Objective assessment of technical skills in cardiac surgery

Julian Hancea,*, Rajesh Aggarwalb, Rex Stanbridgec, Christopher Blauthd, Yaron Munza, Ara Darziab, John Pepperb

aDepartment of Surgical Oncology and Technology, Imperial College London, 10th Floor QEJM Building, St Mary’s Hospital, Praed Street, London, UK
bAcademic Department of Cardiothoracic Surgery, Royal Brompton Hospital, Sydney Road, London SW3 6NP, UK
cDepartment of Cardiothoracic Surgery, St Mary’s Hospital, Praed Street, London, UK
dDepartment of Cardiothoracic Surgery, Guy’s and St Thomas Hospital, Lambeth Palace Road, London SE1 7EH, UK

Received 20 January 2005; received in revised form 10 March 2005; accepted 10 March 2005; Available online 9 April 2005

Abstract

Objective: Reduced training time combined with no rigorous assessment for technical skills makes it difficult for trainees to monitor their competence. We have developed an objective bench-top assessment of technical skills at a level commensurate with a junior registrar in cardiac surgery. Methods: Forty cardiothoracic surgeons were recruited for the study, consisting of 12 junior trainees (year 1-3), 15 senior trainees (year 4-6) and 13 consultants. The assessment consisted of four key tasks on standardised bench-top models: aortic root cannulation, vein-graft to aorta anastomosis, vein-graft to Left Anterior Descending (LAD) anastomosis and femoral triangle dissection. An expert surgeon was present at each station to provide passive assistance and rate performance on a validated global rating scale giving rise to a total possible score of 40. Three expert surgeons repeated the ratings retrospectively, using blinded video recordings. Data analysis employed non-parametric tests. Results: Both live and video scores differentiated significantly between performances of all groups of surgeons for all four stations (P<0.01) (median live and video score for LAD; Junior 19,17; Senior 29,22; Consultant 36,28). Correlations between live and blinded rating were high (r=0.67-0.84; P<0.001) as was inter-rater reliability between the three expert video raters (α=0.81). Conclusions: The use of bench-top tasks to differentiate between cardiac surgeons of differing technical abilities has been validated for the first time. Furthermore, it is unnecessary to perform post-hoc video rating to obtain objective data. These measures can provide formative feedback for surgeons-in-training and lead to the development of a competency-based technical skills curriculum.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Assessment; Competency-based; Operative skills; Technical competence

1. Introduction

Results of surgical interventions are increasingly being scrutinised by the general public with a desire by individual patients to know ‘how good is my surgeon?’. It follows that there needs to be greater transparency in the training of surgeons, with trainees reaching set competencies before achieving independent practice as consultants. It is proposed that the present apprenticeship model of training will be replaced with a shorter structured competency-based program, which will ensure trainees meet required standards throughout their training [1]. By definition, a competency-based curriculum requires valid assessments at regular intervals to examine the surgical trainees’ acquisition of the traits necessary to become a surgeon such as knowledge, technical proficiency and professional attitudes.

Present forms of valid assessment concentrate solely upon examination of knowledge, while assessment of technical proficiency is entirely based on unreliable direct observation of live operating [2] with trainees being labelled with adages such as ‘a safe pair of hands’. Although, methods of objectively assessing technical skills have been developed, they largely remain as research tools and have yet to be applied to current surgical practice. Cardiac surgeons have been amongst the first surgeons to have their practice closely scrutinised [3] in the United Kingdom and are therefore eager to establish rigorous assessments throughout training.

This study describes the development, validation and applications of a ‘bench model’ technical skills assessment for cardiac surgery. Validity refers to the concept of whether the assessment measures what it purports to measure and is divided into the following categories. When applied to technical skills assessment, face validity is a reflection of how closely the assessment resembles the ‘real’ task and content validity considers the extent to which the model measures technical skill as apposed to knowledge. Both these forms of validity can be met by expert consensus. Construct validity is the ability of the assessment to discriminate between different skill levels, whereas concurrent validity compares a new assessment with the current gold standard. Predictive validity determines
whether performance on the assessment corresponds with performance in the operating theatre.

Three specific research questions were investigated: (a) Can cardiac specific bench model discriminate between different technical competences of trainees? (construct validity); (b) Is live assessment equivalent to retrospective blinded video assessment?; (c) Is this assessment feasible as a regular part of training?

2. Methods

2.1. Selection of simulated surgical task

The assessment was aimed to test the broad range of surgical skills required by a junior trainee cardiac surgeon, whilst remaining within the constraints of bench-top simulation. A panel of expert cardiac surgeons established four key surgical tasks, which a trainee could be expected to undertake, with the aid of a single assistant. Each was represented by a separate workstation, with the instruments required presented in standardised manner.

1. Station 1—Aortic Root Cannulation (Fig. 1). Cannulation of the ascending aorta is a fundamental skill, which should be acquired early in a cardiac surgeon’s training. Cadaveric porcine aorta was used to simulate human aorta and was perfused with water by being placed in series with a custom-made circulatory rig (Wetlab, UK) to produce pulsatile pressures of 60–80 mmHg. Candidates were asked to perform two purse-string sutures before performing an arteriotomy. After securing a size 22FG aortic cannula, they proceeded to decannulate and repair the subsequent defect in a standard manner.

2. Station 2—Femoral Triangle Dissection. Cardiac surgeons are expected to gain access to the femoral artery in emergency situations such as locating alternative access for cardio-pulmonary bypass. Utilising a latex model simulating the groin region, candidates were asked to dissect the vascular structures adjacent to the saphenofemoral junction and ligate branches of the saphenous vein to facilitate access to the femoral artery.

3. Station 3—Aorta to Vein-Graft Anastomosis. Cadaveric porcine ureter was used to simulate vein conduit and the subjects were asked to perform an end-to-side anastomosis to a length of cadaveric porcine aorta using 6-0 polypropylene suture.

4. Station 4—Vein-Graft to Left Anterior Descending Coronary Artery (LAD). Candidates were asked to perform an end-to-side anastomosis between the simulated vein conduit and the LAD of a porcine cadaveric heart using 7-0 polypropylene suture.

2.2. Method of assessment

A modified version of the Objective Structured Assessment of Technical Skills (OSATS) was used to mark candidates’ technical performances [4,5]. This scoring method is constructed of eight categories marked on a Likert scale from 1 to 5, tied with descriptive anchors, giving a possible total of 40 marks (Table 1). Scoring was performed ‘live’ by a single expert surgeon, who was also providing ‘passive’ surgical assistance on request. Subsequently, all performances were retrospectively rated using video recordings by three expert surgeons. This was done in a blinded manner with candidates being labelled by an alphabetic code.

2.3. Validation of assessment

Cardiac surgeons of differing seniorities were recruited on a voluntary basis to perform the assessment tasks in order to ascertain construct validity. The surgeons were sub-divided into three groups namely junior specialist registrar (years 1–3), senior specialist registrar (years 4–6) and consultant surgeons.

2.4. Application of assessment

Once initial validation of this multi-station assessment was complete, the next step was to demonstrate the feasibility of applying this examination to the current surgical curriculum. Currently, cardiothoracic trainees in the United Kingdom attend an annual review or a record of in-training assessment (RITA) day at which it was chosen to pilot this assessment. Importantly we wanted to calculate how many candidates could be assessed in a set period of time and the minimum number of faculty required.

2.5. Statistical analysis

Comparison between the different grades of surgeons was performed using the non-parametric Kruskal–Wallis test. The correlations between the live and retrospective scorings were explored by means of Spearman rank correlation. Two types of reliability were explored: internal consistency (inter-station reliability) and variation between expert observers (inter-rater reliability), which were both explored using Cronbach’s alpha.
3. Results

3.1. Validation

Forty cardiac surgeons consisting of 12 junior specialist registrars, 15 senior specialist registrars and 13 consultant surgeons were recruited to perform a total of 160 tasks. Data from the assessments were available from 39 root cannula-tions, 38 femoral triangle dissections, 38 aortic to graft anastomosis and 38 LAD anastomosis. Seven data points were lost due to technical and scheduling difficulties.

Table 2 shows the results for the three groups of surgeons across all four stations including both live and video assessments and indicates significant differences in performance between the three groups of surgeons for all four surgical tasks. This difference is evident whether performances were assessed live by an expert or retrospectively via video recordings as illustrated for LAD anastomosis workstation in Fig. 2. Correlations between the live and video assessments were significant with a correlation coefficient between 0.672 and 0.817 (Fig. 3). A Bland–Altman plot [6] comparing the two methods of rating shows a higher scoring for live marking but this was consistent across the different surgical grades (Fig. 4). Inter-rater reliability between the three expert video raters was high with a Cronbach’s $\alpha = 0.81$, as was internal consistency between stations with a Cronbach’s $\alpha = 0.87$.

3.2. Application of assessment

Fifteen surgeons were assessed in a 6-h assessment session by four members of the faculty. All candidates completed all four tasks within 90 min.

4. Discussion

Cardiac surgery is technically challenging but currently, in a similar manner to other surgical specialties, trainees progress to become independent practitioners without objectively demonstrating competency in basic surgical techniques. Present forms of assessment consist of a combination of log-book assessment and mentor observation. Log-book records simply reflect operative experience and reveal little about a trainee’s proficiency [7]. Furthermore, counting case-loads does not take account of differences in the speed of acquisition of operative skills between individual surgeons. Assessment by mentors of their registrar’s performance is invaluable and will remain a cornerstone of apprentice-based teaching. However, this assessment remains subjective, unreliable and not suitable for summative assessment [8].

Although assessment of technical skills using live cases would provide the ultimate realistic environment for assessment, it is associated with a number of drawbacks. Patient safety remains the over-riding priority in
the operating theatre, which results in appropriate interference from the supervising surgeon at the first sign of the trainee departing from the correct course of action. In addition the operating theatre is a complex team environment in which it is near impossible to isolate the single competency domain of technical skills. Blinding performances within the operating theatre is also not possible. Patient disease variability leads this form of assessment to being inherently unreliable. Finally, the logistics of recruiting external assessors to visit a trainee local hospital would add considerably to the time commitments of an already busy trainee faculty.

Simulation provides a means of delivering a reproducible assessment, which is essential if this process is to be reliable and valid. Bench model assessments have been successfully applied by other surgical specialties such as vascular surgery [9]. Technical performance assessment on a synthetic bench top model has been shown to be as effective as assessment using live animals [10]. In addition performance on bench models has been shown to correlate significantly with performance in real-life operating in other surgical specialties [11]. Despite many developments in the field of surgical assessment, such as virtual reality [12] and motion analysis [13], expert rating with criterion remains the gold standard. The OSATS technique has been extensively validated [14–18]. Furthermore the use of an operation specific checklist in additional to global rating has not been shown to add to the reliability and validity of the assessment process [19]. Cadaveric porcine tissue was chosen for the majority of the assessment stations as it offers high fidelity simulation at a low cost.

The first aim of this work was to demonstrate validity of a cardiac-specific technical skills assessment. Content validity has been largely confirmed by expert agreement when establishing the domain of the examination. Construct validity has been demonstrated for all four assessment workstations with significant discrimination between the three groups of surgeons. The best discriminator was the coronary anastomosis station, which is the most complex task completed by each candidate. In a similar manner to previous research in this field the use of levels of experience as surrogate markers of expertise is a recognised limitation.

Rating performed live by the examiner may offer advantages over retrospective video analysis. The examiner may be able to measure the trainee’s use of assistance with more accuracy and is guaranteed to have an excellent view of the operative field throughout the procedure. The main drawback of live rating is its lack of ability to blind the assessment, which retrospective video recording can deliver by providing only views of the operators’ hands. Results showed that the ability of the assessment to discriminate was not dependent on whether performances were rated live or retrospectively. Furthermore there was a strong correlation between live and blinded scores for each of the four tasks. However, live assessment led to raters awarding consistently higher marks but this was consistent across the range of marks as shown by the Bland-Altman plot.

### Table 2

<table>
<thead>
<tr>
<th></th>
<th>Junior SPR</th>
<th>Senior SPR</th>
<th>Consultant</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aortic cannulation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Live</td>
<td>23.5 (21–26)</td>
<td>34 (26–38)</td>
<td>38 (35–39.75)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Blinded</td>
<td>22.5 (18.25–30)</td>
<td>21 (17–24)</td>
<td>37.5 (29.75–39)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Femoral triangle dissection</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Live</td>
<td>19.5 (16.5–29.5)</td>
<td>27 (18.5–36.25)</td>
<td>38.5 (37.75–40)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Blinded</td>
<td>17 (14–24)</td>
<td>17 (14–22)</td>
<td>28 (19–40)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Aorta-graft anastomosis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Live</td>
<td>19.5 (16.25–23.25)</td>
<td>26 (23–35.25)</td>
<td>37 (32.5–40)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Blinded</td>
<td>15 (10–22)</td>
<td>26 (20–32)</td>
<td>28 (23–35)</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>LAD anastomosis</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Live</td>
<td>19 (15.25–21)</td>
<td>29 (25–35)</td>
<td>36 (33–38)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Blinded</td>
<td>15.5 (13.25–19.5)</td>
<td>19 (16–24)</td>
<td>24 (21–34)</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

*Fig. 2. Box plot illustrating construct validity of the LAD anastomosis station.*
Demonstrating the validity of live assessment is essential if this ‘examination’ is to be applied on an annual basis as it can provide immediate marks and avoid the need for retrospective video analysis which is very labour intensive.

In conjunction with validity it is also necessary to examine the reliability of any form of assessment. Excellent inter-rater reliability was demonstrated achieving a Cronbach’s alpha of greater than 0.8, which is the accepted cut-off for a ‘high stakes’ examination [20]. Internal consistency or inter-station reliability was also demonstrated as high in this skills examination. With a relatively small number of stations possible, repeated attempts at the examination may lead to improved scores or ‘learning bias’. However, there is no reason to suggest why this improvement in simulated performance is not mirrored by an improvement in operating on live patients, which is desirable.

The second part of this study demonstrated the feasibility of running this examination as part of the annual review process of trainees and demonstrated a framework for how it could be applied in the future. Although, this pilot only required four expert observers, one per station, it is envisaged that the final version of this exam would have two examiners per station to make it more robust. This process is labour intensive for faculty but the authors feel it is more efficient that external examiners visiting trainees’ individual centres to perform assessment of live operating. In order to run the assessment it would cost approximately £70 per candidate. At this stage it is not possible to set ‘pass marks’ though once enough data are collected this would be relatively straight-forward. Eventually given enough longitudinal data and surgical outcome results it may be possible to demonstrate predictive validity of this examination.

This study has demonstrated the validity and feasibility of using bench-top simulations to assess technical skills specifically applicable to cardiac surgery. At present the complexity of the assessment is limited by the models available but as simulation improves, it is envisaged that this form of assessment could be applied to more complex procedures. These assessment tools could readily be applied in a summative method to a curriculum facilitating progression from one ‘year’ to the next based on competency rather than simply seniority therefore allowing good candidates to progress faster, whilst slower trainees could...
receive remedial measures. In addition, these measures provide a means of providing formative feedback to trainees to enable them to assess their own progress and rectify weaknesses.

Acknowledgements

The authors are very grateful to Jonathon Hyde, Brian Glenville, Alec Shipolini, Uday Trivedi and Jon Anderson for performing the role of examiners in this study. In addition, we would like to thank the cardiac surgeons who volunteered to participate as subjects. Finally we would like to recognise the contribution of London Postgraduate Deanery, St Jude Medical and Wetlab for their support in this project.

References