Tidsskriftet Tidsskrift for Den norske legeforening

Exam theory for dummies

PERSPECTIVES

PETTER GJERSVIK

E-mail: petter.gjersvik@medisin.uio.no
Petter Gjersvik, professor at the Institute of Clinical Medicine, University of Oslo, where he is Head of Studies in Dermatology and Venereology and head of an exam committee for the medical degree programme. He is also Senior Associate Editor of the Journal of the Norwegian Medical Association.
The author has completed the ICMJE form and declares no conflicts of interest.

The final medical studies exam is intended to test knowledge and skills that the students will need as doctors – not the students' ability to pass an exam. What is the optimal exam type? And what does an exam really test?

Assessment of the students' knowledge and skills constitutes a separate discipline within medical education (1, 2). We could refer to it as *exam theory*, although assessment also takes place in other ways than through an exam. There are thick textbooks, training courses and academic conferences that focus exclusively on assessment methods in education. Procedures and rules have been developed for the design of exam questions as well as for assessment and scoring of answers, and quantitative research methods are used to assess how well exam questions function and how they should be graded in order to be perceived as fair and reliable.

A meaningful discussion of exams is reliant on a good dialogue between teachers, examination experts and students. For teachers as well as internal and external assessors, exam theory is sometimes difficult to relate to, because it involves an established terminology that may appear vague, inapt and confusing, especially when the terms used are in English.

In the following I provide a review of key examination terms, most of which are primarily associated with written exams, as a non-expert attempt to explain them to novices. The strengths and weaknesses of various types of exam are described briefly, because the terms are easier to understand in their proper context. The reader can find more precise and detailed explanations in the specialist literature (2). The topic should

be of interest to students as well as doctors, especially teachers at our medical faculties, because assessment methods impact on the learning behaviour of both students and doctors.

The fundamentals

Formative tests refer to tests that are used to ascertain the students' pre-existing knowledge, enabling the teacher to adapt the teaching to the students' prerequisites (1, 2). Such tests can help improve the motivation for learning and better study habits (3), but otherwise they entail no consequences for the individuals concerned. *Summative tests* refer to tests intended to check what the students have learned and involve grading, either pass/fail or according to a grade scale. In a word, an exam.

Validity and *reliability* are familiar terms for anyone who is engaged in research. High validity means that the test tests what it is intended to test, in our context the students' knowledge and skills in representative parts of the discipline. In other words: are the type of exam and the exam questions adequate and appropriate? High validity is paramount. High reliability implies that the outcome of the test, i.e. the grade, is reliable and replicable. The students should perceive the grade as fair.

Multiple choice questions = box-ticking questions

Today, the testing of medical students' knowledge often takes the form of a digital exam, i.e. with the aid of a computer. Based on experience from the United States and other countries, so-called *multiple choice questions*, often abbreviated *MCQ*, are increasingly used (1, 2). Such questions are provided with multiple response alternatives, usually from three to five, of which only one is the 'single best answer'. The other response alternatives are referred to as *distractors*. These alternatives need to be seen as plausible, should not stand out or be glaringly wrong, but still less correct than the 'single best answer'. *Multiple response questions* are a variant of multiple choice questions, where the candidates must select two to three correct responses from among five to eight alternatives. Other, less frequently used question types also exist.

The main advantage of such box-ticking questions is that they can be answered within a short time. An exam that involves only such questions can therefore include more questions than other types, and thereby encompass larger elements of the discipline. In addition, the scoring is done automatically, since only one response alternative is the 'single best answer' (or more in multiple response questions). By starting with a description of a clinical situation, the exam question can simulate a clinical decision-making process (4). Sets of good multiple choice questions have been shown to differentiate well between strong, intermediate and weak students (2).

> ***«Students learn to identify the correct answer by looking at the wording of the questions – they become test wise»***

However, good multiple choice questions are difficult to formulate, and not all topics are equally well suited for them. The criticism raised against these types of questions also focuses on their failure to reflect clinical realities, that they do not adequately test the candidates' ability to reflect and apply their knowledge, and that they have a negative effect on the students' learning behaviour (2, 5). Moreover, the likelihood of simply guessing the correct answer is high: 25 % for four response alternatives and 33 % and 50 % respectively if the candidate is able to identify one or two incorrect response alternatives. Some candidates will recognise the correct answer when reading the response alternatives given, referred to as *cueing*. Publishing of previous exam questions and experience from sitting previous exams will help students to learn to identify the correct answer by looking at the wording of the questions – they become *test wise*. These disadvantages of multiple choice questions are often underestimated and undercommunicated.

Free-text questions = open-ended questions

Digital exams can also involve questions to which the candidates must respond with a brief text. Such questions should be referred to as *free-text questions* or *open-ended questions* (4).

Many refer to these free-text questions as *essay questions* (or *mini-essay questions*) (7), but this is misleading at best. Essays are a non-fiction genre with long texts that are mainly published in journals and books, virtually a minor thesis (8). This term evokes associations of an old-fashioned type of exam that was abandoned long ago, where the candidates were asked to write a long account about a given topic. In other words, writing an essay is the complete opposite of what a student should do in a digital examination, which is to answer a question by writing a brief, accurate and concise text, preferably in keywords or two to three sentences at most. Using the term *essay* in this context could be seen as an encouragement to write a lengthy answer, which some students unfortunately do, especially when they are uncertain about what to answer.

The advantage of free-text questions is that the candidates must respond without the help of pre-defined response alternatives, in the same way as doctors must act in their clinical work (5). Such questions will often provide a truer and more authentic picture of the candidate's skills (9, 10). However, free-text questions and scoring guidelines can be difficult to prepare. Scoring the responses is time-consuming and can vary according to the scorer's background and prerequisites. Consistent scoring practices can nevertheless be nurtured with the aid of good scoring guidelines, practice in advance and consensual scoring, i.e. the scorers adjust their scores if major discrepancies occur. The number of free-text questions should be kept lower, because answering them may take a little longer. In the UK, a computer program with *very-short-answer questions* has been developed, in which the responses can be scored by a computer program (9), in the same way as the answers to multiple choice questions.

Psychometry

A number of quantitative research methods have been developed to assess how well exam questions function (2). Is a question too easy (nearly everyone answers correctly) or too difficult (hardly anyone answers correctly)? How are the items in a set of questions distributed in terms of their item difficulty? How well does a question differentiate between strong, intermediate and weak candidates? If approximately the same number of students have chosen each of the response alternatives, this will indicate that they have largely guessed. Such methods permit the exam committee to identify questions that have not worked satisfactorily and consider removing them from the grading base (7).

Similarly, methods have been developed to assess discrepancies and precision in scorings of responses to free-text questions (1, 2). Do the scores set by one scorer deviate significantly from those set by others? Such methods to measure *inter-rater reliability* can identify scorers who are 'too lenient' or 'too strict' in one or more questions and possibly adjust scores that deviate too much from the others, i.e. *rater alignment*.

These methods for quality assurance of exam questions and scoring practices are referred to as *psychometry*. In this context, psychometry does not refer to measurement of people's psychological characteristics, as one might think, but to measurement of the way in which examination questions function alone and as a whole, and of the extent to which the candidates' responses have been assessed consistently and reliably.

Standard setting

Grading involves a determination of a final grade, either in the form of pass/fail or according to a grading scale, for example A–F, where A is the best grade and F means fail. The main, and often most difficult task is to determine the pass/fail boundaries. Such processes are referred to as *standard setting* (2).

> ***«Different exam types and types of questions each have their strengths and weaknesses, and they arouse many different and often conflicting opinions»***

Ideally, the degree of difficulty of question sets should be kept stable over the years, but this is difficult to achieve in practice (11). *Relative standard setting* is based on the performance of all the students and an assessment of the difficulty of the questions. *Absolute standard setting* means that the pass grade boundary has been determined in advance. A number of mathematical models for setting the pass grade boundary have been developed, but these are complicated and resource-intensive (2, 11).

In practice, the exam committee will often base its determination of the pass grade boundary on a pragmatic approach that draws on academic discretion. If a grading scale is used, the other levels can be determined based on the pass grade boundary and an equivalent assessment of the boundary to an A, the top grade.

Diversity and totality

An exam is intended to test the knowledge and skills that the students will need as doctors – not the students' ability to pass an exam. Different exam types and types of questions each have their strengths and weaknesses, and they arouse many and often conflicting opinions (2, 12). Rather than choosing one question form over the others, a digital exam should include *both* multiple choice and free-text questions. The practical challenges that this implies are manageable and surmountable. In addition, there is a need for clinical and oral exams that make for a better test of the students' reasoning ability and clinical skills than a digital exam provides. Such exams can be largely standardised to ensure high validity and consistent scoring practices.

The purpose of the final medical studies exam is to provide assurance to society, the health services and the patients that the universities train highly skilled doctors. Moreover, testing and exams help motivate for learning. So whether we like it or not: exams are important.

The author wishes to thank Per Grøttum and Stefan Schauber for their useful contributions.

LITERATURE

1. Epstein RM. Assessment in medical education. N Engl J Med 2007; 356: 387–96. [PubMed][CrossRef]

2. Schuwirth LWT, Ash J. Principles of assessment. I: Walsh K, red. Oxford Textbook of Medical Education. Oxford: Oxford University Press, 2013.

3. Larsen DP, Butler AC. Test-enhanced learning. I: Walsh K, red. Oxford Textbook of Medical Education. Oxford: Oxford University Press, 2013.

4. NTNU. Eksamensoppgaver – medisin – MH. https://innsida.ntnu.no/wiki/-/wiki/Norsk/Eksamensoppgaver+-+Medisin+-+MH Accessed 15.2.2020.

5. Schuwirth LWT, Verheggen MM, van der Vleuten CP et al. Do short cases elicit different thinking processes than factual knowledge questions do? Med Educ 2001; 35: 348–56. [PubMed][CrossRef]

6. NTNU. Eksamensoppgaver – medisin – MH. https://innsida.ntnu.no/wiki/-/wiki/Norsk/Eksamensoppgaver+-+Medisin+-+MH Accessed 15.2.2020.

7. Universitetet i Oslo. Medisin (profesjon). Oppbygging og gjennomføring. https://www.uio.no/studier/program/medisin/oppbygging/ Accessed 15.2.2020.

8. Skei HH. Essay. I: Store norske leksikon. https://snl.no/.search?query=essay Accessed 18.2.2020.

9. Sam AH, Field SM, Collares CF et al. Very-short-answer questions: reliability, discrimination and acceptability. Med Educ 2018; 52: 447–55. [PubMed][CrossRef]

10. Sam AH, Westacott R, Gurnell M et al. Comparing single-best-answer and very-short-answer questions for the assessment of applied medical knowledge in 20 UK medical schools: Cross-sectional study. BMJ Open 2019; 9: e032550. [PubMed][CrossRef]

11. Colberg AB, Vatn D, Standal R et al. How can the examination failure rate be stabilised? Tidsskr Nor Legeforen 2017; 137. doi: 10.4045/tidsskr.17.0025. [PubMed][CrossRef]

12. Hift RJ. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? BMC Med Educ 2014; 14: 249. [PubMed][CrossRef]

Publisert: 11. June 2020. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.20.0142
Received 18.2.2020, accepted 16.4.2020.