GUIDELINES FOR WRITING MULTIPLE-CHOICE QUESTIONS

This document is destined to establish guidelines for writing multiple choice questions. It is destined to be used in all products, services and projects of the European Board of Medical Assessors (EBMA).

***

EBMA is a foundation created by a group of European professionals who have expertise in assessment and/or have leadership roles in universities, or other bodies concerned with medical education and training. Our mission is to promote the quality of the healthcare workforce by providing a series of assessment programmes for individuals and health education institutions. Our vision consists of enhancing public trust and confidence in the healthcare systems in Europe by contributing to the quality of healthcare services provided to patients

***

Medical teachers and other stakeholders are encouraged to make use of these guideline for their own purposes and send us comments and suggestions for improvement.

***

***

The current version of the text has been written by Carlos Fernando Collares, M.D., M.Sc., Ph.D., FACMT, in November 2016, and it has been approved by EBMA Board of Directors in its current form.

***

To obtain more information about EBMA, please visit our website www.ebma.eu or send an e-mail to info@ebma.eu.com

EBMA
P.O. Box 616
6200 MD Maastricht, The Netherlands

INTRODUCTION

There is an unsolved controversy about the appropriateness of using multiple-choice items (multiple-choice questions or MCQs) on medical knowledge tests. Other item response formats have been widely used as well, but none of them has surpassed the popularity of MCQs, which could be attributed to more feasible correction when the number of test takers is high, particularly when compared to open essay questions.

One can argue that the cueing effect caused by recognizing the correct answer among the presented alternatives on MCQs would not occur on an open-ended question asking the most likely diagnosis for a clinical scenario, for example. Nevertheless, research has shown that it is the content of the item stem, i.e. the stimulus, that will mainly determine what is measured, not the item response format, oppositely to what many may believe (Ward, 1982; Swanson, Norcini, & Grosso, 1987). Furthermore, it is noteworthy that different types of item stimuli will elicit different types of cognitive processes (Skakun, Maguire, & Cook, 1994; Schuwirth et al., 2001; Schuwirth & Van der Vleuten, 2004). Scenario-based items tend to measure more clinical reasoning while other types of stimuli tend to measure retrieval of memorized information (Schuwirth et al., 2001; Van der Vleuten et al., 2010).

Scenario-based items offer greater professional authenticity in comparison with "recall items," and thus also expectedly higher degrees of validity. Therefore, one can also expect that greater care must be applied into writing appropriate stimuli rather than the responses (Van der Vleuten et al., 2010).

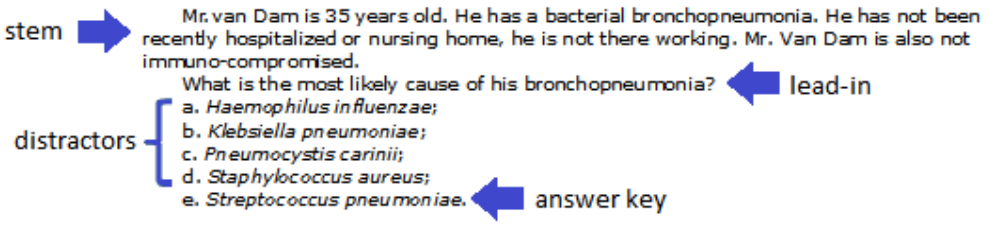
In this short document, we provide concise guidelines for writing relevant, professionally authentic MCQs destined to compose the assessment products and services created and offered by the European Board of Medical Assessors. We hope these guidelines will also be useful for medical teachers not only in Europe but also worldwide.

PRACTICE POINTS

- The number of alternatives should be determined by the educational need.
- Write stems with focused lead-in questions
- Write parallel alternatives regarding length and content, preferably shorter than the stem
- Write positive lead-in questions
- Avoid repeated elements in the alternatives
- Avoid vague and/or absolute frequency terms
- Make sure the alternatives do not overlap
- Do not use alternatives that aggregate other alternatives (e.g. "all/none of the above")
- Make sure the lead-in question and all the alternatives are grammatically and logically congruent
- Avoid absolute and personal wording on the lead-in questions
- Do not repeat words from the stem in the alternatives
- Prefer to measure one attribute at a time
- Aim for relevant, professionally authentic items

COMPONENTS OF A MULTIPLE-CHOICE QUESTION

Ideally, a multiple choice question is composed of a ***stem*** in which the context of the item is described so that a question can be asked. The question that follows the stem is called ***lead-in***. The **options** include a correct alternative (the ***answer key***) and the false alternatives, the ***distractors***. In the item below you can see the parts of an MCQ item:



RECOMMENDATIONS

This section contains recommendations for writing multiple choice questions. Each recommendation can be followed by explanations, examples and suggestions for improvement.

The number of alternatives should be determined by the educational need

Although it is quite frequent that in MCQ-based tests all items have the same number of options (usually four or five), this is not necessary. One might be tempted to think that from a purely psychometric perspective, 4 or 5 alternatives would be superior regarding discrimination.

However, the literature suggests otherwise (Tarrant & Ware, 2010). In fact, certain alternatives have such a high level of implausibility that subsequent analyses reveal them to be non-functional distractors, i.e. distractors with a low percentage of selection (Rodriguez, Kettler, & Elliott, 2014).

Three-alternative items do not have a significant negative impact on **discrimination**, i.e. *the ability of the item to distinguish examinees with superior performance from the underachieving ones*, in comparison to four- or five-alternative items. Concurrently, a significant negative effect of non-functioning distractors on validity and reliability has been shown to occur (Ali & Ruit, 2005; Ali, Carr, & Ruit, 2016).

Therefore, item writers should only use the number of alternatives that are educationally needed, avoiding the creation of "filler alternatives" just for the sake of uniformity with the rest of the test, or to curb successful guessing. In other words, if the MCQ has only three alternatives but all the distractors are functional, this would most likely be a better item than a four- or five-alternative MCQ in which all or most distractors are non-functional.

Write stems with focused lead-in questions

When writing an MCQ, it is a helpful strategy first to conceive the item as a stem ending with an open lead-in question which requires an open short answer. This strategy supports the creation of items focused on the comprehension of a concept or process, or even on the resolution of a problem.

When examinees with enough knowledge can answer an item without looking at the alternatives, it is possible to affirm that the item has a focused stem and lead-in question. If an examinee cannot answer an item without reading the alternatives, this item is considered to be "unfocused" or "stemless" (Case & Swanson, 2000). One can quickly recognize a stemless item as they usually have the lead-in questions like:

- Which of the following alternatives is (in)correct?
- From the alternatives below, which one is (in)correct? or
- About (a certain concept/process/condition), what is it true to affirm?

Although the usage of a question mark is no guarantee of a focused lead-in, it is always a good advice to write the lead-ins using a question mark (?) instead of a colon (:). Sometimes lead-ins ending in a colon might seem to be focused enough, but a closer look might reveal "partial stemlessness." This extra step helps you ensure that knowledgeable examinees can answer the items without looking at the alternatives. A stemless item can be seen on the following example:

Which of the following is correct about rheumatoid arthritis?
A. Is a disease of the articular cartilage
B. Occurs more often in women
C. The spine and elbows are involved

One of the main problems with an item being stemless is that such "stemlessness" is a sort of "Pandora's box" that opens the door for the occurrence of the majority of the other technical item writing flaws. Considering the effect of item writing flaws on item difficulty and item discrimination, and thus fairness and validity (Rush, Rankin, & White, 2016), one

must give paramount importance to writing focused stems as well as avoiding other item writing flaws.

More importantly, stemless items tend to invariably consist of "recall" items, focusing on retrieval of memorized facts rather than the application of knowledge (e.g. clinical reasoning, comprehension of processes). Therefore, stemless items will likely be less professionally authentic than items with focused stems and, consequently, one can infer that validity based on the consequences of testing would likely be diminished.

Write parallel alternatives regarding length and content, preferably shorter than the stem

Even though the correct alternative must be unequivocally right and the distractors must be absolutely wrong, all options should be preferably similar in terms of length, degree of complexity and directed to the same aspect or attribute. *Avoid long, heterogeneous alternatives*. Students without sufficient knowledge might be aware that the correct answer is usually the longest, more elaborate alternative.

Good items have stems that are longer than the alternatives. If the stem is shorter than the alternatives, then the item is likely stemless. See the following example of a stemless item with heterogeneous alternatives in which the answer key is the longest option, with the most "sophisticated" wording:

Regarding cardiovascular risk management:
A. Only medication can reduce the risk of cardiovascular events.
B. Life style modification and environmental control is very important before any attempt to start with medication.**
C. In patients with metabolic syndrome only blood pressure control is important.
D. Dyslipidemias have no role in cardiovascular risk.

Write positive lead-in questions

Negative stems tend to be also stemless. The problem with negative stems derive not only from the commonly associated "stemlessness" but also from the fact that by choosing a negative stem, there is a certain loss in the property of the item in making the "diagnosis" of the examinee's misconceptions. The more implausible the choice was of the student, the more likely it is for him/her to have a high degree of misconception about that particular content.

Conversely, a wrong answer to a more plausible distractor indicates a partial level of knowledge or comprehension of the topic. *If all distractors are correct, the diagnostic function of the distractors is therefore lost*. See below an example of a stemless negative item:

Regarding lacunar infarcts what is NOT correct:
A. Has insidious onset with deficit progression over 2 to 4 days.
B. Relatively typical distribution.
C. Associated with convulsions.*

Negative lead-in questions that deal with crucial, professionally authentic aspects, such as contraindicated drugs and procedures, for example, may deserve an exception. In these situations, we recommend usage of capital letters for the negative part of the stem and the restriction of the length of the alternatives to only one word, if possible. See the following example of an acceptable negative stem:

Which one of the drugs below is NOT recommended in a patient with sinus tachycardia?
A. Atropine*
B. Captopril
C. Losartan
D. Propranolol

Avoid repeated elements in the alternatives

Options with repeated elements lead to confusion, unnecessary workload as well as a higher probability of a successful guess due to a "logical convergence strategy" by the test-wise examinee (Case & Swanson, 2000), leading to the elimination of the least probable alternatives and a higher likelihood of a correct answer due to guessing. Consider the following item:

A 35-year-old man is admitted to the intensive care unit with a severe septic shock. In the first 6 hours after admission, certain parameters are checked and found to deviate from those of a healthy person. Consider these findings:
I. Increased blood pressure
II. Reduced urine production
III. Decreased pulse rate
IV. Warm extremities

Which of the findings above you expect to find in this patient?
A. II and IV
B. I, II and III
C. I, III and IV
D. II, III and IV*
(* = answer key)

The test-wise examinee will look for the elements that appear more frequently. Element I appears twice; II, III and IV appear three times. Using the convergence strategy, the test-wise examinee without knowledge about the topic will quickly exclude B and would likely exclude C as well – even though element I (increased blood pressure) is a highly implausible finding in the patient described in the scenario and quickly eliminated also by the examinee with

sufficient or even partial knowledge. The probabilities of a successful answer will likely similar in all examinees regardless their proficiency levels – something that would be evident if the discrimination of these items is analyzed.

Avoid vague and/or absolute frequency terms

Vague frequency terms such as "commonly", "generally", "usually", "rarely" and "occasionally" should be avoided as they might become targeted by judicial litigations due to their ill-defined nature. On the other hand, absolute frequency terms such as "all", "none", "always" and "never" should be avoided as examinees tend to quickly exclude alternatives which contain them. As the old popular adage among doctors says: "In medicine as in love, no never, no always".

Make sure the alternatives do not overlap

Sometimes item writers might not be aware of implicit overlaps that might occur in their alternatives. This might occur in alternatives containing numbers but also in non-numerical alternatives, such as anatomical locations. See examples of overlapping alternatives in the two following items:

What is the most likely probability of a spontaneous miscarriage in a secundigravida who has also had a spontaneous miscarriage in her first pregnancy?

A. 5%
B. 15%
C. More than 25%
D. 30%
E. 40%

Where must the perforation be done in a needle cricothyrotomy?

A. Below the thyroid cartilage
B. Cricoid cartilage
C. Cricothyroid membrane

While in the first example the alternatives C, D and E clearly overlap, in the second example the detection of the overlap depends on not only attention but also some anatomical knowledge. Both the cricoid cartilage and the cricothyroid membrane are below the thyroid cartilage. In case of a complaint against these items, more than one alternative would eventually be considered correct. In other words, make sure the alternatives are mutually exclusive.

Do not use alternatives that aggregate other alternatives

Not only the options such as "all of the above" or “none of the above” commonly present logical fallacies as they eventually might lead to content overlap, but they also undermine the cognitive diagnostic function of the item. The teacher might have a difficult time to detect the misconceptions that caused the examinee to have a wrong answer. Furthermore, such alternatives are usually distractors and the test-wise examinee without knowledge is aware of that. Therefore, it is common for such items to have less discriminative power.

Make sure the lead-in question and all the alternatives are grammatically and logically congruent

Students will inevitably use grammatical mismatches or logical inconsistencies between the lead-in question and the alternatives to guess the correct answer even if they do not know it. Grammatical and logical connection problems will provide hints for examinees without knowledge. See the following example:

What tissue is the main target of asbestos?
A. Kidneys
B. Liver
C. Pleura
D. Adrenals

In this item, alternatives A and D could be easily discarded by the test-wise examinee due to the fact that they are in the plural form while the lead-in is in the singular. Moreover, options A, B and D are actually organs and not tissues.

It is evident that not only the alternatives are not parallel in terms of content, but the logical, grammatical incongruences will allow the examinee without knowledge to quickly identify alternative C as the correct answer. This is a good example of how easily the technical item flaws can coexist.

Avoid absolute and personal wording on the lead-ins

Although this specific advice will likely not impact item discrimination, it may prevent justifiable complaints and judicial litigations. Instead of “What is the correct diagnosis?”, prefer “Which diagnostic hypothesis is most likely correct?” or simply “What is the most likely diagnosis?”.

Likewise, never ask personal lead-ins such as “Which treatment would you choose?”, for example. In this type of question, all alternatives could be considered correct, as the question is focused on the examinees’ opinion and not the most appropriate answer.

Do not repeat words from the stem in the alternatives

Items that involuntarily provide a tip to examinees due to word(s) of the stem being present in the correct alternative are more common than many test makers think. Such items are prone to be less discriminative. See the following example:

Infection of the orbit is a rare but serious complication of acute sinusitis. From what route will the infection usually spreads to the orbit?
A. Through the roof of the maxillary sinus.
B. Through the bottom of the sinus frontalis.
C. Through the lamina orbital / papyracea of etmoïd.*
D. Through the lateral wall of the sfenoïd.

In this item, not only the correct alternative contains a word that derives from a key word present in the stem (“orbit/orbital”), but it also has heterogeneous alternatives, with the correct answer being the longest one together with an extra explanation, which the other alternatives do not have. This can be considered another example of an item with more than one technical item writing flaw.

Prefer to measure one attribute at a time

If more than one cognitive step is needed to achieve a correct answer on an item, in case of a wrong answer, part of the diagnostic purpose of the item will be lost. For instance, when a scenario-based item does not inform the examinee about the diagnosis of the patient and presents a lead-in question about the appropriate treatment, in case of a wrong answer, there is no possibility for the teacher to identify where the knowledge gap was. The following item is illustrative:

A young, mentally confused patient is admitted in the emergency room with bradycardia, hypotension, sialorrhea, bronchorrhea, diarrhea, vomiting and miosis. What is the most appropriate treatment?
A. Acetylcysteine
B. Atropine
C. Flumazenil
D. Naloxone

Technically, one might not consider the 2-step approach illustrated above a technical item writing flaw, as it is not associated with lower discrimination. One might even argue that such items have a high degree of professional authenticity and might be purposefully used to increase the difficulty of a test with a skewed distribution towards easier items. However, since 2-step items are inherently less "diagnostic", we recommend test assemblers to pay special attention to the relative amount of 2-step items when composing a test.

Aim for relevant, professionally authentic items.

A good advice to write good items is to reflect upon the educational purposes of the item. After the item is written, remember the learning goals associated with it and verify whether the item in its current form is enough fit for its originally intended purposes.

Another good tip is to verify whether the lead-in question can be answered without the clinical scenario from the stem. If the answer is yes, the presence of a scenario is irrelevant and the item cannot be really considered scenario-based.

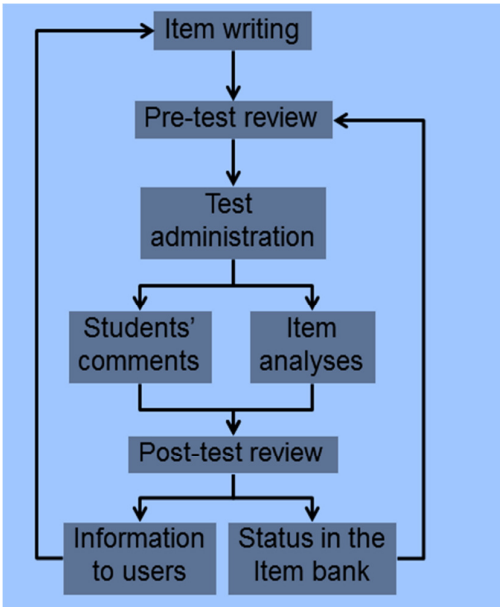
Give preference to relevant concepts, processes and conditions. Test items are not meant to be "tricky" or to cover knowledge about footnotes. Use cases from real-life daily practice as inspiration for scenario-based items. Make sure you offer a clear, sufficient description of the chosen case and the setting where it happens (Schuwirth *et al*., 1999; Schuwirth and Pearce, 2014). We suggest the use of a framework for item relevance as part of the quality management of the test. An example of item relevance framework that we consider quite useful is described on the next section of this paper.

QUALITY MANAGEMENT ACROSS THE ENTIRE TESTING CYCLE

Good quality assurance around test construction is of paramount importance to the validity of test scores. Interdisciplinary review committees are a key component in the management of the quality of the items throughout the entire test cycle.

Item writers with different backgrounds, and coming from different institutions, send their items to the International Progress Test Review Committee. The Committee, composed of medical teachers with broad interdisciplinary expertise, checks the appropriateness of the received items regarding content, language, relevance and absence of item writing flaws, according to the recommended criteria detailed in the following sections of this text. Item writers might receive questions as well as feedback about the items sent to the Committee.

The following graph presents a flowchart in which helps understand the different roles of the International Progress Test Review Committee, which we will use here, as an example of a useful quality management strategy:



After the test is delivered and the data is processed, a post-test review meeting takes place. In this meeting the Committee appreciates the formal complaints and comments by examinees, if there are any. During the meeting, the psychometric analyses are also taken into account to make decisions about answer key changes or item withdrawals. Item parameters such as difficulty and discrimination are used in the process. If the discrimination values are low or if any of the distractors has a higher discrimination than the answer key, the item needs to be reviewed regardless of students' complaints. The same happens with items that have non-invariant parameters across different institutions or countries or even in different tracks within the same institution.

When we speak about non-invariant parameters we are talking about difficulty or discrimination values that are not stable across different groups of examinees. This is important as it can be considered a sign that something else is being measured besides the intended attribute, which, in our case, is "application of medical knowledge". The occurrence of parameter non-invariance can be related to translation problems, differences in protocols and guidelines, differences in legal/ethical codes, differences in therapeutical styles (either more aggressive or more conservative) or even the effect of different curricula. The importance of detecting non-invariant parameters is that they suggest the occurrence of

construct-irrelevant sources of score variance and, as such, they represent a threat to the validity of the test.

If the decision is to retain the item, it can go back to the item bank for review and eventually usage in a subsequent test. Certain items can be rewritten and used again in a new version but some items need to be completely retired. Students receive information about the decisions taken in the post-test review meeting. The Committee might provide feedback to the item writers about such decisions as well.

Schuwirth and Pearce (2014) present a framework for item relevance that we consider quite useful to keep our tests with a high degree of meaningfulness. The framework consists of five criteria ("medical knowledge", "ready knowledge", "incidence in practice", "prevalence or high-risk" and "knowledge foundations in the medical curriculum" and for each criteria, the items can be classified into three categories: "not relevant", "somewhat relevant" and "very relevant". See the following table with the descriptors for the degrees of relevance in the five proposed criteria:

	NOT RELEVANT	SOMEWHAT RELEVANT	VERY RELEVANT
Medical knowledge	Knowledge is an element that is not necessarily specific to a doctor; the baker on the corner knows the answer.	Knowledge is specific to medicine but also known to the interested layperson.	Knowledge is specifically for medicine and requires a proper study and understanding of the subject.
Ready knowledge	The knowledge is not easily recalled but is easy to find. Even specialists in practice cannot remember it.	The knowledge is easy to find, but should be typically recalled when confronted with it in practice.	Any medical doctor has this knowledge at the ready at any time of day. It is a prerequisite for functioning in a practical situation.
Incidence in practice	There is no medical situation (not necessarily clinical) in which this knowledge is important.	While there are medical situations in which this knowledge is important, these situations are not frequent.	This knowledge is important for many practical situations.
Prevalence or high-risk	The knowledge is usually only found in highly specialised centres, is low risk or is rarely found.	The knowledge is found in high-prevalence or high-risk situations in practice, but is not essential for successfully handling the situation.	The knowledge is found in high-prevalence or high-risk situations in practice, and is essential for successfully handling the situation.
Knowledge foundations in the medical curriculum	The knowledge is a fact or an isolated event and is not required for building other concepts in the curriculum.	The knowledge is needed to further understand concepts but the specific knowledge may itself be forgotten (e.g., the Bohr/Haldane effect to understand why haemoglobin releases oxygen into the tissues in the lung).	The knowledge forms the basis for one or more other concepts in the curriculum and it should remain known as explicit knowledge (e.g., Frank-Starling mechanism as a basis for congestive heart failure).

Extracted from Schuwirth and Pearce (2014), with permission from the first author.

A FINAL MESSAGE

This is not an exhaustive guideline. Item writing guidelines with a greater level of detail can be found in Case and Swanson (2000), Haladyna, Downing and Rodriguez (2002), and, more recently, in Schuwirth and Pearce (2014), who gently provided some of the examples used in this document as well as the item relevance framework. Nevertheless, with this succinct document we expect to have provided a simple and feasible framework for quality on item writing and reviewing, ultimately aiming to offer a positive contribution towards better education and assessment.

References

Ali, S. H., & Ruit, K. G. (2015). The Impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. *Perspectives on medical education*, *4*(5), 244-251.

Ali, S. H., Carr, P. A., & Ruit, K. G. (2016). Validity and Reliability of Scores Obtained on Multiple-Choice Questions: Why Functioning Distractors Matter. *Journal of the Scholarship of Teaching and Learning*, *16*(1), 1-14.

Case, S. M., & Swanson, D. B. (2000). *Constructing written test questions for the basic and clinical sciences* (3rd ed). Philadelphia, PA: National Board of Medical Examiners.

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), 309-333.

Rodriguez, M. C., Kettler, R. J., & Elliott, S. N. (2014). Distractor functioning in modified items for test accessibility. *SAGE Open*, *4*(4), 2158244014553586.

Tarrant, M., & Ware, J. (2010). A comparison of the psychometric properties of three-and four-option multiple-choice questions in nursing assessments. *Nurse education today*, *30*(6), 539-543.

Schuwirth, L.W.T., Blackmore, D.B., Mom, E., Van den Wildenberg, F., Stoffers, H., & van der Vleuten, C.P.M. (1999). How to write short cases for assessing problem-solving skills. *Medical Teacher*, *21*(2), 144–150.

Schuwirth, L., & Pearce, J. (2014). Determining the Quality of Assessment Items in Collaborations: Aspects to Discuss to Reach Agreement. Camberwell, VIC: Australian Medical Assessment Collaboration.

Schuwirth, L. W., & Van Der Vleuten, C. P. (2004). Different written assessment methods: what can be said about their strengths and weaknesses? *Medical education*, *38*(9), 974-979.

Schuwirth, L. W., Verheggen, M. M., Van der Vleuten, C. P. M., Boshuizen, H. P. A., & Dinant, G. J. (2001). Do short cases elicit different thinking processes than factual knowledge questions do? *Medical Education*, *35*(4), 348-356.

Skakun, E. N., Maguire, T. O., & Cook, D. A. (1994). Strategy choices in multiple-choice items. *Academic Medicine*, *69*(10), S7-9.

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: Written and computer-based simulations. *Assessment and Evaluation in Higher Education*, *12*(3), 220-246.

Van der Vleuten, C. P. M., Schuwirth, L. W. T., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Best Practice & Research Clinical Obstetrics & Gynaecology*, *24*(6), 703-719.

Verhoeven, B. H., Verwijnen, G. M., Scherpbier, A. J. J. A., & Schuwirth, L. W. T. (1999). Quality assurance in test construction: The approach of a multidisciplinary central test committee/Commentary. *Education for Health*, *12*(1), 49.

Ward, W. C. (1982). A comparison of free-response and multiple-choice forms of verbal aptitude tests. *Applied Psychological Measurement*, *6*(1), 1-11.