Education in Medicine Journal
2012, VOL 4 ISSUE 1
DOI:10.5959/eimj.v4i1.2

ORIGINAL ARTICLE

Discriminant and convergent validity of measurement tools in postgraduate medical education of a surgical-based discipline: Towards assessment program

Shahid Hassan, Mohamad Najib Mat Pa, Muhamad Saiful Bahri Yusoff

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

Abstract

Background: Summative assessment in postgraduate examination globally employs multiple measures. A standard-setting method decides on pass or fail based on an arbitrarily defined cut-off point on a test score, which is often content expert's subjective judgment. Contrary to this a standard-setting strategy primarily practices two approaches, a compensatory approach, which decides on overall performance as a sum of all the test scores and a conjunctive approach that requires passing performance for each instrument. However, the challenge using multiple measures is not due to number of measurement tools but due to logic by which the measures are combined to draw inferences on pass or fail in summative assessment. Conjoint University Board of Examination of Masters' of Otolaryngology and Head-Neck Surgery (ORL-HNS) in Malaysia also uses multiple measures to reach a passing or failing decision in summative assessment. However, the standard setting strategy of assessment is loosely and variably applied to make ultimate decision on pass or fail. To collect the evidences, the summative assessment program of Masters' of ORL-HNS in School of Medical Sciences at Universiti Sains Malaysia was analyzed for validity to evaluate the appropriateness of decisions in postgraduate medical education in Malaysia. **Method:** A retrospective study was undertaken to evaluate the validity of the conjoint summative assessment results of part II examination of USM candidates during May 2000-May 2011. The Pearson correlation and multiple linear regression tests were used to determine the discriminant and convergent validity of assessment tools. Pearson's correlation coefficient analyzed the association between assessment tools and the multiple linear regression compared the dominant roles of factor variables in predicting outcomes. Based on outcome of the study, reforms for standard-setting strategy are also recommended towards programming the assessment in a surgical-based discipline. **Result:** The correlation coefficients of MCQ and essay questions were found not significant (0.16). Long and short cases were shown to have good correlations (0.53). Oral test stood as a component to show fair correlation with written (0.39-0.42) as well as clinical component (0.50-0.66). The predictive values in written tests suggested MCQ predicted by oral (B=0.34, $P<0.01$) and essay predicted by long case (B= 0.23, $p<0.01$). In clinical components long case predicted by oral (B= 0.71, $p<0.05$) and short cases predicted by long case (B= 0.31, $p<0.001$). **Conclusion:** The recorded discriminant and convergent validity evidences conclude that MCQ and essay do not correlate, nor do they predict each other. Long case and short cases significantly correlate with each other however, short cases are predicted by long case. All components though predict the overall performance, long case has the dominant role. The study outcome provides enough evidence to reconsider role of quantitative as well as qualitative evaluation in high stake examination of surgical-based discipline of ORL-HNS.

Keywords

Discriminant validity, Convergent validity, Assessment tools, Postgraduate, Surgery, ORL

How to cite this article?

Hassan, S., Mat Pa, M.N., & Yusoff, M.S.B. (2012). Discriminant and convergent validity of measurement tools in postgraduate medical education of a surgical-based discipline: Towards assessment program. *Education In Medicine Journal*, *4*(1). DOI:10.5959/eimj.v4i1.2

Introduction

The assessment designs in postgraduate examinations have increasingly been using multiple measures as a policy matter universally**.** The reasons that support the use of multiple measures include**.** 1) Precluding dissonance of interpretation. 2) Reducing the measurement error. 3) Providing representative construct of teaching. 4) Allowing fairer opportunities to demonstrate competence. 5) Meeting the objective of assessing all learning domains using relevant measurement tools. Standard setting strategy looks into the issue beyond the standard setting methods. Standard setting strategy has become significantly important in recent years, particularly in postgraduate assessments in which institutions often have no clear policy or consensus at interdisciplinary level like undergraduate medical education. Many disciplines have adopted increasingly complex criteria where as others have followed the traditionally practiced criteria for decision making in summative assessments.

Traditional methods are highly subjective that might lead to monopolize assessment decision. As a consequence of which both candidates and community may suffer. Often the decision made on safe vs. unsafe outcome of a candidate's performance has no documented record for a validity evidence work up. Subjective decision made on one or two responses on critical questions seems to have no defined weightage or structured assessment protocol for factor analysis or correlative studies as a single or multiple linear regression tests to explore discriminate or convergent validity of measurement tools used.

Long case assessment is the most criticized instrument for evaluation of clinical competence in summative examinations in postgraduate medical education for its reliability, validity, standardization and decision-making (1).

Standard setting methods decides on passing or failing of a candidate is based on a cut-off point on a test score, which is predefined. Method used to decide a passing score is based on content expert's subjective judgment, which may be arbitrary. Contrary to this a standard setting strategy primarily practices two different approaches selected as a policy matter. However, to make logical decision in high stake professional licensing examinations in medicine and to avoid its untoward consequences on candidates and community, validity evidences are required.

This becomes even more important in postgraduate examination that uses a battery of tests to cover different aspects of learning domains and multiple items or traits in their content structure, which requires analytic judgment to decide on summative assessments. The compensatory strategy focuses on overall performance, which is a sum of all the test scores. The conjunctive strategy requires passing performance for each instrument in summative assessment list**.**

The need of using multiple measures has also become necessary to meet the accreditation recommendations for quality medical education. However, the challenge using multiple measures is not due to number of measurement tools but due to logic by which the measures are combined to draw inferences that determine the accuracy and appropriateness of decision in high stake examinations (2).

Principles to guide those several approaches that are claimed to represent multiple measures in a standard–setting strategy to reach high stake decisions are as under**.**

1) Conjunctive Approach: Requires the attainment of minimal standard on each of the multiple measures. The reliability of decision that a candidate will receive the practicing license is dependent on least

reliable measure score decided by the faculty experts.

2) Compensatory Approach: Weaker performance on one measure can be offset by stronger performance on another. Two or more measures are required to compensate each other under a predefined mechanism.

3) Complimentary Approach: Requires minimum performance in any of multiple measures of similar construct or any of multiple opportunities using the same measure that can fulfill the requirement. Complementary approach of combining different measures of same construct for high stakes decision in fields like medical education may lead to inappropriate decision and author would not recommend this to be practiced.

4) Mixed Approach: A mixed conjunctive-compensatory approach requires multiple measures for compensatory rule and a predefined minimum performance in multiple measures that fulfill the requirement for compensation by another measure.

It can be of two types: a) No minimum performance set for any one of the multiple measurement tools, which are used for combining the measure to achieve preset standard sore on aggregate for passing the measure; b) a predefined minimum performance in any one-measurement tools, which can be combined with another measurement tool of similar construct or learning domains to achieve a preset standard score to make high stake decision in favor.

Standard setting strategy to make decision on summative assessment using a battery of tests primarily is compensatory or conjunctive or a mixed of both, besides complimentary approach for multiple opportunities intent in medical education. Combining various measurement tools in an assessment design to make decision must follow an appropriate standard-setting strategy for approaches and measures (3). A framework suggested (see figure1) here is modified from design reported earlier (2). Figure essentially illustrates the framework of combining multiple measures to provide the choice of rules that values combining of measures and not exclusively the principles, which determine the fit of object of measurement and the measurement tools. However, insight knowledge of learning domains and the relevant measurement tools to assess those will be needed to determine the principles of combining multiple measures in high stakes decisions.

Rows in this figure represent the forms of multiple measures whereas columns represent approaches to combining the multiple measures (2). The purpose of this work is to determine the applicability of this framework to a program that uses combining of multiple measures to reach a decision on pass or fail in higher medical education. Primarily, the purpose to design multiple measure policy includes**:**

1) To promote student mastery of curriculum content.

2) To facilitate decision making in different measures of same construct using a compensatory approach.

3) To rate the overall performance of a trainee using a conjunctive approach.

4) To ensure the validity of decision for safety of practice of medicine.

5) To incorporate qualitative measures in decision making which is based on rationale of accumulative measures decision rather than couple of mistakes in single measure to plead failing of overall assessment.

For measurement of different construct sufficient proficiency with identification of minimum performance is required in each measure and for this conjunctive approach is the ideal. In case of different measure of same construct one of the four approaches can be utilized as under:

a) Conjunctive-approach, showing minimum performance in each measure.

b) Compensatory-approach, combining scores of two measures by averaging the two scores.

c) Complimentary-approach, achieving credit as passing mark in one of the multiple measures. Complimentary approach serves performance across multiple measures or opportunities in which passing in any one measure or on any one opportunity fulfills the requirement and is practiced by most of the high stake examinations for its later use.

d) Mixed-approach, requiring passing one measure of similar construct by a set standard score and achieving minimum requirement of acceptable performance in another for compensatory decision of result.

High stakes decision on multiple opportunities with same measure allows student to continue taking the course and have multiple attempts. ORL-HNS discipline practices this by allowing 3 attempts in 7 years program. However, all the components whether passed or failed in previous examination will be reattempted in subsequent remedial examinations. Design for combining multiple measures should be consistent with the principles of standard-setting strategy which advocate that the measures of different construct should be combined using a conjunctive approach while the different measure of same construct should be combined through compensatory, complementary or mixed conjunctive-compensatory approaches.

Reliability of high stakes decision concerns the degree of accuracy and consistency with which inferences are reached (2). If measure incorrectly declares a student as failing the requirement he is retained for unnecessary reason (false negative) compared to a student who is declared pass because of illogical decision based on individual measure, is denied further training which may be crucial to his professional practice in future (2). This indicates the importance of multiple test measures as an impact on practice.

The impact of assessment programs have mostly been studied on trainees affected with decisions however, impact of standard-setting strategy using multiple measures also need to be studied on institutions, Program and supervisors as well as examiners. Impact on institution should see that the resources are being rightly used and the policies adopted are dynamic for assessment program.

Institution should be vigilant to bring a change in assessment culture acceptable to faculty. Impact on assessment program should see through selection and setting of the measurement tools, it's vetting and fit of object of measurement. Supervisors for their faculty development and awareness about standard-setting strategy should be undertaken. Finally right choices of examiners should be ensured to conduct the examination using multiple measures in assessment program and to critically evaluate the weaknesses and strengths of measures as a mandatory feedback that can be used for updating the assessment program.

Conjoint examination in Masters' of surgery in medical education in Malaysia practices these strategies in assessment design in which one of the above principles is followed. However, the scope of triangulation in indecisive or borderline cases either has no place in these examinations or it is often ineffectively applied. This is because summative assessment is basically used for high stake decision and continuous assessment, a so-called formative assessment is just a prerequisite to sit the summative examination.

A validity framework is essentially needed to adopt a standard setting strategy. The validity is concerned with the meaningfulness of a test interpretation and validation is a process of collecting the evidences to support the

interpretation of argument of validity of specific test. To argue on validity of a measurement instrument three types of evidences are required (4) and these are: (1) Content structure of the performance being measured in a test. It is essentially required if one wishes to evaluate the reliability and validity using factor analysis across the components of similar learning domains. (2) Reliability of score that is used for making pass or fail decision essentially for independent marking to establish the inter-rater difference or consistency across the scores. (3) Consequences of standard setting strategy that implicate candidates as well as the community as the immediate stakeholders. This important information can be used as fundamental steps to evaluate standard setting strategy that intentionally or unintentionally are practiced in higher medical education under specialty conjoint examination of three major postgraduate medical universities in Malaysia.

In postgraduate examination in medical education the instruments are often criticized for fit between object of assessment and the measurement tools (5). A data collected as the initial critical appraisal of summative assessment in Masters of ORL-HNS (see tables 1 and 2) incited to collect evidences as correlation and predictive value of all instrument and their impact on overall performance that decides on high stake examination in Malaysia (6). This initiated the need to collect validity evidence of measures used in Masters' examination. The initial critical appraisal revealed that long and short cases (see table 1), essay and MCQ tests (see table 2) are not appropriately used for decision-making (7).

To collect the validity further evidences were achieved on individual instrument for its correlative and overall predictive value on performance of candidates in Masters' ORL-HNS examination routinely held under the Specialty Conjoint Board of three major universities (UM, UKM and USM) responsible for postgraduate program in Malaysia**.** The need for this evaluation was in keeping with the notion that the use of multiple measurement instruments does not improve the reliability and validity decisions rather it is the logical use of the measures in triangulation that determines the accuracy and appropriateness of the decisions reached (2).

Education in Medicine Journal
2012, Vol.4 Issue 1
DOI:10.5959/eimj.v4i1.2

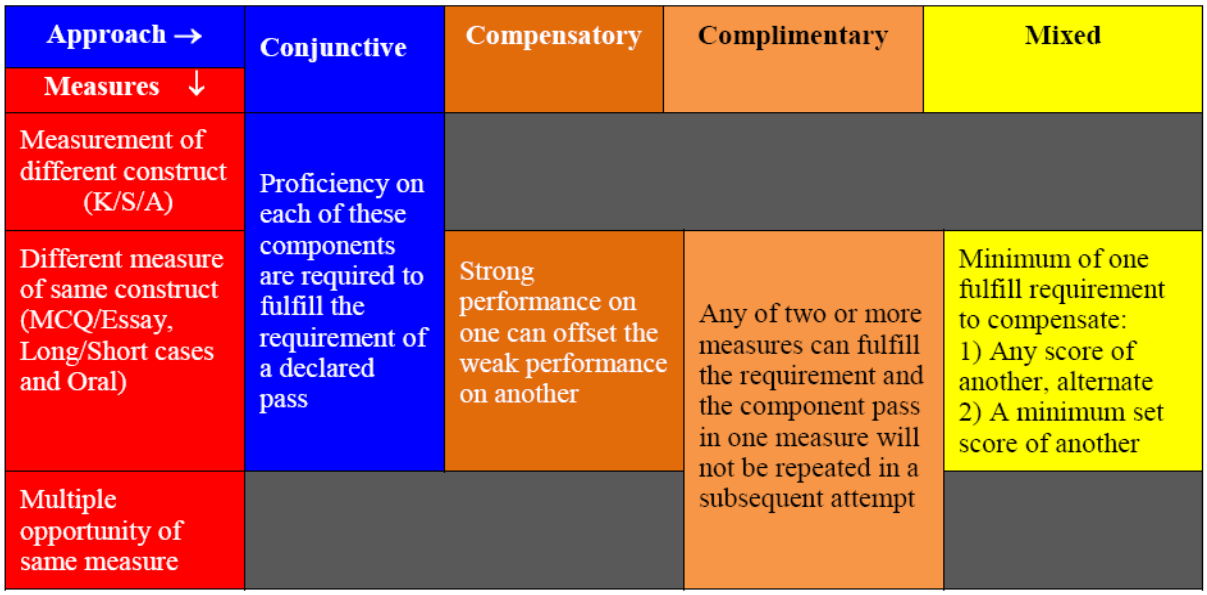

Figure 1: Framework for combining multiple measures with four approaches in a standard-setting strategy.

Method

The retrospective review of 61 candidate's was held to evaluate the summative assessment results of Part II of Masters of Otolaryngology and Head-Neck Surgery in School of Medical Sciences (SMS) at Universiti Sains Malaysia (USM) during 2000-2011 under the conjoint board of three universities (UM, UKM, USM). A critical appraisal of data was carried out (see table 2 and 3) to assess the accuracy and appropriateness of decision compatible with principles of combining multiple measures for decision-making.

To determine sample size, Cohen sample size table 1992 with power of study as 0.8, alpha significant at 0.05 and large effect size for multiple linear regression for five variables was used (8). All subject scores of 61 candidates were taken from the results of USM conjoint summative examinations were analyzed.

Collection of data comprising of results of all measurement tools as individual, as total of two individual measures in each written and clinical component 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 Pearson's correlation coefficient and prediction using unstandardized and standardized Beta coefficient of variables (see table 4) and overall performance.

Finally each measure as an individual assessment tool and as a component of written or clinical test was evaluated for its current practice vs. standard-setting strategy practiced globally. A suggested framework as an outcome of this study is also proposed for future practice of ORL-HNS discipline (see figure 2). The objective was to visualize the appropriateness of decision in theses assessments for a compensatory, conjunctive or mixed conjunctive-compensatory role with a logical standard-setting strategy to practice in a surgical-based discipline.

Recommendations at the end are made to see assessment as a program and the workplace-based assessment for its prospects of future practice.

Result

The correlation coefficients of MCQ and Essay showed no significant correlation (r=0.16). Long case and the short cases have shown good correlations (r=0.53). Essay correlated fairly well with long case (r=0.48), short case (r=0.41) and viva (r=0.39).

Oral stood as a component and showed fair correlation with written (r=0.39-0.42) as well as clinical component (r=0.50-0.66). Correlation of individual instruments and overall performance is significant for all measurement tools with strongest correlation coefficient shown for long case assessment (r=0.86, see table 3).

The linear regression analysis of written component suggests that MCQ is predicted by oral (B=0.34, $P<0.01$) but not by essay question and essay predicted by long case (B= 0.23, $p<0.01$) but not by MCQ. In clinical components long case predicted by oral and essay (B= 0.71 and 0.57 with $p<0.05$ respectively), short cases predicted by long case (B= 0.31, $p<0.001$). Oral predicted by long case and MCQ (B= 0.52 and 0.34 with $p<0.001$ and <0.05 respectively).All components predicted the overall performance with long case predominantly more than the other measures (see table 4).

Table No 1: Number of unsuccessful candidates in summative assessment of Specialty Conjoint Board Examination due to their inability to pass the long case assessment though clearing all other components and an overall score of >50% marks.

Year of exam	Name	Number	Theory (40 marks)			Clinical (40 marks)			Oral (viva) 20% Mark	Total Marks 100	Grades
			MCQ 20% Mark	Essay 20% Mark	Total 40% Mark	Long Case 20%	Short Case 20%	Total 40% Marks			
May 2005	XY	1	<50%	>50%	Pass	<50%	>50%	20.60%	Pass	53.87%	F
Nov 2005	XY	1	<50%	>50%	Pass	<50%	>50%	20.24%	Pass	51.34%	F
Nov 2007	XY	1	<50%	>50%	Pass	<50%	>50%	20.00%	Pass	50.40%	F
May 2008	XY	1	>50%	>50%	Pass	<50%	>50%	20.00%	Pass	53.87%	F
May 2009	XY	1	<50%	>50%	Pass	<50%	>50%	20.00%	Pass	50.07%	F
May 2010	XY	1	<50%	>50%	Pass	<50%	>50%	21.75%	Pass	54.55%	F
Nov 2010	XY	1	>50%	>50%	Pass	<50%	>50%	22.20%	Pass	57.47%	F
May 2011	XY	1	>50%	>50%	Pass	<50%	>50%	20.40%	Pass	57.50%	F
Total	8	All 8 Candidates Failed the Overall Summative Assessment for Failing long case									

Table No 2: Number of declared successful candidates in summative assessment of Specialty Conjoint Board Examination despite of the fact they did not manage to pass the MCQ in written components.

Year of Exam	Names	Number	Theory (40 marks)			Clinical (40 marks)			Oral (viva) 20% Mark	Out-come
			MCQ 20% Mark	Essay 20% Mark	Total 40% Mark	Long Case 20%	Short Case 20%	Total 40% Marks		
May 2004	XY	3	<50	>50	Pass	>50%	>50%	Pass	>50%	Pass
May 2005	XY	1	<50	>50	Pass	>50%	>50%	Pass	>50%	Pass
Nov 2005	XY	1	<50	>50	Pass	>50%	>50%	Pass	>50%	Pass
May 2006	X	1	>50	>50	Pass	>50%	>50%	Pass	>50%	Pass
Nov2006	XY	3	<50	>50	Pass	>50%	>50%	Pass	>50%	Pass
May 2008	XY	5	<50	>50	Pass	>50%	>50%	Pass	>50%	Pass
Nov 2008	XY	4	<50	>50	Pass	>50%	>50%	Pass	>50%	Pass
May 2009	XY	2	<50	>50	Pass	>50%	>50%	Pass	>50%	Pass
Nov 2009	XY	1	<50	>50	Pass	>50%	>50%	Pass	>50%	Pass
Total	21	All 21 Candidates who passed Summative Assessment Despite of Failing MCQ								

Table 3: Correlation between the assessment tools with each other and with an overall performance

ASSESSMENT TOOLS	CORRELATION COEFFICIENT (R) (N=51)					
	MCQ	ESSAY	LONG CASE	SHORT CASE	VIVA	OVERALL SCORE
MCQ	1	0.16	0.26	0.32*	0.43**	0.77***
ESSAY		1	0.48**	0.41**	0.39**	0.62***
LONG CASE			1	0.53***	0.66***	0.86***
SHORT CASE				1	0.50**	0.74***
VIVA					1	0.84***
OVERALL SCORE						1

* <0.05, **<0.01, ***<0.001

DOI:10.5959/eimj.v4i1.2

Table 4: Predictors value for each other and for overall performance predicted by each instrument.

Outcomes Predictors	B	95% CI for B (lower, upper)	β	R^2	F-stat
MCQ					
Viva	0.34	0.12, 0.56	0.423**	0.179	9.37**
Constant	5.67	2.88, 8.45			
ESSAY					
Long case	0.23	0.10, 0.36	0.481**	0.232	12.98**
Constant	10.12	8.57, 11.67			
LONG CASE					
Viva	0.71	0.34, 1.01	0.553***	0.492	20.36***
Essay	0.57	0.06, 1.07	0.268*		
Constant	-4.04	-10.12, 2.12			
SHORT CASE					
Long Case	0.31	0.158, 0.463	0.532***	0.283	16.94***
Constant	8.91				
VIVA					
Long case	0.52	0.333, 0.657	0.657***	0.500	21.04***
MCQ	0.34	0.056, 0.624	0.273*		
Constant	3.510				
OVERALL SCORE					
Long case	2.23	1.813, 2.644	0.855***	0.999	11860.46***
Viva	1.60	1.098, 2.100	0.483***		
Short case	1.37	0.941, 1.831	0.308***		
MCQ	0.99	0.755, 1.231	0.240***		
Essay	1.20	0.966, 1.073	0.186***		
Constant	-0.51				

*<0.05, **<0.01, ***<0.001

Table 5: Another example of decision-making leading to an outcome of summative assessment of three candidates with different overall performance due to inconsistent application of standard setting rules in short and long case clinical measurements.

NO	THEORY			LONG CASE	SHORT CASES					ORAL			OVER ALL
	MCQ 100%	ESSAY 100%	AVG %	100%	CASE 1 100%	CASE 2 100%	CASE 3 100%	CASE 4 100%	SC AV %	1	2	ORAL AVG %	
1	52.65	60	56.3	70	55	45	45	80	56	65	70	67.5	FAIL
2	52.65	62	57.3	40	60	45	65	80	62	55	65	57.5	FAIL
3	39.00	67	53.0	70	70	55	20	60	51	70	55	62.5	PASS

Conjunctive Approach							
Written		**Clinical**			**Attitude** Humanism and Organztnl. efficiency		**Overall performance**
MCQ At or >50%	Essay At or >50%	Long case At >50%	Short cases At or >50%	Oral At or > 50%	Observed in long case assessments		At or >50%
Compensatory Approach							
MCQ </>50%	Essay </>50%	Total At or >50%	Long case </>50%	Short case </>50%	Total At or >50%	Oral At or >50%	Over all performance At or >50%
Conjunctive-Compensatory Mixed Approach							
MCQ Minimal 40-45%	Essay Minimal 40-45%	Total At or >50%	Long case Minimal 40-45%	Short cases Minimal 40-45%	Total At or >50%	Oral At or >50%	Over all performance At or >50%

Figure 2: Framework for combining multiple measures with examples of conjunctive, compensatory and mixed approaches in a standard-setting strateg

Discussion

In conjoint summative assessment of ORL-HNS a framework combining multiple measurement tools are used to reach a passing or failing decision in postgraduate medical education in Malaysia. Standard setting strategy however, is loosely applied to make ultimate decision, which often receives comments challenging the application of principles in standard-setting strategies and its implication on outcome. The visiting examiners recommendations and differences of opinion observed, as controversies arising among the internal examiners during and after the examinations are the other challenges (7). A room for improvement has always been felt in these examinations held twice a year every year. Most of the queries raised during examination board meeting at the end of the examination are either unanswered or defended for being in line with traditionally practiced assessment for many years.

Logics to support unruly behaviors associated with such practices of assessment are often not supported with educational theories or literature evidences. Many of those involved in examination process often wonder, are they practicing an appropriate assessment program in summative examination in postgraduate medical education? and if not then what are the validity evidences to support its appropriateness and accuracy. For example, an initial appraisal recently conducted by the authors suggested that 8 students failed the summative assessment because they failed the long case assessment while passing the all other measures in written and clinical components (see table 1). Long case is the only measurement tool in a battery of assessment, which is done as a single patient work up. On the other hand 21 students pass the summative assessment though they failed the MCQ (see table 2).
MCQ is the only objective tool in summative assessment. The logic for this decision is that long case is judged by conjunctive approach and MCQ is judged by compensatory approach combined with essay questions. In another example (see table 5) a student who failed two short cases out of four that he/she appeared in clinical component of the assessment had failed the short case assessment on aggregate to fail the overall examination, though securing >67% marks on overall performance. Compared to this, another candidate passed the overall examination despite of his/her failing the MCQ and one of the 4 short cases assessment achieving merely 20% marks. The argument to support the decision is that a candidate has to pass three out of four short cases to pass this component and the overall examination.

The question that might be asked is the logic behind this rule? And why a candidate scoring only 20% marks in one of the short cases with an aggregate of 51% gets through the examination compared to another candidate who scores 56% marks on aggregate in the same short case assessment, but is declared unsuccessful. How the two such borderline failures can be worst then one absolute failure in one of the short cases? If qualitatively conceptualized, a candidate with 20% marks in a clinical measure with an aggregate of 51%, a clear fail in MCQ and overall passing marks of 62% can not be better than a candidate with 45% in two of the four short cases with an aggregate of 56%, clear pass in all written measures and overall passing marks of >67% (see table 5).

If the compensatory method is the rule then why is this rule denied in case of first candidate? Only because out of a box additional ruling was introduced to make decision! Appropriate standard-setting strategy if any such rule was to be implemented in short case assessment would have been to fix a minimal performance in every short case (e.g. 40-45%) to be considered for an aggregate passing of short case assessment at 50%. Doesn't this 20%

marks reflects on a candidate's performance as unsafe if safe/unsafe for future practice of medicine is the criteria observed in these examinations? So what are the criteria for declaring a candidate unsafe? No written criteria are produced and often the decision is a subjective feeling of one or two examiners. These and many other related incidents initiated the need to collect validity evidence of measures employed in Masters' of ORL-HNS Specialty Conjoint Board Examination. The initial critical appraisal revealed that long and short cases, essay questions and MCQ tests are not appropriately used for decision-making (see tables 1, 2 and 5).

The challenge using multiple measures is not the number of measurement tools but the logic by which the measures are combined to make high-stakes decision (2). Compensatory strategy considers performance as total score across the multiple measures versus conjunctive strategy, which requires a predefined minimal passing performance in each measure. Evaluation of summative assessment of this surgical based discipline suggests that a mixed conjunctive-compensatory approach has been adopted for decision-making. This approach requires multiple measures for compensatory rule and a pre-defined minimum performance in any one of the multiple measures to fulfil the requirement of passing an examination. Decision-making in complementary strategy is based on minimum performance on any one of multiple measures that fulfil the requirement. Compensatory or conjunctive strategy should be clearly defined for decision making in summative assessment. The present study analyzed the summative assessment results of Part II examination of Masters of Surgery in Otolaryngology and Head-Neck Surgery. The retrospective record was collected for appraisal and evidence of any major impact of an individual measure on a candidate's overall performance in summative examination.

Fundamental steps employed in the standard-setting strategy in these examinations were evaluated for content structure of measurement tools, evidence of reliability and evidence of consequences as under.

Content structure

The content structure of any measurement tool is the main source of validity evidence. Underlying construct of an instrument can be used as holistic (uni-dimensional) or analytic (multi-dimensional) rubric of multiple traits or structured items in a measurement instrument (4). Response to these items or test scores across the raters can be used to determine the test consistency or reliability, which can be evaluated, using factor analysis or correlative statistical study. A reliability coefficient alpha is a useful indicator of reliability. Correlation that requires the item scores only is less complex than factor analysis to study items interrelations. Correlation pattern can be organized with in a multi-traits or multi-methods framework. Correlation between like measures is expectedly high whereas correlation for unlike measures is expectedly low (9).

Content structures particularly of long case and oral tests of clinical component of summative examination were required for factor analysis. No such data was available for factor analysis? Usually the examiners in these summative assessments mark a single unobserved long case or an oral tests with verbal or visual scenarios by face-to-face questions, ultimately rated by consensus of scoring. This reflects 100% agreement or zero inter-rater difference. The only available data therefore was the candidate's score, which was used for both, discriminate and convergent validity utilizing Pearson Correlative test and Multiple Linear Regression tests respectively.

This evaluation reveals unwisely use of long case assessment, which is already less reliable

due to its unobserved encounter, unstructured trait marking and the only one long case used for assessment. Such a monopolized subjective assessment makes the long case assessment an unprecedented measurement tool, which compromises on reliability and validity of tool. A structured, well observed and more than one long case can improve the validity of this assessment tool. (10). Long case assessment and to some extend the oral assessment should be made structured to improve its reliability. The items of structured long case assessment reflective of various skills can be used to accomplish analytic judgment besides providing variables for factor analysis and validity evidence in future (see appendix A). Present practice of assessment in ORL-HNS examination provides no such data for analysis.

Evidence of reliability

Reliability is important evidence as a) reliability of test score and b) inter-rater consistency**.** Cronbach's alpha as correlative coefficient is a useful indicator of reliability of test score and depends on internal consistency of items responses comprising of items score (4). However, in view of lacking such a structured format used in long case assessment and the unprecedented rating done by consensus was relied for establishing reliability as correlative study utilizing Pearson correlative test for discriminate validity and Multiple Linear Regression test for convergent validity. Rater Consistency tells how well the judges rate the same performance. When rater consistency is high reliability is suppose to be high. However, unrecorded inter-rater difference of rating scores for the same construct by the same group of examiners and for the same sample of candidates makes this analysis difficult to perform. Rater consistency is not a direct test of reliability but it is high for high consistency (11).

MCQ and Essay questions showed very poor relationship in present study and suggest that these tests have different construct for extent of learning domains and should not be considered as measures serving the same purpose to be combined for making decision on passing or failing of candidates, as is done in these summative examinations. MCQ as true false items were used to merely test the factual knowledge or recall whereas essay questions were used to test the knowledge for higher cognitive level of comprehension, application, integration and synthesis. MCQ in this assessment program was observed as the only objective assessment tool the outcome of which is ignored by combining it with a very subjective tool, comprising of restricted (short) and extended (long) essay questions.

However, a decision to pass or fail the students on a true/false type of MCQ may not be an appropriate strategy and require a more robust objective assessment tool such as type A MCQ (one best answer), Type R MCQ (extended matching) or key feature questions. The finding of present study and standard practice of assessment suggest that the long case can be used as a compensatory measure for short cases and oral assessment in clinical component (see figure 2). This figure illustrates the framework for combining multiple measures based on principles of different approaches and measures in standard-setting strategy for guidance and adoption of ORL-HNS practice in future.

However, in current practice those assessment measures are independent of each other to decide on candidate's final performance even though they are meant for measuring the same clinical competence. From that notion, combining these assessment tools (short and long cases) for making decisions to judge clinical skills performance can be seen as more logical and valid.

Oral assessment in this analysis stood on its own as a component for it suggested to test both the clinical skills as well as knowledge.

Oral assessment showed significant correlation with clinical as well as written component and more evidently having strongest relationship with long case. Therefore the oral test can also be used as a compensatory tool to help decide on borderline candidates in long case assessment as well. Henceforth relying on long case alone for decision-making seems not to be a very logical decision. The long case was not predicted by the short cases though it did the other way round yet a strong correlation between them indicates that these tools can complement each other on decision-making in clinical performance.

Evidence of consequences

A major contribution of validity of the measurement tool is to establish the effects of test score interpretation and its uses for community and public interest (4). In compensatory standard setting strategy, low performance on one or two items or traits within a measurement tool or between the tools of similar intent or construct in any one component of assessment can be measured by high performance on other items or component respectively. This may well be true for other disciplines but for medical practice very careful judgment is needed as a low performance might not be tolerated in any measurement tool.

Low performance will raise the question of producing unsafe surgeons for community practice. However, the practice of decision making for safe vs. unsafe practice is currently anecdotal and therefore often unrecorded to be used as a document for candidate's feedback. Assessment program needs a) structured protocol for items scoring of each examiner and b) content experts decision to tolerate minimum performance in any given single trait if conjunctive method is used or a minimal score in one construct to be compensated by another if compensatory method is used.

A structured protocol with relevant distribution of marks for each skill and agreement within the examiners must be predefined (see appendix A). Besides, in order to make decision for declaring a candidate unsafe based on his performance should be well recorded using another rating form (see appendix B) by each examiner in the panel. Any inconsistency in scoring the candidate's performance should be available for combined discussion of the examination board. Committee should be empowered to decide reevaluation of candidate if a high rater consistency is observed. A minimum performance in an instrument required for combining with a stronger performance of another measure of same construct must be pre-defined. For example, in a long case it can be set that a candidate must score minimum 40% or 45% marks to be considered for compensatory method in standard setting strategy. However, the total passing mark in that component (clinical in this example) compensated with short cases must be 50% and above to be considered for over all pass if the candidate has cleared all other component as part of a conjunctive approach adopted for making decision in summative examination.

The outcome of the study suggests that the postgraduate assessment should base on quantitative as well as qualitative evaluation for high-stakes decision in summative examination. To achieve logical decisions a drastic revamp of assessment standard-setting strategy for accuracy and appropriateness of summative assessment results is recommended. The study has provided the evidence to consider and preclude an illogical practice of making decisions on individual instrument. The study has also recommended two structured assessment-rating formats (see appendices A and B) to document scoring by individual examiners. 1) A content-structure for marking the long case assessments (see appendix A). 2) A list of criteria for declaring candidate safe

or unsafe for future practice (see appendix B). The content structure and rating done by using proposed forms can provide data for factor analysis of reliability and inter-rater difference in future. Future assessment program based on performance at work (work-place-based assessment) as formative assessment and its triangulation in a qualitative judgment for an accurate and appropriate decision to reach is also recommended to be explored for future practice of conjoint examination of postgraduate medical education in Malaysia.

Towards Assessment Program

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. 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 result alone, rather it should contribute to determine the overall performance of a candidate. The role of individual instrument should be seen in collaboration with other measurement tools for its share to assess the respective learning domains such as knowledge or clinical skills to decide on pass or fail in summative assessment.

Concepts of triangulation in assessment

Triangulation 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 (11).
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.

Justification of triangulation of measures

MCQ results are judged on aggregate with essay questions to evaluate the domain of knowledge is an example of internal triangulation in this 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 especially be useful in situation that demands qualitative approach when assessment of long or short cases 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 blurred 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 work up or a very difficult question asked to analyze a problem solving issue for decision making for his/her level of training.

In such situations, it is not justified to decide on summative assessment for passing or failing 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. This may give us an opportunity to rationalize the nature of mistake committed by the candidate considering the patient's complexity, candidate's intellectual

knowledge and its application, problem solving and therapeutic skills. Candidate's attitude based on internal assessor quick review of candidate's overall performance during 4 years training should also be recorded.

In postgraduate examination of Master of Otolaryngology and Head-Neck Surgery clinical competence is tested directly through short and long case assessments with real patients and oral tests with real or created scenarios shown using slides or video clips. However, in short case assessments there are 3-4 short cases for physical examination and provisional diagnosis, in which candidate's analytic clinical reasoning and problem-solving skills are directly observed and questioned. Similarly there are two rounds of oral test with different sub-specialty patients, different panel of examiners and face-to-face questions. More than one item (3-4 short cases and 2 rounds of oral sessions) in each measure of clinical component, different examiners, different cases and clinical scenarios, direct observation and face-to-face questions improves content as well as context specificities of these two instruments.

Therefore reliability and validity of these measurement tools are also improved. Principles of internal triangulation are also observed in these assessment methods and performance is rated as an aggregate of 3-4 short cases 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 any other measure of same construct if a candidate's performance is not satisfactory for a clear pass due to one or two unexpected responses committed in cross examination. Candidates considered unsafe for medical practice in such cases are not allowed to pass the summative assessment despite of their passing all other components (written, short case and oral) comfortably well. Such incidents though occasional are experienced in these examinations and need to be addressed.

Issues of borderline candidates in long cases

Borderline assessment is though done away with all those written and other clinical components, it is still a matter of concern for long case assessment, as it practically exists in context of borderline assessment. Apparently 45 out of 100 though looks like far away from borderline definition with 50% as the passing marks, but practically it is borderline marks because scoring method used in these assessments are a multiple of five under a close marking system. Situation also points to borderline for a clear pass for a candidate whose one or two major mistakes are viewed as unsafe for a candidate to practice medicine if he/she is allowed to clear the examination. This often leads to controversy among the panelist to decide on rating with consensus especially when it is not a clear fail in someone's opinion.

Interestingly, if a controversy has to be avoided in such circumstances, a decision is rather taken to declare a candidate clear fail by rating him/her for a score of 30 or below. Such subjective judgments affect the performance of a candidate with the other panel of examiners in which only a distinction perhaps can save him/her from an impending disaster created by the currently practiced system. Can this precedent be regarded as good practice of postgraduate assessment in any way? This approach leads to an awkward situation if the candidate happens to clear all other components (see table 1). This unlikely situation is though seen rarely but is never a good experience especially in the presence of an external examiner. This needs a more logical solution and the above proposal offers a possible way out. A fine-tuning of those inventories (appendices A and B) can adjust

them to fit the requirements of any specific assessment.

Options to improve clinical judgment

Long case assessment, which is a single case with varying level of complexity of patients from one candidate to another, unobserved workup and rating of performance achieved by consensus of a panel of examiners, raises the question of context-specificity in validity area. To improve the validity of long case assessment few options are worth considering however, the implementation of any of these options will require a major decision to bring a change in current practice of long case assessment as follows.

1) Increase the number of long cases to make it two long case assessments for each candidate even if that requires reducing the allotted time for one long case, which is currently practiced. Addition of one more long case will improve the reliability of long case assessment due to increase in number, different cases and different panels of examiners for each.

2) Start observing the candidates for their performance on long cases with 20-25% marks reserved for observation. It will be appropriate to observe the candidates for their clinical attributes of communication skills, professionalism, organizational efficiency and humanism. However, until such time that decision to observe the long case assessment is taken, these essential components of clinical skills can be elicited by few leading questions in respective areas as recommended in structured rating form for long case practice (see appendix A).

3) Triangulate the long case assessment with short cases result with an overall passing marks maintained at 50%. However, to achieve a minimum performance, 40-45% marks in long case be considered mandatory for triangulation. Integrated outcome of long and short cases performance as knowledge and skills of clinical competence will add logic to clinical judgment and role of qualitative evaluation in high-stakes decision.

4) The rating of long case assessment should also be reviewed to make it more analytic then holistic in approach (see appendix A). Currently rating in long case assessment is done as close marking using multiple of 5. In this proposed criteria of evaluation using a rating form, a minimum 20% marks is recommended to be allocated for communication skills, professionalism, organizational efficiency and humanism besides, observing for relevance of questions asked or procedures performed in history taking and physical examination respectively.

5) However, if these options are not feasible for immediate practice then authors recommend the implementation of criteria (see appendix B) to decide on a candidate's proficiency if issues of safe or unsafe performance is in question for those scoring less than 50% marks. In such critical review a candidate will have to score 50% marks on point table completed by all examiners involved in long case assessment (see appendix B).

The supervisor from the same center will complete the part reserved for the internal assessor review since it is not justified for other external or internal-external examiners to decide on his clinical and surgical skills in half an hour unobserved encounter of long case assessment.

Criteria will be used for those candidates who have passed all other components and failed the long case assessment scoring close to 50% marks (40 and above) and/or pronounced unsafe for medical practice by any one examiner. Internal assessor from the institution of candidate's belonging can report on organizational efficiency of a candidate by commenting on his organizational behavior in

routine clinical practice during the training. A reflection of organizational efficiency can also be elicited by candidate's current knowledge and its application on given case under examination. Similarly clinical, surgical and interpersonal skills can also be reported. However, all these attributes and communication skills would have been judged better if long case assessments were observed during the patient workup, which is not done in current practice of long case assessment.

Conclusion

The outcome of discriminant and convergent validity evidences concludes that MCQ and essay do not correlate, nor do they predict each other (discriminant). Long case and short cases significantly correlate with each other however, short cases predicted by long case (convergent). Oral significantly correlate with clinical as well as with written component and is predicted by long case.

The validity evidence provided in long case assessment and its impact on overall performance suggest to introducing standard-setting strategy for decision making in summative assessment.

The validity evidences in multiple choice and essay questions suggest that the compensatory approach in written component that allows to combine different measures are not logical and obviate an accurate and appropriate decision in summative assessment. The outcome of present study provides sufficient evidence to review the current assessment method towards an assessment program in which decision on overall performance is based on quantitative as well as qualitative evaluation in high-stakes decision of postgraduate medical education.

Realization of using multiple measures in high-stakes decision is as important as its application for appropriate and logical decision to achieve the fair outcome precluding false negative as well as false positive results. The present study highlights the issues with respect to choosing the right strategy to combine the multiple measures for achieving fair and consistent result as a step forward toward the assessment program. We must understand that merely employing multiple measures will not improve the reliability and validity of decisions, rather it is the logic by which the measures are combined in a decision to make summative assessment more appropriate and accurate.

Reference

1. Teoh NC, Bowden FJ. The case for resurrecting the long case, Views and Reviews. BMJ. 2008; 336: 1250. http://dx.doi.org/10.1136/bmj.39583.596111.94

2. Mitchell DC. Multiple measures and high stakes decisions: A framework for combining measures. Educational measurement, issues and practice; ProQuest Education Journal. 2003; 22 (2).

3. Henderson-Montero D, Julian MW, Yen W M. Multiple measures: Alternative design and analysis models. Educational Measurement: Issues and Practice. 2003; 22: 7–12. http://dx.doi.org/10.1111/j.1745-3992.2003.tb00122.x

4. Haladyna T, Hess R. An evaluation of conjunctive and compensatory standard-setting strategy for test decision. Educational Assessment. 2000; 6 (2): 129-153. http://dx.doi.org/10.1207/S15326977EA0602_03

5. Marchais JE. Learning to become a physician at Sherbrook: A Full Switch to a Student-centered MD Program. Maastricht: Network Publications, 2001.

6. Hassan S. Assessment of postgraduate Program in Otolaryngology and Head-Neck Surgery in Malaysia – Are we Adequate. MSO-HNS News Bulletin. 2011; 4 (1).

7. Hassan S. Postgraduate Assessment in Malaysia: Rationale of Decision Making.

Education in Medicine Journal. 2011; 3 (1): e15.http://dx.doi.org/10.5959/eimj.3.1.2011.e18.
8. Piaw CY. Kaedah dan statistics penyelidikan buku 1: Kaedah penyelidokan. Mc Graw Hill; Malaysia, 2009: pg. 188.
9. Heubert JP, Hauser RM. (Eds.). High stake testing for tracking, promotion and graduation. Washington, DC: National Academy Press. 1999
10. Norcini J. Learning in practice: The death of a long case? BJM. 2002; 324: 408-409. http://dx.doi.org/10.1136/bmj.324.7334.408
11. Neil Jackson, Alex Jamieson and Anwar Khan. Assessment in medical education and training. Oxford: Radcliffe Publishing, 2007

Education in Medicine Journal
2012, Vol.4 Issue 1
DOI:10.5959/eimj.v4i1.2

Appendix A: Structured assessment of long case measurement tool using analytic rubric to avoid subjective influence of any one assessor in panel of examiners.

No	Clinical Attributes of Long Case Assessment	Rating (>70-Excellent, 60-69 Good 50-59-Average <50-Unsatisfactory)	
1	**Information gathering (reflective skills)**	**Total Marks - 25**	**Marks Claimed**
I	Presenting for cross examination on history	13	
II	Presenting for cross exam physical workup	12	
2	**Analytic Reasoning (diagnostic skills)**	**Total Marks - 20**	
I	Differential diagnosis	10	
II	Imaging interpretation	7	
III	Provisional diagnosis	3	
3	**Problem Solving (management skills)**	**Total Marks - 35**	
I	Investigative skills (investigation/diagnosis)	10	
II	Therapeutic skills (curative/palliative)	12	
III	Procedural skills (operative procedure)	8	
IV	Complication of management	3	
V	Prognosis	2	
4	**Interacting (communication skills)**	**Total Marks - 8**	
I	Communication	5	
II	Knowledge and confidence	3	
5	**Attitude and Humanism**	**Total Marks - 12**	
I	Professionalism (values to medical ethics)	3	
II	Humanities (values to patient care)	3	
III	Compassion (demonstration of patient care)	3	
IV	Organizational efficiency	**3**	
	Total Marks vs. Marks Claimed	**100**	
6	**Overall Performance**	**Excellent** / **Satisfactory: Good** / **Average**	**Unsatisfactory: Borderline*** / **Poor**

*Borderline candidates among the unsatisfactory group are those who deserve 40-45 and above but less than 50 out of a total 100 marks for committing some mistakes in long case assessment and considered as unsafe to be allowed independent practice on a qualitative assessment. Assessment is though sought to have consensus is never without reservations by some examiners in the panel.

Education in Medicine Journal
2012, Vol.4 Issue 1
DOI:10.5959/eimj.v4i1.2

Appendix B: Criteria to be observed in long case assessment to decide on safe or unsafe status of a candidate for future practice of medicine if allowed to graduate.

No	Attributes	Remarks (Marking as score or handicap)		
1	**Patient Factor**	**1 Handicap**	**1/2 Handicap**	**0 Handicap**
I	Complexity	Complex case	Less complex	Simple case
II	Language barrier	Sufficient	Partial	No
2	**Intellectual Quality**	**4 Marks**	**2 Mark**	**1 Mark**
I	Knowledge	Excellent	Satisfactory	Unsatisfactory
II	Comprehension	Excellent	Satisfactory	Unsatisfactory
3	**Clinical Skills**	**4 Marks**	**2 Mark**	**1/2 Mark**
I	Problem solving skills	Excellent	Satisfactory	Unsatisfactory
II	Therapeutic skills	Excellent	Satisfactory	Unsatisfactory
4	**Attitude/Professionalism**	**3 Marks**	**2 Mark**	**1/2 Mark**
I	Concern shown for patient	Excellent	Satisfactory	Indifferent
II	Confidence to manage pt.	Excellent	Satisfactory	Unsatisfactory
III	Communication skills	Excellent	Satisfactory	Unsatisfactory
5	**Internal Assessor Review**	**2 Mark**	**1 Mark**	**1/2 Mark**
I	On-job clinical skills	Excellent	Satisfactory	Unsatisfactory
II	On-job surgical skills	Excellent	Satisfactory	Unsatisfactory
III	Organizational efficiencies	Excellent	Satisfactory	Unsatisfactory
IV	Interpersonal skills	Excellent	Satisfactory	Unsatisfactory
6	**Total marks claimed**	**Score.....Out of maximum 35 marks (..... %)**		
7	**Status as safe or unsafe**	**Safe surgeon > 50%**	**Unsafe surgeon < 50%**	

Corresponding author

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

Accepted: July 2011 Published: June 2012