Creating Evidence to Advocate the Validity of Results of Clinical Performance in the Undergraduate Surgery Clerkship

Omayma A.E. Hamed[1,2]; Husain Hamza Jabbad[1]; Hebatullah Alsayed[3]; Asim Alshareef[1]; Mohannad Alzain[1]; Omar I. Saadah[1]; Fatin M. Al-Sayes[1]; Rani Ghazi Ahmad[1]

Faculty of Medicine- King Abdulaziz University [1,] Faculty of Medicine-Cairo University[2], and Faculty of Dentistry- King Abdulaziz University[3]
dr.omayma.aly@gmail.com, hjabbad@gmail.com

Abstract: Background/ Purpose: Evidential bases were not performed en masse to validate assessment results in the undergraduate Surgery clerkship in King Abdulaziz University (KAU). This study aimed at producing a comprehensive package of evidence to prove validity of students' clinical performance assessment results (as defined by Messick's framework). **Method:** Guided by Messick's conceptual framework, the problem was analyzed. Hands-on faculty development on creating an exam blueprint was done: 1. Learning objectives (LOs) revised; 2. Alignment secured; 3. Weight of (LOs) determined; 4. Number of items/topic/domain calculated; and 5. Appropriate assessment methods selected..Quantitative evidences as reliability and correlation coefficients of various validity components were calculated. The underlying values that scaffold validity evidences were explored via a Focus Group Discussion and the results analyzed by content analysis. **Results:** 1. The weight of different domains in the test equally reflected their weight in the curriculum (content validity); 2. Positive unintended consequences resulted from the new assessment approach (consequential validity); 3. There was a statistically significant correlation among various assessment methods that provided evidence for concurrent and predictive validity; 4. Success rates and grades distribution alone could not provide evidence to advocate an argument on validity of results. **Conclusion:** A newly introduced assessment plan with new tools had to be validated by pursuing a comprehensive, unified approach to create evidence from multiple sources of data in order to support the argument of advocating the assessment results.

[Omayma A.E. Hamed; Husain Hamza Jabbad; Hebatullah Alsayed; Asim Alshareef; Mohannad Alzain; Omar I. Saadah; Fatin M. Al-Sayes; Rani Ghazi Ahmad. **Creating Evidence to Advocate the Validity of Results of Clinical Performance in the Undergraduate Surgery Clerkship.** *J Am Sci* 2016;12(3):59-71]. ISSN 1545-1003 (print); ISSN 2375-7264 (online). http://www.jofamericanscience.org. 9. doi:10.7537/marsjas12031609.

Keywords: Construct validity, Messick's framework, undergraduate, assessment, Surgery, evidential bases, ACGME

1. Introduction:

Does a strong evidence of validity from one assessment source obviate the need to seek evidence from other sources? Beckman *et al.* [1] stressed on the inadequate evidence of reported validity of the results of assessment instruments used to assess clinical competence.

From Discrete to Holistic Clinical Approach:

Miller outlined a framework for the development of clinical competence. Assessment of medical students has focused mostly on "knows" and "knows how," which represent the lower two tiers at the base of the pyramid: recall of factual knowledge and the application of this knowledge in problem solving. However, such examinations may fail to document what students will do when faced with a real patient. To determine someone's clinical competence, observing behaviors in action is needed. When this is done in a structured context, it will represent assessment of competence at the third tier of the pyramid, "shows how". This could be assessed using OSCE, which allows evaluation of a diversity of clinical competences with high reliability. On the other hand, actual performance in real workplace environment is represented by the top layer of the pyramid in Miller's model, "does". This is only assessed by direct observation in workplace practice as a physician [2,3].

The level of competence expected from a medical student, however, requires progress from discrete to integrated abilities. This is in line with Benner's Novice to Expert Taxonomy [4]. It is crucial to consider the dynamic nature of the clinical environment, which is characterized by a variety of interacting contextual factors [5]. This presents a limitation of replicating such environment within the OSCE context; thus affecting its predictive validity [6, 7]. OSCE also reduces clinical competence to a number of compartmentalized skills, thus raising the concern that the patient will not be approached holistically by the student [8,9]. To achieve this, integrated objective structured clinical assessment stations are implemented, in order to enable students to integrate both cognitive and clinical skills [10,11].

However, do the integrated stations explicitly assess the high cognitive level processes underlying complex clinical reasoning skills? Methods such as structured viva confined to a specific clinical case, was used to reveal the depth of knowledge, the ability to discuss and defend, as well as, elicit the clinical reasoning process that takes place in the student's mind. Oral discussion of a case with the student explicitly unfolds the schemata pursued by the student to reach a diagnosis. To reduce examiner's subjective judgment on the student's performance, the oral discussion is conducted according to a standardized structured set of questions for all students. Questions are integrated with each other and related to the specific clinical case rather than randomly asked questions by the examiner depending on his personal choice. The set of questions in viva covers all aspects from history and physical examination skills to differential diagnosis, investigation and therapeutic skills as well as the underlying content knowledge. The clinical cases were selected from real patients' case studies; this develops context to the students' learning and make evaluation more authentic to real life practice [(12)]**.**

It could be concluded that it is only by the utilization of a variety of appropriate assessment strategies that assessment results could be obtained as evidence of construct validity. For this reason, a variety of assessment methods should be used to ensure assessment of a holistic approach to a patient including students' communication skills, systematic approach and logical progression through history, examination, differential diagnosis and plan for investigations and management, as well as whether they elicit the correct history and signs. Since each assessment strategy measures a variety of aspects of students' performance, it is assumed that a combined score of all measurements would reflect a more thorough evidence of validity of results [(13)].

Validity and Validation of Clinical Performance Results:

The American Psychological Association (APA) listed four types of validity: (1) Construct validity: which refers to how well a particular test measures the skills or knowledge that it is intended to measure; (2) Content validity: which refers to how well the test scores represent a representative sample of the learning objectives in the domains it intended to measure (knowledge, cognitive skills, psychomotor skills, interpersonal skills, and communication skills); (3) Predictive validity: refers to how well a test can predict later behaviors; (4) Concurrent validity: refers to how close the scores are on two different tests that claim to measure the same construct [(14)]**.**

Zumbo [(15)] stated that validity is a unified concept, and validation is a scientific activity based on the collection of multiple and diverse types of evidence. Validation practice is building an argument based on multiple sources of evidence (e.g. statistical calculations, qualitative data, reflections on one's own values and those of others, and an analysis of unintended consequences) [(16)]. In 1986, Crocker and Algina [(17)] stated that validation practice started with calculation of a single aspect of validity. In 1989, Messick [(18)] described these procedures as fragmented, unitary approaches to validation. Hubley and Zumbo [(19)] described them as "scanty, disconnected bits of evidence to make a two-point decision about the validity of a test". In 1989, there was a shift from many types of validity to a single, but integrated type of validity conceptualized as Messick's Framework (1989). On the same vein, validation practice has also evolved from a fragmented approach to a comprehensive, unified approach in which multiple sources of data are used to support an argument, which might be assessment results in an educational context. This unifying force refers to combining multiple lines of evidence to support the interpretation and use of scores [(20)].

Two categories of evidential basis evolved [(21)]: (1) Evidential basis for test interpretation; and(2) Evidential basis for test use. To be able to provide strong evidential basis for the appropriateness of inferences and actions based on test scores, it is also mandatory to identify the *underlying values* in the course documents, course developers, instructors and learners; to weigh, balance, and compare these values for convergence and make a final judgment as to the extent which these values play themselves in course implementation[(22)]. These evidential bases were not performed en masse to validate assessment results.

Purpose of the Study:

The purpose of this study is to produce a comprehensive package of evidence to prove validity of students' clinical performance assessment results (as defined by Messick's framework).

Context:

In the Faculty of Medicine in King Abdulaziz University, sixth year medical students participate in the Surgery Clerkship over 20 weeks. The required competences of the clerkship matched the National Qualifications Framework and covered the five domains: Knowledge; High Cognitive Skills; Interpersonal and Self-responsibility Skills; Information Technology and Communication Skills; and Psychomotor Skills. The academic reference standards for the required competences were derived from the ACGME six competences, namely: Patient care; Knowledge; Professionalism; Communication skills; Practice-based learning and improvement; and System-based practice (www.ACGME.org) .

In 2012- 2013, the assessment committee in the Surgery Department arrived to the conclusion that the assessment of the sixth year medical students, in the form of MCQ and OSCE, did not measure all the competences required from them at that level. Though the different domains and competences of the Surgery course were covered in the MCQ exam and OSCE, yet the whole construct is not holistically assessed. The committee decided to introduce an assessment plan which encompasses a greater continuum in assessment: an MCQ exam to assess knowledge; an OSCE to assess discrete clinical tasks; an Objective Structured Short Case (OSSC) followed by structured oral examination to assess integrated clinical skills and elicit clinical reasoning process; and case writing/presentation followed by feedback to promote learning throughout the course. However, the committee had to convince the faculty of the causes that triggered the change; and in a departmental meeting, the Head of Department posed a question to faculty: *"What evidence do we have to justify our decision to let the students progress to the next level? What is the evidence which proves that the scores in the records constitute an indicator of students' acquisition of the required outcomes?"* At that time there was no evidential basis for interpreting or using the test results except for the scores of the MCQ and the OSCE exams. None of the faculty members could provide meaning of the scores, or the underlying values and tasks which resulted in these scores.

In 2013-2014, the assessment committee in the department introduced an additional assessment tool to complete the evaluation of the "whole" construct of the course. This was the Objective Structured Clinical Case (OSCC), in which different authentic scenarios of the defined core clinical cases were used to construct multi-staged integrated structured stations which cover the "whole" construct of the case. This was followed by an oral face-to-face exam. To reduce the subjectivity and enhance the reliability of results, the oral exam was structured. This provided students with equal opportunity of fair and standardized assessment while testing their knowledge, clinical skills and attitude. Objectivity of the structured oral exam was determined by laid down questions rather than randomly asked questions by the examiner. Questions unfold the schemata that are followed by the student in performing the clinical skills in the OSCC. They cover the knowledge, history, physical examination skills, investigations, differential diagnosis, and management plans. The questions set, being linked to an authentic OSCC, gives context to the assessment, hence enhances the predictive validity of the ability of the student to transfer the skills to real workplace practice [12]. The only limitation of OSCC is that the student is examined on one case, which affects the content validity of the results. However, it was contained as one of many assessment tools that cover the whole construct of the Surgery course. The aim of the OSCC and structured viva is to compensate for the limitations of the OSCE; yet both complement each other. In addition, during the clerkship, students were exposed to the assigned core clinical cases, and are urged to do clerking and clinical presentations that are assessed through a log book.

The research questions are: (1) How can an ideal assessment practice secure evidence of validity of sixth year students' assessment scores in the Surgery Clerkship?

(2) Does the final success rate alone secure enough evidence for validity to validate sixth year students' assessment scores in the Surgery Clerkship?

Method:

Design:

The study included 326 sixth year medical students, who represented the 2013-2014 cohort. Guided by Messick's framework, the following was performed:

***1.* Analysis of the Problem:** Analysis revealed that the questioned validity of the results of clinical examinations was due to deficient evidences for interpretation and use of assessment results. The framework not only offered a way of analyzing the problem more comprehensively, but also acted as a guide to develop solutions [23].

***2.* Informal Hands-on Faculty Development and Designing a Blueprint:** The clinical examination was built on an examination blueprint and the items were reviewed to provide evidence of fit between the content and validity [24].

A hands-on informal faculty development was set to some faculty members responsible for assessment in order to guide them to refine the objectives. Alignment was revised by two authors who helped faculty develop a table of specifications, and design a blueprint in a way which secured both content and construct validity of results:

***2.1* Revisiting the Learning Objectives:**

The first step in the process was revisiting the "Core Educational Objectives" (CEO) of the Surgery clerkship. The CEOs were confined to ten major statements and a set of potential instruction topics were defined. The "Specific Learning Objectives" (SLO) of each topic were developed.

***2.2* Alignment:**

A table of specifications was designed to display the alignment of the SLOs of each topic with the:

- Accreditation Council for Graduate Medical Education (ACGME) General Competencies and subskills.

- National Commission for Assessment and Academic Accreditation (NCAAA) domains (Knowledge; Cognitive skills; Interpersonal skills; IT/Communication skills; Psychomotor skills).

- Assessment tool(s) appropriate for evaluating the acquisition of each SLO. These tools were derived from the ACGME toolbox accessible online at www.acgme.org.

-Competency level corresponding to each domain in Miller's Pyramid for clinical/procedural skills.

- The appropriate learning material.

This alignment step is one of the "underlying values" which helped in completing the validity evidence package and which was emphasized by Messick. SLOs addressing knowledge and high cognitive skills were assessed through MCQs; SLOs representing clinical, communication, interpersonal, and professional skills were assessed either as discrete OSCE stations, each measuring a particular competency; or as a "whole" clinical case through an OSCC which covers all competencies related to that case (the whole construct) followed by structured oral exam to explore the underlying reasoning process and attitude.

Messick's Framework	Actions Required
1. Ensure assessment of a representative sample of learning objectives that cover the whole construct of the discipline (Content).	1. A table of specifications which proves alignment between learning objectives, teaching/learning tasks and assessment tasks was developed. 2. A blueprint was developed [24]**.**
2. Illuminate the detailed nature of students' performance and reveal the fit between their performance and high-cognitive processes (Response Process).	Objective structured clinical case followed by structured oral exam was used. This ensured measuring a holistic approach to the patient and as well elicited the reasoning process and sequence of thinking behind the students' performance to reach diagnosis and management plan ([25]**.**
3. Ensure the reliability which measures internal consistency, i.e. if all items on an instrument measure the same construct (Internal Structure).	The internal consistency of MCQ and OSCE stations was measured using Cronbach's Alpha [26].
4. Ensure **relations** between various instruments assessing clinical performance (Relations with other variables).	- The correlations between the results of different assessment measures, and as well between each measure and the combined score were calculated [27]. - Also, the scores of the mid-exam and final exam, both of which measured the same construct were correlated' "Concurrent Validity Coefficient" [15]**.**
5. **Consequences:** The results are the effect of many contextual variables other than assessment alone (Unintended results), but they [28]. These effects from a test could be categorized into individual, institutional, systemic and social effects. Sometimes those unanticipated effects are positive and referred to as "positive washback"[30] or "beneficial byproducts"[21, 29].	This is the most controversial evidence of validity [1], and was evaluated using a focus group discussion (FGD) with students.

2.3 Weighing the Learning Objectives Required to be Assessed:

The SLOs intended to be measured in each unit were selected. The total number of selected SLOs in each domain was then used to calculate their weight as a proportion from the total number of SLOs selected in all domains.

2.4 Calculating the number of items:

The number of items in each domain was calculated by using software designed by two of the authors (H.O and A.H). Through this software, if the weight of the selected (SLOs) appears to be equal to the weight of items, then this constitutes evidence that the exam will cover a representative sample of the learning objectives in all topics.

3. Selection of Appropriate Assessment Strategies:

a. The assessment committee introduced an OSCC as one of the assessment measures to assess a holistic approach of integrated tasks, followed by structured oral exam to assess students' detailed cognitive processes that they used to reach diagnosis and management plan. This provided validity evidence of "Response Process". The OSCC ratings were performed using standardized form.

The clinical cases (OSCC) were selected from real patient's case studies, which developed context to the candidate's apprenticeship learning and made evaluation close to workplace-based assessment.

Authenticity also provided an assessment environment in which the cognitive demand i.e. critical thinking (problem solving skills) was consistent with cognitive demands of the situation (real clinical scenario) to which a candidate was exposed during evaluation. This engaged students in an effective clinical reasoning process that helped in the assessment of their overall clinical competence, professional efficiency as well as the communication skills and medical professionalism. Structured oral exam was confined to one specific clinical scenario which triggered the argument of context specificity and low content validity. In the structured oral exam, the questions were integrated with each other, contrariwise to what was seen in traditional oral exam sessions. Questions complemented each other towards problem solving; this was achieved by navigating through history, physical examination, differential diagnosis, investigations, treatment modalities, prognosis, procedural skills and or complications of therapeutic management. Questions that were asked, tested the application of clinical knowledge in the given OSCC and their analytic thinking in order to evaluate their problem solving skills.

b. The OSCE was designed based on the examination blueprint; it assessed discrete clinical, technical, and cognitive skills.

c. The MCQ assessed knowledge and cognitive domains.

d. Case writing and presentation was used as continuous assessment throughout the clerkship to provide feedback and promote learning; marks were assigned through a log book to ensure exposure to all the defined core clinical cases.

***4.* Quantitative Evidence:**

a. The reliability of each of the MCQ and OSCE, was calculated using Cronbach's Alpha correlation coefficient which was considered as one of the evidential basis for test interpretation and use. Although this provided evidence of "internal consistency" of the assessment measurement, yet according to Downing [25], this was of secondary importance for performance ratings. However, we saw that the reliability of test tools eliminated the internal consistency of the test from being a reason for invalid results. Reliability of results of a test reflected that factors other than the construct of the test that might affect the validity were eliminated.

b. A correlation between scores of each assessment instrument was calculated, as well as between each instrument and the combined score. Correlation was computed using Pearson's correlation coefficient. This provided evidence of "relations between different variables". Correlations of scores from the OSCC or the OSCE (both measure application of clinical knowledge) and the MCQ exam or structured viva (both measure content knowledge) provided evidence for predictive validity.

c. A correlation between the mid-exam and final exam scores was calculated. Both exams measured the same construct (cognitive, clinical skills, communication skills, and attitude). This provided evidence of concurrent validity.

d. Content validity coefficient was calculated by comparing the assessed content and skills domains in the course with those that resulted from the blueprint.

e. Item analysis interpretation was performed to reduce construct- irrelevant variance. Factors other than the construct being measured could affect the reliability and validity of results. ; ex, unbalanced difficulty or flawed design of items.

Scores were retrieved from the students' records after obtaining a written approval from the Head of the Surgery Department, who was one of the authors.

Statistical Analysis of Data:

The gathered data was statistically analyzed using the Statistical Package for Social Sciences (SPSS) version 14.0 (SPSS Inc., Chicago, II). Quantitative data (scores) was summarized and presented as mean and standard deviation. The reliability of MCQ and OSCE assessment measures was calculated using Cronbach's Alpha correlation coefficient [26]**.** Pearson's correlation coefficient was calculated to measure the correlations between the scores of each assessment measurement; between each and the combined score; and between the scores in the mid- and final exams. The change rate in the grade distribution was calculated using the equation: New Value – Old Value/ Old Value x 100. The Mann-Whitney Test was used to compare the grade distribution in 2012- 2013 and 2013- 2014. Significance was set at the 95% confidence interval.

***5.* Explore the underlying values that scaffold validity evidence:**

Two focus group discussions (FGD) were conducted: one for faculty and another for students; each group consisted of six participants [36]. Each focus group included a moderator and a recorder who is short-handed in notes-taking. A structured list was prepared by the authors to trigger the conversation, to be clear and focused, and to be open-ended [37]. The list for students included the following questions: (1) To what extent are you satisfied with the surgery curriculum?; (2) How did the OSCC impact your learning?; (3) Was the assessment plan of the course satisfactory? The list for faculty contained two questions: (1) How did the new assessment plan affect your teaching in clinical sessions? (2) What is your opinion on the change in students' behavior during the course after knowing the new assessment plan? Each focus group was scheduled to take an hour. At the end of each focus group, moderators debriefed all authors, and analyzed the gathered information using the content analysis method.

Content analysis began with a set of hypothetical categories which served the evaluation of the experience and explored the values and activities that scaffold the evidences for validity: (1) Departmental/Strategic; (2) Economical/Resources; (3) Social/Behavioral; and (4) Technical/Developmental. Then the set of data resulting from the FGDs of students and faculty were read and any new categories that were explored other than the hypothetical ones were identified. For each category there was evidence in the data in the form of participants' responses to the FGD questions. A grid was plotted which contains the category related to its themes and evidenced by participants' statements. Categories were then validated by giving the identified categories to another researcher in the study. This was then followed by identifying whether or not any one participant in a data set (faculty or student data sets) displayed each category and presented them as frequency and percentage of participants agreeing to each category [36].

***6.* Evaluation of the experience:**

Kirkpatrick's evaluation model was used to evaluate the various stages of the experience [32, 33]**.**The model consists of four levels:

(1) The reaction of the students and their thoughts about the experience: this was evaluated by conducting a FGD with 6th year students.

(2) The student's learning and the increase in knowledge from the experience was evaluated by calculating the change rate in the grade distribution of the 2013-2014 and the 2012-2013 cohorts.

(3) The student's behavioral change and improvement after applying the skills was evaluated by conducting a FGD with faculty.

(4) The impact of the student's performance in workplace practice could not be evaluated since this requires follow-up of students and could be affected by extraneous factors other than the experience that they were exposed to.

Ethical Approval:

Complying to the Faculty bylaws, the students' scores were retrieved from the records after approval from the Head of Surgery Department and the Vice Dean for Development. Ethical approval was obtained from the Research Ethics Committee (REC).

Limitation of the Study:

Many factors might contribute to construct validity of inferences from clinical examinations. These could be categorized into factors which result in construct under-representation, or in construct-irrelevant variance [38]. Using a variety of assessment tools provided one factor as evidence for construct validity; however exam blueprinting, test environment, raters' variability, test psychometrics, and instruction might be other contributing factors. To reduce construct under-representation, the examinations were based on a systematically planned blueprint; the instruction was similar; test environment was controlled; the raters were trained through a series of hands-on faculty development held by international experts in medical education. To minimize construct-irrelevant variance, test items were reviewed pre- and post-test regarding the soundness of their design, relevance, and difficulty and discrimination indices in case of MCQs. However, we could not control the construct irrelevant-variance caused by inappropriate setting of pass/fail score which was standardized centrally by the university bylaws. In addition, the study was conducted on one clerkship, which might affect its generalizability. Future studies are required on other clerkships. Moreover, comparison of students' scores was performed for two different cohorts who sat two different tests which might lead to bias. Hence, this should be repeated for the same cohort of students in further studies but which might carry unethical issue of assessing students using different assessment plans.

3. Results:

(I) Quantitative Evidence of Validity:

1. Table (1) shows that the internal consistency of the MCQ exam and OSCE was close, and represented as reliability coefficients (0.81; 0.80, respectively).This matched positively with the themes in the departmental/strategic category *"Ideal practice in exam preparation provided evidence for reliability and validity of results"*. The percentage of faculty agreeing to this category was higher than students.

2. Table (2) shows that the scores of each assessment measure were significantly positively correlated with the combined score; as well as between each other *($p<0.05$)*. The highest correlation coefficients occurred between the scores of the MCQ exam and all other assessment measures (OSCE, OSCC, Structured oral exam and Writing/Presentation) (0.611; 0.471; 0.570; and 0.474, respectively) thus exploring the common cognitive basis of the used assessment measures. This aligned with the third theme in the technical/developmental category *"Integrated teaching applying knowledge in clinical sessions and reflected in assessment."*

The significant positive correlation between OSCC, Structured oral exam and Writing/Presentation provides evidence on alignment of teaching and assessment, (0.458; 0.416, respectively). This matched with the second theme in the departmental/strategic category *"Alignment between teaching and assessment is mandatory for validity of test results."*

Also an evidence of predictive validity was provided by the significant positive correlation between the OSCE and OSCC both of which measures the application of clinical knowledge and the MCQ exam and structured oral exam, both measures the content knowledge. This matched with the fifth theme in technical/developmental category *"OSCC and structured viva triggered thinking and application of knowledge."*

The highest correlation coefficients occurred between each assessment measure (OSCE; OSCC; Structured oral exam; MCQ, and Writing/Presentation) and the combined score (0.749; 0.658; 0.777; 0.861; and 0.642, respectively). This further provided evidence on construct validity because the assessment measures were complementary and compact.

3. Table (3) provided evidence of concurrent validity whereby it showed a significant positive correlation between the scores in the mid- and final exams which measured the same construct (cognitive, clinical skills, and attitude) (r= 0.672; *p*< 0.0001). The mean values of the scores in each were also close (78.3 and 70.6, respectively).

4. Table (4) reflected the accuracy of results that was evident by a statistically significant difference in the grade distribution between 2012- 2013 and 2013-2014 cohorts *(z=* -3.986, *p< 0.05)*; although the success rates was the same *(z=* -0.577, *p> 0.05).*The grades became more normally distributed reflecting the objectivity and fairness of assessment. This matched with the first theme in social/behavioral category *"Positive impact of assessment plan on students"* whereby students felt the fairness of their grades, being assessed on the same case and asked the same questions. This also related to the first theme in the technical/developmental category *"Objective structured exams ensure fairness".*

(II) Identified categories and themes resulting from the FGDs with faculty and students:

The themes concluded from the FGD with faculty and students were: 1. Ideal practice in exam preparation provided evidence for reliability & validity of results; 2. Alignment between teaching & assessment is mandatory; 3. OSCC & structured oral exam reduced the need for resources; 4. The assessment plan has a positive impact on students; 5. The assessment plan has a positive impact on faculty performance; 6. Appreciation of the effort exerted in improving the curriculum; 7. Objective structured exams ensured fairness; 8. Objective structured exams standardized the line of discussion between faculty & students; 9. Integrated application of knowledge in clinical teaching and its reflection in assessment; 10. Focused structured discussions engaged students and increased attendance; 11. OSCC & structured oral exam triggered thinking and application of knowledge rather than recalling of information only. The resources and departmental planning of assessment constituted the most important positive impact from the new experience (100% agreement), followed by developmental and technical factors concerning the curriculum and its implementation (90% agreement). The development of the curriculum, and its implementation regarding teaching and assessment constituted the highest impact of the new assessment plan for students (agreement 88%).

Table- 1: Descriptive summary and reliability of scores for each type of assessment.

Assessment Type	**Mean ± SD**	**Median**	**Min.- Max. (100)**	**Cronbach Alpha Reliability Coefficient**
MCQ	66 ± 10.7	68	11 - 90	0.81
OSCE	63 ± 14.3	64	24 - 91	0.80
OSCC	77 ± 14.7	80	0 - 100	Not applicable
Structured Viva	76 ± 17.6	80	0 - 100	Not applicable
Combined	68 ± 10.7	70	11 - 89	Not applicable

SD: Standard deviation

Table- 2: Correlation between assessment measures and between each type and the combined score.

	OSCC	**Structured Viva**	**MCQ**	**Writing/Presentation**	**Combined**
OSCE	0.412*	0.498*	0.611*	0.489*	0.749*
OSCC		0.458*	0.471*	0.416*	0.658*
Structured Viva			0.570*	0.456*	0.777*
MCQ				0.474*	0.861*
Writing/ Presentation					0.642*

r: Pearson's correlation coefficient *p-value < 0.05 is significant

Table- 3: Correlation between the scores of the mid- and final exams (Concurrent Validity Coefficient)

	Ẋ± SD	**r**	***p*-value**

Mid- exam scores	78.3 ± 10.2	**0.672***	**<0.0001**
Final exam scores	70.6 ± 11.7		

r: Pearson's correlation coefficient; *p-value < 0.05 is significant

Table- 4: Grade Distribution compared between 2012- 2013 and 2013- 2014 exams

Grade	**2012- 2013 Exam**		**2013- 2014 Exam**		**Change Rate (%)**	**Z**	***p*- value**
	Frequency	**(%)**	**Frequency**	**(%)**			
A	70	18.6	43	13.1	38.5 decrease	-3.986	<0.05*
B	188	50	139	42.4	26 decrease		
C	92	24.4	98	29.9	6.5 increase		
D	26	6.91	38	11.6	46 increase		
F	6	1.56	8	2.4	33 increase		
Success Rate	98.2% (376/383)		97.5% (318/326)		0.7 decrease	-0.577	>0.05

* p< 0.05 is significant

Table- 5: Identified categories and themes resulting from FGDs with faculty and students

Categories	**Themes/ Statements (FS: Faculty) (SS: Student)**	**% Yes**	
		Faculty	**Students**
Departmental/ Strategic	**1.**Ideal practice in exam preparation provided evidence for reliability & validity of results: ***FS1.*** *The new method of preparing blueprint is awesome; items not only covered the content but also the number of items fit into the exam duration* ***FS2.*** *It is mandatory to know exactly what we need from our students to be able to teach and assess them fairly.* ***FS3.*** *Analyzing the exam items post-exam allowed us to evaluate questions and categorize them according to difficulty and discrimination power.* ***SS1.****The exam covered all the topics* ***SS2.*** *The exam was of balanced difficulty; questions stimulate thinking and others needed memorizing.*	100%	70%
	2. Alignment between teaching & assessment is mandatory: ***FS.*** *Students will have no excuse to complain that the exam differed from the way they were taught* ***SS.*** *The exam reflected what we took in class*	100%	100%
Average %		**100%**	**85%**
Economical/ Resources	OSCC& structured viva reduced the need for resources: ***FS.*** *OSCC & structured viva decreased the need for large number of real patients during the exam.*	100%	**-**
Average %		**100%**	**-**
Social/ Behavioral	**1.** Positive impact of assessment plan on students: ***SS1.*** *The end of rotation exam in the new way was very helpful on the long run through moving from one rotation to the other.* ***SS2.*** *All students in my group were examined on the same case and asked the same questions.* ***SS3.*** *Knowing that the exam will stress on history taking and physical examination urged me to do more clerking during the course.* ***FS1.*** *There is less stress and fear among students from the exam.* ***FS2.*** *It is unanticipated that announcing the new way of assessment to students increased their clerking activities. It is a positive point.*	85%	95%
	2. Positive impact of assessment plan on faculty performance: ***FS1.*** *We are now more targeted in our teaching and assessment*		

Categories	**Themes/ Statements**	**% Yes**	
	FS2. *I follow the same path as my colleagues in teaching during sessions* ***SS1.*** *If I miss a session and attend to another tutor, he explains in the same way as my original tutor.* ***SS2.*** *Most of our instructors in clinical sessions are well organized.*	80%	90%
	3. Appreciation of the effort exerted in improving the curriculum: ***SS1.*** *The course highlights great effort that has been put into it.* ***SS2.*** *The assessment plan is well thought out.* ***FS.*** *Reaching consensus among us on the new assessment plan was not easy. We needed evidence that the change we are doing is worth the risk and effort.*	100%	65%
Average %		**88%**	**83%**
Technical/ Developmental	**1.** Objective structured exams ensured fairness: ***FS.*** *The difference between assessors has been reduced.* ***SS.*** *I feel comfortable that I am fairly treated and judged.*	100%	95%
	2. Objective structured exams standardized the line of discussion between faculty and students: ***FS1.*** *I feel now we have the trend of following the same line of discussion during clinical sessions.* ***FS2.*** *Now I know how to prepare my students for the clinical exam.* ***SS.*** *I feel instructors are now more focused on how to teach a topic.*	85%	78%
	3. Integrated application of knowledge in clinical teaching and its reflection in assessment: ***SS1.*** *The curriculum is well structured and organized.* ***SS2.****Clinical sessions were rich in clinical cases, vast amount of knowledge, and evidence-based discussion.* ***FS.*** *We assess what we teach*	100%	100%
	4. Focused structured discussions engaged students and increased attendance: ***SS1.*** *Sessions aroused my curiosity* ***SS2.*** *The benefit of sessions now outweighed that of traditional sessions* ***FS1.*** *Attendance of students increased* ***FS2.*** *Students are very interactive and ask a lot of good questions.*	85%	80%
	5. OSCC & structured viva triggered thinking & application of knowledge rather than recalling of information only: ***SS1.*** *Discussion in the exam lined with the way we are taught.* ***SS2.*** *This is one of the best exams I have ever experienced; it is case-based and involved scenarios.* ***FS.*** *OSCC & structured viva open channels between me and student to explore their way of thinking.*	77%	86%
Average %		**90%**	**88%**

4. Discussion:

The qualities of a good exam include objectivity, validity and reliability of the results, practicability, acceptability, and positive educational impact [40]**.** This study aimed at producing a comprehensive package of evidence to prove validity of students' clinical performance assessment results (as defined by Messick's framework). A newly adopted assessment plan by the Surgery Department was used to build multiple sources of evidence through several statistical procedures that were performed on the test responses. Each test alone yielded scores that were taken to be a measure of each single aspect of validity of the test results (content, consequential, concurrent, predictive and concurrent validity) [25]**.**

The Postgraduate Medical Education and Training Board (PMETB) [39] defined content validity as sampling what the student was expected to achieve and demonstrate. To achieve evidence of content validity, the assessment had to be representative and should cover several categories of competence, a range of patient problems and a number of technical skills. Content validity coefficient was calculated by comparing the list of content and skill areas in the Surgery course with those in the actual test. This was achieved in this study through the table of specifications which showed equal weights of the learning objectives covering the content and the weight of items corresponding to them in each domain. This ensured quantitatively that the weight of different domains in the test equally reflected their weight in the curriculum. This represented the first evidence in the chain of evidence of the unitary construct validity of the results of the newly adopted assessment plan [24]**.** This aspect of validity was the one of greatest concern to the teachers, though they should also pay serious attention to consequential validity.

Consequential validity referred to the effect that assessment could have on learning, and in particular on what students learn and how they learn it. Usually unintended consequences might result from newly adopted assessment measures. Such unintended consequences could be negative, for example students might omit certain aspects of the curriculum because they did not expect to be assessed on them, or they might commit large bodies of factual knowledge to memory without really understanding it in order to pass a test of factual recall and then forget it soon afterwards. Both behaviors would indicate that the assessment had poor consequential validity because both lead to bad learning practices. These unanticipated consequences signaled that the test development had been off-target or incomplete [21,28]**.** In our study, the unintended consequences were positive, whereby the newly adopted assessment plan impacted the students' behavior which was demonstrated as better accountability at clerking, writing and presenting clinical cases. Students' learning also improved as evident by their deep engagement and questioning during clinical sessions. This was unfolded during the FGD with both students and faculty. Both expressed that the highest scrutiny of focusing assessment on history taking and physical examination urged students to more clerking and accountability to pursue better path in history taking and examination. This resulted in better training before the summative exam. Faculty followed almost identical path in clinical teaching after having consensus on the learning outcomes and what were exactly expected of the students. Consequently, students are both taught and assessed equally and fairly. The targeted clinical sessions with clear goals and objectives engaged the students and increased their attendance. The students' learning improved and this was evident by the depth and breadth of their questions during clinical sessions. The positive unintended consequences provided evidence of consequential validity and the impact of the newly adopted assessment plan on students' and faculty's behavior.

According to Zumbo [15] validation is a scientific activity based on the collection of multiple and diverse types of evidence. To complete the validation process of the newly adopted assessment plan, we collected evidence on concurrent validity. This was the degree to which a measurement instrument produced the same results as another accepted or proven instrument that measures the same parameters [41]. The APA [14], also defined concurrent validity as to how close the scores were on two different tests that claim to measure the same construct. For this reason we calculated the correlation coefficient between the scores of the mid and final exams, whereby both measured the same construct using different assessment tools. The scores of both tests significantly positively correlated to each other, which validated the newly used assessment tools, namely OSCC and structured oral exam. Moreover, there was a significant positive correlation between OSCC, Structured oral exam and Writing/Presentation that occurred during the course, which provided evidence on alignment of teaching and assessment and of concurrent validity of scores resulting from these tests. This was emphasized by the comments of faculty and students during the FGDs who expressed their satisfaction of the alignment between teaching and assessment and also how their end of rotation exams and mid exam prepared them well to the final exam.

Predictive validity defined as the degree to which a measure accurately predicted the expected outcomes, so, for example, a measure of attitudes towards preventive care should correlate significantly with preventive care behaviors [41]. In our study, the students were in the undergraduate level, so the predicted behavior would be measured at the third tier of Miller's pyramid "shows" in an OSCE or through an OSCC followed by structured oral exam. We provided evidence of predictive validity by calculating correlation coefficient between the OSCE and OSCC both of which measured the application of clinical knowledge, and the MCQ exam and structured oral exam both measured the content knowledge. Both proved to be significantly positively correlated. The faculty's and students' comments also matched with this evidence whereby both admitted that training on approaching patients holistically triggered their thinking and allowed them to apply clinical knowledge into the OSCC and structured oral exam. This could also be proved by the significant positive correlation between the scores of the MCQ exam and the scores of other clinical tests, which underpinned the common cognitive basis of the used assessment measures. The students also emphasized the application of clinical knowledge into teaching in clinical sessions and how this was reflected in assessment.

We could not advocate an argument on the validation of the new assessment tests by using a fragmented approach of presenting evidence. This was reported by Crocker and Algina (1986), Messick (1989), and Hubley & Zumbo (1996) [17, 18, 19]. Since then, there was a shift from many types of validity to a single, but integrated type conceptualized as construct validity [22]. All the calculated correlation coefficients provided a comprehensive integrated quantitative evidence of construct validity in its unitary concept. Furthermore, there was a significant positive correlation between the scores of the clinical assessment measures and the MCQ scores, although they were assumed to measure different constructs. This denoted that the MCQ measured clinical knowledge which lied within the same construct of clinical skills. We also added to these statistical inferences the highest correlation coefficients which occurred between each assessment measure (OSCE; OSCC; Structured viva; MCQ, and Writing/Presentation) and the combined score. This further provided evidence on construct validity because the assessment measures were complementary and compact. This was proved by Simon *et al*. and Kreiter & Bergus (2007) [41,42]**.** Although there was a statistically significant positive correlation between OSCE and OSCC scores and between each one of them and the combined score; the correlation coefficient between each and the combined score was much higher. This emphasized that the scores of these overlapping interconnected assessment measures when combined can convey information about a more thorough definition of clinical competency. It also made it possible to produce a more reliable valid score for assigning final grades compared to using either measure separately [44].

The reliability of test results was also a component in Messick's framework which added evidence to construct validity of the adopted assessment. In this study, the internal consistency of the MCQ exam and OSCE was close, and was represented as reliability coefficients. This alone did not add much to the evidence of construct validity. However, when related with the high correlation between each measure with the combined score, "relation with other variables", such significant correlations altogether, underpin the construct validity in its broader meaning as defined by Messick [20].

All the quantitative results provided above constituted a strong body of evidence that could be used to advocate the validity of assessment results in its broader meaning, thus answering the first research question.

When we studied the grades between the 2012-2013 and 2013-2014 cohorts, the success rates were the same. Nevertheless, the distribution of grades were significantly different; whereby the grades became more normally distributed reflecting the objectivity and fairness of assessment. This matched with the students' perception of the positive impact of assessment plan and their feelings of the fairness of their grades being assessed on the same case and asked the same questions. However, alone, success rates and grade distribution could not provide solid evidence to advocate an argument on the validity of results. This answered the second research question.

All these done, Messick [16] stated that validity was an evaluative summary not only of scientific evidence, or potential and actual consequences, but also a summary of the underlying values. He explained that there was synthesis between facts and values. We tried to identify diverse underlying values in the Surgery course documents, course designers, instructors and students. We tried through the analyzed results from the FGDs to weigh the balance and compare these values for convergence and made a final judgment as to the extent which these values interplay during course implementation. We found that such underlying values constituted part of the body of evidence of construct validity of assessment results and hence validation of the assessment plan being adopted.

Conclusion:

A newly introduced assessment plan with new tools had to be validated by pursuing a comprehensive, unified approach to create evidence from multiple sources of data in order to support the argument of advocating the assessment results. The sources of evidence could be statistical calculations as correlation coefficients, qualitative data, reflections on one's own values and those of others, and an analysis of unintended consequences [16].

Acknowledgements:

We very much appreciate the Head of Surgery Department, Doctor Hussein Jabbad, for his commitment to provide us with the required data.

Declaration of interest: The authors report no declarations of interest.

References:

1. Beckman TJ, Cook DA, Mandrekar JN. 2005. What is the validity evidence for assessments of clinical teaching? J Gen Intern Med 20:1159-1164.
2. Miller G. 1990. The assessment of clinical skills / competence / performance. Academic Medicine 65 (Suppl.9): S63-S67.
3. Wass V, Van der Vleuten CPM, Shatzer J, Jones R. 2001. Assessment of clinical competence. Lancet 2001; 357:945-949.

4. Benner, P. (1984). From novice to expert: Excellence and power in clinical nursing practice. Menlo Park: Addison-Wesley, pp. 13-34.
5. Barman A. 2005. Critiques on the Objective Structured Clinical Exam. Ann Acad Med Singapore 34: 478- 82.
6. McGrath P, Moxham L, Fox-Young S, *et al.* 2006. Collaborative voices: Reflections on ongoing issues regarding nurse competencies. Contemporary Nurse 22(1): 46-58.
7. Brosnan M, Evans W, Brosnan E, *et al.* 2006. Implementing objective structured clinical skills evaluation (OSCE) in nurse registration programmes in a centre in Ireland: A utilisation focused evaluation. Nurse Education Today 26, 115-122.
8. Major DA. 2005. OSCEs seven years on the bandwagon: the progress of an objective structured clinical evaluation program. Nurse Education Today. 25 (6): 442-454.
9. Benner P, Tanner CA, Chesla CA. 1996. Expertise in Nursing Practice: Caring, Clinical Judgment, and Ethics. Springer Publishing, New York.
10. Bujack L, McMillan M, Dwyer J, *et al.* 1991. Assessing comprehensive nursing performance: the Objective Structured Clinical Assessment (OSCA): Part 1 - Development of the assessment strategy. Nurse Education Today 11, 179-184.
11. Hassan S. 2011. Oral examination as objective structured authentic viva (OSAV). NMJ; 3 (3 & 4): 35-40.
12. Kane MT. 2006. Content-related validity evidence in test development. In: Downing SM, Haladyna TM, editors. Handbook of Test Development. Mahwah, NJ: Lawrence Erlbaum Associates 131-153.
13. Todd MR, Moore CEG. 2010. Medical Finals: Passing the Clinical. 3rd Edition. Carnegie Book Production, Lancaster Printed and bound in the UK by CPI Antony Rowe; xxiv-xxv.
14. American Psychological Association (APA). (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, DC: The Association.
15. Zumbo BD. 2007. Validity: Foundational issues and statistical methodology. In CR Rao & S Sinharay (eds), Psychometrics (Handbook of statistics, vol. 26, pp. 45-79). Amsterdam: Elsevier.
16. Messick S. 1995. Validity of psychological assessment: Validation of inferences from person’s responses and performances as scientific inquiry into score meaning. American Psychologist; 50: 741-749.
17. Crocker L and Algina L. 1986. Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston, Inc.
18. Messick S. Validity. 1989. In: Linn RL, editor. Educational Measurement,3rd Ed. New York: American Council on Education and Macmillan.
19. Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. The Journal of General Psychology, 123, 207–215.
20. Messick, S. (1998). Test validity: A matter of consequences. Social Indicators Research, 45, 35–44.
21. Ruhe V and Zumbo BD. 2009. Evaluation in distance education and e-learning: The unfolding model- Guilford Press; NY 10012. Ch 2: The theory and practice of program evaluation and Ch. 3: Evaluation theory and practice in distance education and e-learning.
22. Cook, DA and Beckman, TJ. 2006. Current concepts in validity and reliability for psychometric instruments: Theory and application. The American Journal of Medicine; 119: 166.e7- 166.e16.
23. Ahmad RG, Hamed OAE. 2014. Impact of adopting a newly developed blueprinting method and relating it to itemanalysis on students’ performance. Med. Teacher 36: S55 – S61.
24. Downing SM. 2003. Validity: on the meaningful interpretation of assessment data. Medical Education37:830-837.
25. Downing SM. 2004. Reliability: on the reproducibility of assessment data. Medical Education 38:1006-1012.
26. Foster SL, Cone JD. 1995. Validity issues in clinical assessment. Psychol Assess 7:248-260.
27. Reckase, M. (1998). Consequential validity from the test developer’s perspective. Educational Measurement: Issues and Practice, 17(2), 13-16.
28. Jones MG; Jones BD; Hargrove T. 2003. The unintended consequences of high-stakes testing. Rowman & Littlefield Publishers, Inc. USA.
29. Messick S. 1996. Validity and washback in language testing. J. Language Testing; 13 (3): 241-256.
30. Bordage G. 2009. Conceptual frameworks to illuminate and magnify. Medical Education 43: 312- 319.
31. Cronbach LJ. 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16:297–334.
32. Kirkpatrick DL. 1998. Evaluating training programs: The fur levels, 3rd edn., Berrett-Koehler Publishers, Inc. San Francisco, CA.
33. Kirkpatrick DL and Kirkpatrick JD. 2005. Transferring learning to behavior: Using the four

levels to improve performance, Berrett-Koehler Publishers, Inc. San Francisco, CA.

34. Bloor M, Frankland J, Thomas M *et al*. 2001. Focus groups in social research. London: SAGE Publications.
35. Kreuger RA. 1994. Focus groups: A practical guide for applied research. 2nd ed. Thousand Oaks, CA: SAGE Publications.
36. Al-Wassia R, Hamed O, Al-Wassia H *et al.* 2015. Cultural challenges to implementation of formative assessment in Saudi Arabia: An exploratory study. Med. Teach.; 37: S9-S19.
37. Schwartz A, editor. 2011. Assessment in Graduate Medical Education: A Primer for Pediatric Program Directors.Chapel Hill, NC: American Board of Pediatrics.
38. Van der Vleuten C. 1996. The assessment of professional competence: developments, research and practical implications. Advances in Health Sciences Education 1: 41-67.
39. PMETB. 2007. Developing and maintaining an assessment system - a PMETB guide to good practice. Available from: http://www.gmcuk.org/Assessment_good_practice_v0207.pdf_31385949.pdf.
40. Muller ES, Harik P, Margolis MJ, *et al*. 2003. An examination of the relationship between clinical skills examination performance and performance on USMLE Step 2. Acad Med 78 (10 Suppl): S27- 29.
41. Simon SR, Bui A, Day S, *et al*.. 2007. The relationship between second year medical students' OSCE scores and USMLE Step 2 scores. J Clin Practice 13 (6): 901- 905.
42. Kreiter CD, Bergus GR. 2007. A study of two clinical performance scores: Assessing the psychometric characteristics of a combined score derived from clinical evaluation forms and OSCEs. Med Educ Online 12: 10. Available from http://www.med-ed-online.org.

2/24/2016