii. Which of the following is correct?
3. **Construct the stem.** This should present a single, clearly formulated problem. The stem should contain enough information to allow candidates to answer without referring to the options
4. **Construct the lead-in** in such a way that it builds on the information in the stem and poses a clear question. Candidates should be able to answer without looking at the options, and should not be able to answer if the information in the stem is masked.
5. **Write the options.** These should be of similar length, along a continuum, grammatically consistent and logically compatible. **All options must differ in a similar dimension**. See common mistakes done in constructing options as below:

Which category of fruit includes grapes, peaches and berries?

A. Spring
B. Winter
C. Summer
D. Tropical

Ineffective distractor – Not homogenous

If appropriate, order the options in a logical order (e.g. numeric, alphabetical, or anatomical). All the distractors must be plausible (think of the educational impact of suggesting that something potentially dangerous is plausible but not the best answer!).

6. **Review the item**
    a. Does it focus on important problems relevant to practice?
    b. Can it be answered without looking at the options?
    c. Are all the relevant facts included in the stem?
    d. Can it be read and answered within approximately one minute?
    e. Are all the options plausible, with one of them standing out as being the best option?
    f. Does the question successfully avoid the pitfalls in Technical Flaws to be Avoided (explained below)?

General Rules in Writing SBAs

1. Each item should **focus on an important concept**, typically a common or potentially catastrophic clinical problem. **Don't waste testing time with questions assessing knowledge of trivial facts**. Focus on problems that would be encountered in real life. Avoid trivial, "tricky," or overly complex questions.
2. Each item should **assess application of knowledge, not recall** of an isolated fact. The item stems may be relatively long; the options should be short.
3. The stem of the item must **pose a clear question**, and it should be possible to arrive at an answer with the options covered. To determine if the question is focused, cover up the options and see if the question is clear and if the examinees can pose an answer based only on the stem. Rewrite the stem and/or options if they could not.
4. All **distractors (ie, incorrect options) should be homogeneous**. They should fall into the same category as the correct answer (eg, all diagnoses, tests, treatments, prognoses, disposition alternatives). Rewrite any dissimilar distractors. Avoid using "double options" (eg, do W and X; do Y because of Z) unless the correct answer and all distractors are double options. Rewrite double options to focus on a single point. All distractors should be plausible, grammatically consistent, logically compatible, and of the same (relative) length as the correct answer. Order the options in logical order (eg, numeric), or in alphabetical order.
5. **Avoid technical item flaws** that provide special benefit to testwise examinees or that pose irrelevant difficulty.

Technical Flaws to be Avoided

1. Issues Related to Testwiseness
    a. Grammatical cues - one or more distractors don't follow grammatically from the stem
    b. Logical cues - a subset of the options is collectively exhaustive
    c. Absolute terms - terms such as "always" or "never" are in some options
    d. Long correct answer - correct answer is longer, more specific, or more complete than other options
    e. Word repeats - a word or phrase is included in the stem and in the correct answer
    f. Convergence strategy - the correct answer includes the most elements in common with the other options

2. Issues Related to Irrelevant Difficulty
    a. Options are long, complicated, or double
    b. Numeric data are not stated consistently
    c. Terms in the options are vague (eg, "rarely," "usually")
    d. Language in the options is not parallel
    e. Options are in a non-logical order
    f. "None of the above" is used as an option
    g. Stems are tricky or unnecessarily complicated
    h. The answer to an item is "hinged" to the answer of a related item

References:

Case SM, Swanson DB: Extended matching items: a practical alternative to free response questions. Teaching and Learning in Medicine 1993, 5:107-15

Featherstone, C., and Hurst, Y. (2014), adapted from Case, S. M and Swanson, D. B. (2001) Constructing Written Test Questions for the Basic Clinical Sciences, NBME: Philadelphia.

Downing SM, Yudkowsky R. Assessment in Health Professions Education. New York, NY: Routledge; 2009.

9 key principles of Single Best Answer question

1. Question can be answered even without the options provided.
2. Identify the **focus** of question.
3. **Avoid rare conditions** and esoteric (very specialised) question that have little clinical application.
4. Questions should **test application of knowledge** to pose medical decisions.
5. **Avoid 'recall questions'** that assess candidate knowledge of definition knowledge of definitions or isolated facts.
6. Provide **detailed scenarios** – mixture positive and negative findings candidate need to sort through patient info, synthesis important findings & conclude the scenario.
7. **Options** should be short, and no additional data should be provided in the options
8. Write **distractors** that are plausible and the same relative length as the correct answer
    - All distractors - should be **homogenous** (same category with correct answer)
    - **Avoid** using absolutes (always, never), vague terms (usually, frequently), negatively phased (except, not)
    - **Clues** to the correctness or incorrectness of one or more options **should be avoided**.
9. For each question, **explain** why the keyed answer is most correct, why each distractors less correct. (give references for supporting evidence if available)

Reference:

Tan, L. and J. McAleer (2008). "The introduction of single best answer questions as a test of knowledge in the final examination for the fellowship of the Royal College of Radiologists in Clinical Oncology." Clinical Oncology 20(8): 571-576.

Guidelines for constructing/reviewing MCQ items

Areas	Guidelines
A. On foundation	1) In order to improve the validity of the test, the objective and domain of the assessment should be defined in as much detail 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 (i.e. arrangement of words) and the semantics (i.e. the meaning of words) should fit with those of the domain 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 ***C1. 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 (i.e. layout) 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 options "All of the above" and "None of the above." 10) The options for each item should appear in order, where applicable, according to some criterion of their content or appearance.
C2. Aspects that should prevent undue (i.e. unnecessary) 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. Do not use terms that may provide (undesirable) information to supplement that given in the stem. 13) It is important to avoid characteristics that, without constituting clear indications of the correctness or incorrectness 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 review the entire test according to the guidelines.

Source: Moreno R, Martinez RJ, Muniz J. (2006). New guidelines for developing multiple-choice items. Methodology; 2 (2): 65-72.