Education in Medicine Journal ISSN 2180-1932

ORIGINAL ARTICLE
Volume 4 Issue 2 2012
DOI:10.5959/eimj.v4i2.107
www.eduimed.com

Construct validity of postgraduate conjoint assessment of master of surgery program of School of Medical Sciences at USM

Shahid Hassan[1], Ahmad Fuad Abdul Rahim[1], Mohamad Najib Mat Pa[1], Mohd Nor Gohar Rahman[2], Muhamad Saiful Bahri Yusoff[1]

[1]Department of Medical Education and [2]Department of Surgery, School of Medical Sciences, Universiti Sains Malaysia

ARTICLE INFO

Received : 17/05/2012
Accepted : 30/08/2012
Published : 01/12/2012

KEYWORD

Construct validity
Reliability
Quality criteria
Assessment
Postgraduate
Medical education

ABSTRACT

Introduction: A clear concept and understanding about the measure and the measuring tools is essential for good practice of assessment. Assessors need to have information about the full range of assessment tools inclusive of psychometric validity and purpose of its use. Subjective inferences drawn from the readily available data as numbers of summative scores over the years and statistical evidences of reliability and validity of assessment tools used to measure student's performance are good sources of feedback for competent assessment program. It also provides meaningful evaluation of learning and teaching in medical education. **Method:** A retrospective study of 119 candidates was carried out to analyze the summative assessment scores of their certifying examination of Masters of Surgery in School of Medical Sciences (SMS) at Universiti Sains Malaysia. Subjective judgment of raw data followed by internal consistency as reliability, convergent validity and discriminant validity as constructs of individual assessment tool was analyzed. Finally each assessment tool as a measure of written or clinical construct was evaluated against six aspects of Messick's criteria for quality control. **Result:** The correlation coefficient for validity and Cronbach's alpha for reliability was evaluated for clinical measures. However, the test of internal reliability was not possible for essay being the only measure in written construct of summative assessment in surgery. All measures of clinical construct were found highly reliable with Cronbach's alpha between 0.962-0.979. Long case and the short cases have shown excellent correlations (r=0.959 at p<0.001). Viva stood on its own and showed good correlation with long case (r=0.933 at p<0.001) as well as with short cases (r=0.926 at p<0.001). The linear regression analysis of essay measure was not predicted by any of the clinical measure. In clinical construct long case was strongly predicted by short case and vice versa (B=0.640 at p<0.001). Viva was predicted by the long case only (B=. 245 at p<. 001). All measures have positively predicted the overall performance however, the long case predominantly more than the other measure of this construct (r^2=0.973 at p<. 001) **Conclusion:** Suggestions to improve the framework of assessment are proposed for future practice of competent assessment program in surgery.

CORRESPONDING AUTHOR: Prof Dr Shahid Hassan, Medical Education Department School of Medical Sciences, Universiti Sains Malaysia 16150 Kota Bharu, Kelantan, Malaysia
Email: shahid@kb.usm.my/gorshahi@yahoo.com

Introduction

Examinee's scores gathered over the years can be utilized as readily available data for statistical analysis to collect evidences on reliability and validity of each instrument in an assessment program. The outcome information can be utilized to review any assessment model for appropriate changes and adjustments both, for selection of new assessment tools as well as for setting up of the standard for taking logical decisions on pass or fail in summative examination. The numbers available as scores not only determine the reliability and validity of assessment but it also guide towards programming the assessment. However, understanding the principles and obtaining the statistical evidences are considered essential to support a change in the current practice of postgraduate assessment.

To clearly understand the concepts of good assessment practice, we need to have information about the entire range of assessment tools and purpose of its use. Subjective inferences drawn from the raw data available as numbers of summative scores and statistical evidences of reliability and validity of instruments can be used as measure of student's authentic performance (1). Individually driven decisions of faculty members for selection of instruments, setting up of standards for strategy and designing of assessment model are less likely to achieve the desired objectives of summative assessment. Every faculty member should be involved in the process of assessment program as much as they are involved in the training program of postgraduate studies. The teaching faculty must be well aware of the various aspects of assessment in a systemic manner such as:

1) Knowing the assessment for its fit between the object of assessment and the measurement tools,
2) Realizing the pitfalls of assessment tools and its impact on current practice,
3) Understanding the assessment to produce competent graduates,
4) Ensuring reasonable reliability of the assessment tools and methods to compute data,
5) Identifying the construct validity for evidence of statistically analyzed data especially when content and criterion validity is in question.
6) Setting up of standards for strategy and methods for a logical decision in summative assessment.

Construct validity of measuring tools in assessment is the most important of four categories into which the recommendations divide validity studies as: predictive validity, concurrent validity, content validity and construct validity (2) Construct validity must ensure that the desired construct is measured. Construct validity in assessment of student's performance is about how well a test or measurement tool measures up to its claim. A written test designed to measure the problem solving skills must measure all attributes of problem solving skills in a given learning domains. Construct validity ensures that the measurement tool in an assessment conforms to the theoretical or conceptual model being tested. Construct validity therefore refers to extend to which inferences can be drawn from the operationalization of theoretical concepts.

In this context construct validity is related to generalizing the outcome of a measure however, towards the same assessment in program. Construct validity essentially can be delivered through two of its subsets as convergent validity and discriminant validity. Convergent validity ensures that if the required theoretical concept predicts that the two measures are correlated are in fact related. Discriminant validity test that the measures that should have no relationship do in fact have no relationship. To estimate the degree to which any two measures are related to each other is determined by correlation coefficient, which in case of similar measures is high while in case of dissimilar measure it is low. Construct validity is needed whenever no criterion or content validity is available to define the measure as adequate in any given assessment.

Faculty's realization and awareness of assessment methods for optimal and logical decision is increasingly growing in postgraduate

training program to produce competence-based graduates. It has been recommended that one should not marry to a single assessment tool in an assessment design (3) to assign pass or fail on performance of candidates in summative assessment, which often is the practice in high stake examination. Assessment now demands practice of a combination of assessment methods in a competent assessment programming in which quality should be derived from both, subjective judgment and psychometrics evaluation of examinee's performance achieved as scores. To develop and evaluate an assessment design of postgraduate program in medical education, a set of criteria to determine quality are needed.

A psychometric framework to address the quality issues have been proposed by Messick (4, 5, 6) and complemented by another 10 point framework with addition to Messick's criteria is the competent assessment program (CAP) proposed by Liesbeth et al. (7). However, the present study will compare the outcome of summative assessment drawn as the subjective judgment and statistically analyzed evidences with six aspects of Messick's construct validity and only those from CAP, which are relevant to six aspects of Messick's criteria (see table 1). Emphasis to periodically review the summative scores provides substantial evidence for readjustments in assessment design to create competent assessment program in postgraduate training has been the author's experience. Periodic review of the raw data compiled as the numbers of summative scores for its subjective judgment and its analysis to achieve statistical evidences is a good practice for faculty development as well as programming the assessment (7). Current review is based on the same practice in which summative assessment scores of individual discipline is first analyzed by the medical educationist and later discussed with teaching faculty involving as many faculty members as possible. Author recommends a 2-day workshop to discuss all the evidences produced as the subjective judgment of the raw data followed by retrospective study of psychometrics for validity of individual measurement tool and its pitfall that subsequently can suggest replacement or addition of an assessment tool if required (8).

Messick's six aspects of quality control criteria provides good framework of construct validity, which can readily be utilized to emphasize the psychometric evaluation of quality control issues in assessment. Messick's concepts of construct validity organized to incorporate content, construct and criterion aspects of validity also include the idea of consequential validity to help determine the effect of assessment on education. This framework can be used in a befitting manner to compare the validity evidences acquired form the summative scores and to suggest new model based on traditional as well as performance based assessment for the future practice. The six aspects of Messick's construct validity include content, substance, structure, generalizability, consequences and externality. The content aspect of this construct validity framework takes into account the competence assessment of knowledge, skills and attitude. The substantive aspect adds the need for thinking process analyzed during the assessment as reflective of the process needed in real life situations by the physicians. The structural aspects of construct validity of the framework are about the fidelity of the summative scoring used for the assessment decision, which should be consistent with the structure of construct domain that is competence.

The generalizability aspect of the framework describes the correlation with other instruments representing the same construct determined across time, occasions and observers. The external aspects of construct validity relates to inter-correlation of the scores obtained in different measurement tools of the same construct and other construct. In this respect assessment tools scores of the same construct will show high correlation with another measures of the same construct (convergent validity), whereas assessment tools of irrelevant construct will show low correlation (discriminant validity). The consequential aspect applies to positive and negative, intended or unintended consequences of assessment procedure on summative result

particularly and on learning and teaching in general.

The current concepts of validity in many reviews have argued validity beyond its psychometric properties to further operationalize its practical use. Two approaches are distinguishable to clarify the concepts of validity for practical use. First approach (9, 10) depends on sources of evidence to demonstrate validity, which is often referred to construct validity and psychometric analysis of summative scores. It provides such an evidence, however the question may arise when Kane (11) describes validity as: " Do the scores yielded by the procedure supply the kind of information that is of interest and are these scores helpful in making logical decision?" Second approach (12) based on this question proposes quality criteria, which is helpful in identifying the issues that deserve attention in validation and clarify how an individual assessment may relate to more global issues of construct validity for a logical decision in assessment comprising of multiple measures. Messick S. (4, 5, 6), Liesbeth et al. (7), Linn et al. (13) and many other authors have described quality criteria of construct validity. Messick's criteria have been chosen to compare the validity evidences collected in this study with aspects of quality criteria described by Messick as well as CAP.

A point to be emphasized in this study is to explore the logic of interpretation of score based on one measurement tool with specific learning domain to combine with another assessment tool of the similar learning domain in a construct to improve generalizability. This may well elaborate the outcome of similar learning domain under assessment such as one best answer MCQ format and extended matching questions added to essay question in written construct of the assessment. The generalizability effect is an important aspect, which is explored in this study to combine different measure with good internal consistency for interpretation of the performance of same construct. For example long and the short cases as the measures of the clinical construct combined in a compensatory approach. To analyze the external consistency of the construct validity, Messick's criteria again provides good framework to determine the relationship between the scores of the measurement tools of the same construct compared to constructs of assessment design, which explores its convergent and discriminant validity. Validity is not just a matter of assessing the right construct but increasingly it pertains to actual and correct use of assessment instruments (6). This essentially is the need of a competent assessment program as well as the desire of an assessor who is keen to practice quality assessment.

Method

This is a retrospective study of summative scores of 119 candidates who underwent exit examination of Masters of Surgery in School of Medical Sciences (SMS) at Universiti Sains Malaysia (USM) during 2006-2012 under the Conjoint Board of Examination of three universities (UM, UKM, USM). A subjective judgment of data was initially made (see table 2, 3 and 4) to evaluate the appropriateness of decision compatible with principles to employ multiple tools and their combined effect for a more precise quantitative as well as qualitative judgment guiding to a logical decision in summative assessment.

All the scores of 119 candidates were included in this study of surgical discipline of School of Medical Sciences at USM however, only the USM results of conjoint summative examinations were analyzed. Collection of data comprising of results of all measurement tools as individual, as combined measures and as overall total to predict the outcome performance was carried out. Data collected was analyzed for validity evidence of test scores for correlation using Spearman's correlation coefficient (R), linear regression as coefficient of determination (R^2) and predictive values using unstandardized and standardized (Beta) coefficient of variables (see table 6) and their overall performance.

Finally each measure as an individual assessment tool and as a component of written or clinical test was evaluated against six aspects of Messick's

criteria framework. The Messick's aspects of construct criterion however, is subjected to readjustments derived from competent assessment program (CAP) criteria proposed by Liesbeth et al. to fit in the need, which incorporate the analogues and differences of the two criteria (see table 1).

Table 1: Quality criteria of construct validity by Messick and CAP utilized here to evaluate various aspects of validity of master of surgery assessment

No	CAP Link of Criterion	Messick's Aspects	Analogies of Messick with competent assessment of performance (CAP) Quality Criterion
1	Authenticity	Content	Messick's criterion prescribes task to include knowledge, skills and attitude. however, CAP's addition to this criterion is the integration of assessment and importance of work environment and social context
2	Cognitive complexity	Substantive	Mesick criteria includes measurement of thinking process while CAP's ensures cognitive complexity during assessment is more or less the same
3	Fairness	Structural	Messick's criteria emphasizes for fidelity of scoring consisted with structure of construct domain. CAP's adds recognition of individual difference of learners with fair chance given to all across the content to demonstrate competence
4	Reproducibility	Generalizability	Messick describe it as correlation with other measure of construct determined across time, occasion and observers. The interpretation of score of one measure should generalize to other tasks in a domain specific construct. CAP focus on combining information sources instead of comparing different tests
5	Education Consequences	Consequential	Messick's consequential to assessment focus on positive and negative effects of assessment on teaching and learning. CAP insist on positive effects than just an effect
6	Comparability	External	Messick external aspect applies to relationship of scores of different measures of the same construct and other construct. For CAP it is a prerequisite for good generalizability

Result

The eyeball judgment of the written assessment suggests that a close marking scheme instead of full range of marks (0-100) is actually practiced, which showed that 91 (77.11%) candidates have been judged between 50-59.9% marks. This can be argued for a very subjective marking (see table 2). Almost similar pattern have been observed in marking the clinical components of long and short cases (see table 3) and assessment by viva (see table 4). This has led to incompatible results of written and clinical measures in which passing rate for essay test is 84.74% vs. clinical 39.80% to reflect overall passing of 27.11% and border line (49.5%-49.9%) failures of 10.16% (see table 5) who were not compensated though the assessment has been so subjective throughout measures of each construct.

The correlation coefficient for construct validity and Cronbach's Alpha for reliability of essay in written construct was not possible due to single tool practiced in summative assessment of surgery. However, all measures of clinical construct were found highly reliable with Cronbach's Alpha between 0.962-0.979 (see table 6). Principle component analysis extracted the measures as 3 factors of similar domains of construct (see table 7). Principle component analysis extracted long and short cases measures as one model while viva component seen on forced extraction established as another model of assessment similar to essay component (see table 7). Long case and the short cases have shown good correlations (r=0.959 at p<0.001). Viva stood on its own and showed fair correlation with long case (r=0.933 at p<0.001) as well as with short cases (r=0.926 at p 0.055). Essay questions expected were poorly correlated with long case (r=0.637), short case (r=0.646) and viva (r=0.640, see table 8). Viva though stood as a component in principle component analysis showed good correlation with long case (r=0.933) and with short cases (r=0.926, see table 8). Correlation of individual instruments with overall performance was noted significant for all measurement tools with strongest correlation coefficient shown for long case assessment. In linear regression analysis, essay as a measure of written test was not predicted by any of clinical measure for obvious reasons. However, measures in clinical construct were highly predicted by each other like the correlation coefficient. In clinical construct long case is strongly predicted by short case and vice versa (B=0.640 at p<0.001). Viva though fairly well, is predicted by the long case only (B=. 245 at p<. 001). All measures have positively predicted the overall performance however, the long case predominantly more than the other measures of this construct (r^2=0.973 at p<. 001, see table 9).

Attempt to estimate borderline marks by calculating the cut off point and discrimination abilities under ROC curve of individual tool were also determined by ROC analysis (see table 10). Discrimination power of individual tool observed as area under ROC curve was significant at < .001 for all four instruments.

Table 2: Outcome distribution of students in varied range of marks secured in written (essay) component of clinical examination (n=118)

No	Range of marks secured (%)	Range of marks secured (out of 40%)	Students within the range, n (%)
1	20-29.9	08-11.9	00 (0%)
2	30-39.9	12-15.9	2 (1.69%)
3	40-49.9	16-19.9	I5 (12.71%)
4	50-59.9	20-23.9	91 (77.11%)
5	60-69.9	24-27.9	9 (7.62%)
6	70-79.9	28-31.9	1 (0.84%)
7	80-89.9	32-35.9	00 (0%)

Table 3: Outcome distribution of students in varied range of marks secured in clinical component (long case and short case) of clinical examination (n=103)

No	Range of marks secured (%)	Range of marks secured (out of 40%)	Students within the range, n (%)
1	20-29.9	08-11.9	00 (0%)
2	30-39.9	12-15.9	00 (0%)
3	40-49.9	16-19.9	42 (40.77%)
4	50-59.9	20-23.9	66 (64.07%)
5	60-69.9	24-27.9	7 (6.79%)
6	70-79.9	28-31.9	3 (2.91%)
7	80-89.9	32-35.9	00 (0%)

Table 4: Outcome distribution of students in varied range of marks secured in clinical component (viva) of clinical examination (n=103)

No	Range of marks secured (%)	Range of marks secured (out of 20%)	Students within the range, n (%)
1	20-29.9	04-5.9	00 (0%)
2	30-39.9	6-7.9	00 (0%)
3	40-49.9	8-9.9	56 (54.36%)
4	**50-59.9**	**10-11.9**	**49 (47.57%)**
5	60-69.9	12-13.9	3 (2.91%)
6	70-79.9	14-15.9	1 (0.97%)
7	80-89.9	16-17.9	1 (0.97%)

Table 5: Students failed with borderline marks of 49.5% and above in clinical component with over all aggregate of 50% and above marks

Total Stdts.	**Written Test (Theory-Essay) N=119**		**Clinical Test (L/Case S/Case, Viva) N=103**		**Over all Pass/Fail Rate N=119**	
	Total Pass	Total Fail	Total Pass	Total Fail	Pass	Fail
119	**100 (84.94%)**	18 (15.25%)	**41 (39.80%)**	62 (60.19%)	**32 (27.11%)**	86 (72.88%)
Borderline Failed Students (49.5 - 49.9)						
12 (10.16%)						

Table 6: Reliability as internal consistency between the measures of same construct Cronbach's Alpha

No	Correlative Number of Items = 2	Cronbach's Alpha
1	Essay the only measure in written construct	Cannot be analysed
2	Long case and short cases	0.979
3	Long case and viva	0.965
4	Short cases and viva	0.962

Table 7: Principle component analysis shown to have extracted the measures as 3 factors of similar domains of construct.

	2 factors extracted		3 factors extracted		
Assessment tool	Factor		Factor		
	1	2	1	2	3
Essay	0.079	0.309	0.090	0.316	0.053
Long case	0.820	0.311	0.808	-0.145	-0.042
Short case	0.807	0.563	0.815	0.142	-0.034
Viva	0.288	0.102	0.297	-0.093	0.194

Table 8: Correlation between assessment tools with each other and with overall performance of the candidates

Variable	**Pearson's (r)**	**Essay**	**Long Case**	**Short Case**	**Viva**
Essay	Coefficient p-value	1.000	0.637 < 0.001	0.646 < 0.001	0.640 < .001
Long Case	Coefficient p-value		1.000	0.959*** < 0.001	0.933 < 0.001
Short Case	Coefficient p-value			1.000	0.926 < 0.001
Viva	Coefficient				1.000

Table 9: Individual instrument predicted by other instrument/s and the overall performance predicted by each instrument

Outcome **Predictor**	**Unstandardized B**	**95% CI for Unstandardized B**	**Standardized B**	**R^2**	***P* value**
Essay		No variable predicts the essay question			
None					
Long Case		Constant = 4.139			
Short Case	0.617	(0.468, 0.766)	0.640	0.409	< 0.001
Short Case		Constant = 3.225			
Long Case	0.644	(0.503, 0.824)	0.640	0.409	< 0.001
Viva		Constant = 8.057			
Long Case	0.197	(0.040, 0.354)	0.245	0.060	< 0.001
Overall Score		Constant = 1.593			
Essay	0.966	(0.894, 1.037)	0.973	0.973	< 0.001
Long Case	0.934	(0.823, 1.045)	0.373		< 0.001
Short Case	0.944	(0.837, 1.045)	0.335		< 0.001
Viva	1.045	(0.937, 1.154)	0.516		< 0.001

Table 10: Borderline marks estimated and discrimination abilities determined under ROC curve

No	Assessment tool	Borderline Marks estimated by ROC Analysis			Discrimination
		Cut off Point	Sensitivity	Specificity	Area under ROC curve (*p*-value)
1	Essay	20.9 (Out of 40)	87.5%	78.9%	0.88 (< .001)
2	Long Case	9.1 (Out of 20)	62.5%	73.7%	0.85(< .001)
3	Short Case	9. 7 (Out of 20)	85.0%	89.5%	0.89(< .001)
4	Viva	9.7 (Out of 20)	63.8%	57.9%	0.71(< .001)

Discussion

Multiple assessment tools are employed in summative examination of Master of Surgery, which is held twice a year under the Conjoint Board of Examination of three universities, namely UM, UKM and USM. Candidate's performance is evaluated both in written and clinical constructs with multiple measures and a mixed compensatory-conjunctive approach is employed for logical decision on pass and fails using a clear standard of 50% and above to declare pass. Performance of continuous assessment by each university is considered as the prerequisite to sit the summative assessment in written examination. Written examination is held earlier and those who pass this component are only allowed to sit the clinical examination. The overviews of seven years of summative scores of 119 candidates represent two examinations held every year in May/June and October/November. In written examination only one assessment tool, which is essay question format as two papers are utilized in written construct and it is mandatory to pass this component to sit the clinical test. The clinical test uses three assessment tools that are long case, short case and viva (oral). There are two long cases, which are observed through the workup by the respective assessors however, candidates are not marked exclusively for their observed performance and rather a holistic scoring rubric accommodates those observed professional behaviours? There are three short cases and four viva sessions (2 stations on principles of surgery, one station on surgical pathology and one station on operative surgery) with different panels of examiners.

The questions in viva are randomly asked and no structured format is followed. All the clinical examinations are supervised by a number of examiners set out to represent multiple panels in order to accomplish the examination in allotted time, which usually has big turnout of candidates from three conjoint universities.

It has been observed with concern that the written test for the assessment of knowledge is obtained by using only one instrument which is essay question format, held as two different papers, comprising of two extended essay questions and five short notes (restricted essay questions) in each of the two papers. There is no demarcation of contents or the subjects for each paper. Used as single assessment tool, essay test raises the issues of content specificity and authenticity of evaluation of student's cognitive performances across the entire content undertaken in four years of structured training.

Besides, arguments on reliability, validity, standardization and generalizability of assessment of written construct by one method are open to many questions. The subjective evidence to this concern is provided by the eyeball judgment of summative scores of student's performance in written (see table 2) versus clinical evaluation (see table 2 and 3). The score obtained are not reflective of a realistic assessment of the cognitive performance and is evident from the scoring pattern and the failing/passing rate observed in two assessment tools that is essay versus long and short cases and viva respectively (see table 2, 3 and 4). Marking scores though range from 0-100 are confined to 50-60% in written measure and 40-69% in clinical measures, which reflects unrealistic marking not without subjective bias.

Point of concern is the candidature of 12 (10.16%) borderline candidates that is those scoring between 49.5%-49.9% and not 45-49.9% marks claimed by 51 (43.22%, see table 5) candidates. The candidates were not considered for assessment on qualitative basis though the assessment has been very subjective through the entire range of measures used in each construct and this is arguable. A good standard setting strategy that allows qualitative evaluation and triangulation with formative assessment (14) could have addressed the issue of borderline candidates to get the benefit of doubt in an assessment, which uses highly subjective assessment tool for written test.

Extended essay questions (so called long essay) were based on clinical scenarios, which are generally structured well to test student's analytic clinical thinking and problem solving

skills though within the constraints of the time allowed and the feasibility granted to mark these questions. Extended essay questions are marked by two different examiners who are not provided with structured marking scheme and the process is time consuming and lack standardization (15). The so called short notes (restricted essay questions) can also be questioned for its reliability and context validity since the direct questions on selected topics are not indicated as to how detailed an answer is required or what aspect of the topic has to be answered. Carefully setting up of the restricted essay questions and training of assessors to use a systemic marking schedule can improve reliability (15).

However, overemphasis in restricted essay questions should be avoided as it may lead to fragmentation and trivialization of question. The advantages lie in their flexibility of response, creativity and lateral thinking. But the basic strategy of restricted essay question is a well-defined subject, which invites candidate's response in declared time for each such answer. A good alternative may be that the extended question is replaced by scenario-based multiple short essay questions (SEQ) or modified essay questions (MEQ), key feature questions/problem (KFQ/P) or added to it another format of interpretive objective item of extended matching questions (EMQ) to test candidate's problem solving abilities. Although these are reasonably valid and reliable, they are time consuming to produce and in addition require good faculty development in writing these formats of questions besides, large numbers needed to be written. One best answer (OBA) as objective multiple-choice questions may be another option to add on to essay questions to improve the reliability of written assessment.

Statistical evidences on reliability was obtained by estimating Cronbach's Alpha to establish the internal consistency in clinical construct**,** however it was not possible in written construct of current practice of surgery in which essay question format is the only instrument practiced in the assessment method. It does not fulfill the assumption to calculate the Cronbach's Alpha, which is not only a matter of concern for being unable to practice multiple instruments for a valid assessment but also for making evaluation of program impossible for reliability. Developing a theoretical concept of reliability is important (discussed above) before it is operationalize by Cronbach's Alpha and further explored by extracted matrix in principle component analysis. A Cronbach's Alpha of 0.8 and above is considered good for reliability (16).

Utilizing at least two assessment tools for any construct is well achieved in case of clinical assessment in which all three measures (long case, short case and viva) are analyzed for internal reliability in pair. The analysis showed good internal reliability with Cronbach's Alpha between long case and short cases, between long case and viva and between short case and viva (see table 8). Looking at the principle component analysis, essay and rest of the three measures (long case, short case and viva if at all it is clinical by practice) were indicated to be the separate models. Nevertheless analysis of viva scores showed fairly good correlation with long case and short cases is further verified by extraction matrix (see table 7), which showed long case, short cases and viva as one model and essay as a different model. This suggests that the long case and short cases at least can be considered for compensatory approach conveniently in the assessment of clinical construct.

However, viva can be conjunctively combined with rest of the two clinical measures in same construct (clinical) and essay in a separate measure of written construct. When high stake decisions are based on measure of different construct and different measure of same constructs, a mixed approach involving conjunctive and compensatory approach should be employed (17). For the validity a theoretical conceptual model of assessment should be able to operationalize well to achieve good face validity and construct validity (6). Statistically construct validity is analyzed by conjunctive and discriminant validity as correlative coefficient (R) and bivalent linear regression of coefficient discrimination (R^2) along with standardized (B) with significant p value at <0.05. In current

analysis Pearson's correlation coefficient suggested excellent correlation between long case and short cases, good correlation between long case and viva, viva and short case and fair correlation between essay and rest of all three measures of clinical construct (see table 8). It was also seen on principle component analysis in which long case and short cases, essay and viva are extracted in separate models. Bivalent linear regression coefficient (R^2) as expected has been the same as Pearson's correlation (r) with long case predicted well by short cases and vice versa and viva predicted by long case. However, coefficient determination (R^2) suggested strong prediction of long case by short cases, which is also explained by the slope of regression line (un standardized B) both of which are significant (see table 8).

ROC analysis for determining cut off marks was indicated for long case, short cases, essay and viva (see table 10) based on standard setting method of norm reference. But the analysis may suggest the importance to address the issue of borderline candidates in a summative assessment of high stake decision that predominantly comprise of a number of subjective measurement tools. Qualitative evaluation in addition to quantitative assessment therefore becomes necessary to be considered for a logical decision in certifying examinations in postgraduate medical education. Best of all observed under the ROC is the short case for its discriminative ability between good and the poor performing students in order to determine the cut-off point for deciding on passing marks. Keeping optimal sensitivity and specificity in view, the long case is found to have 9.1 out of 20 marks (see table 10) and for short cases 9.7 out of 20 marks (see table 10). However, viva was established not to be a good discriminating tool under ROC line (see figure above). Students for declaring pass as logical standard setting method in assessment program of general surgery currently practiced, short cases showed best discriminative ability under ROC line followed by the essay component of assessment in this analysis (see table 10).

Showing poor correlation, essay is not analyzed correlating with any of the clinical measures. Long and short cases have shown good correlation and can be considered for compensatory approach in logical decision making on pass or fail. However, long case and viva and short case and viva have shown correlation, which is fairly ok. It has been recommended to consider viva for a conjunctive approach towards logical decision on pass or fail. Essay was not predicted by any of the clinical measures and it was confirmed on principle component analysis as well in which essay stood alone as extracted separate model in the matrix. Good convergent validity was observed for long case and short cases while essay and viva showed high discriminant validity with moderate correlation between them.

Current assessment tools versus Messick's and CAPs aspects of quality criteria

The content validity, construct validity and predictive validity of assessment tools in a conceptual theoretical model must meet operationalization in which structural match of criteria with construct is also measured**.** Evaluation to compare the outcome analysis of measures in summative assessment of surgery against the six aspects in Messick's with comparable analogies and differences advocated by competent assessment program (CAP) proposed by Liesbeth et al. was also carried out. This comparative study as a quality control exercise for construct validity yielded interesting finding as following.

1. Content aspect:

Messick's criterion of validity on content aspect prescribes task to include knowledge, skills and attitude. However, as per CAP's recommendations to this criterion is the integration of assessment and importance of work environment with social context, called authenticity. It expects to reflect encounter as realistic as possible. The structured extended essay question with clinical scenario and real patient with observed long case and short cases do encompass authenticity. However,

unstructured restricted essay the so-called short notes and unstandardized viva can be argued for content aspect. Scenario-based short essay and structured viva with visual exhibits can improve the validity. Both concepts in current assessment are unintentionally overlooked in its basic principle and selection of assessment tools for authenticity of assessment towards competent assessment program. Knowledge is tested by a single measure of essay questions in written construct, which has its own pitfall of content specificity, reliability, generalizability and standardization.

Authenticity and reliability can be improved by adding another assessment tool however, the one, which is more objective and content valid besides, introducing the structured marking scheme across the entire range of marks allocated for each question. This will improve the standardization of assessment of essay question. Clinical skills though evaluated through two long cases, three short case and three viva sessions are not compatible with training in workplace-based assessment. This can be improved by structuring the long case and viva and by observing the long case for student's performance of medical professionalism, attitude and organizational efficiency while he is undertaking the patient for workup before reflecting before the panel of examiner for cross-examination. Otherwise an unobserved long case becomes test of "knows how" then "shows how" for competent assessment as per Miller's criteria.

2. Substantive aspect:

Messick criteria insist to include measurement of clinical thinking process as it is used by practitioners in the field while CAP looks at it as the cognitive complexity ensured during assessment relevant to the level of training. Both criteria Messick and CAP demand analysis of critical thinking and problem solving skills. Scenario based essay and long and short cases do provide opportunity to test analytic clinical reasoning and problem solving skills. Haphazard questions that lack context specificity in viva and ambiguous short notes are argued for lack of substantive aspect. Scenario based structured viva and scenario-based short essay question (SEQ) will promote assessment of thinking skills.

Cognitive complexity of CAP criteria, which includes measurement of thinking process as substantive link to Messick's aspect as well as ensuring the inclusion of all knowledge, skills and attitude to be measured though fairly accounted for, may also be argued. Essay question, particularly extended essay are recently modified to scenario based real clinical cases with sufficient input to test analytic clinical reasoning skills on complex cases. However, content validity remains in question with two extended and five unstructured restricted essay questions. On the other hand, clinical assessments are carried out with varied cases of comparatively simple to highly complex cases produced by the center of the venue. Standardized test of clinical performance is compromised by a lottery draw among the candidates, raising questions on inter-case reliability. This can be improved by structuring the marking scheme to accommodate handicaps marks for complex cases (1)

3. Structural aspect:

Messick's criteria emphasizes for fidelity of scoring consisted with structure of construct domains as competence. All knowledge, skills and attitude matching with the construct have to be measured. CAP's adds recognition of individual differences between learners, which it links to fair chance given to learners to demonstrate their competence across the content. Emphasis is to cover entire domain. The lack of scoring scheme in essay and unrealistic use of close-marking system with notion of global rating is not well defined for its structural aspects to all the assessors. Marks given for poor, average, good or very good may sound different to different assessors involved unless meaningful elaboration of these terms are well announced. A format with elaboration of structured performance consistent with score in each construct is needed.

4. Generalizability aspect:

Messick describe it as correlation with other measure of construct determined across time, occasion and observers. The interpretation of score of one measure should generalize to other tasks in a domain specific construct. CAP refers this to reproducibility of decision in which outcome of the assessment should be able to apply to other setting and task. Generalizability is increased when large sample across content is used. This provides warranty to check consistency of candidate's performance across cases, assessors and time across the summative assessment. This implies to know the inter-rater and inter-case difference in long and short cases and four viva sessions. A statistical workup will be required to provide evidences to see implementation of this aspect. However, it requires obtaining individual assessors marks. Assessor's calibration and training will improve the situation.

Reproducibility in Messick's criteria looks at generalizability, standardization and internal consistency of assessment tools involved (3) while CAP focus on combining information sources instead of comparing different tests (6). The Messick's concern of increasing reproducibility and reliability has been violated in many aspects for generalizability and inter-rater reliability. Multiple panel of examiners though a constraint of huge number of candidates in each session of summative assessment, varied complexity of cases and lack of structured marking scheme are not compatible with Messick's criteria of increasing reproducibility.

The CAPs concepts of combining information sources instead of comparing different tests needs appropriate standard setting strategy, which has been subjectively judged to be illogically implemented in clinical construct. Long and short cases shown to be strongly correlated and predictive of each other could have been easily be done with compensatory approach towards decision making and altered the result of those 12% borderline candidates (see table 5) to make it a logical decision. Statistical analysis of summative scores of examination in surgery strongly recommend that the decision on long and short cases should take compensatory approach towards overall grade rather than conjunctive approach for standard setting strategy in clinical construct that affects the outcome of result.

5. Consequential aspect:

Messick's consequential effects of assessment focus on positive and negative effects of assessment on teaching and learning. CAP also explores the impact of assessment on educational outcome. However, the competent assessment program insists on positive effects than just an effect. To achieve this aspect relevant assessment tools should be purposefully used to guide learning. Effects of measurement tools on acquisition of competence in learning process in written test may not motivate students to prepare with know and know how across the content. However, it is well achieved with self-directed clinical learning to practice independent patient's workup. This promotes knowledge, skills and attitude. Structured viva can also motivate similar learning.

To evaluate this criterion need evaluation of training program for concurrent and predictive validity. Stakeholder's feedback from the institution that the passed out trainees will serve during and after their gazetment will be valuable for such evaluation. Issues identified from those feedbacks can be addressed to bring positive effects of educational consequences. However, a process of consistent feedback on graduating surgeon's performance in real world would be mandatory for improving the educational consequences with positive outcome.

6. External aspect:

Messick's external aspect as criteria of quality assessment and construct validity applies to relationship of scores of different measures of the same construct and other construct. Messick's comparability has recommended multitrait-multimethod analysis. For CAP it is a prerequisite for good generalizability, in which relationship between different measures reflects

the competence generalization. It also guides to design standard setting strategy. Correlation and prediction provided evidence of high correlation between long and short cases while poor correlation between essay and clinical measure, which is relevant. Viva correlates fairly well with both and indicates session a mixture of recall of knowledge-based versus demonstration of skills based questions. To have logical good correlation with clinical measure, viva should have problem solving real patient scenario shown as slides, videos and relevant exhibits.

Conclusion

It has been recommended to employ multiple measures in high-stakes decision in postgraduate assessment of medical education. However, for a logical decision in summative assessment of a certifying examination, the tools used must find a good fit of subjective as well as objective measures and a framework to guide the selection of approaches to combine those multiple measures.

A well-established compensatory and conjunctive approach for combining measures should ensure quantitative as well as qualitative measures to make logical decisions appropriate for higher education. Written construct assessed with single tool of essay questions, which is not followed by the standard marking scheme raise the question of reliability as well as content validity of summative assessment in surgery. However, clinical assessment with multiple measures has been encouraging for its statistical evidence of high internal reliability. Two observed long cases with different set of examiners provide fair opportunity to candidates to demonstrate competence. Evidence also suggest that validity of assessment can further be improved by combining these two measures with good correlative and predictive values in a compensatory approach for making logical decision on pass and fail in summative assessment in surgery.

To improve the current assessment framework, changes are imperative to follow the set of principles to guide a test culture of competent assessment program. It is recommended to continue the current practice of essay question format however, with addition of another more objective assessment tool such as one best answer (OBA) or key feature question/problem (KFQ/P) in written construct. A compensatory approach is strongly recommended to combine the scores of long and short cases while maintaining a conjunctive role for viva towards a logical decision in a well-observed standard-setting strategy to be practiced in the discipline of surgery.

Acknowledgement

Authors would like to thank Research Committee of School of Medical Sciences, USM for allowing the short-term grant and Ethic Committee for allowing analysing the scores of summative assessment of School of Medical Sciences, USM obtained from the conjoint examination during 2006-2012.

Reference

1. Hassan S, Mat Pa MN and Yusoff MSB. Discriminant and convergent validity of measurement tools in postgraduate medical education of a surgical-based discipline: Towards assessment program. Education in Medicine Journal, 2011; 3 (2), e1-e18
2. Anastasi A. The concepts of validity in the interpretation of test scores. Educ. Psychol. Measmt. 1950; 10, 67-681.
3. Van der Vleuten CPM. (1996) the assessment of professional competence: theoretical developments, research and practical implications. Advances in Health Sciences Education, 1, 41-67
4. Messick S. The psychology of educational measurement. Educational Measurement, 1984; 21, 215-237.
5. Messick S. The interplay of evidences and consequences in the validation of performance assessment. Educational Researchers, 1994; 23, 13-23
6. Messick S. Validity of psychological assessment: Validation of inferences from persons' response and performances as scientific inquiry into score meaning.

American Psychologist, 1995; 50, 741-749
7. Baartman LKJ, Bastiaens TJ, Kirschner PA, Van der Vleuten CPM. Evaluating assessment quality in competence-based education: A qualitative comparison of two frameworks. Educational Research Review 2007; 2 114-129
8. Hassan S. Postgraduate Assessment in Malaysia: Rationale of Decision Making. Education in Medicine, 2011; 3 (1): e1-5.
9. Hassan S. Assessment of postgraduate Program in Otolaryngology and Head-Neck Surgery in Malaysia – Are we Adequate. MSO-HNS News Bulletin, Vol. 4, Issue 1, May 2011.
10. Kane MT. Current concerns in validity theory. Journal of Educational Measurement, 2001; 38, 319-342.
11. Shepard LA. Evaluating test validity. Review of Research in Education, 1993; 19, 405-450.
12. Kane MT. Certification testing as an illustration of argument-based validation. Measurement: Interdisciplinary Research and Perspectives, 2004; 2, 135-170.
13. Crooks TJ, Kane MT. Threats to the valid use of assessments. Assessments in Education: Principles, Policy & Practice, 1996; 3, 265-285.
14. Linn RL, Baker J, Dunbar SB. Complex performance–based assessment: expectations and validation criteria. Educational Researcher, 1991; 20, 15-21.
15. Jackson N, Jamieson A, Khan A. Assessment in medical education and training. Oxford: RadcliffePublishing, 2007.
16. Bryman A, Cramer D. Quantitative data analysis with SPSS release 10 for windows: Concepts and their measurement. Routledge Taylor and Francis Group Publication, Hove and New York, 2003; 55-68
17. Mitchell DC. Multiple measures and high stakes decisions: A framework for combining measures. Educational Measurement Issues and Practice, 2003; 22, 2.