Postgraduate Assessment: Rationale of Logical Decisions, *Education in Medicine Journal*, 2011 Vol.3 (1):e1-e5
doi:10.5959/eimj.3.1.2011.e1

EDITORIAL

Postgraduate Assessment: Rationale of Logical Decisions

Shahid Hassan

Department of Medical Education/Otorhinolaryngology, School of Medical Sciences, Universiti Sains Malaysia

Traditional assessment is now often criticized for measuring the clinical skills as the true reflection of competency acquired by a postgraduate trainee and tested in summative examination. The commonly used measurement tools in surgical-based discipline's summative assessment held as conjoint examination of three universities (UM, UKM and USM) in Malaysia include: MCQ (True/false), Short and Extended Essays (Questions), Short and Long Cases and Oral (Viva). Decision on pass and fail often adopts a mixed conjunctive-compensatory approach. Long case assessments are more criticized for reliability of assessment tool being unobserved and inconsistent across raters and across cases. The tool used by many disciplines is a single long case for assessment therefore faces the question of context-specificity and subjective bias. Reliability in long case assessment is difficult to achieve besides feasibility and standardization (1) as the other problems. Lack of sufficient testing time, unstructured random questions and inconsistent scoring are some other causes of low reliability. These unobserved long cases are often relied on candidate's reflection on patient's medical history and physical examination. This makes the long case an assessment of "knows how" rather than "shows how" of Miller's pyramid of competence learning. The validity challenged for its context specificity also faces the problem of lacking content structure and rating by consensus of examiners.

Reliability and validity somewhat better in short case assessment is primarily due to content-specificity, performed on a variety of observed encounters between students and the patients. Integrated scoring averaging across the cases further improves the reliability of this measurement tool that speaks of triangulation within the clinical component of short-case examinations. However, there is a similar issue of rating the performance with consensus in individual short cases like the one practiced in oral and long case examinations. Another problem in short case assessment relates to poor familiarization of principles and philosophy of short cases examination. It has been observed that short cases are sometimes over enthusiastically done by structuring a case into a number of clinical attributes, e.g. bit of history, investigative, diagnostic and therapeutic skills. This denies the objectives of short case assessment, which is primarily meant to judge the competence and critical thinking skills of a candidate on physical examination and diagnosis.

Observing the workup and adding other long cases in assessment may improve both, reliability and validity to some extent. Other options to test clinical competency in postgraduate medical education are OSCE (objective structured clinical examination) or WPBA (workplace based assessment), which may not be feasible due to constraints of logistics and requirement of intensive faculty development. These methods of assessment are not only resource intense but also require a major decision in readjustments of practice of assessment culture, which is globally moving from competence to performance-based assessment such as Mini-CEX and DOPS (2).

Long case examination in Specialty Conjoint Board of Postgraduate Medical Education for its summative assessment in Malaysia often allows only one case for one candidate due to time constraints. The panel of examiners usually reviews the short listed cases to select the right cases for this encounter on the morning of examination. However all the examiners involved may not necessarily be examining all the cases selected for this encounter. In practice students reflect on their medical interviewing and physical examination skills by presenting the case before the panel of 3-4 examiners for cross-examination after 20-30 minutes of patient's work up. These unobserved cases turn the long case assessment to an oral examination especially for an evaluator who somehow has not been able to examine the case. Observing the long case encounter between candidate and patient can improve the validity besides, providing opportunity to examiners to

evaluate the candidates for their overall competency of soft skills including communication skills, patient-doctor interpersonal skills, organizational efficiency and humanistic qualities. Humanism is the important aspect of postgraduate training, which needs to be observed to assess the meta-skills of a candidate's professionalism for which we have no other instrument than short-case assessment in our summative examination. Individual rater assessment, confining to its clinical attribute of physical examination and increasing the time of this examination from 7 minutes to 10 minutes per case will help to improve its validity. This indirectly will bring short case assessment close to Mini-CEX with exception of inability to provide feedback and rather carrying it out in a more obvious test situation of competence testing than performance testing.

The outcome of measures in written component as essay questions and MCQ is decided using a compensatory approach however, the same approach is denied in clinical component. Long and short cases stand on its own for a decision, which uses a conjunctive approach to make decision on pass or fail. In short case assessment the decision is even beyond the rule of a conjunctive method. Rule set in short cases assessment says that candidates have to pass this measure to pass the summative examination (conjunctive approach).

However, 3-4 short cases though considered items of one measure (part of clinical component) will be allowed to compensate each other only if the candidate passes a minimum of more than two out of four short cases. This is like applying the principle of conjunctive method twice in decision-making on short cases assessment. A short case assessment to stand on its own as an independent measure is required to pass (1^{ST} application of conjunctive rule) separately. Besides, more than two of the four short cases are essential to pass (2^{nd} application of conjunctive rule) as independent clinical measures before considered for an aggregate 50% to pass this measure in clinical component. This amended conjunctive approach is practiced in ORL-HNS discipline. Other disciplines have adopted different rules on their own to practice conjunctive approach.

If the same principle is applied to pass the essay as a measure in written component for example, then a candidate might be required to pass two long essay questions separately before it is considered to be combined with third essay question for overall 50% score to pass the measure. It sounds ridiculous that somewhere (clinical component) we rigidly follow the rule of conjunctive approach or even beyond the recommendations of conjunctive method. However, in another measure (written component) we loosely apply the principle of compensatory method to make decision on pass or fail. Contrary to any standard setting strategy for making decision on pass or fail, the weak performance of MCQ is compensated by the strong performance of essay questions, which exactly do not test the same extent of educational taxonomy of learning domains as the essay questions. Here the question may arise that why the long and short case assessments are not used for decision making even though they belong to the same educational taxonomy of assessment domains. Strength of short case assessment can be utilized to compensate the weakness of long case assessment and vice versa. Especially when the short cases are observed, multiple and examined by different panel of assessors compared to a single, unobserved and one set of examiners in long case assessment.

The evidence to make decision on pass or fail should base on satisfactory or unsatisfactory performance of trainees both on quantitative scoring as well as on qualitative observation (1). A student's bad day in one assessment tool should not exist in isolation. An ideal instrument is not the one that stands out to decide on summative assessment outcome alone, rather the one, which contributes to determine the overall performance of a candidate (3)**.** The role of individual instrument should be seen in collaboration with other measurement tools for its due share to assess the respective learning domains such as knowledge or clinical skills to decide on pass or fail in summative assessment. After all staging a summative assessment is not about keeping up the sanctity of individual tools or the rule set by individuals, conversely at the cost of right decision on a candidate's career. More important in assessment is the process, which guides assessors to reach to a logical decision that genuinely allows a candidate to pass or fail the examination. Triangulation may be a good choice and it refers to making a qualitative judgment based on best-practice evidences on assessment gathered over different time, under different circumstances, by different evaluators and using different methods (1). Triangulation can be called upon for a right qualitative judgment utilizing the complementing role of assessment tools, at least in the same component (MCQ and essay in written or short and long cases in clinical), also called internal triangulation. This adjustment will provide the benefit of doubt to candidates especially if the quantitative judgment score of an individual instrument is in question. This indirectly will inculcate the concept of quantitative assessment in postgraduate examination rather than utilizing individual instrument in isolation to decide on summative results.

MCQ results are judged on aggregate with essay questions to evaluate the domain of knowledge is an example of internal triangulation loosely applied in postgraduate assessment. The question is why long case result cannot be integrated with short cases if not with oral test results to evaluate the domain of clinical skills? This concept of triangulation can specially be useful in situation that demands qualitative approach when evaluation of long case implicates the overall result of summative assessment. Decision making in such cases often leads to controversy of safe versus unsafe surgeon when one or two unexpected answer from a candidate is considered blunder by one or more examiners in the panel. The question may arise that should a single or couple of mistakes be allowed to determine the fate of a candidate on his bad day which may have other influencing factors like undue stress of examination, very complex case allotted for long case work up or a very difficult out of box question asked to analyze a problem solving issue for decision making for his/her level of training. In such a situation it is not justified to decide on overall summative assessment for passing or failing of a candidate on pretext of his/her safe or unsafe performance in long case assessment without a rationale. The judgment needs to consider a number of factors with its due weightage to decide on pass or fail and this needs a structured format for documented decision that can subsequently be used for feedback. This will also give an opportunity to examiners to rationalize the nature of mistake considering the patient's complexity, candidate's intellectual knowledge and its application, problem solving and therapeutic skills and attitude based on briefing from the internal assessor's quick review of candidate's overall performance in 4 years of his training.

In postgraduate examination of Master of Surgical-based Disciplines, clinical competence is tested directly through short and long case assessments with real patients and oral assessment with real or created scenarios using slides or video clips. However, in short case assessments there are 3-4 cases for physical examination and provisional diagnosis, which are directly observed. Similarly there are two rounds of oral test with different sub-specialty patient's scenarios, different panel of examiners and face-to-face questions. More than one item (3-4 short cases and 2 rounds of oral) in each component, different examiners, different cases and clinical scenarios, direct observation and face-to-face questions improves content as well as context specificities, therefore reliability and validity of these instruments. Principles of internal triangulation are also observed in these assessment methods and performance is rated as an aggregate of 3-4 cases in short case assessment and 2 rounds of viva in oral assessment respectively. However, long case assessment is carried out through a single and unobserved patient's workup, which is not analyzed in triangulation with another long case or any other tool in clinical component if a candidate's performance is not satisfactory for a clear pass due to one or two unexpected responses committed in cross examination. A candidate considered unsafe for medical practice in such case is not allowed to pass the summative assessment despite of his passing all other components comfortably well. Such incidents though occasional are experienced in these examinations and need to be addressed. Long case assessment based on a single case with varying level of complexity of patients from one candidate to another, unobserved workup and rating of performance achieved by consensus will keep raising the question of its validity as an outcome of summative assessment.

To improve the validity of measurement tools in these summative assessments few options are worth considering however, the implementation of any of these options will require a major decision to bring a change in standard setting strategy currently practiced in summative assessment as follows.

1. Essay questions should use structured clinical scenarios and short answer written format to improve context specificity and standardization. Consistency of marking can be improved by using model answers and at least two assessors mandatory for marking each question. In case of wide disparity in scores of two assessors 3rd assessor should be invited to decide the score. Though resource intensive, these steps will improve the reliability of written format.

2. Oral questions should be structured using clinical scenarios or scripts with laid down questions same for every candidate to standardize assessment and improve validity besides, attempting to reduce inter-rater and inter-case differences to improve reliability.

3. Every patient selected for long and short case assessments should have been examined by assessors' who are suppose to use those cases for assessment. At the same time assessors must also decide on nature of complexity of cases categorizing as simple, less complex and most complex cases. Allowances of mistakes should accordingly be allowed in quest of borderline candidates assessed on complex cases. This will bring some order to standardization if not entirely.

4. Long cases must be observed during the clerking process without interfering in work up by the candidates.

However, rating on observation using a checklist and a fixed percentage (20-30%) of marks out of total allocated marks for long case should preferably be used. Candidates' apprehension on observation is linked to increasing the time allotted from 20-25 minutes to 30-45 minutes to obviate the stress of examination and observation.

5. Rating of long and short cases and orals should avoid consensus marking and a logical method of individual rating by each examiner subsequently averaging the marks by the chief examiner should determine the candidate's score in long case assessment.

6. A long case assessment can be logically decided along with short cases adopting a compensatory approach. There are ways to reassure clinical competence as safe versus unsafe if a decision on pass or fail is logically made on conjunctive or compensatory or for that matter a mixed conjunctive-compensatory approach (4)**.** For example a candidate must score minimum 40% in each short cases or a long case before it is considered for compensation within the short cases or short and the long cases.

7. Alternatively two long cases should be used for clinical assessments if long and short case assessments are to be considered as separate clinical domains. If this is not possible due to the time constraints than at least a second long case assessment should be practiced in those borderline cases who fail the summative assessment due to their failing the long case. A panel of examiners that must include external examiner besides, internal and external examiners other than those involved in the first long case assessment should carry out the second long case assessment.

8. If long case assessment is decided with a conjunctive approach in current standard setting strategy then MCQ and essays questions are ought to be decided with similar strategy else the educational philosophy of assessment will be considered violated. In current practice MCQ and essay questions cannot be compensated for each other for their absolute difference in reliability as well as difference in object of measurement. Although MCQ is considered a reliable instrument for its content specificity compared to essay questions but the two items differs in its philosophy of learning domains. It is not a logical decision, simply because the objects of learning domains tested in MCQ and essay questions are not the same. The essay questions test the knowledge for its factual recall, comprehension, analysis, application and synthesis compared to true/false MCQ format, which tests the factual knowledge alone or at the most comprehension (5)**.**

Besides, MCQ is the only objective instrument in the entire battery of summative assessment and failing this assessment tool should be looked at failing the entire written component and this sound logical. Alternatively, to improve the validity of MCQ and its rationale for compensation by essay questions in decision-making of written test it can be improved by changing the true/false format of MCQ to a single best answer or extended matching multiple choice questions.

9. A content structure to do rating of individual candidate in long case assessment used by all examiners for marking the various clinical attribute as the process goes on, will make the measure more standardized and less subjective when summed up to an average mark of all examiners. This content structure can also be used for feedback of the students on their performance and for evaluation if required.

10. Options should be provided to rate the borderline students by examiners involved in long/short case assessments to use a structured format when it comes to decide on issues of candidate's safe or unsafe status as future surgeons in practice. This will produce documented evidence and will also help to avoid influence of one examiner over other, which is apparently observed in marking with consensus. This document can also be used for providing feedback to borderline students or to brief examiners in examination board meeting.

After critically evaluating the measurement tools employed in summative assessment of postgraduate medical education particularly in surgical-based disciplines it is concluded that a number of criteria used in decision-making on pass or fail is nowhere close to principals of educational theories or taxonomies. Who set these rules, how authentic these are and what context they have to principles of learning and assessment, are the mind-boggling questions that nobody takes the responsibilities to answer. Inherited from the elders and practiced over the years with minor adjustments made by the individuals however, without looking at the impact it has on standard setting and subsequent decision-making is all that we are executing. Whether we like it or not we must now follow the best-practiced medical education consistent with principles of measures and approaches in standard setting strategy in decision-making. This will enable us to make a logical and appropriate decision in summative assessment for its quantitative as well as qualitative role in judgment of postgraduate examination in medical education.

References

1. Jackson N, Jamieson A and Khan A. Assessment in medical education and training. Oxford: Radcliffe Publishing, 2007.
2. Hassan S. Clinical skills authentic learning with workplace-based assessments. Kota Bharu: SARANA Publishing, 2009.
3. Mitchell DC Multiple measures and high stakes decisions: A framework for combining measures. Educational measurement, issues and practice; summer 2003; 22, 2; ProQuest Education Journal.
4. Haladyna T and Hess R. An evaluation of conjunctive and compensatory standard setting strategy for test decision. Educational Assessment, 6 (2), 129-153. Lawrence Arlbaum Associates, Inc.1999-2000
5. Jacques EM and collaborators. Learning to become a physician at Sherbrooke. Maastricht: Network Publications, 2001.

Corresponding author:

Associate Professor Dr Shahid Hassan
Medical Education/Otorhinolaryngology Department, School of Medical Sciences, Universiti Sains Malaysia, Kubang Kerian, 16150 Kota Bharu, Kelantan, Malaysia
Email: shahid@kb.usm.my, gorshahi@yahoo.com