Psicothema

Psicothema 2015, Vol. 27, No. 4, 388-394
doi: 10.7334/psicothema2015.110

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

# Guidelines based on validity criteria for the development of multiple choice items

Rafael Moreno[1], Rafael J. Martínez[1] and José Muñiz[2]
[1] Universidad de Sevilla and [2] Universidad de Oviedo

## Abstract

**Background**: Many different guidelines have been presented for the construction of multiple choice items. Those guidelines have been based on the observation of errors when constructing items but not on any clear scientific criterion. Our main objective was to draw up guidelines for the development of multiple choice items based on validity criteria. **Method:** We used the properties of adjustment, precision, and differentiation, applying them to three basic phases of instrument construction: the definition of the objective and its context; their expression in the instrument and item stem; and the elaboration of response options. **Results:** We have combined these properties and phases to give nine general guidelines with a firm theoretical footing. **Conclusions:** Finally, we have written a checklist with twenty-four points to check how far the measurement instruments comply with the proposed guidelines.

***Keywords:*** Validity, Test development, Multiple-choice items, Item-writing, Guidelines.

## Resumen

***Directrices para el desarrollo de ítems de elección múltiple basadas en validez.*** **Antecedentes:** se han propuesto diferentes directrices para la construcción de ítems de elección múltiple, basadas sobre todo en la observación de errores al construir los ítems pero no en algún criterio científico claro. El objetivo central del presente trabajo es generar directrices para el desarrollo de ítems de elección múltiple basadas en criterios de validez. **Método**: se utilizan las propiedades de ajuste, precisión y diferenciación, aplicándolas a tres etapas fundamentales del desarrollo de instrumentos de evaluación: definición del objetivo y su contexto, su implementación en el instrumento y enunciado de los ítems, y elaboración de las opciones de respuesta. **Resultados**: la combinación entre tales propiedades y etapas da lugar a nueve directrices generales que, además de quedar fundamentadas, permiten resolver cualquier duda que surja a quienes desarrollan ítems de elección múltiple. **Conclusiones:** para facilitar esa labor, las directrices son complementadas con una lista de veinticuatro cuestiones con la que comprobar el grado en que los instrumentos de medida cumplen las directrices propuestas.

***Palabras clave:*** validez, desarrollo de instrumentos, ítems de elección múltiple, construcción de ítems, directrices.

Multiple choice items are used in a wide range of assessment settings including ability, aptitude, attitude and personality instruments. Their construction has to be systematic to avoid defective items which are all too common (Downing, 2005; Tarrant & Ware, 2008). This was the purpose behind the taxonomies of guidelines designed by Haladyna and Downing (1989a), Osterlind (1998), Haladyna, Downing and Rodríguez (2002) and Haladyna and Rodríguez (2013).

Haladyna et al. (2002) reorganized the guidelines presented by Haladyna et al. (1989a), which in turn summarized above 40 taxonomies. Drawing on the guidelines of Haladyna et al. (2002), Moreno, Martínez and Muñiz (2004) produced a shortened version summarizing those guidelines and excluding others that were irrelevant or repetitive. After expert assessment and modifications (Moreno, Martínez & Muñiz 2006), they were then used in two courses: one on exam construction for university lecturers and the other on the writing of questionnaires in psychology for PhD students. This practical application highlighted certain flaws as well as advantages. The main flaw with the current and previous guidelines is that they do not come explicitly from common criteria. Thus, the person learning these guidelines may assume them as reasonable, but be unable to understand their foundations. This may lead to a mechanical rather than an autonomous use where any doubt not explicitly included in them could be dealt with. This could even lead to some guidelines being left out, as happens with other test development tasks such as the adaptation to other languages (Rios & Sireci, 2014).

In general the existing guidelines come from common sense and the observation of errors in constructed items, but there is no body of empirical studies supporting them sufficiently, with the exception of the number of response options. This was highlighted by Haladyna and Downing (1989b) and Haladyna et al. (2002) and little has changed since then. On the other hand, the absence

Received: April 28, 2015 • Accepted: September 11, 2015
Corresponding author: Rafael Moreno
Facultad de Psicología
Universidad de Sevilla
41018 Sevilla (Spain)
e-mail: rmoreno@us.es

of a common notion behind those guidelines was considered by Osterlind (1990; 1998) and Haladyna and Rodriguez (2013). These authors pointed to *Scientific Validity* as a basic notion to generate new guidelines or summarize existing ones although they have not used it explicitly.

As validity is the criteria which instruments and their items must comply with, it should form the basis of guidelines for the development of instruments with multiple choice items. Many terms and notions have been used to refer to validity in different phases of research and in studies covering a wide range of methodological issues (American Psychological Association, American Educational Research Association & National Council on Measurement in Education, 2014; Cronbach, 1982; Kane, 2006; Lane, 2014; Padilla & Benítez, 2014; Rios & Wells, 2014; Shadish, Cook & Campbell, 2002; Sireci & Faulkner-Bond, 2014). A review of all of these notions has enabled us to narrow them down to three properties of validity (Martínez & Moreno, 2014): adjustment of each of the elements considered to their respective referents, precision of all of them, and differentiation among them.

Adjustment or representativeness of each and every element in relation to a specific referent implies that no element is surplus to, or short of, what is required. Adjustment therefore implies the exhaustiveness of relevant elements but also parsimony because of the absence of unnecessary ones. Precision or clarity means being able to understand without ambiguity each element of the instrument and its elaboration, which is proven by the consistency or reliability with which it is understood and used. Finally, differentiation or control of the elements which are understood as different and independent means that they are mutually exclusive. The opposite case, overlapping or confusion between two or more elements which are meant to be different and independent, occurs when they appear to be associated when they should not, or the differences between them are not clear enough.

These three properties should be taken into account in the several tasks involved in the development of instruments and their items. The first task deals with the definition of aims and then the writing of the items, while the others concern content validity studies and pilot testing (Downing, 2006; Muñiz & Fonseca, 2008; Schmeiser & Welch, 2006). This study focuses on the initial task that we have separated into the following three phases: (a) the delimitation of the objective being assessed and its context, (b) their expression in the stem of each item and in the instrument as a whole, and (c) the writing of the options for each item.

Given the above, the objective of our study is to draw up guidelines for the construction of instruments with multiple choice items by applying explicitly the three validity defining properties in each of the three construction phases. This combination of validity properties and test construction phases provides us with nine guidelines.

The nine guidelines would be more efficient if they were specified at an operational level, as happens with those existing in the literature. This would enable us to check whether the existing guidelines have been entirely subsumed by the new ones. To do this, two aspects of each guideline were specified: the content being assessed and the elements which contextualize them in their delimitation and in their subsequent expression in the instrument.

## *Guidelines*

The following nine guidelines are proposed for the construction of multiple choice items.

### *a) Valid delimitation of the referent of the instrument to be developed*

The assessment goal for which the instrument is being developed must dictate the content and the context which are going to be considered, and these will be the referents for the construction of the instrument. Hence, content and context must be defined with the three validity properties as explained below:

1. List the content and the context of the assessment carefully without missing anything out or including anything surplus.
   *The contents*. The referent contents are usually those expressed by the constructs proposed in any subject area or field. They are normally specified as lists of concepts, subject syllabi, conceptual maps with their contents and interrelations, or specification tables with dual entries, one for the contents and another for the possible tasks to be performed with each one according to taxonomies such as Bloom's (Krathwohl, 2002). In any of those modes, one should make a full list of all the contents marked by the objective. At the same time, contents that do not appear should not be added. Accepted ethical values should form part of the content objectives.
   *The context*. To meet the objective, one must clarify the contextual characteristics that may influence the contents and the assessment. Some of these have to do with the people the instrument is being designed for: age, cultural level, languages or disabilities. Other characteristics refer to their setting, such as socio-economic status, type of culture and social and ethical norms. Other features refer to the assessment itself, such as the instrument and how it is administered. All relevant characteristics should be specified.
2. Delimit assessment content and context precisely, with no ambiguity.
   *The contents*. Each content of the assessment subject matter must be clearly delimited, so that any professional would be able to understand it consistently or reliably at any time. When appropriate, it should also be clear which relations are being considered between the contents.
   *The context*. Each aspect of the people under assessment, their setting and the assessment itself must be precisely delimited. This should allow any professional or evaluator to understand them in the same way.
3. Differentiate assessment contents and context, with no overlaps between them.
   *The contents*. The delimitation of the diverse referent contents should ensure that they are all clearly different; each with characteristics which differentiate them from the others. Likewise, one should avoid overlapping between categories or levels which are meant to be different; they should all be mutually exclusive.
   *The context*. One should clearly differentiate the elements of the people being assessed, their setting and the assessment itself. Their respective categories or levels should not be associated or overlap unless this occurs in the objective.

### *b) Valid expression of referent contents and context in the instrument*

After the adjusted, precise and differentiated delimitation of the reference contents and context, the following step is to decide how to build and administer the instrument. This process has to follow the three properties of validity as indicated below.

4. Adjust the assessment content and context in the stem of each item and in the instrument as a whole to their referents.
   *The contents*. The content of the stem for each item must be an indicator of one or more of the referents. All of the referents should be included in the test if possible, or at least a representative sample if there are too many. In all cases, one should respect the specific weight in the referent of those included in the test, together with the categories or levels which are to be considered. Care should be taken to avoid introducing contents which do not fit or which leave out others which are wanted but have not been properly expressed.
   *The context*. The articulation of context elements in the instrument should fit the referents. That fit should dictate how to formulate item content: as a question, a sentence stem to be completed or an indication to follow; at empirical level, including data or not, at abstract level with a definition or substituting it with a simple label or term, or using both possibilities; with verbal, numeric-formal, iconic-graphic codes, or any combination of these; and describing defining characteristics, highlighting one or more categories or levels these characteristics may occur in, or combining both modes. In all cases, the syntax and semantics of the codes used must adjust to their referents.
   The instrument should contain the number of items necessary to achieve content fit, no more no less. When the referent contents involve certain sequences, they must be respected in the instrument. A decision also has to be made about whether the instrument is going to be administered individually or as a group, whether there is to be a time limit and what instructions are to be given. Care should be taken to exclude what is surplus but include what is necessary. For example, one should make sure people cannot cheat because they are sitting too close to each other; they should have enough time to answer all the questions; and one must not leave out vital information such as the instructions. In the same way, the instrument needs to be ethically in tune with the people being assessed and their setting; it should not offend anyone.
   When sampling of the contextual elements is needed, certain procedures should be followed to facilitate fit or representativeness in relation to their respective referent. These could specify the referents in terms of simple or compound units, and use random or intentional criteria for the choice of units for the sample (Lohr, 1999; Martínez & Moreno, 2014). If, for some reason, the sample obtained cannot be representative, the results should be considered with that limitation, without generalizing about what is not represented in the sample.
5. Express the content and context of the stem of each item precisely.
   *The contents*. People should be able to identify precisely the content covered in each question and the relations between them when this is the case. That includes accuracy in the categories or levels considered for each content and relation.
   One might need to test consistency in the understanding of one or more items and their responses. This can be done by repeating the same item content in a specific instrument with different appearance. This is done in the "two halves" and "parallel tests" procedures for the study of instrument reliability (Crocker & Algina, 1986; Muñiz, 2000) and in so-called "isomorphic" items (Bejar, 1993).
   *The context*. The rest of the elements implemented in the instrument should also be precise and unambiguous. The grammar and wording used in each stem are crucial in this sense. Care should be taken to choose the words which best express what is wanted, respecting the precise meaning of the technical terms used and being aware of possible double meanings of the terms used in everyday language. Ambiguity due to excessively short, long, verbiages or complicated sentences should be avoided. Whenever possible, negative expressions should not be used because they often lead to problems of understanding.
   Numbering each item, placing them on the same page and column if possible and using adequate font type and size help written instruments to be more precise. This also applies for the instructions for answering the test.
6. Differentiate the content and context of the stem of each item.
   *The contents*. When the stem of an item includes diverse contents, they should all be differentiated, without overlaps or confusion between them. One should also clearly differentiate the contents of the different items of the instrument. To avoid such confusion, the stems of different items must be independent of each other, even when their respective contents are related in the referent.
   *The context*. In each item, semantics, syntax or other elements should be clearly differentiated. Care should also be taken to ensure these elements do not overlap or confuse their effects with those of the content being assessed: for example, when inappropriate wording in the stem induces a specific response.
   In the instrument as a whole, one needs to differentiate between the assessed contents and elements of its articulation such as the number and sequence of items. If this number is very high or the sequence is unsuitable, it may tire people or confuse them. In such cases, one has to reach the best possible balance between differentiation and adjustment.
   The layout on the page is important for people to perceive the differences between the items, with enough space between them on each page; this is also true for the instrument's instructions.

*c) Valid writing of response options*

In the format of multiple choice items, the steps we have outlined above must be completed with the construction of the options. This process also has to follow the three properties of validity as indicated below.

7. Adjust the content and contexts of each option to the stem.
   *The contents*. The options and stem must work together in order to fit: both should express the intended content. Hence it is preferable to avoid stems which just ask people to: "Choose the correct option from the following". In these cases the item content is determined by the options, which

can introduce a variety of subjects which are an obstacle to the fit with the referent contents.

On the other hand, the options will fit the stem when their contents are plausible and not trivial in relation to the stem. There are two ways to build this plausibility: conceptually by using contents which are thematically close to each other, and empirically by drawing on the experience of the subject. In achievement tests, the incorrect options can be constructed with common errors made by students learning the assessed contents. The presence of trivial content in the options may mean that other relevant or plausible content is left out.

*The context*. The plausibility or fit of all the options with the stem content helps to decide how many to have. Too many options may be an obstacle to the fit because some contents may lack sufficient variety. Three options is considered a suitable number. Many studies have recommended this, considering the balance between the minimum number of options and reducing the probability of people guessing (Rodriguez, 2005). Special care should be taken when constructing the last option as this is often filled with trivial contents when there is no relevant content left.

The options and stem should also fit in terms of the syntax and semantics of the codes used, which must be appropriate for the assessed contents. If the stem is a question, each option must be an answer; and if it is the first half of a statement or indication, each option should complete the sentence properly and make sense. In general, the stem as a question favors its fit with the response options because they usually express content completely. Options which do not fit the stem by denying or contradicting what it expresses should be avoided. This occurs in the last option of the following item.

Inappropriate item:
How is the index of kurtosis in a leptokurtic distribution of values?

1. Greater than 0
2. Less than 0
3. Such an index does not allow this assessment

The fit between options and stem in syntax includes their congruence in punctuation, gender and number. Each option should start with a capital letter if the stem is a question, but not when the option completes the sentence. And if the stem includes most of the item's content, the options will tend to be short.

8. Express the content and contexts of each option precisely.

   *The contents*. The content of each option must be clear. Where appropriate, this includes accuracy in the criterion which makes one of the options correct. For that reason, when one option is more correct than the others, the criterion of that hierarchy must be clearly identifiable.

   *The context*. The contextual elements of the options, such as their wording, letter size or spatial disposition, must be clear. As is true for stems, ambiguous wording and negative expressions should be avoided in the options because they complicate understanding.

9. Differentiate clearly between the options and the stem, in terms of both contents and contexts.

   *The contents*. There must be a clear difference between the contents of the different options, making it easier to discriminate between them. When there are several correct responses, offering only one correct option will help this differentiation.

   Thus all options should be mutually exclusive in terms of content, with no overlaps as sometimes occurs in quantitative values with some of them being included in two or more intervals. Hierarchical options that include defining elements from previous ones should not be considered as overlaps because each option adds its own elements.

   Mutual exclusion or exclusivity of the options also implies that they all refer to the stem and not to one or more of the other options. Thus, options such as "None of ..." or "All the above are correct (or incorrect)", or "Options A and B are correct" should be avoided.

   There should also be differentiation between the contents of each option and the stem. Not doing so may lead people towards certain options or away from others, as in the following item because of the first option.

Inappropriate item:
What type of validity is a synonym for generalization of a set of data?

1. External validity
2. Homocedasticity
3. Accuracy

*The context*. The differentiation between the options must also be clear in the layout of the options. Grammar and wording are important in this sense, as are more perceptive aspects. Careful use of wording and terms, indentations and bold script, and font type and size are recommended. Attention should also be paid to the spatial layout which best distinguishes the different options. Vertical layout tends to be the most appropriate except when the set of options is a numerical scale for graded responses.

If care is not taken over those elements, they may induce or hamper the choice or rejection of options, with that response overlapping with what the person under assessment would answer if such confusion did not exist. This can happen with adverbs such as "sometimes" or "occasionally" which are often true, and "always" or "never" that are often false, which provides crucial information in many subject areas. In the following item, the correct answer is the only one that 'sounds French'.

Inappropriate item
Which poet wrote his work in French?

1. Walt Whitman
2. Aleksei Koltsov
3. Charles Baudelaire

An option which stands out from the others in terms of content or length, grammar or wording is another feature which can induce answers. The extra attention paid to the option may overlap with

the response a person might give, had this difference not existed. This occurs in the last option of the following item.

Inappropriate item:
Who was the author of the controversial "Little Albert" experiment?

1. Abraham H. Maslow
2. Burrhus F. Skinner
3. John B. Watson, who established behaviourism

Options should be clearly organized or ordered. Overlaps may occur when options which include dates, quantities or contents in two or more options do not appear in order. This applies to the item below, corrected in the version that follows it.

Inappropriate item:
Of the following, which denominations of measurement scales were proposed by S. S. Stevens?

1. Ratio, ordinal, and nominal
2. Harmonic, nominal, and topographical
3. Large, topographical, and harmonic

Of the following, which denominations of measurement scales were proposed by S. S. Stevens?

1. Ratio, ordinal, and nominal
2. Topographical, harmonic, and nominal
3. Topographical, harmonic, and large

Overlaps which induce responses through options may also occur in the instrument as a whole, especially if the items contain different numbers of options or the placing of correct options gives clues. The first overlap may be avoided by homogenizing the number of options in each item, and the second by distributing the position of each correct option randomly throughout the instrument, with people being told about this.

These derived guidelines are summarized in Table 1. The first in each phase of test construction refers to adjustment, the second to precision and the third to differentiation.

To make the verification of the extent of compliance with the proposed guidelines easier for test designers, we have drawn up a checklist of twenty-four questions (see Table 2) organized by phases and properties. It includes specifications referring to contents and context.

*Table 1*
Guidelines based on validity criteria for the development of multiple-choice items

1. List the content and the context of the assessment carefully without missing anything out or including anything surplus.
2. Delimit assessment content and context precisely, with no ambiguity.
3. Differentiate assessment contents and context, with no overlaps between them.
4. Adjust the assessment content and context in the stem of each item and in the instrument as a whole to their referents.
5. Express the content and context of the stem of each item precisely.
6. Differentiate the content and context of the stem of each item.
7. Adjust the content and contexts of each option to the stem.
8. Express the content and contexts of each option precisely.
9. Differentiate clearly between the options and the stem, in terms of both contents and contexts.

*Table 2*
Checklist: Questions to check guidelines

*Valid delimitation of the referent of the instrument to be developed*

Adjustment

1.1. Have you included all the assessment contents, making sure there are no irrelevant ones?

1.2. Have you specified the main characteristics of the people being assessed, including their setting?

Precision

2.1. Have you accurately delimited the assessment contents and their context?

Differentiation

3.1. Have you clearly differentiated the distinct assessment contents and their context, with no overlaps?

*Valid expression of contents and referent context in the instrument*.

Adjustment

4.1. Does the content of each item correspond to one of the referents?

4.2. Does the set of contents included in the instrument represent all or a valid sample of the referents?

4.3. Do the quantity and sequence of items, the expression of the contents, and the context fit their referents?

Precision

5.1. Can you identify the content of each stem precisely?

5.2. Are the grammar and wording of each stem and the instructions correct and clear?

Differentiation

6.1. Have you avoided overlapping content between two or more items that could influence people's choice?

6.2. Have you made sure no element in the instrument, its items and administration induces a response to any item?

*Valid elaboration of response options*

Adjustment

7.1. Have you chosen a number of options which ensures that all option content is plausible?

7.2. Does each option fit the stem in terms of grammar, wording and any other aspect of expression?

Accuracy

8.1. Is the content of each option precisely identifiable?

8.2. Do the grammar, wording and other elements of expression help the clarity of each option?

Differentiation

9.1. Have you differentiated the correct option adequately, allowing examinees to rule out the rest?

9.2. Are the contents of the different options mutually exclusive?

9.3. Have you avoided overlaps between the contents of each option and the stem which could influence response choice?

9.4. Do the grammar, wording and layout of the options help to differentiate them from the stem?

9.5. Have you made sure your grammar and wording do not influence the choice of an option?

9.6. Have you made sure that no option is clearly different from the rest in content or any other way, such as grammar, wording or length?

9.7. Have you ordered or organized the options appropriately?

9.8. Have you made sure there is the same number of options in all the instrument's items?

9.9. Have you made sure that no aspect of the test induces a given response, such as the position of the correct option in each item?

## Discussion and conclusions

We have drawn up a set of guidelines for the construction of instruments with multiple choice items by explicitly applying the three constitutive properties of validity to the three phases considered in this construction. These guidelines have been specified in terms of the content being assessed and other elements which contextualize them in each of the mentioned phases.

The sequence in which the guidelines are presented is not always that of their application. When there are no well specified referents, the test builder starts from generic ideas or a more or less structured set of items which help delimit the objective and its context. This process of delimitation of referents, items and instruments is normally iterative until a valid final construction is achieved. Once the instrument has been constructed it will require testing by experts and pilot data collection.

In the application of the three validity properties as criteria for the construction of the instrument one has to bear in mind that these properties influence each other mutually. Precision facilitates differentiation and adjustment; while inaccuracy means that some elements will be implicit, hampering the fit and they may overlap with others. In turn, differentiation and adjustment help each other, and both favor precision, just as their deficiencies hamper accuracy and undermine each other; thus, the lack of fit due to an excess of elements facilitates overlap between them or with others.

The first advantage of this new set of guidelines is that there are a lot fewer than the 31 previous guidelines proposed by Haladyna et al. (2002). The second is that they are better organized than the 15 guidelines proposed by Moreno et al. (2006). Furthermore, the specifications of these new guidelines exhaustively subsume the previous ones presented by Moreno et al. (2006) as shown in Table 3. Previous guidelines 1 and 2 referring to the delimitation of referents are condensed into new guideline 1. Those referring to the expression of the referents in the instrument and its items are now organized as follows: previous 3 is in new 4, previous 4 in new 1, 4 and 5, and previous 5 in new 4 and 6. The remaining guidelines, from 6 to 15, referring to the options, are included in the new version in numbers 7, 8 and 9 as shown in Table 3.

Guidelines 2 and 3 in the current version, referring to accuracy and differentiation in the delimitation phase of the referents of instrument construction, do not appear in Table 3 because they add recommendations which were absent in the guidelines of the reviewed literature. Likewise, new guidelines 4, 5 and 6 refer to elements in the item stems, in particular their expression and layout, which are not normally mentioned in the existing guidelines in the literature.

*Table 3*
Relation between the two versions of guidelines

| Phases | Guidelines in Moreno et al. (2006) | Guidelines in this paper |
|---|---|---|
| a) Delimitation of referents | 1 | 1 |
| | 2 | 1 |
| b) Expression in items and instrument | 3 | 4 |
| | 4 | 1, 4 and 5 |
| | 5 | 4 and 6 |
| c) Elaboration of response options | 6 | 7 |
| | 7 | 8 and 9 |
| | 8 | 9 |
| | 9 | 9 |
| | 10 | 9 |
| | 11 | 7 |
| | 12 | 7 and 9 |
| | 13 | 9 |
| | 14 | 7 |
| | 15 | 9 |

In turn, the conceptual foundation laid for the generated and existing guidelines makes them more predictable for someone who understands validity properties and the instrument elaboration phases. The combination of these properties and phases should also make it possible for test designers to resolve autonomously any doubt arising during construction and which is not explicitly addressed in the guidelines.

These guidelines could also be used in a systematic plan of empirical assessment of all existing guidelines, something which is needed to fill this gap in the field of multiple choice item construction. The convergence of this conceptual foundation and empirical research should allow substantial progress to be made in this area.

## Acknowledgments

This work was funded by the Spanish Ministry of Economy and Competitiveness, Project reference PSI2014-56114-P.

## References

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. L. Mislevy & I. I. Bejar (eds.), *Test theory for a new generation of tests*. Hillsdale: Erlbaum.

Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*(2), 133-143.

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

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

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Westport, CT: American Council on Education/Praeger.

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory into Practice, 41*(4), 212-218.

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*, 127-135.

Lohr, S. L (1999). *Sampling: Design and Analysis*. Pacific Grove, CA: Duxbury.

Martínez, R. J., & Moreno, R. (2014). *¿Cómo plantear y responder preguntas de manera científica?* [How to raise and answer questions in the scientific manner]. Madrid: Síntesis.

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

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

Muñiz, J. (2000). *Teoría clásica de los tests* [Classical Test Theory]. Madrid: Pirámide.

Muñiz, J., & Fonseca, E. (2008). Construcción de instrumentos de medida para la evaluación universitaria. *Revista de Investigación en Educación, 5*, 13-25.

Osterlind, S. J. (1990). Establishing criteria for meritorious test items. *Educational Research Quarterly, 14*(3), 26-30.

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

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*, 136-144.

Rios, J. A., & Sireci, S. G. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing, 14*(4), 289-312.

Rios, J. A., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*, 108-116.

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.

Schmeiser, C.B., & Welch, C. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Hougton Mifflin.

Sireci, S.G., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*, 100-107.

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education, 42*(2), 198-206.