

Are multiple-choice questions a good tool for the assessment of clinical competence in Internal Medicine?

Flavio Tangianu,¹ Antonino Mazzone,² Franco Berti,³ Giuliano Pinna,⁴ Irene Bortolotti,⁵ Fabrizio Colombo,⁶ Carlo Nozzoli,⁷ Micaela La Regina,⁸ Antonio Greco,⁹ Cristina Filannino,¹⁰ Mauro Silingardi,¹¹ Roberto Nardi¹¹

¹Internal Medicine, S. Martino Hospital, Oristano; ²Medical Department, Internal Medicine, ASST Ovest-Milanese, Legnano (MI); ³Internal Medicine, S. Camillo Forlanini, Roma; ⁴Internal Medicine, Asti; ⁵Managing Director-Planning Congressi, Bologna; ⁶Internal Medicine, Niguarda Ca' Granda Hospital, Milano; ⁷Internal Medicine and Emergency, Careggi Hospital, Firenze; ⁸Internal Medicine, S.S. Risk Management, Presidio Ospedaliero Unico del Levante Ligure, La Spezia; ⁹Geriatric Department, Casa Sollievo della Sofferenza, San Giovanni Rotondo (FG); ¹⁰Department of Government, Health and not for profit, Luigi Bocconi University, Milano; ¹¹Internal Medicine, Bologna, Italy

ABSTRACT

There are many feasible tools for the assessment of clinical practice, but there is a wide consensus on the fact that the simultaneous use of several different methods could be strategic for a comprehensive overall judgment of clinical competence. Multiple-choice questions (MCQs) are a well-established reliable method of assessing knowledge. Constructing effective MCQ tests and items requires scrupulous care in the design, review and validation stages. Creating high-quality multiple-choice questions requires a very deep experience, knowledge and large amount of time. Hereby, after reviewing their construction, strengths and limitations, we debate their completeness for the assessment of professional competence.

*It is important to ask the right question
at the right time in the right way
and then listen for the answer*

Introduction

There are many feasible tools for the assessment of clinical practice, but there is a wide consensus on the fact that including in this process the simultaneous use of several different methods could be strategic for a comprehensive overall judgment of clinical competence. Multiple-choice questions (MCQs) are a

well-established reliable method for knowledge assessment,¹ increasingly used in postgraduate exams, owing to their higher validity and easy scoring.^{2,3} Using MCQs could be the first step for the assessment of clinical competence.⁴

The type *Multiple choice, best answer* has been widely used as an assessment tool in continuous medical education.⁵ If appropriately constructed, MCQs could be efficient, objective, discriminative in combination with other tools to get a comprehensive competence assessment strategy.⁶ Cognitive knowledge assessed by MCQs is well related to overall competence and performance, but examinees and examiners alike often perceive this tool as *unfair*.⁷ MCQs could be designed to evaluate the extent of the candidate's knowledge and clinical judgment in the areas in which an internist should demonstrate a high level of competence and thinking but, likewise other methods of assessment, they have some advantages and limitations, requiring a high level of discriminating judgment.⁸⁻¹⁰

Correspondence: Roberto Nardi, Internal Medicine, Bologna, Italy.

E-mail: nardidoc48@gmail.com

Key words: Multiple-choice questions; clinical competence; pass mark criteria.

Conflict of interest: the authors declare no potential conflict of interest.

Received for publication: 7 January 2018.

Accepted for publication: 27 March 2018.

This work is licensed under a Creative Commons Attribution NonCommercial 4.0 License (CC BY-NC 4.0).

©Copyright F. Tangianu et al., 2018

Licensee PAGEPress, Italy

Italian Journal of Medicine 2018; 12:88-96

doi:10.4081/ijm.2018.980

Writing multiple-choice questions according to the Bloom's taxonomy levels of cognition

Higher-level thinking refers to the own internist's ability to look at increasingly complex cognitive skills, from the most basic knowledge - at the bottom - to

evaluation - at the top.¹¹ The 2001 corresponding revised Bloom's taxonomy levels of cognition included remembering, understanding, applying, analyzing, evaluating and creating¹² (Figure 1). According to Bloom's taxonomy, higher level of thinking requires some specific skills¹³ (Table 1).¹⁴

How to structure the best multiple-choice questions?

Structuring multiple choice items in order to test higher order thinking, we should have to purpose problems requiring multi-logical thinking, defined as *thinking that requires knowledge of more than one fact to logically and systematically apply concepts to a... problem*.⁹ As general guidelines, before formulating an assessment strategy, you should define the scope of the examination. Cognitive levels being tested should be decided, as the format of MCQ may depend on that. A sufficient number of items is required for a valid MCQ examination that covers a specific topic; probably 60-100 items are optimal for an examination of 60-90 min duration.¹⁵ MCQs include three components: i) the *stem* (the question or scenario); ii) the *distractors* (the wrong options); iii) the *key* (the right answer).

Constructing an effective stem

An MCQ is composed of two parts. The first is the problem, known as the stem. It poses the question, with a list of suggested solutions, as options or possible answers. The answers/options contain a key that is the best answer to the question, with some distractors that are plausible - but incorrect - answers to the question.¹⁰ The stem is not always a question in the strict sense of the word: it could be presented as a brief scenario or as a case study. It should be meaningful by itself, presenting a definite problem, without containing irrelevant material which could decrease the reliability and the validity of the test.¹⁶ When creating the item, express the full problem in the stem; phrase your questions as simply as possible. Ask yourself if the candidates would be able to answer the question without looking at the options. In order to make options easier to read, understand and answer the question quickly, put all relevant information in the stem, without repeating it in each of the options that could be included. Eliminate excessive wording and irrelevant information from the stem: they could be confounding and leading to waste of time.¹⁷ It is better to avoid negative statements or questions, such as which is not correct, because negative phrasing could be a source of misunderstanding.¹⁸ Negative questions are alike statements as which of the following is NOT true, and the majority of testing

experts recommend against using them. Avoid incomplete statement format: the stem would be presented in the form of a clear and complete question, focusing on getting a clear answer rather than holding a partial sentence.¹⁹

The alternatives: distractors and the key (the *right-best* answer)

The options consist of one correct (*the key*) or best alternative, which is the right answer. The incorrect alternatives serve as distractors, but they must be plausible. Distractors are inferior alternatives, *i.e.* incorrect answers in the list of options. It is important to limit the number of alternatives. Three-choice items are about as effective as four or five-choice items, mainly because it is difficult to come up with plausible distractors. Implausible alternatives should not be used. Alternatives should be stated clearly and concisely, mutually exclusive, avoiding overlapping answers. They have to be presented in a logic order (alphabetical or numerical), homogenous in content, grammatically consistent with the stem, using simple, precise and unambiguous wording, congruent in form, similar in length. Language must be similar in all items, and the placement of the correct answer balanced. Make sure there is only one best answer. Avoid *easy (non-functioning)*²⁰ distractors, as well as having two or more options that are correct. In placing the issues, you have to instruct all participating doctors

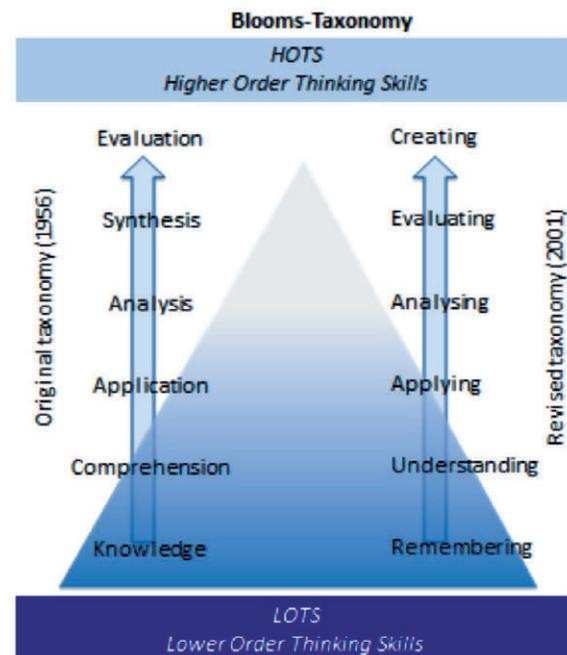


Figure 1. Bloom's level of taxonomy (revised by Anderson): from lower to higher thinking skills.

to select the *best answer* rather than the *correct answer*. Avoid *categoric* words, such as *never* or *always*. Use familiar language, avoiding trick questions and negative wording. The alternatives *all of the above* and *none of the above* should not be used. Complex multiple choice with different combinations of options in the answer and excessively wordy items should be avoided. According to the goal of testing the

higher-order thinking, alternatives should be selected among those requiring a high level of discrimination. Finally, we recommend to use only one correct option, being sure that the best option is clearly the best option, remembering that: i) the item should include one and only one correct or clearly best answer; ii) with one correct answer, alternatives should be mutually exclusive and not overlapping; iii) using MC

Table 1. Skills requested for higher levels of thinking according to Bloom's taxonomy.

Taxonomy level 1956	Representative words	Sample behavior <i>Example: CHF and COPD</i>	Taxonomy level 2001	Key terms
<i>Evaluation</i>	Critique Summarize	Assess the effectiveness of the protocol	<i>Creating</i>	Designing, constructing, planning, producing, inventing, devising, making, programming, filming, animating, blogging, video blogging, mixing, remixing, wiki-ing, publishing, videocasting, podcasting, directing/producing, creating or building mash ups
<i>Synthesis</i>	Organize Design	Develop a new protocol for treating CHF and COPD	<i>Evaluating</i>	Checking, hypothesizing, critiquing, experimenting, judging, testing, detecting, monitoring (blog/vlog) commenting, reviewing, posting, moderating, collaborating, networking, reflecting, (α and β) testing
<i>Analysis</i>	Compare Categorize	Compare and contrast progression of CHF and COPD Determine if a patient has a CHF or COPD	<i>Analyzing</i>	Comparing, organizing, deconstructing, attributing, outlining, finding, structuring, integrating, mashing, linking, reverse-engineering, cracking, mind-mapping, validating, tagging
<i>Application</i>	Organize Solve	Describe the standard process for determining if a patient has a CHF or COPD	<i>Applying</i>	Implementing, carrying out, using, executing, running, loading, playing, operating, hacking, uploading, sharing, editing
<i>Comprehension</i>	Distinguish Match	Match symptoms with their associated ailments	<i>Understanding</i>	Interpreting, summarizing, inferring, paraphrasing, classifying, comparing, explaining, exemplifying, advanced searching, Boolean searching, blog journaling, twittering, categorizing and tagging, commenting, annotating, subscribing
<i>Knowledge</i>	Identify Label	Identify three symptoms of CHF and COPD	<i>Remembering</i>	Recognizing, listing, describing, identifying, retrieving, naming, locating, finding, bullet pointing, highlighting, bookmarking, social networking, social bookmarking, favorite-ing/ local bookmarking, searching, Googling

CHF, chronic heart failure; COPD, chronic obstructive pulmonary disease. *Modified from Churches, 2010.*¹⁴

with questions containing more than one right answer lowers discrimination between candidates.²¹ In Table 2 are reported the main issues for writing MCQs.²²

Qualitative item analysis

Properly constructed MCQs are very useful if consonant with defined analysis tools in their assessment.²³ Item analysis includes qualitative and quantitative criteria. Qualitative analysis looks at whether the content matches the information, attitude,

characteristic or behavior being assessed. A quick checklist for questions is reported in Table 3.²⁴

Quantitative item evaluation: validation is an important step

Quantitative evaluation implies the measure of item difficulty, as the percentage of participants who get that item correct: its score can range from 0 to 100%, with low value ($\leq 30\%$) corresponding to high difficulty and high value ($\geq 80\%$) to low difficulty.²⁵

Table 2. Tips for writing multiple-choice questions.

General rules

- 1) You have to be sure that each item reflects a clearly defined field of assessment
- 2) Choose items of appropriate and different levels of difficulty
- 3) Have the test reviewed by someone who can find mistakes, clues, grammar and punctuation problems before you administer the exam to candidates
- 4) Consider the time required for examination: different item formats require different time duration (single response - 40 s, whereas case history type - 60 to 90 s)

Stem

- The stem of the item should be written in clear, concise, unambiguous and precise language
- Aim to write as a complete sentence
- Avoid *trigger* words (e.g. pin-rolling tremor)
- Avoid use of unnecessary content
- Avoid the use of negatives, such as NOT or EXCEPT; if used, be sure that the word appears capitalized and boldface. Avoid double negatives.
- Avoid exceptions, absolutes and qualifiers in question stems
- Avoid vague expressions like fairly high, considerably greater, etc.
- Avoid clues suggestive of the right answer
- Avoid veering away from the content/cognitive level that you aimed to address

Responses: correct answer (*the key*) and distractors

- Keep choices homogeneous in content (distractors are in the same category as the correct answer)
- Make all distractors plausible
- Avoid repeating the same content in the options - move it to the stem if possible
- Provide a sufficient number of distractors
- Use words for distractors in a similar way to the correct answer, keeping to similar length
- All answers should be plausible and homogenous. Items need to be independent of one another
- Answer choices should be similar in length and grammatical form
- List answer choices in alphabetical or numerical order
- Avoid *all of the above/none of the above* as a response
- Avoid negatives such as NOT
- Avoid the use of specific determiners such as always, never, completely, and absolutely
- Avoid technical flaws (tense or plurality for example)
- A correct statement but not the answer to the question is a good distractor

Table 3. A quick checklist for questions.

Does the question?	Does the question avoid?
Test the: <i>knowledge</i> <i>comprehension</i> <i>application</i> <i>analysis</i> <i>evaluation</i> <i>synthesis</i>	Repeating text in the question and the answer
Test the level that was intended	Double negatives
Have a clearly worded stem	Use of <i>all of the above/none of the above</i>
Have only one clearly correct answer	Giving clues in the question
Have a range of a/b/c/d ... correct answers	
Consider inclusivity	
Have a distractor that is similar in wording and length to the correct one	
Have plausible distractor	

In the validation process we have to define our goal standard in formulation of MCQs, mostly as validity and reliability. For validity the test must measure what the examinees are expected to know. This is accomplished by writing test questions that align with the objectives. Reliability deals with the ability of a test to measure a test-taker's ability consistently. Exams may have different levels of difficulty, owing to the fact that questions are randomly pulled from a test bank. Several tools have been performed to increase MCQ validity.²⁵ The quality of MCQs is determined by three parameters such as difficulty index, discrimination index, and distractor efficiency. Their definitions and details, are reported in Table 4.²⁶

Fixation of pass level in multiple-choice questions examination: the *pass mark* concept

Fixation of pass level in MCQ examination is more difficult than in other formats. Though several formats are available, MCQs function much better for the purpose of ranking. A pass mark is a special score that serves as boundary between those who perform well enough and those who do not²⁷ *absolute* pass mark. The purpose in MCQs examination is to select the candidates able to perform well enough (*pass*) and to eliminate the group of candidates that do not perform well enough (*fail*). In order to achieve this goal, a (limited) number of questions are presented to the candidates. The discriminative power of the examination will depend on the *validity* of the questions used. The validity of the test is influenced by two factors: i) the *degree of difficulty* of questions;

it can be assessed by calculating the P-value (*i.e.* percentage of candidates answering correctly); if all candidates answer the item correctly the P-value is 1.00 and if nobody answers the question correctly the P-value is 0.00: as a rule of thumb, ranges between 0.25-0.30 and 0.70-0.75 are used, avoiding questions with P-value above 0.90 or below 0.10;²⁸ ii) the *degree of discriminative power* of questions, as the objective measurement of the degree to which the question is able to discriminate strong from weak candidates. It can be assessed by calculating the Rit value (item-total correlation or discrimination/point bi-serial): it is the correlation between the question score to total examination score. In other terms whether the item was answered correctly mainly by those candidates who also had a high score on the test (and incorrectly by those with a low total test score). As a thumb rule you have to avoid questions with Rit-value below 0.20. Types of pass marks are divided in:

- *absolute* pass mark: i) expressed as a number (*e.g.* 70 correct responses) of test questions; ii) expressed as a percentage (*e.g.* 70% correct responses) of test questions; iii) no influence of the caliber of the total candidate population; iv) it is possible that all candidates pass with high distinction/fail; v) a fair amount of experience is demanded to set the pass marks.
- *relative* pass mark: i) expressed as a number (*e.g.* 50 best performers) of examinees; ii) expressed as a percentage (*e.g.* top 20% performers) of examinees; iii) very easy to use for examiners; iv) appropriate in large candidate populations (>40) to be sure that the candidate population is representative; v) stimulation of competition between candidates.

Table 4. Difficulty index, discrimination index and distractor effectiveness.

Parameter	Formula for calculation	Categories and cut offs	Meaning of the categories
Difficulty Index	$(H+L) \times 100 / T$	<i>Difficult</i> : Difficulty index less than 30%	Less than 30% of the candidates could answer the question correctly
		<i>Medium</i> : Difficulty index 30% to 70%	Between 30% and 70% of the candidates could answer the question correctly
		<i>Easy</i> : Difficulty index above 70%	More than 70% of the candidates could answer the question correctly.
Discrimination Index	$(H-L) \times 2 / T$	<i>Good discriminator</i> : Discrimination index more than or equal to 0.2	H is at least 10% more than L
		<i>Poor discriminator</i> : Discrimination index less than 0.2	H does not exceed L by more than 10%
Distractor effectiveness	Percentage of candidates having marked the distractor as the right answer	<i>Functional distractor</i> : distractor effectiveness more than or equal to 5%	At least 5% of students marked that distractor as the right answer
		<i>Non-functional distractor</i> : distractor effectiveness less than 5%	Less than 5% of students have marked the distractor as the right answer

H, number of high achievers who have answered the question correctly; L, number of low achievers who have answered the question correctly; T, total number of candidates considered for analysis.

The Angoff method

The Angoff method was proposed in order to set pass marks.²⁹ It is the most commonly used tool, convenient to use, well-researched, easily explained and customized, applicable to several response formats. Prior to the examination, the probability of the questions to be answered correctly by a *minimally competent candidate*, whose knowledge, skills and abilities are just enough, have to be estimated by several judges. In practice: judges assign probabilities that a hypothetical *minimally competent borderline candidate* will be able to answer each item correctly. For each judge, probabilities are summed to get a minimum performance level (MPL). MPLs are averaged to get a final passing score. Judge training and calibration are essential. It is critical, before starting the exercise, that the board of judges can develop a standard, common picture of skills, abilities and attitudes,³⁰ in order to accurately define the baseline features of the *minimally competent, borderline candidate*. A simplified example of the Angoff method for the calculation of MPLs and passing score based upon the *yes/no* predicted probability of questions to be answered correctly is reported in Table 5. Based on the averaged estimates, an arbitrary (absolute) pass mark is set prior to the examination.

Angoff cut-off score can be calculated as predicted percentage of correct answers: the mean of everyone's judgment is calculated for each item; this is often referred to as the *predicted difficulty* and each expert's judgment for an item should be the same or within a close and defined range (around 10%). Each predicted difficulty (mean) is added together and divided by the

total number of items in the exam to get the cut-off percentage. This percentage of the total marks for the exam indicates the cut-off mark. Table 6 shows how a cut-off mark may be calculated using the % Angoff method.

In conclusion, Angoff is a standard method requiring expert's judgement about how difficult each item is, by predicting the percentage of borderline candidates that would get the question correct. It is the most widely used method of standard setting, able to *reflect the difficulty of the content and focusing* the level at which candidates should be performing to meet the desired standard. On the other hand, this process is time consuming and labor intensive, as the judges must look at every test item. They have to be confident with their definition of a *borderline candidate*, not just assuming an *average candidate*. This method requires a large sample of judges for accuracy and reliability, with a wide range of different ages, genders, and levels of seniority.³¹

The evidence-based practice competence

Competence can broadly be defined as a concept that incorporates a variety of domains including knowledge, skills and attitudes.³² The evidence-based practice (EBP) competence involves making clinical decisions informed by the most relevant and valid evidence available. Understanding the principles of EBP is fundamental to provide the *best practice*.³³ Teaching EBP skills to health professionals with respect to their EBP competence could be able to increase participants' EBP knowledge and skills.³⁴ The practice of EBP consists of the following five key steps, each one of them requiring a different level of

Table 5. Predicting passing score with yes/no (1/0) calculations.

Item	Expert 1	Expert 2	Calculation of <i>passing score</i>
1	1	1	Average of MPLs = = (3+2)/2 = = 2.5 items corrected for passing the exam
2	1	0	
3	1	1	
4	0	0	
5	0	0	
MP _{E 1, 2, ...}	3	2	

Table 6. Predicting passing score as percentage: an example.

Item	Expert 1 (%)	Expert 2 (%)	Expert 3 (%)	Mean (%)
1	60	55	60	58.3
2	50	60	55	55
3	45	50	45	46.6
4	50	50	50	50
5	40	50	40	43.3
<i>Cut off percentage:</i>				50.64

knowledge and skill (*i.e.* competence):³⁵ i) *converting clinical scenarios into a structured answerable question requiring*: knowledge to construct a question using the PICO (*patient, intervention, comparison-control-comparator, outcome*) mnemonic; ii) *searching the literature to identify the best available evidence to answer the question requiring*: acquisition and application of literature searching skills across a variety of databases; iii) *critically appraising the evidence for its validity and applicability requiring*: a certain level of expertise in epidemiology and biostatistics; iv) *applying the results of the appraisal into clinical practice requiring*: ability to synthesize and communicate the results to relevant parties (*i.e.* health professionals, patients); v) *evaluation/assessment of the EBP process requiring*: the health professional to evaluate the EBP process and assess its impact within the clinical context in which it was implemented.

Some limitations in its approach have to be emphasized. First, randomized controlled trials (RCTs) do not reflect the real world of clinical practice: complex patients are often not considered in the admission/exclusion criteria in studies.³⁶ RCTs are not yet validated assessment tools in distinguishing, when applied, the effects of pre/post training in EBP. Participants are prone to experiencing recall bias, believing that their baseline ability was much poorer

than it actually was, therefore increasing their perceived improvement following the training intervention. Demonstrating EBP competence is a complex task, despite several potentially available tools.³⁷⁻⁴³ Still today, no single assessment method can adequately provide all of the necessary data to assess complete EBP competence.⁴⁴ For demonstrating EBP competence and exploring how EBP competence could be best assessed, we need further researches.⁴⁵

Strengths, advantages and limitations of multiple-choice questions

Like other methods of assessment, MCQs have their strengths and weaknesses⁴⁵⁻⁴⁷ (Table 7).

Multiple-choice questions: a possible double-cut weapon, if taken individually

Once we have defined what methodologies are meant to create MCQs, the next question is about doubts on results. For instance, how is it possible that some candidates with a brighter curriculum study go worse and beyond in their performances compared to less good competitors? And more: why candidates who studied for one year can achieve a lower score than those who worked only one month?⁴⁸ A MCQs

Table 7. Strengths and weaknesses of multiple-choice questions.

Strengths of MCQs	Limitations of MCQs
- Useful for the assessment of lower/higher-level thinking (in simple/complex vignettes and/or scenarios)	- Constructing good items/questions is difficult: Take a long time and are difficult to write Constrain creative responses from learners
- It can cover a lot of material very efficiently	
- Easy to understand	- May have more than one correct answer
- Questions are easy to mark	- Scores can be influenced by reading ability
- Time-efficient to administer	- It is frequently difficult to find plausible distractors
- Scoring - by hand or electronically - is easy, objective, and reliable	
- Questions can be scored by a computer, which makes them an attractive assessment approach for large classes	- Skill and scrutiny still required in the setting process
- Banks of questions created and can be used for reuse/sharing	- Partial knowledge not recognized in more straight forward questions
- Extended matching	- Can only be used to test the cognitive domain
- More complex with multiple questions and answers on a common theme	- Can be ineffective for assessing some types of problem solving and the ability to organize and express ideas
- The major advantage over MCQs is that the larger number of response options reduces the likelihood that the question will cue the candidates to the correct answer; they are less likely to recognize the correct answer	- Can be ineffective for assessing some types of problem solving and the ability to organize and express ideas
- They can also be used to test problem-solving skills such as clinical diagnosis or patient management	
- Reliable (if set correctly...)	- The closed-ended nature of MCQ tests makes them particularly inappropriate for assessing originality and creativity in thinking at a high cognitive level
- Marking is time efficient, no second marking required	- A possible <i>double-cut weapon</i> , if taken individually
- Rapid feedback to candidates and staff	
- Enables assessment of large range of curriculum	

MCQ, multiple-choice questions.

test implies a preparation and skills other than those required to develop any written work or to support an oral interview. In order to perform a test, knowledge alone is not enough: reasoning, choosing and using reading and resolution techniques skills are crucial in passing the test.⁴⁹ In practice, the elements to be taken into account to best deal with the test are: i) a good basic preparation; ii) an appropriate method of study; iii) an appropriate technique for reading and answering the questions; iv) an effective test management, including timing at the exam; v) an excellent *psychological* management of the exam.⁵⁰ In other words, the skills required in the approach to MCQs are more specific than those that should expect. The study methods and didactic goals differ from the school or university preparation paths to which we are accustomed. Abstract reasoning is a difficult test. It requires both logic and speed. To pass this test candidates must find the analogies between shapes and combinations of different shapes very rapidly.⁵¹

Conclusions

Factual knowledge underpins competence in all clinical skills. There is a growing need to test knowledge and skills in the assessment of professionals in hospital Internal Medicine. We reviewed literature about their strengths, advantages and limitations and we debated if they can provide a full judgment of the overall skills of professionals. Multiple-choice questions will have a part to play, but only as one component of the overall assessment of clinical competence. Constructing effective MCQ tests and items requires scrupulous care in the design, review and validation stages. Creating high quality multiple-choice questions requires a very deep experience, knowledge and large amount of time. For this reason, rather than constructing a test from scratch, it may be more efficient to search what other validated tests, already existing, can be useful to this task. MCQs are a good basic tool in assessing knowledge and reasoning, but, for a full judgment of the overall skills of professionals, they can be only the starting point. A good doctor has *to know, to know how to be and how to do it*.

References

1. Abdel-Hameed AA, Al-Faris EA, Alorainy IA, Al-Rukban MO. The criteria and analysis of good multiple choice questions in a health professional setting. *Saudi Med J* 2005;26:1505-10.
2. Case S, Swanson D. Constructing written test questions for the basic and clinical sciences. 3rd ed. The incorrect alternatives serve as distractors, but must be plausible. The incorrect alternatives serve as distractors, but must be plausible. Philadelphia: National Board of Medical Examiners; 2003.
3. Tarrant M, Ware J. A framework for improving the quality of multiple-choice assessments. *Nurse Educ* 2012;37:98-104.
4. Posey B. Improving your odds of answering multiple choice questions correctly; 2012. Available from: <http://www.pearsonitcertification.com/articles/article.aspx?p=1960217>
5. Vyas R, Supe A. Multiple choice questions: a literature review on the optimal number of options. *Natl Med J India* 2008;21:130-3.
6. Brady AM. Assessment of learning with multiple-choice questions. *Nurse Educ Pract* 2005;5:238-42.
7. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. *Med Teach* 2004;26:709-12.
8. Aiken LR. Writing multiple-choice items to measure higher-order educational objectives. *Educ Psychol Measure* 1982;42:803-6.
9. Morrison S, Walsh KF. Writing multiple-choice test items that promote and measure critical thinking. *J Nurs Educ* 2001;40:17-24.
10. Brame JC. Writing good multiple choice test questions; 2013. Available from: <https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/>
11. Bloom BS, Engelhart M, Furst EJ, et al. Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. New York: David McKay Company; 1956.
12. Anderson LW, Krathwohl D. A taxonomy for learning, teaching and assessing: a revision of Bloom's taxonomy of educational objectives. New York: Longman; 2001.
13. Guilford JP. The nature of human intelligence. New York: McGraw-Hill; 1967.
14. Churches A. Bloom's digital taxonomy; 2010. Australian School Library NSW incorporated, 2010. Available from: <http://burtonslifelearning.pbworks.com/f/BloomDigitalTaxonomy2001.pdf>
15. Kar SS, Lakshminarayanan S, Mahalakshmy T. Basic principles of constructing multiple choice questions. *Indian J Commun Family Med* 2015;1:65-9.
16. Haladyna TM, Downing SM. Validity of a taxonomy of multiple-choice item-writing rules. *Appl Measure Educ* 1989;2:51-78.
17. Centre for Teaching Excellence, University of Waterloo. Designing multiple-choice questions. Available from: <https://uwaterloo.ca/centre-for-teaching-excellence/teaching-resources/teaching-tips/developing-assignments/assignment-design/designing-multiple-choice-questions>
18. Rodriguez MC. The art and science of item writing: a meta-analysis of multiple choice item format effects. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, 1997.
19. Statman S. Asking a clear question and get a clear answer: an enquiry into the question/answer and sentence completion formats of multiple choice items. *System* 1998;6:367-76.
20. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Med Educ* 2009;9:40.
21. Brigham Young University. 14 rules for writing multiple-choice questions - Annual University Conference; 2001.

- Available from: <https://testing.byu.edu/handbooks/14%20Rules%20for%20Writing%20Multiple-Choice%20Questions.pdf>
22. McMahon G, Peyre SE. Educational Research Methods Program, Introduction to test development. Available from: http://ecommons.med.harvard.edu/ec_res/EF232EA2-FCD2-483B-93CC-CA36114F05
 23. Mehta M, Banode S, Adwal S. Analysis of multiple choice questions (MCQ): important part of assessment of medical students. *Int J Med Res Rev* 2016;4:199-204.
 24. Vuolo J. Multiple Choice Questions (MCQs), Ensuring the question fits the purpose. Available from: <http://slideplayer.com/slide/2558182/>
 25. Ramakrishnan M, Sathe AB, Vinayak A, Item analysis: a tool to increase MCQ validity. *Indian J Basic Appl Med Res* 2017;6:67-71.
 26. Sathe AB, Bhadre RB, Bhalkar MS, Mosamkar SP. Subjective versus objective item analysis of multiple choice question. *Int J Healthcare Biomed Res* 2017; 5:89-94.
 27. Mathysen DGP. Setting pass marks for examinations, EBO - European Board of Ophthalmology - CESMA meeting, November 28th, 2009, Brussels.
 28. Schuwirth L, Pearce J. Determining the quality of assessment items in collaborations: aspects to discuss to reach agreement developed by the Australian medical assessment collaboration; 2014. Available from: http://research.acer.edu.au/higher_education/42
 29. Angoff WH. Scales, norms and equivalent scores. Educational Measurements, Washington, DC: American Council on Education; 1971.
 30. Harley D. Making appropriate pass-fail decisions. Available from: <https://wiki.usask.ca/download/attachments/3997707/Passing%20Score%20U%20of%20S.ppt?api=v2>
 31. Maxinity Co, UK. Standard setting simplified - Angoff. Available from: <https://www.maxinity.co.uk/blog/2016/7/22/standard-setting-simplified-angoff>
 32. Holmboe E, Hawkins R. Methods for evaluating the clinical competence of residents in internal medicine: a review. *Ann Intern Med* 1998;129:42-8.
 33. Dawes M, Summerskill W, Glasziou P, et al. Sicily statement on evidence-based practice Second International Conference of Evidence-Based Health Care Teachers and Developers. *BMC Med Educ* 2005;5:1.
 34. Parkes J, Hyde C, Deeks J, Milne R. Teaching critical appraisal skills in health care setting. *Cochrane Database Syst Rev* 2001;3:CD001270.
 35. Rosenberg W, Donald A. Evidence based medicine: an approach to clinical problem-solving. *BMJ* 1995;310: 1122-6.
 36. Nardi R, Berti F, Greco A, et al. Complexity in hospital internal medicine departments: what are we talking about? *Ital J Med* 2013;7:142-55.
 37. Fritsche L, Greenhalgh T, Falck-Ytter Y, et al. Do short courses in evidence based medicine improve knowledge and skills? Validation of Berlin questionnaire and before and after study of courses in evidence based medicine. *BMJ* 2002;325:1338-41.
 38. Ramos K, Schafer S, Tracz S. Validation of the Fresno test of competence in evidence based medicine. *BMJ* 2003;326:319-21.
 39. Wood E. What are extended matching questions? *Biosci Educ Electron J* 2003;1:1-8.
 40. Leung W. Multiple choice questions in evidence based medicine. *Postgrad Med J* 2000;76:594-5.
 41. The College of Family Physicians of Canada. Pearls. Available from: <http://http://www.cfpc.ca/Pearls/>
 42. University of Sydney. Sydney Medical Program: EBM PEARLS. Available from: <http://www.gmp.usyd.edu.au>
 43. Anderson M, Stickley T. Finding reality: the use of objective structured clinical examination (OSCE) in the assessment of mental health nursing students interpersonal skills. *Nurse Educ Pract* 2002;2:160-8.
 44. Ilic D. Assessing competency in evidence based practice: strengths and limitations of current tools in practice. *BMC Med Educ* 2009;9:53.
 45. Zimmario DM. Writing good multiple-choice exams. Austin, TX: University of Texas; 2004.
 46. Moss E. Multiple choice questions: their value as an assessment tool. *Curr Open Anaesthesiol* 2001;14:661-6.
 47. McAllister D, Guidice RM. This is only a test: a machine-graded improvement to the multiple-choice and true-false examination. *Teach Higher Educ* 2012; 17:193-207.
 48. Bonora M, Storti D. Superare la prova a test - Tecniche e metodi per risolvere i test psicoattitudinali di logica e di cultura generale - Con software di simulazione. Napoli: EdiSes; 2017.
 49. Bonora M, Cordua F, Storti D. Il nuovo concorso a cattedra. Tecniche e metodi per la preselezione. Metodologia per una preparazione efficace. Napoli: EdiSes; 2012.
 50. Bonora M, Storti D. L'ammissione al TFA. Tecniche e metodi per superare le prove del concorso. Napoli: EdiSes; 2011.
 51. Hetru E, Crepin B. Abstract Reasoning MCQ 2013: For European Institution Competitions Paperback, Orseu, 2013.