REVIEW ARTICLE

Deliberations on the contemporary assessment system: A Narrative Review

Usman Mahboob

There are different apprehensions regarding the contemporary assessment system. Often, I listen to my colleagues saying that multiple-choice questions are seen as easier to score. Why can't all assessments be multiple-choice tests? Some others would say, whether the tests given reflect what students will need to know as competent professionals? What evidence can be collected to make sure that test content is relevant? Others come up with concerns that there is a perception amongst students that some examiners are harsher than others and some tasks are easier than others. What can be done to evaluate whether this is the case? Sometimes, the students come up with queries that they are concerned about being observed when interacting with patients. They are not sure why this is needed. What rationale is there for using workplace-based assessment? Some of the students worry if the pass marks for the assessments are 'correct', and what is the evidence for the cut-off scores? All these questions are important, and I would deliberate upon them with evidence from the literature.

Deliberating on the first query of using multiple choice questions for everything, we know that assessment of a medical student is complex process as there are multiple domains of learning such as the cognition, skills and behaviors (Norcini and McKinley, 2007)(Boulet and Raymond, 2018). Each of the domain further has multiple levels from simple to complex tasks (Norcini and McKinley, 2007). For example, the cognition is further divided into six levels, starting from recall (Cognition level 1 or C1) up to creativity (Cognition level 6 or C6) (Norcini and McKinley, 2007). Similarly, the skills and behaviors also have levels starting from observation up to performance and practice (Norcini and McKinley, 2007). Moreover, there are different competences within each domain that further complicates our task as assessor to appropriately assess a student (Boulet and Raymond, 2018). For instance, within the cognitive domain, it is not just making the learning objectives based on the Bloom's Taxonomy that would simplify our task because the literature suggest that individuals have different thinking mechanisms, such as fast and slow thinking to perform a task (Kahneman, 2011). We as educationalist do not know what sort of cognitive mechanism have, we triggered through our exam items (Swanson and Case, 1998).

Multiple Choice Questions is one of the assessment instruments to measure competencies related to cognitive domain. This means that we cannot use multiple choice questions to measure the skills and behaviors domains, so clearly multiple-choice questions cannot assess all domains of learning (Vleuten *et al*, 2010). Within the cognitive domain, there are multiple levels and different ways of thinking mechanisms (Kahneman, 2011). Each assessment instrument has its strength and limitations. Multiple choice questions may be able to assess few of the competencies, also with some added benefits in terms of marking but there always are limitations. The multiple-choice question is no different when it comes to the strengths and limitations profile of an assessment instrument (Swanson and Case, 1998). There are certain competencies that can be easily assessed using multiple choice questions (Swanson and Case, 1998). For example, content that requires recall, application, and analysis can be assessed with the help of multiple-choice questions. However, creativity or synthesis which is cognition level six (C6) as per Blooms' Taxonomy, cannot be assessed with closed-ended questions such as a multiple-choice question. This means that we need some additional assessment instruments to measure the higher levels of cognition within the cognitive domain. For example, asking students to explore an open-ended question as a research project can assess the higher levels of cognition because the students would be gathering information from different sources of literature, and then synthesizing it to answer the question. It is reported that marking and reading the essay questions would be time consuming for the teachers (McLean and Gale, 2018). Hence, the teacher to student's ratio in assessing the higher levels of cognition needs to be monitored so that teachers or assessors can give appropriate time to assess the higher levels of cognition of their students.

Hence, we have to use other forms of assessment instruments along with multiple choice questions to assess the cognitive domain. This will help to assess the different levels of cognition and will also incite the different thinking mechanisms.

Regarding the concerns, whether the tests given reflect what students will need to know as competent professionals? What evidence can be collected to make sure that test content is relevant? It is one of an important issue for medical education

Corresponding Author Details:
Usman Mahboob
Director and Assistant Professor
Institute of Health Professions Education & Research
Khyber Medical University, Peshawar
usman.mahboob@kmu.edu.pk

Funding Source: NIL; Conflict of interest: NIL
Received: July 20th ,2019; Accepted: August 24th, 2019

and assessment directors whether the tests that they are taking are reflective of the students being competent practitioners? It is also quite challenging as some of the competencies such as professionalism or professional identity formation are difficult to be measured quantitatively with the traditional assessment instruments (Cruess, Cruess, & Steinert, 2016). Moreover, there is also a question if all the competencies that are required for a medical graduate can be assessed with the assessment instruments presently available? Hence, we as educationalists have to provide evidences for the assessment of required competencies and relevant content.

One of the ways that we can opt is to carefully align the required content with their relevant assessment instruments. This can be done with the help of assessment blueprints, or also known as table of specifications in some of the literature (Norcini and McKinley, 2013). An assessment blueprint enables us to demonstrate our planned curriculum, that is, what are our planned objectives, and how are we going to teach and assess them (Boulet and Raymond, 2018).

We can also use the validity construct in addition to the assessment blueprints to provide evidence for testing the relevant content. Validity means that the test is able to measure what it is supposed to measure (Boulet and Raymond, 2018). There are different types of validity but one of the validity that is required in this situation to establish the appropriateness of the content is the Content Validity. The content validity is established by number of subject experts who comment on the appropriateness and relevance of the content (Lawshe, 1975).

The third method by which the relevance of content can be established is through standard setting. A standard is a single cut-off score to qualitatively declare a student competent or incompetent based on the judgement of subject experts (Norcini and McKinley, 2013). There are different ways of standard setting for example Angoff, Ebel, Borderline method, etc. (Norcini and McKinley, 2013). Although, the main purpose is the establish and decide the cut-off score but during the process, the experts also debate on the appropriateness and relevance of the content. This means that the standard setting methods also have validity procedures that are in-built in their process of establishing the cut-off score. These are some of methods by which we can provide an evidence of relevance of content that is required to produce a competent practitioner.

The next issue is the perception amongst students that some examiners are harsher than others and some tasks are easier than others. Both these observations have quite a lot of truth in them and can be evaluated following the contemporary medical education evaluation techniques. The first issue reported is that some examiners are harsher than others. In the terms of assessment, it has been reported in the literature as ‘hawk-dove effect’ (McManus et al, 2006, Murphy et al, 2009). There are different reasons identified in the literature for some of the examiners to be more stringent than others such as age, ethnic background, behavioral reasons, educational background, and experience in number of years (McManus et al, 2006). Specifically, those examiners who are from ethnic minority and have more experience show more stringency (McManus et al, 2006). Interestingly, it has been reported elsewhere how the glucose levels affect the decision making of the pass-fail judgements (Kahneman, 2011).

There are psychometric methods reported in the literature, such as Rasch modelling that can help determine the ‘hawk-dove effect’ of different examiners, and whether it is too extreme or within a zone of normal deviation (McManus et al, 2006, Murphy, et al, 2009). Moreover, the literature also suggests ways to minimise the hawk-dove effect by identifying and paring such examiners so the strictness of one can be compensated by the leniency of the other examiner (McManus et al, 2006).

The other issue in this situation is that the students find some tasks easier than others. This is dependent on the complexity of tasks and also on the competence level of students. For example, a medical student may achieve independent measuring of blood pressure in his/her first year but even a consultant surgeon may not be able to perform complex surgery such as a Whipple procedure. This means that while developing tasks we as educationalist have to consider both the competence level of our students and the complexity of the tasks. One way to theoretically understand it is taking help from the cognitive load theory (Merrienboer 2013). The cognitive load theory suggests that there are three types of cognitive loads; namely, the Intrinsic, Extraneous, and Germane loads (Merrienboer 2013). The intrinsic load is associated with the complexity of the task. The extraneous load is added on the working memory of students due to teacher who does not plan his/her teaching session as per students need (Merrienboer 2013). The third load is the germane or the good load that helps the student to understand the task and is added by using teaching methods that helps students understand the task (Merrienboer 2013). The teachers can use different instructional designs such as 4CID model to plan their teaching session of the complex tasks (Merrienboer 2013). One of the ways to understand the difficulty of task can be to pilot test the task with few students or junior colleagues. Another way to determine the complexity of the task can be through standard setting methods where a cut-off score is established after the experts discuss each task and determine its cut-off score based on their judgements (Norcini and McKinley, 2013). However, it is important that the experts who have been called for setting standards have relevant experience so as to make credible judgements (Norcini and McKinley, 2013). A third way to evaluate the complexity of tasks is by applying the post-exam item analysis techniques. The difficulty of task is evaluated after the performance of students in the exam. Each item's difficulty in exam can be measured. The items can be placed from extremely easy (100% students correctly answered the item) to extremely difficult (100% students failed on that specific item). The item analysis enables the teachers to determine which tasks were easier in exams as compared to more difficult tasks.

Another concern that comes from students is about their observation when interacting with patients. Health professions training programmes require interaction of students with patients. The student-patient interaction is not very often in initial years of student's training due to the issues of patient safety, and due to heavy workload on clinical faculty. However, with passage of time in the training programme, these student-patient interactions increase. There is also a strong theoretical basis for better learning when the students are put in a context or a given situation (Wenger, 1998). For example, infection control can be taught through a lecture however the learning can be more effective if the students practically learn it in an operation theatre. Moreover, the undergraduate students or foundation year house job doctors are yet not competent enough to practice independently and require supervision for the obvious reasons of patient safety. Although, some of the students may not like being observed but it is one of the requirements for their training. The examiners observing them can give them constructive feedback to further improve their performance (Etheridge and Boursicot, 2013). Feedback is one of the essential components of workplace-based assessments, and it is suggested in the literature that the time for feedback to the student should be almost equal to one third of the procedure or task time (Etheridge and Boursicot, 2013), that is, for a fifteen minutes tasks, there should be at least five minutes for the feedback hence having a total of twenty minutes time on the whole.

Further, it is important for the examiners and senior colleagues to establish trust in the competence of their students or trainees. The 'trust' is one of the behavioral constructs that also starts initially with an observation (Etheridge and Boursicot, 2013). Hence, observation of students or house officers by senior colleagues or teachers during clinical encounters is important to establish trust on student's competence levels.

Additionally, in the workplace, there are different skills that are required by the students to demonstrate, and each skills are quite different to other. There are different workplace-based assessment instruments and each of them assess only certain aspects of student's performance during clinical practice. For instance, the Mini Clinical Evaluation Exercise (Mini-CEX) can primarily assess the history taking and physical examination skills of students (Etheridge and Boursicot, 2013). Similarly, the Directly Observed Procedural Skills (DOPS) is required to assess the technical and procedural skills of students (Etheridge and Boursicot, 2013). More so, the Case-based Discussion (CbD) is required to assess the clinical reasoning skills, decision making skills, ethics and professionalism (Etheridge and Boursicot, 2013). Further, multi-source feedback (MSF) or 360-degree assessment collects feedback about a student on their performance from multiple sources such as patients, senior and junior colleagues, nursing staff, and administrative staff (Etheridge and Boursicot, 2013). All these workplace-based assessments require observation of students so they can be given appropriate feedback on their technical and non-technical skills (Etheridge and Boursicot, 2013). Hence, clinical encounters at workplace are quite complex and require training of students from different aspects to fully train them that cannot be accomplished without observation.

Some students also worry whether the pass marks for the assessments are 'correct', and what is the evidence for the cut-off score in their exams? A standard is a single cut-off score that determines the competence of a student in a particular exam (Norcini and McKinley, 2013). The cut-off score is decided by experts who make a qualitative judgement (Norcini and McKinley, 2013). The purpose is not to establish an absolute truth but to demonstrate the creditability of pass-fail decision in an exam (Norcini and McKinley, 2013).

There are certain variables related to standard setters that may affect the creditability of the standard setting process; such as age, gender, ethnicity, their understanding of the learners, their educational qualification, and their place of work. Moreover, the definition of competence varies with time, place and person (Norcini and McKinley, 2013). Hence, it is important that the standard setters must know the learners and the competence level expected from them and the standard setters must be called from different places. This is one of the first requirement to have the profile of the standard setters to establish their credibility.
Moreover, the selection of method of standard setting is important, and how familiar are the standard setters with the method of standard setting. There are many standard setting methods for different assessment instruments and types of exams (Norcini and McKinley, 2013). It is essential to use the appropriate standard setting method, and also to train the standard setters on that method of standard setting so they know the procedure. The training can be done by providing them certain data to solve it following the steps of the standard setting procedure. The record of these exercises is important and can be required at latter stages to show the experience of the standard setters.

Further, every standard setter writes a cut-off score for each item (Norcini and McKinley, 2013). The mean score of all the standard setters is calculated to determine the cut-off score for each item (Norcini and McKinley, 2013). The total cut-off score is calculated by adding the pass marks of each individual item (Norcini and McKinley, 2013). The cut-off scores for items would also help in differentiating the hawks from doves, that is, those examiners who are quite strict from those who are lenient (McManus et al, 2006). Hence, it is important to keep the record of these cut-off scores of each item for the future records, and to have a balanced standard setting team for future exams (Norcini and McKinley, 2013). Additionally, the meeting minutes is an important document to keep the record for the decisions made during the meeting.

Lastly, the exam results and post-exam item analysis is an important document to see the performance of students on each item and to make comparisons with the standard setting meeting (Norcini and McKinley, 2013). It would be important to document the items that behaved as predicted by the standard setters, and those items that would show unexpected response; for example, majority of the borderline students either secured quite high marks than the cut-off score or vice versa (Norcini and McKinley, 2013). All the documents mentioned above would ensure the creditability of the standard setting process and would also improve the quality of exam items.

There are many other aspects that could not be discussed in this debate on contemporary assessment system in medical education. Another area that needs deliberations is the futuristic assessment system and how it would address the limitations of the current system?

Disclaimer: This work is derived from one of the assignments of the author submitted for his certificate from the Keele University.

References:

Boulet, J. and Raymond, M. (2018) 'Blueprinting: Planning your tests. FAIMER-Keele Master's in Health Professions Education: Accreditation and Assessment. Module 1, Unit 2.', *FAIMER Centre for Distance Learning, CenMEDIC*. 6th edn. London, pp. 7–90.

Cruess, R. L., Cruess, S. R., & Steinert, Y. (2016). 'Amending Miller's pyramid to include professional identity formation'. *Acad Med*, *91*(2), pp. 180–185.

Etheridge, L. and Boursicot, K. (2013) 'Performance and workplace assessment', in Dent, J. A. and Harden, R. M. (eds) *A practical guide for medical teachers*. 4th edn. London: Elsevier Limited.

Kahneman, D. (2011) Thinking, fast and slow. New York: Farrar, Straus and Giroux.

Lawshe, CH. (1975) A quantitative approach to content validity. *Pers Psychol*, 28(4), pp. 563–75.

McLean, M. and Gale, R. (2018) Essays and short answer questions. *FAIMER-Keele Master's in Health Professions Education: Accreditation and Assessment*. Module 1, Unit 5, 5th edition. FAIMER Centre for Distance Learning, CenMEDIC, London.

McManus, IC. Thompson, M. and Mollon, J. (2006) ' Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling' *BMC Med Educ.* 42(6) doi:10.1186/1472-6920-6-42

Merrienboer, J.J.G. (2013) 'Instructional Design', in Dent, J. A. and Harden, R. M. (eds) *A practical guide for medical teachers*. 4th edn. London: Elsevier Limited.

Murphy, JM. Seneviratne, R. Remers, O and Davis, M. (2009) 'Hawks' and 'doves': effect of feedback on grades awarded by supervisors of student selected components, *Med Teach*, 31(10), e484-e488, DOI: 10.3109/01421590903258670

Norcini, J. and McKinley, D. W. (2007) 'Assessment methods in medical education', *Teaching and Teacher Education*, 23(3), pp. 239–250. doi: 10.1016/j.tate.2006.12.021. Norcini, J. and Troncon, L. (2018) *Foundations of assessment. FAIMER-Keele Master's in Health Professions Education: Accreditation and Assessment. Module 1, Unit 1.* 6th edn. London: FAIMER Centre for Distance Learning CenMEDIC.

Norcini, J. and McKinley, D. W. (2013) 'Standard Setting', in Dent, J. A. and Harden, R. M. (eds) *A practical guide for medical teachers*. 4th edn. London: Elsevier Limited.

Swanson, D. and Case, S. (1998) *Constructing written test questions for the basic and clincial sciences*. 3rd Ed. National Board of Medical Examiners. 3750 Market Street

Philadelphia, PA 19104.

Van Der Vleuten, C. Schuwirth, L. Scheele, F. Driessen, E. and Hodges, B. (2010) 'The assessment of professional competence: building blocks for theory development', Best Practice & Research Clinical Obstetrics and Gynecology, pp. 1-17. doi:10.1016/j.bpobgyn.2010.04.001

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge university press.