JOURNAL OF
Surgical Simulation

Journal of Surgical Simulation (2018) **5,** 8–23
DOI: https://doi.org/10.1102/2051-7726.2018.0002

ORIGINAL ARTICLE

Lap Mentor-based assessment of laparoscopic surgical skills: a validation study

Khalid Munir Bhatti,[a],* Lubna Baig,[b] Syed Moyn Aly,[c] Kamran Ahmad Malik,[d] Hafiz Amjad Hussain,[e] Zainab Nasser Al-Balushi,[d] Mahmoud Hatem Sherif,[d] Khoula Saud Al Harrasi,[f] Kadhim Mustafa Taqi,[f] Houd Al Abri[g] and Hani Al-Qadhi[d]

[a]Royal Derby Teaching Hospital, Derby, Derbyshire, UK; [b]Institute of Public Health, Jinnah Sindh Medical University, Karachi, Pakistan; [c]Department of Medical Education, Jinnah Sindh Medical Education, Karachi, Pakistan; [d]Sultan Qaboos University Hospital, Muscat, Oman; [e]Allied and DHQ Hospital, Faisalabad, Pakistan; [f]Oman Medical Speciality Board, Ministry of Health, Muscat, Oman; [g]Medical Simulation Center, Oman Medical Specialty Board, Muscat, Oman

**Corresponding author at: Department of Medical Education, Royal Derby Teaching Hospital, Uttoxeter Road, Derby, Derbyshire DE22 3NE, UK. Email: drkhalidmunirbhatti@yahoo.com*

Date accepted for publication: 5 February 2018

Abstract

Objective: To validate the scores of assessments of basic laparoscopic surgical skills obtained through performance on LAP Mentor. **Design**: Cross-sectional validation study. **Setting**: Oman Medical Specialty Board (OMSB) skills lab, Muscat, Oman. **Participants**: Twenty-three surgical residents, registered with OMSB, at different years of residency, underwent assessment of basic laparoscopic surgical skills on the LAP Mentor. A construct validation model was used to validate the scores. **Results**: 35% of the candidates, all belonging to the senior group, passed the assessment. Cronbach's alpha was 0.87 with a standard error of measurement of 1.53 (95% confidence interval). The intra-class coefficient varied from 0.88 to 0.95 for different scales. Factor analysis revealed two underlying constructs, i.e. technical skills and patient safety skills, that explained competency in laparoscopy surgery. A review of the literature supported content validity evidence. Relationships with other variables were documented through convergent and divergent evidence. A correlation for the data revealed that actual residency year and total score achieved were significantly related ($r = 0.51$, N = 23, $P = 0.01$, two-tailed). Senior residents scored significantly higher than junior residents on overall performance. Fisher's exact test showed that more seniors than juniors passed on overall performance using specific criteria determined by factor analysis and Angoff's method for standard setting. The majority of residents and raters agreed or strongly agreed on feasibility, conducting such tests regularly, meaningful feedback, and objectivity and reuse of the tools. **Conclusions**: Assessment on LAP Mentor using different rating scales and construct-based standard setting methods provides meaningful scores. Periodic summative assessment is acceptable to residents.

Keywords: *LAP Mentor; laparoscopic skills; simulation; construct validation; assessment*

Introduction

Laparoscopy has revolutionized the field of surgery and it is now necessary for surgeons in training to achieve competency in this important skill. Due to the steep learning curve and increased incidence and severity of complications in the early phase of learning,[1] it seems imperative that surgeons in training should be able to demonstrate competency in laparoscopic surgery. Hence, assessment is necessary during and at completion of training. However, assessment is a complex task and requires consideration of many factors.

Any assessment process involves presentation of a stimulus, response of the examinee, collection of data, transformation of data to a score, and interpretation of scores.[2] For assessment of laparoscopic surgical skills, the stimulus is a task that could be completed using a laparoscope. Many models can be used for stimulus presentation and completion (real patients, cadavers, animals, and simulators).[3] The response of the examinee depends on many factors, including differences in competency levels.[2] Different tools have been used previously for data recording, such as in-training evaluation reports (ITER), procedural logs, objective structured assessment of technical skills (OSATS) and global operative assessment of laparoscopic skills (GOALS).[4] Data collected through these tools are transformed into a score. For interpretation to be reasonable and the meaning of the data to be legitimate, this process should be error free and defensible. This approach ensures the validity of scores and the meanings associated with them.

LAP Mentor is a high-fidelity simulator. It has primarily been developed to train surgical residents in laparoscopic surgical skills. A review of the literature shows that simulation-based assessment using LAP Mentor as a model is feasible.[5,6] However, studies on the validation of assessment are scarce. Even the available studies are more oriented towards the psychometric properties of the assessment tool, i.e. validity and reliability in isolation. Moreover, validity has been divided into different categories such as content, construct, criterion-related and concurrent. Conversely, according to contemporary theory, validity is a unitary concept. It requires collection of multiple pieces of evidence to support or refute the meaning associated with assessment data.[2] With this in mind, the current study aimed to define the role of LAP Mentor as a model for assessment of laparoscopic surgical skills by providing validity evidence for laparoscopic surgical skills assessment. If its role is well defined, we would expect this to help assessors in making informed choices when assessing candidates' competency in laparoscopic surgery. Moreover, mandatory certification through assessment on this model before the qualifying final board examination in surgery would help ensure the achievement of this important competency. This in turn would improve patient care and patient satisfaction.

We hypothesized that for assessment scores to be valid, senior residents should achieve higher scores and their pass rate should be higher than that of the junior residents.

Material and methods

LAP Mentor was used as a model for stimulation presentation and task completion (a five-task test: Appendix 1). A task-specific assessment tool comprising a checklist, a global rating score (GRS), and a year in training (YIT) score was used for collection of assessment data (Appendix 2). Pass/fail decisions were based on the borderline method for individual tasks and Angoff's method for the overall result.[2] A test report form was given to students for meaningful feedback (Appendix 3). The validation process involved collection of evidence on content, response process, internal structure, relation to other variables, and consequences of testing.

Study design and setting

A cross-sectional validation study was used. The study was conducted at Oman Medical Specialty Board (OMSB) skills lab, Muscat, Oman.

Study duration

After approval from the departments of Dow University of Health Sciences (DUHS), Karachi, Pakistan; Sultan Qaboos University Hospital (SQUH), Muscat, Oman; OMSB, Muscat Oman; and the OMSB skills lab, this study was conducted over a period of 6 months from November 2015 to May 2016.

Sample size

Twenty-six surgical residents, registered with the OMSB, from different residency year groups, were invited to take part in the study, of whom 23 participated. Informed consent was given by the participants, after the purpose, procedure, benefits and costs of the study had been explained by the principal investigator. No formula was used to calculate sample size because the entire population of surgical residents in the institution was included. Moreover, there were no available data on mean scores and no effect size was expected. Nevertheless, sample size was in accordance with the number of participants in similar studies available in literature.[7,8]

Inclusion and exclusion criteria

All the male and female postgraduate general surgical residents (PGY1-5), who were registered with OMSB for training in general surgery and had completed basic modules of laparoscopic training on LAP Mentor, were invited to participate in the study. Residents who did not agree to be part of the study were excluded.

Assessment process

A five-task test was devised based on basic and essential modules of LAP Mentor (Appendix 1). The tasks included grasping and clipping, use of electrocautery, peg transfer, placement of a ligating loop, and pattern cutting. The data collection instrument was devised after consultation with experts in the field, who were senior registrars or above with at least 10 years' experience in laparoscopic surgery. The instrument consisted of a checklist, GRS and YIT scale (Appendix 2). Checklists and GRS are valid tools for data collection. The YIT scale has recently been validated by Pugh et al.[9] Checklist descriptors and assigned scores for each task were developed through consensus of experts.

The data collection instrument containing all three scales was pilot tested. Participants were blinded because they were not aware that they were being assessed. This testing was carried out during their routine practice on the LAP Mentor. Assessment was by two clinical supervisors assigned for this purpose. Feedback was taken from the assessors on the data collection instrument. This strategy avoided induction of bias. After pilot testing, the instrument did not require any modification.

Participation in the study was voluntary. The assessment procedure on LAP Mentor was explained to the participants. To avoid any bias due to lack of orientation on the LAP Mentor tasks, all the participants had already completed basic modules. On the stations where no prior practice had been done, orientation was given to the participants. Before the start of the test, every resident completed a questionnaire outlining their previous laparoscopic experience in terms of procedures completed (under supervision or independently) and any advanced laparoscopy courses (Appendix 4). Due to recall bias, previous laparoscopy experience was not taken into account for the purpose of inferential statistics. Residents were divided into two groups based on their actual residency years (ARY). Residents of years 1–3 were grouped as juniors and those of years 4 and 5 were grouped as seniors.

The five-task test was administered at the OMSB skills lab (Appendix 1). Each resident was issued with an examination number and was asked to complete the five tasks. They were allowed only one attempt. Performance was recorded by an external camera for later review by two external raters who were senior surgeons, well trained in the assessment of laparoscopic surgical skills. During video recording the identities of the residents were masked to avoid rater bias. There was no time restriction to complete the tasks; however, the time taken to complete each task was taken into account when marking on all rating scales. Testing continued over a period of 8 weeks until the last candidate completed the test. Participants were not directly observed and no feedback was provided at the time of the performance.

Data analysis

SPSS version 19 was used for data entry and analysis. Descriptive statistics were also used to analyze minimum, maximum, and average score with standard deviation (SD) on each station and overall performance.

Cronbach's alpha was used to determine reliability by internal consistency. Inter-examiner correlation was determined using intra-class correlation (ICC) for checklist, GRS and YIT scale scores. Factor analysis was done to document internal consistency in terms of the extent to which individual items on the checklist explain the construct of interest. Validity evidence was collected by determination of the correlation of ARY to the checklist, GRS and YIT scale scores using Pearson's correlation coefficient. Comparison of mean scores between junior and senior residents was done using a Mann-Whitney U test (Levene's test of homogeneity was negative). Regression analysis was done to determine the predictor of the scores. Gender was assumed as a co-variable to year in training, whereas age was not, because higher age is associated with higher year in training. An additional source of validation evidence was collected by comparison of pass/fail ratios between senior and junior groups of residents. As two cells (50.0%) had an expected count of less than 5, Fisher's exact test was used for this analysis rather than a chi-squared test. A P value <0.05 was taken as significant.

Standard setting

Each station was marked by two independent examiners, and the average total score for each station was calculated. A category result for each station based on a global rating scale was assigned. Total marks for the whole five-task test were also calculated. As determined by factor analysis, two underlying constructs were identified, and the total score on each construct was assigned. The borderline method was used to determine the cut-off score for each station. For this method, checklist scores of the borderline candidates were used to calculate a cut-off score.[11] Standard error of measurement (SEM) for the whole test was calculated by taking reliability and SD into account. However, no adjustments in the score were made. The minimum level of competence on each construct was decided by a group of four examiners, using Angoff's method. They were asked to examine the items of components of underlying constructs and then to predict how many minimally qualified candidates would be able to perform adequately on those items.[11] So to pass the test overall, the candidate had to a achieve score above the cut-off for total scores determined by the borderline method and also above the cut-off for each construct determined by Angoff's method.

Score reporting and collection of feedback

Each candidate was given detailed feedback showing their mark on each task, task cut-off with SD, overall mean, mean on each task, total score on each construct, cut-off on each construct, and overall decision. Results were sent to candidates approximately 4 weeks after the test in the format shown in Appendix 3.

Participants were asked to fill in a feedback form to give their opinions about both the test and the test report form (Appendices 3 and 5A). The test report forms for the candidates showing their scores were given to the raters. The raters were then asked to fill in feedback forms (Appendices 3 and 5A).

Collection of validity evidence

After completion of the study and reporting of the results, evidence for validity was collected. A construct validation model was used for this purpose. The process included documentation of five major sources of validity based on the Standards for Educational and Psychological Testing Standards as described by Downing and Yudkowsky.[11]

Results

Of 26 general surgical residents invited, 23 participated in the study (12 female, 11 male). Age ranged from 23 to 32 years with a mean of 27.39 ± 2.17 years (SD). Almost all were right handed (22 versus 1). Depending on their experience, 16 residents were classified as juniors and seven as seniors. Seventeen (74%) had previous training on a box trainer. Ten (43.5 %) residents had advanced laparoscopic training and were certified on an advanced laparoscopic surgery course.

Table 1 presents the minimum, maximum, and average score for each task with SD. The minimum score for the five stations collectively was 27.99/50, and the maximum was 44.75/50 with an average of 32.21 ± 5.14 (SD). The cut-off value for these stations together was 28.84. Although, 22 students achieved scores above the cut-off value, only eight students were given a pass (34.78%) using the specific criteria described in the methodology section.

Cronbach's alpha value for this five-task test with 28 measurements marked by two examiners was found to be 0.87 with an SEM of 1.53 (95% CI). Application of "Cronbach's alpha if item deleted" did not improve the value of Cronbach's alpha. A high degree of reliability was found between two raters on all three scales, i.e. checklist (ICC, 0.88; CI 95%, 0.72–0.97; $P = 0.00$), GRS (ICC, 0.95; CI 95%, 0.88–0.98; $P = 0.00$), and YIT (ICC, 0.86; CI 95%, 0.67–0.94; $P = 0.00$).

Thematic analysis identified eight items from the checklists relating to the five tasks. These included smooth movement, both-hand coordination, hand–eye coordination, no drop of object, appropriate time to complete, safe application, no collateral damage and appropriate tissue traction. Scores on these items were combined and subjected to factor analysis. A Kaiser-Meyer-Olkin value of 0.71 and Bartley test values (chi-squared 94.185, df 28, significance 0.00) showed adequacy of the sample.

Results of the factor analysis showed that five factors loaded to component A and three factors loaded to component B (see Table 2). Careful analysis of these factors showed that

Table 2. Factor analysis with item loadings to two components

Factor analysis		Component	
		A	B
Factor loadings	Smooth movement	0.779	0.503
	Both-hand coordination	0.533	
	Hand–eye coordination:	0.868	
	No dropping of object	0.855	
	Appropriate time to complete	0.564	
	Safe application		0.945
	No collateral damage		0.870
	Appropriate tissue traction		0.918
Initial eigenvalues		3.27	2.31
Extraction sums of squared loadings (% of variance)		40.97	28.82
Rotation sums of squared loadings	(% of variance)	35.96	33.84
	Cumulative %	35.96	69.79
	Extraction method: principal component analysis Rotation method: varimax with Kaiser normalization		

Table 1. Descriptive statistics

		Minimum	Maximum	Mean	Standard deviation	95 % confidence interval	Range	Cut-off score (borderline method)
1. Grasping and clipping	Average score of two examiners (checklist)	3.25	10.00	6.35	1.67	±0.68	5.67–7.03	6.00
2. Use of electrocautery	Average score of two examiners (checklist)	.50	9.50	5.19	2.63	±1.07	4.12–6.26	5.10
3. Peg transfer	Average score of two examiners (checklist)	5.50	10.00	8.98	1.70	±0.69	8.29–9.67	5.94
4. Application of loop	Average score of two examiners (checklist)	2.00	9.50	6.83	1.90	±0.78	6.05–7.61	5.93
5. Pattern cutting	Average score of two examiners (checklist)	5.00	9.50	6.84	1.33	±0.54	6.3–7.38	5.87

the five factors were related to equipment handling, whereas the three factors were related to safe surgical approach. These were named technical skills and safe laparoscopic surgical practice, respectively.

Correlation for the data revealed that residency year and total score achieved were significantly related ($r = 0.52$, N = 23, $P = 0.01$, two-tailed). A strongly positive and statistically significant correlation was found among the three scales used to collect data on performance of the resident to assess laparoscopic surgical skills ($r = 0.95–0.96$, N = 23, $P < 0.01$). A Mann-Whitney U test was used to compare the scores on total performance between the junior residents and senior residents. Senior residents scored significantly higher than junior residents on overall performance (N = 7, mean rank 18.00, sum of ranks 126.00 versus N = 16, mean rank 9.38, sum of ranks 150; Mann-Whitney U = 14, $P = 0.00$) (see Table 3).

Regression analysis showed that total scores on the checklists could be predicted by actual residency year (β = 1.24, $P = 0.04$), whereas gender (β = −0.30, $P = 0.86$) was not a significant predictor (see Table 4).

The pass rate was high (96%) when the borderline method was used, and the total cut-off score was calculated by combining cut-off scores on individual stations. However, using Angoff's method, and determining minimal competency on individual constructs rather than on tasks, produced more meaningful results: a pass rate of 35% and a statistically significant difference between junior and senior residents (see Table 5).

Table 3. Results of Mann-Whitney U test comparing mean and rank sums of total scores for junior and senior residents

	Ranks					
	Level of residency	**N**	**Mean rank**	**Rank sum**	**Mann-Whitney U**	***P* value**
Overall score	Junior	16	9.38	150.00	14.000	0.005
	Senior	7	18.00	126.00		
	Total	23				

Table 4. Results of regression analysis

		Coefficients				
Model		**Unstandardized coefficients**		**Standardized coefficients**	**t**	**Significance**
		B	**Standard error**	**Beta**		
1	(Constant)	32.23	2.97		10.85	*0.00*
	Actual residency year	1.24	0.56	0.45	2.22	*0.04*
	Gender	−0.30	1.70	−0.04	−0.18	0.86

Dependent variable: total score (checklist)

A total of 20 residents, four raters, and the program director completed feedback forms. Most of the residents and raters agreed or strongly agreed on the feasibility of the tests and that they should be conducted regularly. Similarly, they agreed that the feedback form provided meaningful feedback compared with the performance metrics produced by LAP Mentor: 75% of the raters and 45% of the residents did not agree that performance metrics produced by LAP Mentor should be included in the test report form (Appendix 3). Regarding the different scoring system, 100% of the raters agreed on reuse of the checklist and GRS, whereas 75% agreed to reuse YIT. The response of residents was variable: 80% agreed on reuse of the checklist, 70% on reuse of GRS, and 60% on reuse of the YIT scale.

Discussion

The aim of the current study was to validate the scores of assessment of basic laparoscopic surgical skills obtained using LAP Mentor. A construct validation approach recommends collection of supporting data, based on five types of evidence, i.e. content, internal structure, response process, correlation to other variables, and consequences.[10,11] The current study documents the evidence on most, if not all, of these aspects.

Regarding content validation, answers to certain questions can support validity. For example, were the tasks included in the test representative of the construct of interest? Were the tasks sufficient in number to avoid the threat to validity of construct under-representation? Other areas that can support content validity include documentation of the evidence that there was no induction of error during the process of

Table 5. Results for Fisher's exact test comparing pass/fail ratio between junior and senior residents according to two standard setting techniques

Standard setting technique	**Level of residency**	**Pass/fail**		**Total**	**Fisher's exact test (*P* value)**
		Pass	**Fail**		
Borderline method	Junior	15	1	16	0.696
	Senior	7	0	7	
	Total	22	1	23	
Angoff's method	Junior	2	14	16	0.002
	Senior	6	1	7	
	Total	8	15	23	

construction of test tasks, and that the checklists included items that were truly representative of the underlying construct.[12]

The current study included a five-task test. The tasks comprised peg transfer, application of a loop, pattern cutting, grasping and clipping, and use of electrocautery. The first three tasks are the same as those included in the FLS course.[5] There is a consensus that these tasks are representative of the domain of laparoscopic surgical skills and have transfer validity.[5,13–15] However, in our assessment for basic laparoscopic surgical skills certification, the two tasks of intra- and extra-corporeal suturing were replaced. The reasons for doing so included limited exposure of residents to these skills on real patients. Moreover, to document construct validity, we needed tasks for which differences in construct existed according to level of residency. Hence, these two tasks were replaced with two other tasks on which senior residents had previously had enough practice while doing surgery on real patients. Grasping and clipping are skills commonly used while performing laparoscopic cholecystectomy, one of the most common surgical procedures. Similarly, use of electrocautery is a common task, because this skill is used when performing cauterization of the mesoappendix during laparoscopic appendectomy, and while removing the gall bladder from its bed during laparoscopic cholecystectomy. There was a consensus among local experts in the field that these two tasks can better assess competency in basic laparoscopic surgery than the rarely used intra- or extra-corporeal suturing modules. The most commonly used basic laparoscopic surgical skills include these five tasks, and hence by including all of these tasks the issue of construct under-representation was dealt with. Errors in checklist construction were minimized because these were created by experienced surgeons and were also pilot tested. Moreover, items on the checklists (see Appendix 2, tasks 1–5) were more realistic than those provided by LAP Mentor as objective metrics feedback. Limited construct validity of LAP Mentor metrics has been reported by Andreatta et al.[16] The process of test construction and content representation mentioned in the current study meets the standard criterion of content validation as described by Beckman et al.,[12] according to whom the standard of content validation is documentation of either a well-defined process for developing instrument content or a reference to a previous study on an assessment instrument that meets these criteria.[12] In our study. both of these approaches were used.

The standard criterion for response process validation requires supportive data for documentation of one or more of the following: thought processes, avoidance of halo error, and demonstration of low response error.[12] In other words, response validation should provide evidence that the only difference among the residents was that of level of competency in basic laparoscopic surgical skills and all the sources of possible error were controlled. Critical review of the methodology section of the current study reveals that the conditions were standardized during the assessment process. Assessment prompts were clearly communicated. All the residents were familiar with the format of the examination. All candidates were assessed on three previously practiced and two previously un-practiced tasks. They were assessed at the usual time of their practice on LAP Mentor. The technical performance of LAP Mentor itself remained consistent. On one occasion, when one of the laparoscopic instruments was broken, the study was discontinued for 2 weeks and was restarted only after the problem was resolved. The halo effect was avoided as resident identity was masked during the recording procedure. Calculation of scores for each candidate was double checked. Pass/fail decisions were based on defensible methods. All of these measures helped to deal with construct-irrelevant threats to validity. Low SEM (1.53 with 95% CI) is evidence that there was low response error. Similarly, indirect evidence on the response process can be inferred from the finding that senior residents had higher scores on patient safety skills, probably because of consideration of this aspect while completing assessment tasks. Hence, the current study meets the standard criterion for response process validation.

The standard for internal structure evidence requires documentation of the extent to which individual items within the instrument match the underlying construct. The methods to support this evidence include reporting internal consistency reliability and factor analysis. Internal structure evidence is not unique to the current study. It has been reported by many authors previously in different forms.[17–19] However, supporting data consisted of reliability analysis and inter-rater correlation coefficients.[17–19] The current study is distinctive in that we have reported factor analysis to support internal structure evidence along with other evidence.

A Cronbach's alpha value of 0.87 on a five-task test with 28 measurements is supportive of the homogenous structure of the items on the test.[20] Small SEM on total score (1.53) with 95% CI is evidence of low error in measurement, suggesting high reliability. Additional supportive evidence is the high degree of reliability between the two raters on all three scales, i.e. checklist, GRS, and YIT scales, calculated by ICC. Other studies have reported some of the internal structure data individually. Vassiliou et al.[21] reported a Cronbach value of 0.86 and an inter-rater reliability coefficient for the total scores of 0.998. These values are close to the values reported in the current study. Most of the literature on the validity of assessment on virtual reality has not reported SEM, which is considered to be very important evidence.[2,20]

As mentioned earlier, in the current study, internal structure evidence was supported by factor analysis, which carries a high value as supporting evidence.[20,22] Eight-item factor analysis identified two underlying constructs, i.e. technical skills and safe laparoscopic surgical practice. The constructs are related but are also distinct in that the competency in one does not automatically translate into competency in the other. In other words, basic laparoscopic surgical skill requires technical skills and safe laparoscopic surgical practice. However, having good technical skills does not necessarily mean that a surgeon will apply the principles of safe practice while performing laparoscopic surgery. Identification of these two underlying constructs is important because in order to become certified in basic laparoscopic surgical skills, the candidates must show minimal competency on both constructs.

Relation to other variables is another important standard of validity evidence. To satisfy this standard, one must document that correlation existed between the scores and another measure of the same construct.[12] Ideally, the score on the current instrument must have been correlated to scores on an existing gold standard instrument. As assessment on LAP Mentor in the past was not checklist based, there were no data available. Hence, we used existing differences in residents as a measure of the construct. At the start of the study, we hypothesized that senior residents would have better laparoscopic surgical skills than junior residents and hence would have better scores than junior residents (convergent validity evidence). We also predicted that scores would not differ based on the gender of the residents (divergent validity evidence).

Numerous findings in the current study supported convergent evidence. Senior residents performed better than junior residents in many ways. First, a good positive and statistically significant correlation was found between the total scores and ARY. Similar findings have been reported by many authors.[23–25] Second, scores on all three scales correlated significantly, positively, and strongly with one another, suggesting that these measured the same construct. When a group comparison was made between the senior and junior residents, it showed that the mean scores of the seniors were statistically higher than those of the juniors. Another way to look at the convergent validity was to consider the pass/fail ratio between these two groups. Results for Fisher's exact test showed that this ratio was higher for senior residents. All of the above findings support convergent validity for relationship to other variables.

Divergent evidence was supported by the finding that total scores on the checklist could be predicted by actual residency year and not by gender as determined by regression analysis. As gender is unrelated to competency in laparoscopic surgery, according to expectation, it was not found to be predictive of total performance score.

Consequences is the least well-addressed standard of validity.[26,27] To satisfy this standard, a description of the consequences of assessment that would affect the validity of score interpretations is required.[12] These consequences may be positive or negative. Hence, relevant areas that need consideration include the possible role of testing in improving educational outcomes, acceptability to the residents, avoidance of false positives and negatives and unanticipated harms, and defensibility of standard setting methods.

Based on the Kirkpatrick model, the current study addressed the 'reaction level' of outcomes only.[11] More long-term studies are needed to evaluate higher levels of outcomes. However, our survey from the students showed that most of the residents and raters agreed on conducting such tests regularly. Similarly, they agreed that the test report form (Appendix 3) provided meaningful feedback compared with the objective performance metrics produced by LAP Mentor. This type of feedback from residents and raters demonstrates that they valued this assessment process.

The current study is distinctive in many ways. The use of checklists for assessment on LAP Mentor is unique to this study. Historically, task-specific checklists have been used to calculate scores in assessment of open and laparoscopic surgical skills. Examples include OSATS and GOALS. After the introduction of virtual reality simulators, such a score could not be calculated. Until recently, no method of calculating total scores incorporating measured metrics existed. For instance, Rosenthal et al.[28] have described a method of calculating total scores using a four-step procedure. It includes standardization of metrics, calculation of mean summary measures per dimension, re-standardization and unification of directionalities, and calculation of a weighted average as a total performance score. Our study documents that task-specific checklists can be used reliably and with more feasibility to calculate overall score. These are required to make pass/fail decisions by the borderline method on individual tasks and to provide meaningful feedback. Decisions on overall pass/fail have been based on findings of factor analysis. This has formed the basis for the development of a generic tool for certification of basic laparoscopic surgical skills that requires further psychometric studies (Appendix 6).

Reporting performance on virtual reality simulators is also not an easy task. Three different methods have been described. These include reporting all measured metrics produced automatically by the simulator; reporting some of the metrics (risking partial reporting); and summarizing multiple outcomes to mean summary measures or into a total score. We selected a method that we believe is superior. We provided the candidates with scores on each task, mean scores with SD for each task, cut-off scores for each task, pass/fail decision for each task, performance on GRS for each task, performance on year in training score, criterion for assessment on each task, criterion for overall pass and result on complete performance. This method has been previously validated by Pugh et al.;[9] however, we have modified the test report form (Appendix 3) because our standard setting method was based on findings of factor analysis.

The current study has some limitations. Like most studies on the validity of assessment of LAP Mentor, the number of participants was low. For evidence of relationships to other variables, a built-in difference in construct was used for correlation. Objective metric scores produced by LAP Mentor were not used for correlation or comparison purposes. Item differential is another way to provide evidence of internal structure, which was not reported in the current study. Divergent evidence should come from other assessment scores unrelated to the construct of interest. Similarly, multi-trait multi-model evidence for discriminant validity is considered more valuable. The current study has been validated in our setting and its generalizability is not known, because validity is considered to be a property of the assessment process in the local context. Moreover, the assessment process described in our study is more resource intensive, requiring surgical simulators for assessment and personnel support for marking. The standard setting method requires factor analysis.

Recommendations and further investigation

Validation of scores and interpretation of the scores obtained through assessment on LAP Mentor has been established by this study. The following recommendations should direct future investigations:

(a) A proposed generic instrument (SQUH Tool) containing a construct-based checklist and two other scales can be validated by further studies and, once validated, should be used for assessment on LAP Mentor.

(b) Electrocautery and grasping and clipping should be included in assessments for basic laparoscopic surgical skills certification.

(c) The standard setting method should include Angoff's method based on achievement of minimal competency in both underlying constructs, i.e. technical skills and patient safety skills.

(d) To avoid a halo effect, a recording feature should be added to LAP Mentor, which will also help in giving feedback at a later date in the event that a simulation supervisor is not available.

Conclusions

Assessment on LAP Mentor using different rating scales and construct-based standard setting methods provides meaningful scores. Periodic summative assessment is acceptable to residents. Use of an external camera increases the feasibility and acceptability. It also provides an opportunity for both trainees and trainers to review performances. Moreover, use of recording improves reliability because any halo effect is removed. However, tasks should be chosen appropriately and should reflect content validity. Pass/fail decisions should be criterion based and the criterion should be decided by experts in the field taking into account the two underlying constructs. The test report form (Appendix 3) should contain scores of multiple assessment tools for more constructive feedback.

Conflict of interest

None declared.

Acknowledgements

We would like to thank Mr. Isehaq Marzouq Salim Al-Raqmi for his backing in providing technical support for this study on LAP Mentor. This research has been conducted as a requirement for Masters in Health Profession Education from Dow University of Health Sciences, Karachi, Pakistan. The study has not been funded by any company/ institute.

References

1. Von Websky MW, Vitz M, Raptis DA, Rosenthal R, Clavien PA, Hahnloser D. Basic laparoscopic training using the Simbionix LAP Mentor: setting the standards in the novice group. J Surg Educ 2012; 69: 459–467. https://doi.org/10.1016/j.jsurg.2011.12.006.

2. Downing S. Validity: on the meaning ful interpretation of assessment data. Med Educ 2003; 37: 830–837. https://doi.org/10.1046/j.1365-2923.2003.01594.x.
3. Feldman LS, Sherman V, Fried GM. Using simulators to assess laparoscopic competence: ready for widespread use? Surgery 2004; 135: 28–42. https://doi.org/10.1016/S0039.
4. Ahmed K, Miskovic D, Darzi A, Athanasiou THG. Observational tools for assessment of procedural skills: a systematic review. Am J Surg 2011; 202: 469–480. https://doi.org/10.1016/j.amjsurg.2010.10.020.
5. Fried GM. FLS assessment of competency using simulated laparoscopic tasks. J Gastrointest Surg 2008; 12: 210–212. https://doi.org/10.1007/s11605-007-0355-0.
6. Kim TH, Ha JM, Cho JW, You YC, Sung GT. Assessment of the laparoscopic training validity of a virtual reality simulator (LAP Mentor TM). Korean J Urol 2009; 50: 989–995. https://doi.org/10.4111/kju.2009.50.10.989.
7. Hove PD, Tuijthof GJ, Verdaasdonk EG, Stassen LP DJ. Objective assessment of technical surgical skills. Br J Surg 2010; 97: 972–987. https://doi.org/10.1002/bjs.7115.
8. Fried GM, Feldman LS. Objective assessment of technical performance. World J Surg 2008; 32: 156–160. https://doi.org/10.1007/s00268-007-9143-y.
9. Pugh D, Touchie C, Wood TJ, Humphrey-Murto S. Progress testing: is there a role for the OSCE? Med Educ 2014; 48: 623–631. https://doi.org/10.1111/medu.12423.
10. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. 1st ed. Washington DC: AERA Publications; 1999.
11. Downing SM, Yudkowsky R. Assessment in health professions education. New York: Routledge; 2007.
12. Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? J Gen Intern Med 2005; 20: 1159–1164. https://doi.org/10.1111/j.1525-1497.2005.0258.x.
13. Arikatla VS, Ahn W, Sankaranarayanan G, De S. Towards virtual FLS: development of a peg transfer simulator. Int J Med Robot Comput Assist Surg 2014; 10: 344–355. https://doi.org/10.1002/rcs.1534.
14. Steigerwald S. Do fundamentals of laparoscopic surgery (FLS) and LapVR evaluation metrics predict intra-operative performance? 2014. http://hdl.handle.net/1993/23199.
15. McCluney AL, Vassiliou MC, Kaneva PA, Cao J, Stanbridge DD, Feldman LS, et al. FLS simulator performance predicts intraoperative laparoscopic skill. Surg Endosc Other Interv Tech 2007; 21: 1991–1995. https://doi.org/10.1007/s00464-007-9451-1.
16. Andreatta PB, Woodrum DT, Gauger PG, Minter RM. LapMentor metrics possess limited construct validity. Simul Healthc 2008; 3: 16–25. https://doi.org/10.1097/SIH.0b013e31816366b9.
17. Gallagher AG, Satava RM. Virtual reality as a metric for the assessment of laparoscopic psychomotor skills: learning curves and reliability measures. Surg Endosc Other Interv Tech 2002; 16: 1746–1752. https://doi.org/10.1007/s00464-001-8215-6.
18. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Papasavas P, Dosis A, et al. An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. Ann Surg 2007; 245: 992–999. https://doi.org/10.1097/01.sla.0000262780.17950.e5.
19. Paisley MAM, Baldwin PJ, Paterson-Brown S. Validity of surgical simulation for the assessment of operative skill. Br J Surg 2001; 88: 1525–1532. https://doi.org/10.1046/j.0007-1323.2001.01880.x.
20. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ 2011; 2: 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd.
21. Vassiliou MC, Ghitulescu GA, Feldman LS, Stanbridge D, Leffondré K, Sigman HH, et al. The MISTELS program to measure technical skill in laparoscopic surgery: evidence for reliability. Surg Endosc Other Interv Tech 2006; 20: 744–747. https://doi.org/10.1007/s00464-005-3008-y.
22. Vallevand A, Violato C. A predictive and construct validity study of a high-stakes objective clinical examination for assessing the clinical competence of international medical graduates. Teach Learn Med 2012; 24: 168–176. https://doi.org/10.1080/10401334.2012.664988.
23. McDougall EM, Corica FA, Boker JR, Sala LG, Stoliar G, Borin JF, et al. Construct validity testing of a laparoscopic surgical simulator. J Am Coll Surg 2006; 202: 779–787. https://doi.org/10.1016/j.jamcollsurg.2006.01.004.
24. Sánchez-Peralta LF, Sánchez-Margallo FM, Moyano-Cuevas JL, Pagador JB, Enciso-Sanz S, Sánchez-González P, et al. Construct and face validity of SINERGIA laparoscopic virtual reality simulator. Int J Comput Assist Radiol Surg 2010; 5: 307–315. https://doi.org/10.1007/s11548-010-0425-8.
25. Giannotti D, Patrizi G, Casella G, Di Rocco G, Marchetti M, Frezzotti F, et al. Can virtual reality simulators be a certification tool for bariatric surgeons? Surg Endosc Other Interv Tech 2014; 28: 242–248. https://doi.org/10.1007/s00464-013-3179-x.
26. Fraser SA, Klassen DR, Feldman LS, Ghitulescu GA, Stanbridge D, Fried GM. Evaluating laparoscopic skills, setting the pass/fail score for the MISTELS system. Surg Endosc Other Interv Tech 2003; 17: 964–967. https://doi.org/10.1007/s00464-002-8828-4.
27. Stunt J, Wulms P, Kerkhoffs G, Dankelman J, van Dijk C, Tuijthof G. How valid are commercially available medical simulators? Adv Med Educ Pract 2014; 5: 385–395. https://doi.org/10.2147/AMEP.S63435.
28. Rosenthal R, von Websky MW, Hoffmann H, Vitz M, Hahnloser D, Bucher DC, et al. How to report multiple outcome metrics in virtual reality simulation. Eur Surg 2015; 47: 202–205. https://doi.org/10.1007/s10353-015-0327-7.

Appendix 1: Test for assessment of laparoscopic surgical skills

Candidate instructions

(1) *Follow the instructions of the LAP Mentor to complete the tasks given below.*

(2) *Your performance will be recorded by LAP Mentor as a score on different metrics and also by an external camera for later review by external raters.*

(3) *During video recording your identity will not be visible.*

(4) *There is no time restriction to complete the tasks; however, it will contribute to overall performance.*

Tasks

(1) **Clipping and grasping:** Grasp a leaking duct, stretch it until the red segment turns green and then place a clip on the green segment

(2) **Use of electrocautery:** Use electrocautery to divide the structures visible on the screen

(3) **Peg transfer:** Lift peg with the dominant hand, transfer to other hand and then spot on the pegboard. Repeat the task with the non-dominant hand.

(4) **Pattern cutting:** Retract the form and cut the fibers in a circle

(5) **Placement of ligating loop:** Place a ligating loop on the base of an appendix-like structure and then divide the loop.

Appendix 2: Data collection instrument

Task 1: clipping and grasping

Marking schedule

ID number:

Construct: This station tests the student's ability to apply endoclips

Item	Performed completely	Performed but not fully completed	Not performed
Smooth movement	2	1	0
Uses both hands	1	0.5	0
Jaws visible before clip application	1	0.5	0
Applies appropriate traction on the tissue/no damage to the tissue	2	1	0
No clips are dropped	1	0.5	0
Applies clips accurately well across the vessel	2	1	0
Completes the task in the desired time	1	0.5	0

Total score: (maximum score: 10)

A. Global rating scale at the level of final year post-graduation

(1) Inferior
(2) Poor
(3) Borderline unsatisfactory
(4) Borderline satisfactory
(5) Good
(6) Excellent

B. Year in training scale

In your opinion, this resident is functioning at the level of a:

PGY 1	PGY 2	PGY 3	PGY 4	PGY 5
□	□	□	□	□

Name of Examiner: ____________________________

Signature of Examiner: ____________________________

Task 2: use of electrocautery

Marking schedule

ID number:

Construct: This station tests the student's ability to use electrocautery safely

Item	Performed completely	Performed but not fully completed	Not performed
Smooth movement	2	1	0
Applies cautery after isolating the tissue to be burnt	2	1	0
Applies appropriate traction on the tissue	2	1	0
No collateral tissue damage	2	1	0
Completes the task in desired time	2	1	0

Total score: (maximum score: 10)

A. Global rating scale at the level of final year post-graduation

(1) Inferior
(2) Poor
(3) Borderline unsatisfactory
(4) Borderline satisfactory
(5) Good
(6) Excellent

B. Year in training scale

In your opinion, this resident is functioning at the level of a:

PGY 1	PGY 2	PGY 3	PGY 4	PGY 5
□	□	□	□	□

Name of Examiner: ____________________________

Signature of Examiner: ____________________________

Task 3: peg transfer

Marking schedule

ID number:

Construct: This station tests the student's ability to transfer pegs

Item	Performed completely	Performed but not fully completed	Not performed
Smooth movement	2	1	0
Demonstrates hand–eye coordination	2	1	0
Demonstrates both-hand coordination	2	1	0
Completes task without dropping the objects	2	1	0
Completes task within the time allowed	2	1	0

Total score: (maximum score: 10)

A. Global rating scale at the level of final year post-graduation

(1) Inferior
(2) Poor
(3) Borderline unsatisfactory
(4) Borderline satisfactory
(5) Good
(6) Excellent

B. Year in training scale

In your opinion, this resident is functioning at the level of a:

PGY 1	PGY 2	PGY 3	PGY 4	PGY 5
□	□	□	□	□

Name of Examiner: ____________________________

Signature of Examiner: ____________________________

Task 4: pattern cutting

Marking schedule

ID number:

Construct: This station tests the student's ability to cut the tissue appropriately

Item	Performed completely	Performed but not fully completed	Not performed
Smooth movement	2	1	0
Demonstrates hand–eye coordination	1	0.5	0
Demonstrates both-hand coordination	1	0.5	0
Applies appropriate traction, no tissue damage	2	1	0
Cuts on the line marked	2	1	0
Completes task within the time allowed	2	1	0

Total score: (maximum score: 10)

A. Global rating scale at the level of final year postgraduation

(1) Inferior
(2) Poor
(3) Borderline unsatisfactory
(4) Borderline satisfactory
(5) Good
(6) Excellent

B. Year in training scale

In your opinion, this resident is functioning at the level of a:

PGY 1	PGY 2	PGY 3	PGY 4	PGY 5
□	□	□	□	□

Name of Examiner: ______________________________

Signature of Examiner: ______________________________

Task 5: placement of ligating loop

Marking schedule

ID number:

Construct: This station tests the student's ability to apply a ligating loop

Item	Performed completely	Performed but not fully completed	Not performed
Smooth movement	2	1	0
Demonstrates hand–eye coordination	1	0.5	0
Demonstrates both-hand coordination	1	0.5	0
Applies appropriate traction, no tissue damage	2	1	0
Applies loop on the desired location (base of the appendix)	2	1	0
Completes task within the time allowed	2	1	0

Total score: (maximum score: 10)

A. Global rating scale at the level of final year postgraduation

(1) Inferior
(2) Poor
(3) Borderline unsatisfactory
(4) Borderline satisfactory
(5) Good
(6) Excellent

B. Year in training scale

In your opinion, this resident is functioning at the level of a:

PGY 1	PGY 2	PGY 3	PGY 4	PGY 5
□	□	□	□	□

Name of Examiner: ______________________________

Signature of Examiner: ______________________________

Appendix 3: Test report form

Candidate name: ______________________ **Candidate number:**______________________

Date of test: 09.11.2015 **Date of result:** 20.12.2015

	Tasks					
	Grasping and clipping	**Use of electrocautery**	**Peg transfer**	**Placement of ligating loop**	**Pattern cutting**	**Over all**
Means of overall group ± (SD)	6.35 ± (1.67)	5.19 ± (0.63)	8.97 ± (1.70)	6.82 ± (1.90)	6.84 ± (1.33)	35.05 ± (4.24)
Your score *(out of 10)*	***6.60***	***8.00***	***10.00***	***6.00***	***6.75***	***37.35***
Cut-off score	6.00	5.10	5.94	5.93	5.87	28.84
Pass/fail	***Pass***	***Pass***	***Pass***	***Pass***	***Pass***	***See overall decision****
Examiner rated you on the global rating scale as……	Borderline satisfactory	Borderline satisfactory–good	Good	Borderline satisfactory–good	Borderline satisfactory–good	Borderline satisfactory–good
Examiner rated your training at residency level……..	PGY2–PGY3	PGY3	PGY4	PGY2–PGY3	PGY3	PGY3
Score on construct 1 (technical skills) (out of 29)	24.72 (cut-off 19.86)					
Score on construct 2 (safe laparoscopic surgical practice) (out of 21)	13.84 (cut-off 12.33)					
Overall decision*	***Pass***					

You were assessed on the following parameters

(1) **Construct 1 (technical skills)**
- (a) Smooth movement
- (b) Both-hand coordination
- (c) Hand–eye coordination
- (d) No dropping of object
- (e) Appropriate time to complete

(2) **Construct 2 (safe laparoscopic surgical practice)**
- (a) Safe application
- (b) No collateral damage
- (c) Appropriate tissue traction

***Passing criteria: all of the following**
- Overall score > 28.84
- Score on technical skills >19.86
- Score on safe laparoscopic surgical practice >12.33

Signature and Stamp of Director of training

Appendix 4: Questionnaire given to study participants before taking the test

Identification number:

Date:

(1) Residency year:

(2) Age:

(3) Gender: F/M

(4) Dominant hand: L/R

(5) How much experience do you have with laparoscopic procedures?

(a) Number of laparoscopic procedures operated under supervision:

(b) Number of procedures operated independently:

(6) Have you practiced on laparoscopic box trainers: Y/N

(7) Have you attended any advanced course in laparoscopy: Y/N

(8) Are you certified in any advanced laparoscopy course: Y/N

Appendix 5: A. Survey about the test and the test report form (Appendix 3) (for candidates)

		Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**
1.	Testing on LAP Mentor is feasible					
2.	Such tests should be conducted on a regular basis					
3.	Test report form (Appendix 3) provides meaningful feedback					
4.	Score measured by checklist reflects my competency in laparoscopic skills					
5.	I would like to be rated on the checklist for assessment on LAP Mentor in future					
6.	Global rating scale is fair					
7.	I would like to be rated on global rating scale for assessment on LAP Mentor in future					
8.	Year in training scale is fair					
9.	I would like to be rated on year in training scale for assessment on LAP Mentor in future					
10.	Test report form (Appendix 3) is incomplete as it does not contain performance metrics					

Appendix 5. B. Survey about the test and test report form (Appendix 3) (for raters and program directors)

		Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1.	Testing on LAP Mentor is feasible					
2.	Such tests should be conducted on a regular basis					
3.	Test report form (Appendix 3) provides meaningful feedback					
4.	Score measured by the checklist reflects the capabilities of the residents					
5.	I would like to use the checklist for assessment on LAP Mentor in future					
6.	Global rating scale is fair					
7.	I would like to use global rating scale for assessment on LAP Mentor in future					
8.	Year in training scale is fair					
9.	I would like to use year in training scale for assessment on LAP Mentor in future					
10.	Test report form (Appendix 3) is incomplete as it does not contain performance metrics					

Appendix 6: Proposed tool for LAP Mentor-based basic laparoscopic surgical skills certification (SQUH tool)

Task:
ID number:
Construct:

A. Checklist
Candidate's score: **Maximum score available for task:**

Construct (laparoscopic surgical skills)		**Not applicable x**	**Not demonstrated 0**	**Partially demonstrated 1 mark**	**Fully demonstrated 2 marks**
Component A (technical skills)	Smooth movement				
	Both-hand coordination				
	Hand–eye coordination:				
	No dropping of object				
	Appropriate time to complete				
Component B (safe practice)	Safe application				
	No collateral damage				
	Appropriate tissue traction				

B. Global rating scale at the level of final year post-graduation

(1) Inferior

(2) Poor

(3) Borderline unsatisfactory

(4) Borderline satisfactory

(5) Good

(6) Excellent

C. Year in training scale

In your opinion, this resident is functioning at the level of a:

PGY 1	**PGY 2**	**PGY 3**	**PGY 4**	**PGY 5**
□	□	□	□	□

Name of Examiner: ____________________________

Signature of Examiner: ____________________________