Validity of an Exam Assessment in Surgical Skill: EBSQ-VASC Pilot Study

V. A. Pandey, J. H. N. Wolfe,* A. K. Lindahl, J. A. Rauwerda, D. Bergqvist,
On Behalf of the European Board of Vascular Surgery

Regional Vascular Unit, 2nd Floor Mary Stanford Wing, St Mary’s Hospital, Praed Street, Paddington, London W2 1NY, UK

Presented at the XVII Annual Meeting of the European Society for Vascular Surgery (ESVS), Dublin (September 2003) and awarded the prize of the best clinical article presentation

Key Words: Vascular; Surgery; Training; Assessment; Validity; Examination; EBSQ-VASC; Dissection; Anastomosis; Knot-typing; Video; OSATS; ICSAD; ICEPS.

Introduction

Background

Surgical examinations vary from country to country but all are in place to safeguard patients and ensure a high standard of surgeon. Surgeons are assessed at all stages of their career. The medical student is expected to know the essential facts required of any medical practitioner. The basic surgical trainee is expected to have the knowledge to move onto higher surgical training, the higher surgical trainee must demonstrate the knowledge and judgement and experience to progress to independent practice.

The different systems of surgical practice between European countries were not of great importance in the past, but the implementation of EC Directive 75/362 in 1975 allows doctors to migrate between the countries of the European Union.

The Union Europeene des Medicin Specialistes (UEMS) Division and Board of Vascular Surgery sought to harmonize vascular surgical training requirements by the introduction of the European Board of Surgery Qualification in Vascular Surgery (EBSQ-VASC). The examination was initiated in 1996 and to apply to sit the exam, the candidate must be in possession of an European CCST (Certificate of Completion of Specialist Training). It consists of two components: part one involves accreditation of Log Book data. Performing a prerequisite number of index vascular surgical procedures allows passage to part two, the viva voce examination. This tests the candidates’ clinical ability including knowledge of vascular disease and vascular surgery, a critical appraisal of a scientific paper in vascular surgery and an overall assessment including the candidates’ involvement in clinical audit and morbidity and mortality meetings.

Some would argue that the current methods of surgical assessment are flawed by a failure to assess one of the critical aspects of surgery—technical competence and there is concern that the limitations imposed by the European Working Time Directive will lead to surgeons finishing training with less experience. This combined with a growing demand by the general public for ‘quality assurance’ in surgery has prompted surgical educators worldwide to look for a method of assessing technical skill.

An increasing number of institutions are evaluating technical competence during the in-training assessment of surgeons. Such assessments have not found their way into current postgraduate surgical examinations, however, this was not always the case. Until 1946 the Royal College of Surgeons of England assessed technical competence by asking candidates to perform an operative procedure on a cadaver. Due to shortage of cadavers, this component of the examinations was dropped and reliance was placed on the trainers assessment of the trainee.

Questions regarding the validity of technical
Competency evaluation preclude their use in postgraduate surgical examinations. The purpose of this pilot study was to assess the validity of a technical skills exercise for incorporation into the EBSQ-VASC exam.

Methods

We tested the following aspects of validity:

Content validity—Does the test assess the content domain it purports to? In the case of written exams, does it test the appropriate breadth of knowledge, in an assessment of technical skill; does it test the critical aspects of technical competence.

Construct validity—Is the test a valid discriminator between surgeons of varying skill? With respect to a technical skills assessment the differences between an experienced surgeon and a novice surgeon should be measurably different.

Internal consistency—This reflects the consistency of candidates’ performance in all aspects of the examination. In an examination with high internal consistency candidates perform uniformly in all areas of the examination.

Inter-observer reliability—The level of agreement between assessors. Do all the assessors agree on the performance on any one individual? The level of inter-observer correlation acceptable for a ‘high-stakes’ examination is greater than 0.8.23,24

Content validity

The examination had three stations testing three skills essential to the vascular surgeon.

Dissection technique

This was tested using a previously validated model of the saphenofemoral junction (Lims & Things, UK). This model has surface landmarks imprinted on the mould that hold it in place. The model itself contains a skin layer, superficial fatty layer, deep fascial layer and the long saphenous vein with its four groin tributaries draining into the common femoral vein. The vessels were fluid filled with simulated blood. Each candidate was asked to perform a saphenofemoral junction ligation on this model, as they would do in the operating theatre. Standard vascular surgical instruments, sutures and ties were available for each candidate. There was a time limit of half an hour for this procedure. Each examiner or candidate was offered passive assistance by an EBSQ-VASC examiner.

Anastomotic technique

The technique of distal anastomosis was assessed using a leg model simulating an anastomosis to the anterior tibial artery (Stiftung Vascular International, Switzerland). The models are used at the European Vascular Course in Pontresina, Switzerland as well as workshops held during the annual scientific meeting of the European Society for Vascular Surgery. A simulated wound insert is placed into the appropriate anatomical position for the desired anastomosis. The walls of the insert are retractable and the simulated artery is exposed. The candidate has to demonstrate control of the vessel using slings and clamps, create an arteriotomy within the vessel and anastomose a PTFE graft to this simulated artery. The choice of anastomotic technique was left to each individual candidate. The rating scales used to assess anastomotic technique takes account of the variation in anastomotic methods. Once again, an examiner provided passive assistance. The time limit for this station was 25 min.

Knot tying

This was assessed using a commercially available knot-tying assessor (Lims & Things, UK). The task involved tying a standard reef knot with three throws over a hook. The hook lies in the bottom of a beaker to simulate difficult access and is attached by means of a magnet. Excessive tension in the suture disrupts the magnet. This is equivalent to disrupting a blood vessel. Performance at this station was objectively assessed using the Imperial College Surgical Assessment Device (ICSAD). ICSAD consists of commercially available hardware (Isotrak II, Polhemus, USA) and patented software. Motion trackers worn on the dorsum of both hands over the third metacarpal give exact x, y and z coordinates in relation to a fixed coordinate. This allows exact quantification by four measurements.

Firstly, number of hand movements. Tremor artefact is removed by means of software filters. Secondly the path length—this is the total distance moved by the operators’ hands during the procedure. Thirdly procedural time and fourth is the average speed of each hand.

Datta et al. demonstrated the construct validity of the ICSAD model in laparoscopic and open surgery25,26 and also the correlation between motion analysis and objective assessments27 using global rating scales of operative performance. Motion analysis also correlates with procedural outcome in a bench model simulation.28,29

In addition to motion analysis data, we observed each participant for magnetic disruption in the knot tying rig as well as testing each knot for slippage.
Construct validity

We used two distinct subgroups to establish the construct validity for the examination.

Eight experienced surgeons (examiners from the EBSQ-VASC board of assessors) participated in the exercise. The majority of candidates volunteered to participate in this assessment and eight were selected at random. Both groups were assessed at each of the stations.

Internal consistency

We assessed internal consistency by correlating the performance of each participant between the two operative procedures, as well as correlating operative scores with knot tying skill.

Inter-observer reliability

The level of agreement between examiners at each of the operative stations was assessed for each participant.

Two separate rating scales were used to assess the procedures:

(1) A global rating scale of operative performance as used in the objective structured assessment of technical skill (OSATS) examination. Developed by the University of Toronto, it is a five-point rating scale testing eight generic aspects of technical skill. It has been validated in numerous studies. They also demonstrated the transfer of skill from bench to the operating theatre and have established that this form of examination is portable. More recent work using the rating scale has demonstrated a significant correlation between performance on bench models and in the operating theatre.

(2) Imperial college evaluation of procedure-specific skill (ICEPS). This tests the procedural content of a particular operation and was developed following studies that demonstrated poor inter-observer reliability and construct validity of checklists of procedural steps. Global marking (using the OSATS rating scale) in conjunction with ICEPS scoring has demonstrated a high inter-observer reliability as well as construct validity. The scores from OSATS and ICEPS were therefore added to derive an overall score for the operative procedure.

Live marking

Examiners marked each other so that only one mark was available for each examiner. Candidates were marked by two examiners; one observing the candidate and the other assisting passively. The examiners were asked to mark independently and not discuss the performance of the candidate.

Video tape marking

The blinded video tape footage of all the participants (examiners and candidates) performing both procedures was sent to three separate members of the EBSQ-VASC board of assessors (D.B., J.A.R., A.K.L.). The three video-assessors were not involved with the running of the examination.

Each examiner and candidate therefore had a live and video-tape mark for their generic (OSATS) and procedure-specific (ICEPS) performance. For the purpose of a ‘high-stakes’ postgraduate medical examination the inter-observer reliability should be greater than 0.8.

Statistical analysis

Non-parametric tests were used in the statistical analysis. Inter-observer reliability was assessed using Cronbach’s alpha reliability coefficient (α). This is the standard statistical test used when assessing the reliability of a rating scale. The Mann–Whitney U test evaluated the differences between the two groups performing the three tasks. Bland and Altman plots were used to show agreement between live and video assessment scores. Spearman’s Rank correlation coefficient (rs) was used to assess the correlation between a surgeon’s dissection versus anastomotic skills. In addition Spearman’s test was used to correlate the results of operative skills with knot-tying skills.

SPSS v10.0 for Windows was used in the statistical analysis and P < 0.05 was considered statistically significant.

Results

Data from the live assessments were analysed shortly after the examination. The blinded footage was distributed to D.B., J.A.R. and A.K.L. who did not communicate their results with each other. The order of procedures on the videotape footage was randomized and the identity of the surgeon was concealed. ICSAD data were analysed using patented software.
Construct validity

Dissection
Adding generic and procedure specific marks derived a total operative score. There was a significant difference between the examiners and candidates performing this task when assessed live ($P = 0.001$, Fig. 1) however, such differences were not evident on blinded video assessment (Fig. 2). Bland and Altman plots were used to assess the agreement (Fig. 3). There was a mean difference of $-9.5$ for video versus live marking (range $-31.9$–$12.9$) suggesting lower marking for video assessment. All participants’ marks were within the limits of agreement.

Anastomosis
Significant differences were seen in the live marking and videotape assessment of both examiners and candidates ($P = 0.003$ and $P = 0.021$, Figs. 4 and 5, respectively). Bland and Altman plots (Fig. 6) showed all participants marks were within the limits of agreement, the mean difference was $-7.2$ for video versus live assessment (range $-30.8$–$15.2$), again lower marking was seen with videotape assessment.

Knot tying
Dominant hand motion analysis showed evidence of construct validity by comparing examiners and candidates. Non-dominant hand motion analysis did not discriminate between the two groups. Of the parameters measured, the number of dominant hand movements required to tie the knot discriminated significantly between the two groups ($P = 0.027$—Fig. 7), procedural time to tie the knot was not statistically significant ($P = 0.065$) and the distance travelled and average speed of the surgeons hands did not discriminate between the two groups.

We derived a composite score taking into account the important factors of disruption of the magnets and knot slippage ($100 - [\text{dominant hand moves}] - [25 \text{ for magnet disruption}] - [25 \text{ for knot slippage}]$). This composite score also discriminated between the two groups ($P = 0.03$—Fig. 8).

Internal consistency

The overall live marks of each participant performing a saphenofemoral junction ligation were correlated with similar marks for a distal anastomosis. There was
a highly significant correlation for the participants’ marks \((r = 0.891, \ P < 0.001—\text{Fig. 9})\) demonstrating evidence for internal consistency of the examination for live scoring. The correlation was not as good for videotape scoring but still significant \((r = 0.66, \ P < 0.001)\).

We assessed the relationship between assessment of operative performance and the results of electromagnetic motion analysis using ICSAD. There was a significant correlation between knot tying and live assessment scores \((r = -0.77, \ P < 0.01—\text{Fig. 10})\), there was no such correlation with videotape assessment \((r = -0.13, \ P = \text{NS})\).

Live marking offered a high inter-observer correlation for both saphenofemoral junction ligation and anastomotic technique \((\alpha = 0.83\ \text{and} \ 0.80, \ \text{respectively})\). The correlation was much lower for videotape assessment \((0.74\ \text{and} \ 0.38, \ \text{respectively})\).

**Discussion**

The examination of technical skill is not a recent development but developing a robust and objectively valid method is a challenge. It once played an integral part of the examination for fellowship to the Royal College of Surgeons of England.\(^{21}\) Examiners assessed surgeons performing operations on cadavers but unfortunately the Royal Colleges discontinued this assessment process (due to an ongoing shortage of cadavers) in the 1940’s.\(^{22}\) Current assessment of operative skill relies on information from the trainers as well as extrapolated data from operative log-books and viva voce examinations in operative surgery. Some consider that these measures of assessment are invalid when considering an exam candidates technical skill.

Mentors work closely with their trainees and are the most reliable source of feedback on the progress of that particular trainee since direct observation is the best method of assessment. Mentors do this in day-to-day practice, however, without structured criteria such assessments may be unreliable and invalid.\(^{35}\) This method of assessment is also seen as subjective\(^{12}\) and may on rare occasion be affected by the mentor-trainee relationship. This method is irreplaceable,
however, as the opinion of the trainer remains invaluable.

Mentor assessment is supported by the information derived from operative log-books. Although log-book assessment reflects the relative experience of the surgeon the technical performance of a particular surgeon cannot simply be extrapolated by the number of procedures performed.\(^{36}\) Furthermore this method does not reflect the varying numbers of procedures a surgeon has to perform before gaining competence.\(^{37}\)

Written and viva voce examinations have content validity in that they have the power to assess the entire content domain of vascular surgery. But it is unreason-

able to think that a candidates' factual knowledge has any bearing on their technical skill.

Several institutions have developed generic and procedure specific criteria in order to rate a surgeons' technical competence. The global rating scale used in OSATS assessments in Toronto has been extensively validated.\(^{14,15,19,20,24,30,31}\) This rating scale tests a surgeons' generic technical skill. Procedure-specific content is usually assessed using a checklist. Regehr et al.,\(^{31}\) however, showed a global rating scale demonstrates better inter-observer reliability discrimination between surgeons of differing skill (construct validity). The Imperial College group demonstrated a high inter-observer reliability and construct validity of a procedure-specific rating scale.\(^{32}\) They now employ

---

**Fig. 7.** Box and whisker plot showing significant differences in dexterity (dominant hand movements) between candidates and examiners. There is a lower score with increasing experience demonstrate experienced surgeons have greater economy of movement ($P = 0.027$).

**Fig. 8.** Box and whisker plot showing composite knot-tying scores for candidates and examiners demonstrating significant differences between the groups. The composite score showed a significant correlation with total operative score ($P < 0.03$).

**Fig. 9.** Scatterplot demonstrating internal consistency for live assessment scores. Total score for saphenofemoral junction ligation on the X-axis and total score for distal anastomosis on Y-axis ($r_s = 0.891, P < 0.001$).

**Fig. 10.** Scatterplot showing the correlation between dexterity and live assessments ($r_s = 0.891, P < 0.01$).
the ICEPS in the formative assessment of surgical trainees.

Due to the ongoing shortage and variability of cadavers, bench model simulations were used in this assessment. This allows for standardisation of the assessment process and removes any ethical or storage concerns. The Toronto group demonstrated that performance on bench models correlates with performance on animal models, more recently Datta and Beard assessed surgical trainees performing a saphenofemoral junction ligation on bench model simulation, as well as in theatre using the validated global rating scale and established a highly significant correlation between the two.

This pilot assessment of surgical skill of post-CCST surgeons demonstrates the aspects of validity necessary for a 'high-stakes' summative examination. Content validity was established using three separate models each attempting to test a separate aspect of the technical skill necessary for a vascular surgeon. The surgical instruments used were those used in clinical practice and the models used had face-validity (degree of realism) that was high according to questionnaire feedback from the trainees.

Construct validity was excellent for live assessment but was less reliable using video assessment. There may be a number of reasons for this:

1. Examiners bias. This is suggested by assessing raw data from the examination. A number of examiners were given the highest marks possible for some of the operative stations. In the experience of the authors this is rarely seen in surgical assessment and corresponds to a perfect procedure with no faults. The Bland and Altman plots also suggest positive bias towards examiners with live assessment, the degree of bias is nevertheless within the statistically acceptable range.

2. The quality of video footage was criticised. This was recorded blinding the identity of the surgeon (only gloves and gowns were visible), however, feedback from the three video assessors agreed that there were points in the video footage that were difficult to assess. Indeed some crucial errors picked up in live assessment were not evident in blinded videotape assessment (i.e. ligation of a long saphenous vein tributary rather than the saphenous vein itself, knots slipping from a tributary etc.). The footage may have been improved by using multiple video cameras but this is not feasible in the context of a surgical examination.

3. The exercises did not discriminate between the two groups. This has been seen in various studies when performing a saphenofemoral junction ligation.

Beyond a certain level of experience most surgeons produce a technically competent performance on the model. The anastomotic exercise, however, showed good discrimination for both live and video assessment.

Internal consistency was good since each candidate scored similar marks for the saphenofemoral junction ligation and the distal anastomosis. Again the level of internal consistency was higher with the live assessment than for the videotape assessment.

In this pilot study, we conclude that live assessment offered greater validity than videotape assessment. Internal consistency and inter-observer correlation were inadequate for the video assessment.

Bench model testing assessed live by two examiners appears to offer greater discriminatory power than video assessment and the problems with the creation, distribution and review of video footage are all avoided. Critics may suggest that this method is inherently biased, however, it is worth noting scores from live assessment also show a significant correlation with objective motion analysis. This correlation was not seen in videotape assessment.

The three bench models used have sufficient overall construct validity and inter-observer reliability for a 'high-stakes' examination when live assessment is used. Discriminating between levels of competence depends on the complexity of the test used. Knot tying is a basic skill and surgeons with their CCST should all be able to perform this with ease. The hand movements used nevertheless discriminated between surgeons at the end of their training and experienced surgeons. Saphenofemoral junction dissection and ligation is learnt at an early stage in training and one would expect the majority of surgeons to perform this with competence. This was the case but again it remains a useful discriminator and it is reasonable to argue that a surgeon having difficulty with this procedure should not practice independently. The most sophisticated skill we tested was the anterior tibial artery anastomosis and this exercise was the best discriminator between examiners and candidates. If a technical skill exam based on these tests were introduced at CCST level, perhaps the marks should be weighted towards the more sophisticated test.

Increasing public demand and the freedom to work throughout the European Community necessitates quality control measures. These measures should ensure all surgeons have the necessary experience, decision-making qualities, and technical competence to work within any country in the European Union. The European Board of Vascular Surgery seeks to
demonstrate a robust method of testing technical skill that may be integral in future surgical examinations.

Acknowledgements

The authors would like to acknowledge the following individuals and organisations: Professor Christos Liapis for his considerable assistance regarding the set-up of the examination. Professor Ara Darzi for his support and loan of the Imperial College Surgical Assessment Device.

We would like to thank the volunteer candidates and examiners who participated in this pilot technical skills exercise:

EBSQ-VASC 2002 Technical skills examiners

M. Cairoles (Spain), M. Lachat (Switzerland), P. Lamont (UK), J. Largiader (Switzerland), L. Mendes Pedro (Portugal), J.-B. Rico (France), J. Sandermann (Denmark), J. Scott (UK).

EBSQ-VASC 2002 Technical skills exam candidates

S. Debus (Germany), P. Cuenuod (Switzerland), M. Matsagas (Greece), A. Mansilha (Portugal), D. Mayer (Switzerland), J. Poniewierski (Belgium), R. Pulli (Italy) and E. Yilmaz (Netherlands/Turkey).

Mr Mark Jackson FRACS, Mr Colin Bicknell MRCS and Ms Vivienne Rose for their significant contribution towards the smooth running of the examination. Mr Jonathan Beard for his work on developing the surgical knot assessor used in the examination. The Vascular International Foundation for the loan of surgical instruments, distal anastomosis models and technical expertise that was available on the day of the examination. Limbs & Things, UK for providing the saphenofemoral junction models.

References


2. **EC Directive 75/362, Publ Eur Communities 1975; L167(30–6):1.**


Accepted 17 December 2003