INTRODUCTION
Training and assessment are complementary practices. Assessment of medical technical training is evolving and there are moves towards competency based outcomes for quality assurance. Technological proficiency of trainees is considered to be of paramount importance. Typically skill learning is based on an apprenticeship model, in which the assessment of technical proficiency is the responsibility of the trainers and has produced excellent clinicians and surgeons whose attainment of clinical skills has been based on long-hours of training programs. However, traditionally assessment is largely subjective. In subjective assessment there is no pre-established measure or standard and is thus based solely on the opinion of the evaluator and carries a low reliability. Low reliability means that high stake decisions cannot be based solely on their results. Objective assessment that is more measurable is essential because deficiencies in training and performance are difficult to correct without objective feedback.

With regards to this, methods such as OSCE (objective structured clinical examination) have been introduced across the medical world as objective tools for assessment of clinical examination at all educational levels. In a similar manner, OSATS represents a quantitative and reproducible method that allows structured observation of technical skills marked against a predetermined checklist. Since its introduction in 1997, it has been used both on bench work models as well as simulation. It has, however, been seen that OSATS has performed less well in simulation and bench settings as compared to a real situation (Cronbach's alpha 0.33 and 0.61 respectively).

Workplace-based assessment (WPBA) is the assessment of a trainee's professional skills and attitude and provides evidence of appropriate everyday clinical competence. It has the advantage of high content validity as the assessment is based on actual performance in the workplace. WPBA assessment is set to become an essential and significant element of specialist medical training over the next few years in Pakistan.

ABSTRACT
Objective: To estimate the reliability of an objectively structured assessment of technical skills (OSATS) in patients for Endotracheal Intubation (ETI), using task-specific checklists during administration of general anaesthesia.
Study Design: Co-relational reliability study.
Place and Duration of Study: Department of Anaesthesia and Medical Education, Pakistan Institute of Medical Sciences, Islamabad, from September to November 2010.
Methodology: Ten first year residents of the department of anaesthesia who had successfully completed a one-day course of ETI on mannequins and had performed 25 ETI in actual patients under supervision, were asked to perform ETI on anaesthetized patients in the operation theater after consent. The procedure was directly observed by two expert observers (ER1 and ER2) and one inexperienced observer (novice, NR) using a validated task oriented checklist. Each trainee repeated the procedure on another patient and was reassessed using the same checklist. Inter-rater for test and retest, and expert to novice reliabilities along with internal consistency were calculated.
Results: The inter-rater reliability was 0.92 in test and was 0.98 in retest (Ebel's calculation). Spearman Rank Correlation Coefficient (SRCC) for ER 1 and NR was 0.76 in test and 0.89 in retest. Cronbach's alpha for internal consistency was 0.93 for first test and 0.97 for retest.
Conclusion: The OSATS for ETI in real patients demonstrated excellent inter-rater test reliability, inter-rater retest reliability, expert-novice reliability and internal consistency. This instrument, therefore, proves promising as a tool for competency-based evaluations.

Key words: OSATS. Inter-rater reliability. Task specific checklists. Internal consistency.
The various assessment tools used at the workplace are mini-clinical evaluation exercises, direct observation of procedural skills, objective structured assessment of technical skills, mini-PAT (mini-peer assessment tool) and case based discussion.

All these workplace-based assessment tools including OSATS have shown acceptable reliabilities when tested in simulation or on bench model.

The current best evidence suggests use of task specific validated checklists for assessment of technical skills in medical specialties. OSATS is widely used to assess the technical skills but no previous study on OSATS has tested this assessment tool on real patients.

This study was conducted to estimate the reliability of results of OSATS using task specific checklists for Endotracheal Intubation performed by first year anaesthesia residents in real patients.

**METHODOLOGY**

The study was conducted in the Department of Anaesthesia, Pakistan Institute of Medical Sciences, Islamabad, from September to December 2010. A sample size of 10 trainees was selected with an overall 60 observations. The participants were first year resident trainees of the department of anaesthesia who had attended a one day course on ETI in mannequins and had performed ETI in 25 real patients under supervision soon after attending the workshop.

An informed consent was obtained from the participants and patients. The hospital ethics committee (HEC-PIMS) approved the study after reviewing the synopsis and interviewing the principal investigator.

The participants were observed as they performed intubation on real patients in the operation theater. Three raters, two experts including a senior anaesthetist (ER1) and a senior surgeon (ER2) and a novice rater (NR, surgical resident from a different department who never used checklist for assessment) acted as observers. A checklist was developed after consulting anaesthetists, general surgeons, emergency room specialists and instructors running Advanced Trauma Life Support and Advanced Cardiac Life Support courses. The checklist was pilot tested and validated before actual implementation.

Each participant was required to give general anaesthesia to a patient on routine elective list. All difficult intubations based on LEMON (Look, Evaluate, Mallampati, Obstruction, Neck) criteria score more than 5 were excluded. The process of intubation performed on anaesthetized or paralyzed patients undergoing elective surgery with predicted “easy” intubation patients (LEMON less than 5 and Mallampati class I) was observed and marked on a task specific checklist. There were three major tasks with 17 essential components in the check list. Each participant was observed on two performances on the same day. For inter-rater reliability, 10 participants were rated by the three observers. For retest reliability, participants were chosen on the basis that their performance had reached a plateau after prior practice, and each had achieved a total score of more than 16 when tested. Ten pairs of scores were used to assess retest reliability. All scores were entered in Statistical Package for Social Sciences (SPSS) 17 data sheet for analysis.

For inter-rater reliability for all raters in test and retest; Ebel intraclass correlation was used. This formula approximates an intraclass correlation and is very flexible in that it requires no assumptions about the number of judges rating each person or object rated nor which judges rate each person or object. A value of > 0.8 was considered a very good agreement among the raters.

Inter-rater reliability between expert and novice raters was determined using Spearman's rank correlation coefficient.

Internal consistency was measured through Cronbach's α (alpha). A value of 0.5-0.8 was considered moderately reliable and more than 0.8 as ideal.

**RESULTS**

The results are outlined in Tables I-V. The inter-rater reliability in the first test was 0.92 and was 0.98 in the retest based on Ebel's calculations. Spearman Rank Correlation Coefficient (SRCC) for ER 1 and NR was 0.76 in test and 0.89 in retest. The value showed marked improvement in the retest. This also demonstrated a strong positive correlation between data of expert and novice rater. Cronbach's alpha for internal consistency was 0.93 for all ratings (an ideal reliability coefficient) and 0.97 for retest depicting a significant improvement in the performance of the raters and participants during retest.

**Table I: Inter-rater reliability of three observers rating trainees performing ETI in first test.**

| Number of trainees rated | 10.0 |
| Harmonic mean number of raters per trainee rated | 3.0 |
| Total number of observations | 30.0 |
| Mean rating† | 15.57 |
| Standard Deviation of the ratings† | 1.13 |
| Inter-rater correlation | 0.92 |

**Table II: Inter-rater reliability of three observers rating trainees performing ETI in retest.**

| Number of trainees rated | 10.00 |
| Harmonic mean number of raters per trainee rated | 3.00 |
| Mean rating† | 16.83 |
| Standard deviation of the ratings† | 0.42 |
| Inter-rater correlation | 0.98 |
Table III: Spearman rank correlation coefficient (SRCC) ER1 to NR for test and retest.

<table>
<thead>
<tr>
<th>SRCC: 1 - ((6 x 67.5) / (123-12)) = 0.76</th>
<th>SD 0.93</th>
</tr>
</thead>
<tbody>
<tr>
<td>ER 1 to NR in test</td>
<td></td>
</tr>
<tr>
<td>SRCC: 1 - ((6 x 22.5) / (113-11)) = 0.89</td>
<td>SD 5.03</td>
</tr>
</tbody>
</table>

DISCUSSION

There is continued debate about whether a good performance on a simulated exercise equates with real life clinical skills. In this study for the first time, the technical skill of ETI was observed in real patients in operating rooms, that is in the actual workplace of an anaesthetist.

ETI is one the most important technical skill which every anaesthetist should know and be proficient in. Three important aspects of reliability that is inter-rater for test and retest and expert-novice were assessed in this study. All these reliabilities and internal consistency showed excellent correlation between different observers. These results indicate that OSATS for ETI meets the standards required for high-stake assessments and also validates the strengths of pilot tested checklist for ETI. Bould et al. argue that the best evidence for a valid, reliable, feasible, and comprehensive assessment tool to assess procedural skill in anaesthesia lies in the use of checklists and Global Rating Scales. There is good evidence for the use of a combination of a checklist and GRS in the setting of medical education research and this combination of tools could be considered the ‘gold standard’ in this setting, however, we only used checklist for this study. It is recognized that a reliability more than 0.8 is ideal, especially if the results of the assessment will form part of a ‘high stakes’ examination. The excellent inter-rater reliability in this study compares to a study carried out by Vasillieu et al. on OSATS of laparoscopic surgical skills performed on MISTEL system which showed an excellent inter rater reliability of 0.998 and consistency of 0.89. They used two raters to observe the trainees directly and their video to mark score on a task oriented checklists. Whereas, we used three raters to mark trainees on real patients. The results of these studies exceed the threshold level of 0.8 required for high-stake evaluations. This study showed a good internal consistency for OSATS checklist which was consistent with a previous study in Seattle on 24 residents in a seven station assessment, including laparoscopic and open procedures, where the internal consistency was high for both the checklist and the global rating scale (0.84-0.97). In this study we assessed only one skill, but used more raters. A study by Martin on OSATS for surgical skills showed a mean inter-rater reliability across stations from 0.64 to 0.72. Internal consistency was moderate to high (α: 0.61 – 0.74).

A small study of surgical trainees using ADEPT (Advanced Dundee Endoscopic Psychomotor Testing) showed good correlation with independent clinical assessment (concurrent validity). A recent study by Saleem et al. in which objective assessment was compared between direct observation and video recording. The comparison of two video performances and direct observations have shown high concordance in terms of ICC as well as categorical Kappa. There was a good correlation between the two video assessments (ICC 0.957) as well as among the observed assessment and the two video assessments (ICC 0.973 with 95% CI; 0.959 – 0.982 and 0.957 with 95% CI; 0.934 - 0.971). Use of video will remove the bias due to presence of an observer which affects the performance of the trainees.

A study by Morgan et al. carried out on anaesthesia simulation demonstrated a better inter-rater reliability with checklists compared to global rating scale. The correlation was 0.77 in checklists and 0.66 in GBSs. Kim et al. also reported similar results in crisis resource management skills performed in simulated environment. The majority of work regarding objective assessments of surgical skill used simulation. The benefits of simulation are evident. Most importantly, it provides a reproducible tool for assessment. It enables the assessment to be distanced from the operative setting and reduces intrusion into operating lists, also allowing for simultaneous examination of students, thus reducing time costs. It is also possible to ‘blind’ examiners by using examiners from outside a unit.

The use of real life settings for assessment of surgical skills has obvious problems: potential for variation in patient anatomy, cancellations of lists, unexpected findings or complications and the inability to examine more than one trainee at any session. It is not possible to ‘blind’ the examiner unless videos of procedures are used.
In this study, real patients in operating theater setting were used and found feasible, practical and cost effective. Only one major skill was used in selected patients who were relatively easy from intubation point. We recommend more skills to be assessed in real patients to address above mentioned issues related to real patients in operating rooms.24

The reliability coefficient of non-structured viva voce (< 0.4) examinations is clearly less than that of either essay examinations (around 0.65) or multiple-choice tests (usually > 0.8).25 However, the use of the structured rating procedure elevated the reliability coefficient of the viva voce examination from between 0.3 and 0.4 in the case of the unstructured examination to between 0.7 and 0.8 in the case of the structured examination.

The study may have limitations in terms of the Hawthorne effect on the performances when they are informed beforehand about being observed.

Potential limitations of the study are highlighted by the small sample population used in this instance. For future studies, performing a power calculation before beginning will enable adequate samples to be recruited. There may be an element of selection bias in that all trainees were appropriately selected for the task. The use of a fabricated environment suggests that the same results may not be obtained had the trainees been subjected to other external stresses and pressures (such as time, for example). For such a test to be extrapolated to a real life situation, future studies would need to mimic such circumstances as closely as possible.

CONCLUSION

The current methodology used to teach ETI in the competence based curriculum of anaesthesia and its assessment using task specific checklist have excellent reliability, which exceeds the threshold level of 0.8 required for high-stake evaluations. These findings support introduction of OSATS using checklists in other residency programs running competence based curricula. To further enhance the reliability, global rating scales may be added.

Acknowledgement: Author is thankful to Dr. Amina Ahmad, Director MCPS (HPE) for her guidance and help in this research. Also to Dr. Shebaz Khan and Mr. Intisar Siddiqui (Statistical Officer) for their help as well as PIMS for this cooperation.

REFERENCES


