Assessing the surgical skills of trainees in the operating theatre

Assessing the surgical skills of trainees in the operating theatre: a prospective observational study of the methodology

JD Beard,[1]* J Marriott,[2] H Purdie[3] and J Crossley[4]

[1]Sheffield Vascular Institute, Northern General Hospital, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK
[2]Department of Reproductive and Developmental Medicine, University of Sheffield, Sheffield, UK
[3]Clinical Research Facility, Royal Hallamshire Hospital, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK
[4]Academic Unit of Medical Education, University of Sheffield, Sheffield, UK

*Corresponding author

Executive summary

Health Technology Assessment 2011; Vol. 15: No. 1
DOI: 10.3310/hta15010

Health Technology Assessment
NIHR HTA programme
www.hta.ac.uk

Executive Summary

Background

Until recently, surgical training in the UK was based upon an apprenticeship model. Trainees undertook many years of training and were required to pass knowledge-based exams before becoming consultants. Surgical skills were not formally assessed. The Postgraduate Medical Education and Training Board now requires all postgraduate medical specialties to provide comprehensive curricula, in which the competencies defined in the syllabus are blueprinted to an assessment system. The introduction of the European Working Time Directive (EWTD), a shorter duration of training and UK NHS service pressures also demand the development of more efficient surgical training methods, in which supervised training opportunities are maximised.

Surgical specialties have introduced workplace-based assessment (WBA) to assess the surgical skills of trainees in the operating theatre. Some, including the Royal College of Obstetricians and Gynaecologists, have adapted an existing method, called Objective Structured Assessment of Technical Skills (OSATS). The Orthopaedic Curriculum and Assessment Project (OCAP) and the Intercollegiate Surgical Curriculum Programme (ISCP) have developed a new method called procedure-based assessment (PBA), which also predominantly assesses technical skills. The University of Aberdeen, in collaboration with the Royal College of Surgeons of Edinburgh, has developed a behavioural rating system called Non-technical Skills for Surgeons (NOTSS). This is designed to rate non-technical skills including situation awareness, communication and teamwork, decision-making and leadership.

The purpose of assessment can be to aid learning (assessment *for* learning, or formative) and/or to demonstrate achievement (assessment *of* learning, or summative). Whatever the purpose, the principal consideration of a well-designed and -evaluated assessment system is to ensure that the assessment methods adopted are valid, reliable and acceptable and have educational impact.

Objective

The primary aim of the study was to compare the user satisfaction and acceptability, reliability and validity of three WBA methods for assessing the surgical skills of trainees in the operating theatre (PBA, OSATS and NOTSS) across a range of different surgical specialties and index procedures.

Methods

This was a prospective, observational study conducted over 2 years within the operating theatres of three teaching hospitals in Sheffield.

The methods selected for study were PBA, OSATS and NOTSS as these address different aspects of surgical performance (technical and non-technical skills) and are used in differing assessment and training contexts in the UK. The specialties selected were obstetrics and gynaecology (O&G), upper gastrointestinal surgery, colorectal surgery, cardiac surgery, vascular surgery and orthopaedic surgery. Two to four typical index procedures were selected from each specialty.

Surgical trainees in the chosen specialties were directly observed performing typical index procedures on patients who, along with the trainees, had given their informed consent to participate in the study. Trainees were assessed using a combination of two of the three methods (OSATS or PBA and NOTSS for O&G as this specialty was the only one using OSATS; PBA and NOTSS for the other specialties) by the clinical supervisor (or consultant supervisor) for the case and the anaesthetist and/or scrub nurse, as well as one or more independent assessors from the research team.

The aim was that at least two assessors would assess each surgical trainee undertaking at least two different index procedures in his or her specialty on at least two occasions, equating with a minimum of eight assessments per trainee. This sampling strategy was designed to allow the estimation of variation in trainee performance between individual cases and types of index procedure and differences in case complexity, as well as variability in assessor stringency and subjectivity. Furthermore, the procedures would be assessed as close together as possible for an individual trainee to avoid any significant training effect. In this way, the study methodology was orientated to provide performance-focused assessments most suited to reliability analysis.

Generalisability theory provides a reliability estimate. It is not a hypothesis test and does not therefore include an accepted approach for power calculation. However, to produce dependable reliable estimates it is essential to sample each relevant factor, principally trainees, cases and assessors, as widely and representatively as possible. An overall target of 450 cases was set, of which 150 were intended to be within O&G to allow comparison of PBAs and OSATSs within that speciality.

User satisfaction and acceptability data for each assessment method from clinical supervisor and trainee perspectives were obtained from structured questionnaires. The reliability of each method was estimated using generalisability theory. Evidence of validity for the methods included internal tool structure, correlation between tools, construct validity, predictive validity, interprocedural differences, the effect of assessor designation and the effect of assessment on performance.

Results

Information about the study was sent to 832 patients but 274 were not approached to give consent because of lack of availability of an inpatient bed on the day of surgery, alteration or cancellation of the operating list or known non-availability of a trainee. Of the 558 patients who were given consent, a total of 437 (78%) cases were included in the study. The most common reasons for non-recruitment after consent were lack of availability of a trainee to perform the case (25%), the clinical supervisor personally performing the case despite a trainee being present (20%) and no list time available for training (12%).

Fifty-one clinical supervisors, 56 anaesthetists, 39 scrub nurses, two surgical care practitioners (SCPs) and four independent assessors provided 1635 assessments on 85 trainees undertaking the 437 cases. A total of 749 PBAs, 695 NOTSS and 191 OSATSs assessments were performed.

The PBA possesses high reliability ($G > 0.8$ for three assessors judging one different case each) for assessing the same index procedure. However, good PBA reliability for a mix of index procedures can be achieved only by using large numbers of cases and assessors owing to strong procedure-specific variance in scores.

Objective Structured Assessment of Technical Skills was evaluated only within O&G. It had lower reliability than PBA ($G > 0.8$ for five assessors judging one different case each) for assessing the same index procedure. OSATS also requires large numbers of cases and assessors for reliability over a mix of index procedures because of strong procedure-specific factors. A post hoc comparison of PBA reliability between O&G and non-O&G cases shows a striking difference. Within O&G, a good level of reliability ($G > 0.8$) was not obtained using a feasible number of assessments. Conversely, within non-O&G cases the reliability of PBA was exceptionally high, with only two assessor judgements for a given index procedure being required. These findings reveal that both tools perform differently within O&G. The most likely reason for this is the higher proportion of O&G trainees with training concerns (42% vs 4% for all other specialties).

The reliability of NOTSS was lower than that for PBA or OSATS ($G > 0.8$ for six assessors judging one different case each) for assessing the same index procedure. However, as procedure-specific factors exert a lesser influence on NOTSS, reliability for a mix of procedures can be achieved using eight assessor judgements.

Construct validity for PBA was demonstrated by the significant correlation of scores with age, specialty training (ST) level, total years and UK years of surgical training, total and recent experience of relevant index procedure ($r = 0.31–0.71$). The OSATS tool did not demonstrate any evidence for construct validity, which may also be explained by the O&G cohort-specific factors. All of the four NOTSS categories demonstrated construct validity for many of these measures: all significantly predicted decision-making and situation awareness scores ($r = 0.22–0.57$); only ST level, UK years of surgical training and recent experience of relevant index procedure predicted communication and teamwork and leadership scores ($r = 0.25–0.46$). NOTSS also demonstrated a valid internal structure, with the observed factor structure of scores almost perfectly matching the intended structure of the tool.

The scores for the three WBA methods correlated strongly and statistically significantly. The strongest correlations were within each tool ($r = 0.73$ between checklist and global ratings for PBA; $r = 0.84$ between task-specific and generic ratings for OSATS; $r = 0.74–0.76$ between the four categories within NOTSS), which is an indication of the good internal content validity of each tool. The correlations across the three methods ($r = 0.40–0.67$) were strongest between NOTSS and PBA or OSATS in the 'decision-making' domains in which there was the greatest overlap of assessment items. This provides evidence for criterion validity as tools measuring the same construct should correlate.

There is some evidence of predictive (outcome) validity for the study's assessment methods. We found twice as many significant correlations between case outcomes and scores than could be expected by chance alone. This is the first time that such an effect has been demonstrated for trainee assessments.

Except for OSATS, there was little variation in scoring between different designations of assessor (0%–4%). Our independent assessor ratings using PBA were as reliable as the clinical supervisor ratings, and our independent assessor ratings using NOTSS were as reliable as the anaesthetist and scrub nurse ratings.

There were only 27% of cases in which either the trainee or the clinical supervisor felt that performance had been affected by assessment, although there was little agreement between their judgements for individual cases. Although the cases judged to have been affected by direct observational assessment had lower scores, the video-recorded cases did not. This suggests that neither affect performance a priori but rather that a poor performance may be attributed to the assessment conditions.

User satisfaction and acceptability results are presented as descriptive statistics and as a proportion of those who responded to the questionnaires. The response rates were 85% for clinical supervisors and 78% for trainees regarding PBA, 85% for both clinical supervisors and trainees regarding OSATS, and 67% for scrub nurses and 54% for anaesthetists regarding NOTSS.

Clinical supervisors and trainees provided predominantly positive responses about the use of PBA, although the most positive responses were from trainees. The majority of clinical supervisors agreed that PBA was valuable for providing feedback (77%), as an assessment for learning (72%) and for a summative purpose (68.5%). Most clinical supervisors felt that PBA was important in surgical education (78%) and were likely to use PBA in the future, even if it were not mandatory (68%). The majority of trainees agreed that the PBA was valuable for providing feedback (88%), as an assessment for learning (72%) and for a summative purpose (64%). Most trainees also felt that PBA was important in surgical education (82%) and were likely to use PBA in the future (84%).

Clinical supervisors in O&G provided predominantly positive responses about the use of OSATS, but with a greater number of negative responses regarding its summative use. The majority of O&G clinical supervisors agreed that OSATS was valuable for providing feedback (88%) as an assessment for learning (76%) and for a summative purpose (59%). O&G trainees provided far less positive responses than clinical supervisors, again with the greatest number of negative responses regarding its summative use. While a majority agreed that OSATS was valuable for providing feedback (83%), 50% were of the opinion that OSATS was valuable as an aid to learning and only 39% for a summative purpose. For those clinical supervisors and trainees who used both methods, their overall satisfaction with OSATS was less than for PBA.

The vast majority of scrub nurses agreed that NOTSS allowed them to easily rate interpersonal skills such as communication, teamwork and leadership (92%), and cognitive skills such as situation awareness and decision-making (85%), although there was much less agreement from anaesthetists (60% and 27% respectively). A majority of scrub nurses and anaesthetists agreed that NOTSS was valuable for reflective practice (91% vs 73%) and as an adjunct to the assessment of technical skills (81% vs 60%). The majority of scrub nurses agreed that NOTSS would enhance safety in the operating theatre (65%), although there was less agreement from anaesthetists (27%).

Discussion

The PBA tool possesses good overall utility as an assessment method given the good evidence for high reliability, validity and user satisfaction/acceptability. Our results indicate that PBA is highly suitable as an assessment *for* learning and as an assessment *of* learning. Furthermore, the ISCP and OCAP can be reassured about the continued use of PBA as their main WBA method for surgical specialty trainees. However, the high reliability results for PBA are procedure specific and therefore trainees must be adequately assessed on each individual index procedure. We have no reason to believe that PBA would be less valid or reliable in other surgical specialties, although further evaluation within other specialties may be useful.

OSATS is a less reliable method than PBA although good reliability ($G > 0.8$) remains achievable using feasible numbers of assessor judgements. OSATS failed to demonstrate construct validity and there was lower overall user satisfaction than for PBA, especially among trainees. Owing to the lower overall utility, specialties that use OSATS might wish to consider altering the tool design or switching to PBA. However, there were fundamental cohort differences within O&G, with a higher proportion of senior trainees with training concerns. This is likely to have reduced

estimated reliability and undermined construct validity for OSATS. Furthermore, our post hoc analysis of the PBA within O&G and non-O&G specialties revealed that even the reliability of PBA was reduced in O&G.

Whether PBA or OSATS are used to assess surgical skills within a training programme, the purpose, timing and frequency of WBA require detailed guidance for both trainees and clinical supervisors to ensure that they are used correctly and provide maximum educational effectiveness. Even if relatively low numbers of assessments are required for good reliability, this should not detract from their primary purpose as an assessment for learning, which requires frequent assessment. Furthermore, user satisfaction/acceptability for a summative purpose is lower. Clinical supervisors would benefit from continued training in assessment and feedback techniques to maximise the educational potential of WBA.

Non-technical Skills for Surgeons is a promising tool for the assessment of non-technical skills. Good reliability ($G > 0.8$) can be achieved using feasible numbers of assessor judgements, without intensive assessor training. Given the prerequisite that reliable assessment using PBA/OSATS demands adequate assessment of individual index procedures, there would be no difficulty obtaining an adequate sample of a mix of procedures to permit a reliable assessment of non-technical skills using NOTSS. NOTSS may complement the more technical assessment methods, especially for trainees who have mastered the technical aspects of a procedure. Surgical specialties may wish to consider the inclusion of NOTSS into their assessment framework and/or consider integrating elements of NOTSS into their 'technical' WBA tools.

The analyses used to estimate reliability for the study's assessment methods highlighted that assessor designation contributed little variation (0%–4%) to scoring using PBA and NOTSS. These findings suggest that WBA could be completed by alternative assessors, such as anaesthetists, SCPs and scrub nurses.

The reliability of PBA and NOTSS was just as good for those assessors who had received less rigorous training. This has important implications for the routine implementation of WBA. However, training of clinical supervisors is required for good supervision and feedback.

Our difficulties with recruitment have shed light on the challenges faced by clinical supervisors and trainees in undertaking WBA. If the obstacles to recruitment that we have identified were addressed, we estimate that trainees might gain access to at least twice as many training cases within the same timeframe. These findings have important implications for training and assessment, given the requirement for surgical training to be more efficient within shorter training schemes with fewer hours for training. We have identified three levels of obstacles to achieving systematic supervised training in the operating theatre:

1. *Organisational-level obstacles* These may be amenable to change by successful lobbying for improved training conditions, e.g. allocation of more theatre time per case, ring-fenced beds for elective admissions, establishment of training opportunities at local diagnosis and treatment centres.
2. *Professional-level obstacles* These are amenable to change by forward planning and reorganisation of workload by the key stakeholders (clinical supervisors and trainees), e.g. rota design including taking trainees off on-call at night, voluntary use of the 8-hour EWTD opt-out for additional training.
3. *Individual-level obstacles* These are amenable to direct change by individual groups of clinical supervisors and trainees, with the intention of improving their working relationship for training, e.g. better matching of suitable trainees to appropriate surgical cases, consultant

commitment to regularly supervise trainees performing suitable cases, active trainee involvement in identifying and requesting opportunities for WBA.

Implications for practice

In summary, the main implications for the assessment of surgical trainees in the operating theatre are:

- Evaluating assessment methods using the utility index provides a comprehensive estimate of their validity, reliability and user satisfaction/acceptability to guide successful implementation.
- PBA has high utility for assessing predominantly technical skills, and relevant stakeholders can be assured of its suitability for assessment within surgical training.
- NOTSS has high utility for assessing non-technical skills. Surgical specialties may wish to consider including the method into their assessment framework.
- OSATS has more limited utility, but we do not know whether this finding extends to other cohorts of O&G trainees/assessors and other specialties that use OSATS.
- Trainees need to be adequately assessed using PBA/OSATS for each individual index procedure in order to achieve good reliability ($G > 0.8$).
- Assessment methods, such as PBA, that use anchored and well-defined standards for performance assist assessors in making highly reliable judgements.
- The primary purpose of WBA should be as an assessment *for* learning. Although we demonstrated good reliability using relatively low numbers of cases, regular assessment is desirable for maximum educational impact.
- User satisfaction/acceptability for PBA/OSATS methods is highest for providing feedback and as an aid for learning. These purposes need re-emphasising through training to improve overall utility and implementation.
- The good reliability of PBA and NOTSS across different assessor groups supports the use of non-surgeon assessors, e.g. anaesthetists, SCPs and scrub nurses.
- If the organisational-, professional- and individual-level obstacles to training that we have identified were addressed, trainees could access twice as many training cases within the same timeframe.

Limitations of this study

The main limitation of our study is that it was conducted in one city. However, there were three hospital sites that had a range of working and training cultures, and trainees rotated into these teaching hospitals from surrounding district general hospitals during the study period. Our sampling strategy for cases and almost total participation from potential trainees and clinical supervisors also increases our confidence in the generalisability of our conclusions. The study could not include all surgical specialties or all index procedures, but we believe that the methods have been adequately evaluated for interspecialty and interprocedural differences.

A high proportion of senior trainees with training concerns were found in our O&G cohort, which made this population more homogeneous. Under these conditions, the reliability of a tool may be reduced. These O&G cohort-specific differences may have been responsible for the lower OSATS validity/reliability and lower O&G than non-O&G PBA reliability. It is not known whether OSATS would demonstrate better reliability using different cohorts of O&G trainees or other specialties that use OSATS.

Ongoing and future research

We are continuing our research into the following areas using the wealth of data produced by the study. These include:

- The effect of assessment on performance: does qualitative analysis of our questionnaires suggest that WBA captures authentic performance?
- The nature of WBA feedback: what are the characteristics and quality of the feedback that we have observed and documented clinical supervisors providing?
- The ongoing validation of NOTSS: can our NOTSS data and DVDs be used to develop NOTSS for use in surgical training programmes?
- The influences on WBA user satisfaction and acceptability: can qualitative analysis of our questionnaires shed more light on this?
- The video assessment of surgical skills: can our DVDs be used to provide reliable assessments of trainees?
- The WBA training of clinical supervisors: can our DVDs be used in workshops/courses to improve training for assessors?

Further research is required into the following areas:

- the educational effectiveness of WBA
- the relationship between surgical experience, performance and outcomes
- the use of non-surgeon assessors to assess the surgical skills of trainees
- the value of using DVDs of operations for additional trainee feedback.

Conclusion

We believe that this is the largest study of the assessment of surgical skills in the workplace to have been undertaken. Despite the difficulties with recruitment, the primary aims of the study – to evaluate the reliability, validity and user acceptability and satisfaction of PBA, OSATS and NOTSS – were achieved.

Funding

Funding for this study was provided by Health Technology Assessment programme of the National Institute for Health Research.

Publication

Beard JD, Marriott J, Purdie H, Crossley J. Assessing the surgical skills of trainees in the operating theatre: a prospective observational study of the methodology. *Health Technol Assess* 2011;**15**(1).

How to obtain copies of this and other HTA programme reports

An electronic version of this title, in Adobe Acrobat format, is available for downloading free of charge for personal use from the HTA website (www.hta.ac.uk). A fully searchable DVD is also available (see below).

Printed copies of HTA journal series issues cost £20 each (post and packing free in the UK) to both public **and** private sector purchasers from our despatch agents.

Non-UK purchasers will have to pay a small fee for post and packing. For European countries the cost is £2 per issue and for the rest of the world £3 per issue.

How to order:

– fax (with **credit card details**)
– post (with **credit card details** or **cheque**)
– phone during office hours (**credit card** only).

Additionally the HTA website allows you to either print out your order or download a blank order form.

Contact details are as follows:

Synergie UK (HTA Department)
Digital House, The Loddon Centre
Wade Road
Basingstoke
Hants RG24 8QW

Email: orders@hta.ac.uk

Tel: 0845 812 4000 – ask for 'HTA Payment Services' (out-of-hours answer-phone service)

Fax: 0845 812 4001 – put 'HTA Order' on the fax header

Payment methods

Paying by cheque
If you pay by cheque, the cheque must be in **pounds sterling**, made payable to *University of Southampton* and drawn on a bank with a UK address.

Paying by credit card
You can order using your credit card by phone, fax or post.

Subscriptions

NHS libraries can subscribe free of charge. Public libraries can subscribe at a reduced cost of £100 for each volume (normally comprising 40–50 titles). The commercial subscription rate is £400 per volume (addresses within the UK) and £600 per volume (addresses outside the UK). Please see our website for details. Subscriptions can be purchased only for the current or forthcoming volume.

How do I get a copy of *HTA on DVD*?

Please use the form on the HTA website (www.hta.ac.uk/htacd/index.shtml). *HTA on DVD* is currently free of charge worldwide.

The website also provides information about the HTA programme and lists the membership of the various committees.

NIHR Health Technology Assessment programme

The Health Technology Assessment (HTA) programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The research findings from the HTA programme directly influence decision-making bodies such as the National Institute for Health and Clinical Excellence (NICE) and the National Screening Committee (NSC). HTA findings also help to improve the quality of clinical practice in the NHS indirectly in that they form a key component of the 'National Knowledge Service'.

The HTA programme is needs led in that it fills gaps in the evidence needed by the NHS. There are three routes to the start of projects.

First is the commissioned route. Suggestions for research are actively sought from people working in the NHS, from the public and consumer groups and from professional bodies such as royal colleges and NHS trusts. These suggestions are carefully prioritised by panels of independent experts (including NHS service users). The HTA programme then commissions the research by competitive tender.

Second, the HTA programme provides grants for clinical trials for researchers who identify research questions. These are assessed for importance to patients and the NHS, and scientific rigour.

Third, through its Technology Assessment Report (TAR) call-off contract, the HTA programme commissions bespoke reports, principally for NICE, but also for other policy-makers. TARs bring together evidence on the value of specific technologies.

Some HTA research projects, including TARs, may take only months, others need several years. They can cost from as little as £40,000 to over £1 million, and may involve synthesising existing evidence, undertaking a trial, or other research collecting new data to answer a research problem.

The final reports from HTA projects are peer reviewed by a number of independent expert referees before publication in the widely read journal series *Health Technology Assessment*.

Criteria for inclusion in the HTA journal series

Reports are published in the HTA journal series if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The research reported in this issue of the journal was commissioned by the National Coordinating Centre for Research Methodology (NCCRM), and was formally transferred to the HTA programme in April 2007 under the newly established NIHR Methodology Panel. The HTA programme project number is 06/92/05. The contractual start date was in April 2007. The draft report began editorial review in September 2009 and was accepted for publication in May 2010. The commissioning brief was devised by the NCCRM who specified the research question and study design. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the referees for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

The views expressed in this publication are those of the authors and not necessarily those of the HTA programme or the Department of Health.

ISSN 1366-5278

Published by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk), on behalf of NETSCC, HTA.
Printed on acid-free paper in the UK by the Charlesworth Group.