Health Technology Assessment 2011; Vol. 15: No. 1

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

JD Beard, J Marriott, H Purdie and J Crossley

January 2011
10.3310/hta15010

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

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.

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

Declared competing interests of authors: Jonathan Beard is a member of the Curriculum Development and Assessment Group of the Royal College of Surgeons of England.

Published January 2011
DOI: 10.3310/hta15010

This report should be referenced as follows:

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).

Health Technology Assessment is indexed and abstracted in *Index Medicus*/MEDLINE, *Excerpta Medica*/EMBASE, *Science Citation Index Expanded* (*SciSearch*®) and *Current Contents*®/Clinical Medicine.

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. MR

Abstract

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

Objectives: To compare user satisfaction and acceptability, reliability and validity of three different methods of assessing the surgical skills of trainees by direct observation in the operating theatre across a range of different surgical specialties and index procedures.
Design and setting: A 2-year prospective, observational study in the operating theatres of three teaching hospitals in Sheffield.
Methods: The assessment methods were procedure-based assessment (PBA), Objective Structured Assessment of Technical Skills (OSATS) and Non-technical Skills for Surgeons (NOTSS). The specialties were obstetrics and gynaecology (O&G) and upper gastrointestinal, colorectal, cardiac, vascular and orthopaedic surgery. Two to four typical index procedures were selected from each specialty. Surgical trainees were directly observed performing typical index procedures and assessed using a combination of two of the three methods (OSATS or PBA and NOTSS for O&G, PBA and NOTSS for the other specialties) by the consultant clinical supervisor for the case and the anaesthetist and/or scrub nurse, as well as one or more independent assessors from the research team.
Outcome measures: Information on user satisfaction and acceptability of each assessment method from both assessor and trainee perspectives was obtained from structured questionnaires. The reliability of each method was measured using generalisability theory. Aspects of validity included the internal structure of each tool and correlation between tools, construct validity, predictive validity, interprocedural differences, the effect of assessor designation and the effect of assessment on performance.
Results: Of the 558 patients who were consented, a total of 437 (78%) cases were included in the study: 51 consultant clinical supervisors, 56 anaesthetists, 39 nurses, 2 surgical care practitioners and 4 independent assessors provided 1635 assessments on 85 trainees undertaking the 437 cases. A total of 749 PBAs, 695 NOTSS and 191 OSATSs were performed. Non-O&G clinical supervisors and trainees provided mixed, but predominantly positive, responses about a range of applications of PBA. Most felt that PBA was important in surgical education, and would use it again in the future and did not feel that it added time to the operating list. The overall satisfaction of O&G clinical supervisors and trainees with OSATS was not as high, and a majority of those who used both preferred PBA. A majority of anaesthetists and nurses felt that NOTSS allowed them

to rate interpersonal skills (communication, teamwork and leadership) more easily than cognitive skills (situation awareness and decision-making), that it had formative value and that it was a valuable adjunct to the assessment of technical skills. PBA demonstrated high reliability ($G > 0.8$ for only three assessor judgements on the same index procedure). OSATS had lower reliability ($G > 0.8$ for five assessor judgements on the same index procedure). Both were less reliable on a mix of procedures because of strong procedure-specific factors. A direct comparison of PBA between O&G and non-O&G cases showed a striking difference in reliability. Within O&G, a good level of reliability ($G > 0.8$) could not be obtained using a feasible number of assessments. Conversely, the reliability within non-O&G cases was exceptionally high, with only two assessor judgements being required. The reasons for this difference probably include the more summative purpose of assessment in O&G and the much higher proportion of O&G trainees in this study with training concerns (42% vs 4%). The reliability of NOTSS was lower than that for PBA. Reliability for the same procedure ($G > 0.8$) required six assessor judgements. However, as procedure-specific factors exerted a lesser influence on NOTSS, reliability on a mix of procedures could be achieved using only eight assessor judgements. NOTSS also demonstrated a valid internal structure. The strongest correlations between NOTSS and PBA or OSATS were in the 'decision-making' domain. PBA and NOTSS showed better construct validity than OSATS, the year of training and the number of recent index procedures performed being significant independent predictors of performance. There was little variation in scoring between different procedures or different designations of assessor.

Conclusions: The results suggest that PBA is a reliable and acceptable method of assessing surgical skills, with good construct validity. Specialties that use OSATS may wish to consider changing the design or switching to PBA. Whatever workplace-based assessment method is used, the purpose, timing and frequency of assessment require detailed guidance. NOTSS is a promising tool for the assessment of non-technical skills, and surgical specialties may wish to consider its inclusion in their assessment framework. Further research is required into the use of health-care professionals other than consultant surgeons to assess trainees, the relationship between performance and experience, the educational impact of assessment and the additional value of video recording.

Funding: The National Institute for Health Research Health Technology Assessment programme.

Contents

Glossary

Annual review of competence progression (ARCP) A postgraduate school (deanery) process that scrutinises each trainee's suitability to progress to the next stage of, or to complete, a training programme. It is usually held annually, but some specialties have more frequent reviews in the early years of training. Foundation programmes have a similar annual review process. The review panel, which includes the programme director, bases its recommendations on evidence in the trainee's portfolio of experience and competencies gained, together with the reports of the supervisor(s). The ARCP is not in itself an assessment exercise.

Appraisal An individual and private planned review of progress between trainee and clinical supervisor that focuses on achievements, future learning and career guidance. Appraisal forms part of the initial, interim and final meetings that trainees have with their educational or clinical supervisor during a placement.

Assessment The process of measuring a trainee's knowledge, skills, judgement or professional behaviour against defined standards. Assessment should be as objective and reproducible as possible. A reliable test should produce the same or similar score on two occasions or by two assessors. The validity of a test is determined by the extent to which it measures what it sets out to measure and its educational impact. Assessments can be referenced in two ways:

- *Criterion referenced* refers to an absolute standard, i.e. the trainee's performance against a benchmark. Such a benchmark might be the ability to perform a procedure competently without help from the assessor.
- *Norm referenced* ranks a trainee's performance against all the others in the same cohort, i.e. satisfactory for that level of training. Norm-referenced assessments are inherently more difficult to determine and, whenever possible, should not be used.

Assessment can have different and multiple purposes, including determining a level of competence, aiding learning through constructive feedback, measuring progress over time or certifying competence. Assessments can be categorised as *for* or *of* learning, although there is a continuum between these two poles.

Assessment *for* learning Is primarily aimed at aiding learning through constructive feedback that identifies areas for development. Alternative terms are formative or low-stakes assessment. Lower reliability is acceptable for individual assessments as they can and should be repeated frequently. This increases their reliability and helps to document progress. Such assessments are ideally undertaken in the workplace.

Assessment *of* learning Is primarily aimed at determining a level of competence to permit progression of training or certification. Such assessments are undertaken infrequently (e.g. examinations) and must have high reliability as they often form the basis of pass/fail decisions. Alternative terms are summative or high-stakes assessment.

Assessment system An assessment system (or assessment programme) is designed to ensure that trainees learn the knowledge, skills, judgement and professional behaviours required by a training syllabus. The combination of an assessment system and a syllabus are the key components that specifically address assessment practice within a curriculum. Contemporary best practice favours assessment systems that are multifaceted and assess an appropriate spectrum of a syllabus in a reliable way. This is done through a blueprint.

Assessor An experienced health-care professional (HCP) who undertakes an assessment. Assessors require training in the relevant assessment methodology and should normally be competent (preferably expert) in the knowledge, skill, judgement or professional behaviour that is being assessed. Training is not required for HCPs who provide ratings for multisource feedback.

Blueprint A template used to define the content of a syllabus or an assessment in terms of key competencies. This can help to ensure that the assessments used in the assessment system cover all the competencies required by the syllabus.

Certification The process by which governmental, non-governmental or professional organisations or other statutory bodies grant recognition to a trainee who has met certain predetermined standards specified by the organisation and who voluntarily seeks such recognition.

Clinical supervisor A senior doctor (trainer) responsible for overseeing a trainee's clinical work and providing constructive feedback during a training placement. Some training schemes appoint an educational supervisor for each placement. The roles of clinical and educational supervisor may then be merged.

Competence A trainee's ability to perform a particular activity to the required standard (i.e. that required for patient safety), while being observed in the workplace or in a controlled representation of the workplace (e.g. in simulation). Competence comes from experience combined with constructive feedback and reflective practice (self-assessment/insight). Competence is a prerequisite for satisfactory performance in real life, although many doctors progress to a higher level of excellence during their career. A competent doctor may perform poorly for many reasons including tiredness, stress, illness or a lack of resources.

Competencies A set of abilities that includes knowledge, skills, judgement and professional behaviours.

Construct A construct is an attribute, proficiency, ability or skill that exists in theory and has been observed to exist in practice, such as 'surgical skill'. Constructs are vital within assessment theory as they provide the underpinning framework for establishing assessment design and validity.

Curriculum A curriculum is a statement of the aims and intended learning outcomes of an educational programme. It states the rationale, content, organisation, processes and methods of teaching, learning, assessment, supervision and feedback. If appropriate, it will also stipulate the entry criteria and duration of the programme.

Educational agreement A mutually acceptable educational development plan drawn up jointly by the trainee and his or her educational supervisor. The content of the educational agreement will depend upon the aspirations of the trainee (as laid out in their personal development plan), the learning outcomes required by the curriculum and the opportunities available during the placement. A structured learning plan is an alternative term. The learning outcomes that have been achieved should be signed off by the educational or clinical supervisor at the end of each placement.

Educational impact *See* Consequential validity (under **Validity**).

Educational supervisor A senior doctor (trainer) responsible for the overall supervision and management of a trainee's educational programme during a training placement or series of placements. The educational supervisor is responsible for the trainee's educational agreement.

Experience Exposure to a range of medical practice and clinical activity.

Formative assessment *See* **Assessment *for* learning**.

Generalisability theory Generalisability theory was developed by Cronbach as an extension of classic reliability theory and of his own procedure for calculating Cronbach's alpha. The theory holds that none of the variation across scores is random, but all the variance can be attributed to one or other factor (e.g. the stringency of the assessor, the ability of the trainee, etc.). The *G*-study uses variance component analysis to estimate how big an effect each relevant factor has on the assessment score. The *D*-study then combines these variance components using equations that express the reliability of the assessment. For example, if the score varies greatly across trainee identity (and therefore, presumably, ability), but varies little across the cases that a trainee performs or across judge identity (and therefore, presumably, stringency), then the assessment will be calculated to be reliable. Where reliability is calculated for the observed population of trainees, judges and cases the figure is called a reliability coefficient. Where the variances are used to extrapolate beyond the observed sample by mathematical modelling, the figure is called a generalisability coefficient. This kind of analysis supersedes the classic estimation of 'reliability' on the basis of single sources of variation such as 'inter-rater' reliability or 'test, retest' reliability because it evaluates all these sources of error simultaneously in an overlapping experiment that uses all the available data. For this reason, too, the coefficients are lower than those produced by classic reliability tests because all the sources of error are combined.

High-stakes assessment *See* **Assessment *of* learning**.

Learning outcomes The competencies to be acquired by the end of a period of training.

Low-stakes assessment *See* **Assessment *for* learning**.

Multisource feedback An important tool for obtaining evidence about interpersonal and communication skills, judgement, professional behaviour and clinical practice. All those working with a trainee (including trainers, fellow trainees and senior nurses) are asked to rate the trainee's performance in various domains such as teamwork, communication, decision-making, etc. towards the end of a training placement. These ratings are collated and fed back to the trainee by his or her supervisor. This forms an important part of the appraisal process. Alternative terms are peer review or 360° feedback (often incorrectly called 360° appraisal)

Peer review See **Multisource feedback**.

Performance The application of competence in real life. In the case of medicine, it denotes what a trainee actually does in his or her encounters with patients, their relatives and carers, colleagues, team members, other members of staff, etc. Performance is not the same as knowing or being able to do everything. On the contrary, it may well be about knowing what you do not or even cannot know – in other words, knowing your own limitations.

Personal development plan (PDP) A prioritised list of educational needs and intended learning outcomes compiled by a trainee prior to meeting with the educational supervisor. The PDP is an integral part of reflective practice and self-directed learning.

Placement The period of postgraduate medical training in one specialty at one training institution. In the early years of training there is often more than one placement per year (e.g. three 4-month placements in the foundation programme). There may be a different educational supervisor for each placement or one for the whole year. In the latter case, day-to-day supervision will be overseen by a clinical supervisor.

Portfolio A collection of evidence documenting a trainee's learning and achievements during his or her training. The trainee takes responsibility for the portfolio's creation and maintenance. Portfolios have traditionally been paper based but many training programmes are moving to electronic (web-based) portfolios. In the UK, portfolios are used routinely as a Record of In-Training Assessment (RITA), which forms the basis for the annual review of progress. This process is now termed the ARCP.

Professionalism Adherence to a set of values comprising statutory professional obligations, formally agreed codes of conduct, and the informal expectations of patients and colleagues. Key values include acting in the patient's best interest and maintaining the standards of competence and knowledge expected of members of highly trained professions. These standards will include ethical elements such as integrity, probity, accountability, duty and honour. In addition to medical knowledge and skills, medical professionals should present psychosocial and humanistic qualities such as caring, empathy, humility and compassion, social responsibility and sensitivity to people's culture and beliefs. Professionalism is demonstrated by professional behaviour.

Programme director A senior doctor with overall responsibility for a postgraduate training programme (foundation or specialty), which includes a number of trainees and their respective trainers.

Reliability Expresses a trust in the accuracy or provision of the correct results. In the case of assessments, it is an expression of precision and discrimination. There are several important dimensions of reliability. These include:

- *Equivalence* or alternate-form reliability is the degree to which alternate forms of the kind of assessment produce congruent results.
- *Homogeneity* is the extent to which various items in an assessment legitimately link together to measure a single characteristic.
- *Inter-rater* reliability refers to the extent to which different assessors give similar ratings for similar performances.
- *Intra-rater* reliability is concerned with the extent to which a single assessor would give similar marks for almost identical performance.

Review Consideration of past events, achievements and performance. This may be either a formal or an informal process and can be an integral part of appraisal, assessment and feedback.

Reflective practice A process of evaluating one's own achievements, behaviour, professional performance and competencies. Reflective practice is an important part of self-directed and lifelong learning and requires insight into one's own areas of development. An alternative term is self-assessment.

RITA Record of In-Training Assessment. A portfolio of assessments that are carried out during training, which is used throughout UK postgraduate medical education. It is important to note that the RITA is not an assessment in its own right, nor is it a review of progress, although it is likely to be used as a source of evidence, gained through assessment, that informs the ARCP.

Self-directed learning The method of learning used by successful adult learners who take responsibility for their own learning. Such learning is usually goal motivated and relevant, i.e. applicable to their work or other responsibilities. Adult learners may not be interested in knowledge for its own sake.

Skill The ability to perform a task to at least a competent level. A skill is best (most efficiently) gained through regular practice (experience) combined with reflective practice (self assessment/ insight) and constructive feedback.

Standards In medical education standards may be defined as 'a model design or formulation related to different aspects of medical education and presented in such way to make possible assessment of graduates' performance in compliance with generally accepted professional requirements'. Thus, a standard is both a goal (what should be done) and a measure of progress towards that goal (how well it was done).

Summative assessment *See* **Assessment *of* learning**.

Syllabus A list, or some other kind of summary description, or course contents or topics that might be tested in examinations. In modern medical education, a detailed curriculum is the document of choice and the syllabus would not be regarded as an adequate substitute, although one might usefully be included as an appendix.

360° feedback *See* **Multisource feedback**.

Trainee Any doctor participating in an educationally approved postgraduate medical training programme (foundation or specialty).

Trainer A senior doctor who provides educational support for a more junior doctor (trainee). Trainers include clinical and educational supervisors. All trainers require training in teaching and assessment methods, including giving constructive feedback. Educational supervisors require additional training in appraisal and career guidance.

Training The ongoing, workplace-based process by which experience is obtained, constructive feedback provided and key competencies achieved.

Triangulation The principle, particularly important in workplace-based assessment (WBA), that whenever possible evidence of progress, attainment or difficulties should be obtained from more than one assessor, on more than one occasion, and if possible using more than one assessment method.

Utility Utility refers to an evaluation, often in cost–benefit form, of the relative value of using an assessment, or using one kind of assessment rather than another. An assessment with good utility must have high *reliability*, *validity* and *educational impact*. It must also be *acceptable* to assessors and trainees (covert surveillance may be reliable but it is probably unacceptable in most cases) and *feasible* (there is no point in developing a 'perfect' assessment that is too difficult or expensive to use).

Validity In the case of assessment, validity refers to the degree to which a measurement instrument truly measures what it is supposed to measure. It is concerned with whether the right things are being assessed, in the right way, and with a positive influence of learning. There are many different dimensions of validity including:

Content validity An assessment has content validity if the components reflect the abilities (knowledge, skills or behaviours) that it is designed to measure.

Face validity Related to content validity. Face validity can be described from the perspective of an interested lay observer. If he or she feels that the right things are being assessed in the right way, then the assessment has good face validity.

Construct validity The extent to which the assessment, and the individual components of the assessment, test the professional constructs on which they are based. For instance, an assessment has construct validity if senior trainees achieve higher scores than junior trainees.

Predictive validity This refers to the degree to which an assessment predicts expected outcomes. For example, a measure of attitudes (behaviour) towards preventive care should correlate significantly with preventive care behaviours.

***Consequential validity* (educational impact)** This is an important aspect of the validity of assessment. It refers to the effect that an assessment has on learning, and in particular on what trainees learn and how they learn it. For example, they might omit certain aspects of a syllabus because they do not expect to be assessed on them, or they might commit large bodies of factual knowledge to memory without really understanding them in order to pass a test of factual recall, and then forget them soon afterwards. Both these behaviours would indicate that the assessment has poor educational impact because both lead to poor learning behaviours.

WBA Workplace-based assessment. The assessment of performance based on what a trainee actually does in the workplace. The main aim of WBA is to assess those aspects of real day-to-day performance that a remote-controlled assessment of competence cannot assess. It is very well suited to the purpose of aiding learning (assessment *for* learning) by providing trainees with constructive feedback. Trainees can use the same methodology to assess themselves (reflective practice). The assessments help the supervisor to chart a trainee's progress during a placement. Although the principal role of each assessment is *for* learning, the entire collection can be used to inform the ARCP.

In most of its UK implementations, WBA is trainee led, the trainee choosing the method, timing, activity and assessor under the guidance of the supervisor according to the learning outcomes laid out in the educational agreement. Trainees are encouraged to use as many different assessments and assessors as possible, as this improves reliability.

Most WBAs are designed to help the assessor provide objective, constructive feedback immediately after the activity. Although many WBAs are web based, the forms can be downloaded and a paper copy used for the assessment and feedback. The trainee can then upload the results on to the website for authorisation by the assessor.

Multisource feedback is a unique form of WBA in that it uses a collection of untrained raters, and the feedback based on the collated ratings is subsequently fed back to the trainee by the supervisor. Thus, it has aspects of assessment *of* and *for* learning.

List of abbreviations

ANTS	Anaesthetist's Non-technical Skills
ARCP	annual review of competency progression
ASA	American Society of Anesthesiologists
AVR	aortic valve replacement
ATIS	adjusted total item score (PBA)
ATTS	adjusted total task score (OSATS)
ATGS	adjusted total generic score (OSATS)
CBA	competency-based assessment
CCT	Certificate of Completion of Training
CDS	communication/teamwork domain score (NOTSS)
CUSUM	cumulative sum
DDS	decision-making domain score (NOTSS)
EWTD	European Working Time Directive
GMC	General Medical Council
GI	gastrointestinal (surgery)
GS	global score (NOTSS)
HDU	high-dependency unit
IA	independent assessor
ICU	intensive care unit
ISCP	Intercollegiate Surgical Curriculum Programme
LDS	leadership domain score (NOTSS)
Mini-PAT	Mini-Peer Assessment Tool
MTAS	Medical Training Application Service
NOTSS	Non-technical Skills for Surgeons
O&G	obstetrics and gynaecology
OCAP	Orthopaedic Curriculum and Assessment Project
OpComp	Operative Competency (form)
OSATS	Objective Structured Assessment of Technical Skills
PMETB	Postgraduate Medical Education and Training Board
PBA	procedure-based assessment
RITA	Record of In-Training Assessment
RCOG	Royal College of Obstetricians and Gynaecologists
SAC	Specialty Advisory Committee
SCP	surgical-care practitioner
SDS	situation awareness domain score (NOTSS)
ST	specialty training
WBA	workplace-based assessment

All abbreviations that have been used in this report are listed here unless the abbreviation is well known (e.g. NHS), or it has been used only once, or it is a non-standard abbreviation used only in figures/tables/appendices, in which case the abbreviation is defined in the figure legend or in the notes at the end of the table.

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.

Chapter 1

Introduction

> The art of medicine is to be learned only by experience, 'tis not an inheritance; it cannot be revealed. Learn to see, learn to hear, learn to feel, learn to smell, and know that by practice alone can you become an expert.
>
> Sir William Osler, 1919[1]

Background and rationale for the study

Sir William Osler was only partially correct. Experience is vital, but, as Halsted observed,[2] 'Experience can mean doing the wrong thing over and over again'. Becoming an expert also requires feedback, which is informed by assessment. Assessment is therefore the cornerstone of education and training, driving both teaching and learning, and shaping the overall nature of a curriculum. Within the context of postgraduate medical education, an assessment system may assume a regulatory role, by ensuring the quality of training delivered and educational standards for the purposes of professional regulation, clinical governance and patient safety.

The principal consideration of a well-designed and -evaluated assessment system is to ensure that the assessment methods adopted are valid, reliable, acceptable and cost-effective and have educational impact. Evidence of validity and reliability are essential characteristics of fair and defensible assessments and a prerequisite for making the high-stakes assessment decisions that allow progression in training and certification. All assessment systems must provide evidence of assessment rigour, in particular for identifying underperforming doctors who could compromise patient safety. The development of robust methods of assessment is axiomatic as they underpin the current competency-based assessment systems and curricula of all UK postgraduate training programmes.

The evolution of surgical training in the UK

Surgical training in the UK has been in a state of constant evolution since the 1990s. Radical changes to the regulation of surgical training and reforms in educational policy, with fundamental shifts in the delivery of surgical care and public attitudes towards surgery, have collectively driven the modernisation towards competency-based surgical curricula.

Until the Calman report[3] initiated changes in the structure of postgraduate medical training, the Halstedian model of 'surgical preceptorship' had been used with modifications for over a century.[2] The craft specialties, including surgery as well as medical and interventional specialties, taught technical and surgical procedures through clinical exposure and experience within lengthy training programmes. Trainees were required to complete a set number of years of training and pass knowledge-based exams in their specialty to achieve their Certificate of Completion of Training (CCT). Technical skills and non-technical skills including teamwork, decision-making and communication were not formally assessed in exams or in the workplace.

There has been growing pressure to introduce regular assessment of practical skills or competence in the interests of public, political and professional accountability. The assessment

of practical skills was initially encouraged by the Joint Committee for Higher Surgical Training (JCHST) in 2001. This was closely supported by the 2002 consultation paper *Unfinished Business*,[4] which set out the case for major reforms in postgraduate training, concluding that 'a new Postgraduate Medical Education and Training Board [PMETB] will be required to ensure that, throughout training, all assessments and examinations … are appropriate, valid and reliable'.

The PMETB assumed statutory responsibility in 2005, its remit being to establish and maintain standards for all postgraduate assessment programmes and curricula.[5] The PMETB has required all postgraduate specialties to provide comprehensive curricula, in which the competencies defined in the syllabus are blueprinted to the assessment programme. All postgraduate assessment programmes required urgent reform to be able to assess those competencies that could not be assessed adequately by examinations, in particular technical skills and professional behaviours. For the craft specialties, the formalised assessment of technical skills demanded different methods of assessment. Workplace-based assessments (WBAs) have been implemented to address this gap in assessment programmes.

Workplace- and competency-based assessment

Although competency-based assessment (CBA) and WBA have been relatively recently introduced within medicine, accelerated through changes to policy, these originate and are now well established within the education field. Since the 1970s, educationalists have been concerned with defining aspects of learning, in order to clearly define the content of training or education periods and to align these with assessment processes. Bloom's taxonomy[6] was one of the first frameworks developed, which divided learning into 'knowledge', 'skills' and 'attributes'. This led to the development of *learning objectives* for defining educational content and *learning outcomes* for defining assessment content. The outcome-based model of assessment remains the predominant approach to assessment within higher education.[7]

The competency-based approach to workplace assessment originates from occupational and vocational sectors of education. During the 1980s, there was a political drive to make the UK workforce more competitive globally. Parallel attempts were made to divide aspects of vocational learning into *competencies*, using functional analysis of job roles, to serve the purpose of assessing occupational competence within vocational training.[8] The National Vocational Movement[9] produced 'standards of competence' and work-based competencies were assessed upon these clearly defined outcomes.

The assessment of competencies is highly relevant to medical practice. Competencies define job-related tasks or roles, using applied and integrated aspects of knowledge, skills and attributes. CBA is concerned with the assessment of essential competencies designed to ensure that health professionals perform their job to an acceptable standard of clinical competence. From the late 1990s, this approach to assessment has been adopted widely in health education including nursing,[10] undergraduate medical training[11,12] and postgraduate medical training.[13]

Within UK postgraduate training, CBA was implemented under the umbrella of 'Modernising Medical Careers'.[14] This started with the Foundation Programme in 2005,[15] which marked the introduction of WBA into postgraduate medical training. That same year, the Orthopaedic Curriculum and Assessment Project (OCAP) (www.ocap.org.uk) was introduced. In 2007 the other surgical specialties participating in the Intercollegiate Surgical Curriculum Programme (ISCP), together with obstetrics and gynaecology (O&G), launched their new competency-based surgical curricula (www.iscp.ac.uk, www.rcog.org.uk). Central to these new curricula are the formal, structured assessments of surgical skill in the workplace, to provide an authentic

assessment of day-to-day working practice and to maximise the educational impact of the experience.

It is important to acknowledge that the adoption of WBA has not been solely a response to changes in educational policy. There are many strengths to WBA, as well as limitations, as with any approach to assessment.[16] WBA offers a method of assessing job-related competencies that cannot be fully assessed by other assessment methods. Focusing on competencies achieved rather than time served allows for more individualised training and the opportunity to identify those trainees who may need additional support. It is potentially highly valid, assessing what doctors actually do in practice (performance) as well as their ability to modify their performance under different clinical circumstances. The evidence for the reliability of WBA is emerging, although there is a current paucity stemming from the difficulties involved in collating suitable and sufficient assessments for reliability evidence.

The success of an assessment method is determined by its effectiveness and WBA is no exception. It is important that WBA is shown to translate positive learner reactions and learning outcomes into improvements in clinical performance and patient/health outcomes. Using a modified version of Kirkpatrick's model,[17] described in Freeth *et al*.,[18] four levels of assessment effectiveness can be evaluated. The lowest level, level 1, concerns learners' reactions and satisfaction with the assessment experience; level 2 concerns a change in learning outcomes; level 3 concerns a change in behaviour (divided into self-reported changes for level 3a and measured changes for level 3b); and level 4 concerns a change in patient outcomes. A review of performance-based assessment, including the use of peer assessment, portfolio, appraisal report and medical audit, highlighted that there are 19 studies providing evidence of positive assessment effectiveness at levels 1, 2 and 3.[19] One of these studies reported the audit loop for using SAIL (Sheffield Assessment Instrument for Letters), a WBA method designed to improve the communication between secondary and primary care using referral letters, with every doctor improving their mean scores 3 months after receiving feedback on the quality of their clinic letters.[20] It is acknowledged that empirical evidence is lacking for WBA supporting an improvement in the routine practice of doctors. It is a significant challenge to design and implement research that is able to demonstrate changes in patient outcomes for any given assessment method, although it is increasingly recognised that evaluation should focus on programmes rather than on methods.[21] This is because assessing performance requires integrated assessment methods within a broad assessment programme. As with clinical evidence guidelines, much of the evidence for WBA is circumstantial. For example, there is evidence that the educational principles of WBA, including one-to-one competency-based instruction[22] and giving feedback on performance,[23,24] are effective educational interventions. Improvements in training, including changes in attitudes and behaviours towards supervised training with feedback, should be viewed as the first necessary step towards long-term improvements in patient care and surgical safety. Furthermore, a significant outcome of using WBA is that trainees in difficulty and underperforming trainees, not previously identified, can more easily have their training needs identified by WBA with appropriate remediation.

The introduction of WBA is a challenge for all stakeholders, and issues of implementation, including the time and resources required, constitute potential limitations. However, there are several compelling reasons that make WBA more suited to the current surgical training climate than the examinations and time-served model.

The opportunity to gain surgical training and experience in the operating theatre has decreased significantly since the Calman report initiated shortened surgical training time.[3] We, and others, have shown a reduction in the number of operations undertaken and the level of surgical competence achieved by surgical trainees following the Calman reforms.[25] Furthermore, the European Working Time Directive (EWTD) was enacted into UK law in 1998 and has legislated

for a step-wise reduction in the working hours of doctors in training. This directive undoubtedly has beneficial objectives, both in the interests of patient safety and quality of care and in terms of providing doctors with a better work–life balance and good training. The maximum number of hours trainees can stay in the workplace is 48 hours per week from August 2009, with a voluntary option of working 56 hours.[26] To balance training with service provision, it has been necessary to institute full shift rotas for the majority of training posts, requiring trainees to work regular night shifts with a loss of daytime supervised surgical training. These changes in working practices equate to an overall reduction in surgical training time in the operating theatre, with increasing amounts of elective surgery being performed by consultants. Under these conditions, the traditional apprenticeship model is no longer appropriate for the current training structure.

The demonstration of surgical competence ensures that the dual concerns of patient safety and training quality are satisfied. It must be emphasised that WBA is not a substitute for procedural experience. The two are complementary in terms of achieving surgical competence. The questions of how many hours or number of procedures are required to train a surgeon are yet to be answered.

The adoption of CBA and WBA tools is designed to improve the efficiency of surgical training, by directing both clinical supervisor (or consultant supervisor) and trainee to use each supervised operation (or other clinical encounter) as an opportunity for objective assessment and constructive feedback. Direct supervision and feedback are integral to WBA practice, and should take place as a formalised part of the assessment process, in contrast to the more ad hoc nature of supervision and feedback often observed within the apprenticeship training model. In this way, surgical training using WBA becomes a focused educational activity with joint trainee and clinical supervisor responsibility, to meet the challenge of an overall reduction in training time. Specific to the operating theatre, WBA tools can be completed with feedback in the time between surgical cases.

Principles of assessment

Assessment theory and practice can differ substantially. To address this gap between theory and practice, we illustrate the key concepts from assessment theory to provide a relevant and applied background for WBA.

Assessment is concerned with a process of measuring a trainee's knowledge, skills, judgement or professional behaviour against defined standards. The WBA tools used within the current surgical curricula use explicit judgements against defined performance-based criteria. Assessment is distinct from appraisal. The latter is designed to review the progress and performance of an individual trainee. It is a planned process, focusing on achievements, future learning and career guidance, in which the criteria are usually internal and individual.

This report is focused on the assessment of surgical skills, whereby the skills of surgeons in training are judged against a defined reference. Assessments can be referenced in two ways:

1. Criterion-referenced assessment compares a trainee's performance to an absolute standard. Such a benchmark might be the ability to perform a procedure competently and independently.
2. Norm-referenced assessment compares a trainee's performance with other trainees in the same cohort. Such a reference might include a below average, average or above average performance within a particular cohort.

Criterion referencing, rather than norm referencing, is used within WBA, i.e. a trainee's performance is not compared with their peers but with a fixed standard. Criterion-referenced assessments assist assessors in making consistent judgements by setting absolute standards of performance and a clear description of the standard expected. Recent consensus statements on WBA support the use of criterion-referenced assessments using rating scales with clear text descriptors.[27]

At best, norm-referenced assessments imply that the 'average' level will increase with time and experience. However, this is inherently more difficult for assessors to determine, as it depends on the performance of that particular cohort and relies on intuitive assessor judgements.

Various benchmarks are used by current WBA tools. Those used by the Foundation Programme and core medical and core surgical training use the standard expected of a trainee at that level of training.[28] The procedure-based assessment (PBA) used by both OCAP and ICSP for the specialty index procedures uses the standard expected for certification of completion of training[29] with a summary judgement based on the ability of the trainee to perform the procedure independently. The Objective Structured Assessment of Technical Skills (OSATS) used by the Royal College of Obstetricians and Gynaecologists (RCOG) during the course of this study (version in use before February 2009) adopts a pass/fail judgement. The standard against which this judgement is made is not explicit on the form,[30] although there are benchmarks within the curriculum and training portfolio that define the standard as that required for independent practice.

Purpose of assessment

Assessment has been shown to drive both learning[31,32] and teaching.[33] In view of this, it is essential when designing an assessment system that the object and purpose of the assessment are clearly defined to all stakeholders from the outset.

The purpose of an assessment should determine every aspect of its design.[34] These aspects include:

1. the choice of assessment methods
2. the selection of assessment tools
3. the way in which the above are combined
4. the number of assessments required
5. the timing of assessments
6. the way in which the outcomes are used to make decisions regarding progression or certification.

The purposes of assessment will be different for the individual trainee being assessed, the training programme, the employer and the public,[35] and are presented in *Table 1*. Some purposes may be shared by different stakeholders whereas there may also be conflicting assessment purposes between different stakeholder groups.

Assessment has multiple purposes. However its educational purpose can be broadly divided into:

1. Assessment *for* learning (alternative terms are formative or low-stakes assessment). This is primarily intended to aid a trainee's learning through the provision of constructive feedback, identifying good practice and areas for development.
2. Assessment *of* learning (alternative terms are summative or high-stakes assessment). This is primarily aimed at determining a level of educational achievement relative to a defined

TABLE 1 Purposes of assessment: adapted from van Sickle *et al.*[24]

For the trainee
Provide feedback about strengths and weaknesses to guide future learning
Foster habits of self-reflection and self-remediation
Promote access to advanced training
For the curriculum
Respond to lack of demonstrated competence (targeted training)
Certify progression in training over time
Certify achievement of curricular outcomes
Foster curricular change
Create curricular coherence
Cross-validate other methods of assessment in the curriculum
Establish standards of competence for trainees at different levels
For the institution
Discriminate among trainees for progression in training or access to subspecialty training
Guide a process of institutional self-reflection and self-remediation
Develop shared educational values among a diverse community of educators
Promote faculty development
Provide data for educational research
For the public
Certify competence of doctors in training
Identify unsafe or poorly performing doctors

standard. Such assessments are infrequent and usually take place at set times to permit progression in training or certification.

Some assessments can serve both purposes and there is a continuum between these two poles. *In-training assessment* may seek to integrate both assessment purposes into an overarching assessment framework, in recognition of the fact that they may be complementary in reinforcing feedback and self-directed learning.[36]

Workplace-based assessment has a particular strength for formative assessment (i.e. assessment *for* learning), through the direct observation of trainees by trained assessors and the provision of immediate feedback. However, an assessment *of* learning for progression still needs to be made to inform the annual review of competency progression (ARCP), using all sources of evidence including WBA. Conflict exists with the use of WBA to serve both educational purposes, and this is carefully considered within the most recent policy documents on WBA.[16,27] The importance of making the purpose of assessment explicit to stakeholders is now recognised as fundamental to the successful implementation of WBA.

Levels of assessment

There are different levels of assessment that can be targeted by a given method of clinical assessment. Miller's assessment pyramid[37] describes a simple hierarchy for the development and assessment of clinical skills. The four levels of assessment are illustrated here with reference to surgical skill assessment methods (*Figure 1*). They describe fundamentally different assessment *constructs*, in terms of both the nature of learning they require and also the situational context

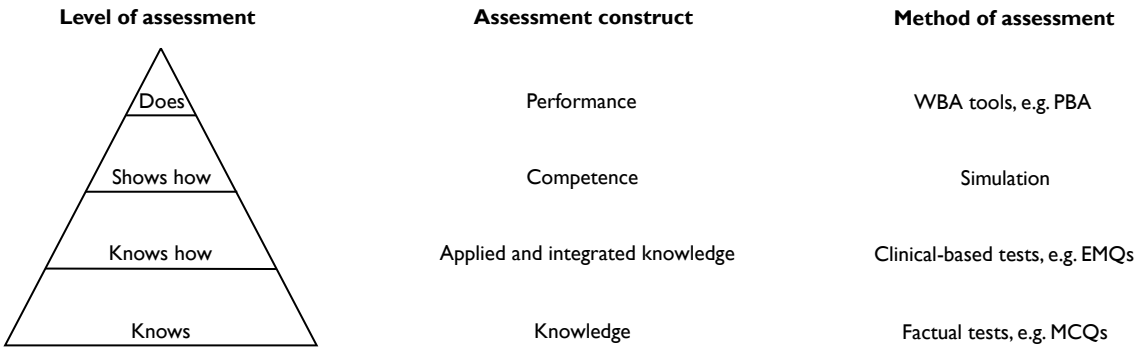


FIGURE 1 Relationship between Miller's pyramid and methods of assessment. EMQs, extended matching questions; MCQs, multiple choice questions.

of the learning. The strength of Miller's model has been for guiding curricular design and the selection of assessment methods to target the appropriate and intended levels of assessment.

The lower two levels of Miller's pyramid describe the cognitive domains of knowledge ('knows') and integrated and applied knowledge ('knows how'). In terms of surgical skills assessment, the first level could encompass the simple recall of anatomy and physiology facts, with a suitable method of assessment being a factual test of knowledge. The second level could relate to applied anatomy and physiology, and this would be appropriately targeted by clinically based tests (e.g. problem-based scenarios, extended matching questions) to assess the deeper nature of applied knowledge. These levels of assessment are mostly addressed within both undergraduate and postgraduate assessment systems using written examination methods. The top two levels of the pyramid are the behavioural domains that Miller termed 'shows how' and 'does', distinguishing competence from performance. Competence can be defined as 'what a person does in a controlled representation of professional practice' whereas performance is 'what a person does in actual professional practice' within the workplace.[38] Both of these assessment constructs are addressed within postgraduate assessment programmes using a variety of assessment methods. Improvements in competence and performance usually come from experience (practice) combined with constructive feedback,[39] with feedback aimed at providing 'an informed, non-evaluative, objective appraisal that is intended to improve clinical skills'.[40]

A complex relationship exists between competence and performance. Competence cannot necessarily predict performance, and this has been demonstrated in a number of medical education contexts using a variety of assessment methods.[41–43] Therefore, the use of CBAs is inappropriate for the assessment of performance.

The Cambridge model described by Rethans *et al.*[38] is an extension of Miller's pyramid. It conceptualises performance as a window to competence, illustrating that competence is a prerequisite for performance with several additional factors that influence the normal day-to-day performance of doctors (*Figure 2*). These were classified in the model as *individual-related influences* (e.g. physical and mental health of the doctor, state of mind at the time of assessment, relationships with peers) and *system-related influences* (e.g. time pressures, guidelines, facilities). In the case of surgical skill assessment there could equally be added *case-related influences* (e.g. case complexity or type of procedure).

Workplace-based assessment aims to target the level of performance, assessing most closely the actual behaviour of doctors in the workplace. Individual-, system- and case-related influences cannot be fully controlled for in the context of WBA. However, WBA moves beyond

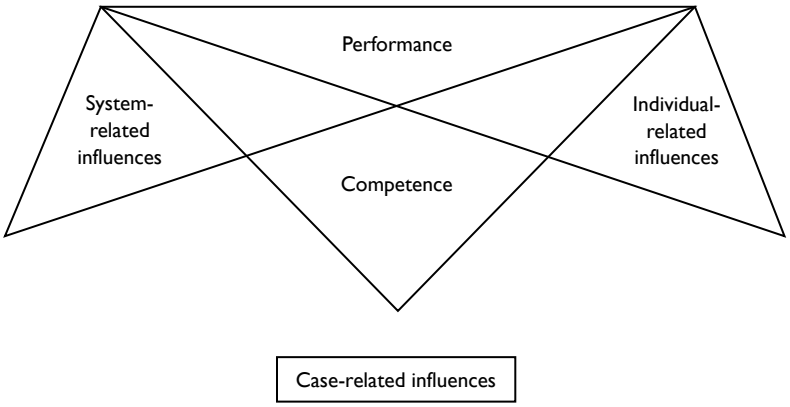


FIGURE 2 The Cambridge model of performance.

the theoretical construct of assessing professional practice under 'controlled' conditions (competence) and seeks to assess authentic performance in the workplace. In practice, the distinction between competence assessment and performance assessment is becoming less clear cut as practising doctors are subject to continuous assessment in the workplace using methods such as multisource feedback and analysis of patient records/letters. One important consideration when using WBA is that case-related and judge-related effects are more dominant influences than for other performance-based assessments.

Performance-based assessment

Performance-based assessment is concerned with assessing complex, 'higher order' knowledge and skills in the workplace context in which they are used, generally with open-ended tasks that require substantial assessor time to complete.[44] The adoption of WBA tools to provide structured performance-based assessment of surgical skills in the operating theatre is a relatively recent development within surgical training programmes. However, there is considerable experience in using performance-based assessment in other medical education contexts, including covert standardised patients[45] and peer assessment.[43] The lessons from these experiences[44,46] are summarised here to highlight the particular considerations of performance-based assessments:

1. The behaviour and performance of a doctor are highly dependent on the nature of the problem or task undertaken. The level of performance achieved for one problem or task is not a good predictor for subsequent ones – a finding termed *case specificity*.[47,48] Therefore, performance-based assessments need to sample widely and consider both context (situation/task) and construct (knowledge/skill/attitude), as complex interactions exist between these dimensions. For example, to adequately assess surgical performance would require assessment of different operations that demand different decision-making and technical skills.
2. Assessors make subjective judgements even when using assessment methods that include clear and objective descriptors as the performance criteria. It is vital to ensure that assessment criteria are rigorously developed and well understood by assessors through training. Performance-based assessments should also draw upon the judgements of as many assessors as feasible in order to limit the impact of assessor bias on assessment scores.[34] For example, an assessment system that adopts WBA needs to stipulate that sufficient numbers of assessors are involved in assessing the performance of an individual doctor.

3. The most complex aspects of performance (e.g. judgement and decision-making) are the most difficult of all for which to devise observable and/or meaningful criteria that will allow adequate assessment. However, they remain a vital part of assessing overall performance. It is important that performance-based assessments seek to assess broadly, rather than focusing assessment on the easy-to-assess aspects. Global rating scales have the ability to assess these more complex dimensions of performance.[49]
4. The use of systematic methods to select assessment content (e.g. task analysis and Delphi processes using specialty experts) is a good approach for designing valid assessment methods. However, it may not be possible or feasible to develop assessment methods to include all of the validated content.
5. Performance is a unified assessment construct and it is difficult to identify obvious planes of cleavage. There have been systematic efforts directed towards defining components of performance,[50] and in developing checklists to allow assessment of specific competencies there is evidence that expert judgement using global rating scales provides a superior assessment of performance. Comparing the use of detailed checklist scales with global rating scales in a variety of contexts, assessors produce more reliable assessments using global scales.[51,52] There is also evidence that trainees find checklists helpful both for training and for informing feedback.[53] Therefore, the roles of checklists and global ratings are complementary within performance-based assessment.
6. Assessing performance requires the triangulation of assessment methods to facilitate an overall judgement. There is no one assessment method that can sample across all relevant contexts and constructs or that should be relied upon singularly for the assessment of performance.[54] This is the approach that has been adopted by the General Medical Council's (GMC's) Performance Procedure, for which several assessment methods have been selected to assess poorly performing doctors.[55] These target levels of both competence and performance in the light of their complex and non-linear relationship. It is also the approach used within the summative annual review process for doctors in training (termed ARCP), which judges the suitability of each trainee to progress or complete training. The ARCP panel uses evidence from multisource feedback – WBA, educational supervisor reports and examinations – to judge if the trainee's performance is satisfactory or otherwise. The breadth of the assessment methods also allows for specific deficiencies in performance to be highlighted and addressed within the ARCP, for example there may be specific concerns about a trainee's operative skill and a collection of WBAs over the year of training will fail to show progression in training.

These unique considerations need to be fully appreciated when designing and implementing performance-based assessment for both research and training purposes. Our research methodology, outlined in *Chapter 2*, is aligned to these considerations for researching performance-based assessment methods.

Designing and evaluating assessment methods

All assessments methods need to balance rigour (reliability and validity) against practicality (feasibility, cost and acceptability).[34] van der Vleuten's utility index[56] offers a useful conceptual framework for assessment design and evaluation. Ensuring that our assessment methods are valid, reliable, acceptable and cost-effective and have educational impact are the principal considerations of a well-designed and -evaluated assessment system. The ideal assessment method would possess all these essential measurement characteristics. However, the choice of assessment methods adopted within any training programme should be determined by balancing these conflicting considerations to fit the purpose of the assessment (*Figure 3*). For example,

$$\text{Utility} = V_w \times R_w \times A_w \times E_w \times C_w$$

$$\text{Validity}_{\text{weighted}} \times \text{Reliability}_{\text{weighted}} \times \text{Acceptability}_{\text{weighted}} \times \text{Educational}_{\text{weighted}} \times \text{Cost}_{\text{weighted}}$$

FIGURE 3 The utility of assessment methods.[56]

a high-stakes examination will need to weight reliability and validity, whereas an in-training assessment may focus upon educational impact. All assessment methods demand acceptability.

Evidence of validity and reliability are essential characteristics of fair and defensible assessments, particularly in identifying underperforming doctors who could compromise patient safety.[48] The PMETB standards of assessment, set in 2004[57] and recently updated,[5] have stipulated that postgraduate training programmes provide reliability and validity evidence for their assessment methods. By design, the focus of interest for UK training programmes has been in ensuring that their assessment methods are defensible. However, there is a more recent consensus of opinion that successful implementation of WBA demands a redress of acceptability and feasibility issues.[27]

The psychometrics of assessment

Validity and reliability are metric properties, termed psychometrics when they concern the quantitative measurement of psychological variables such as behaviour and cognition. Psychometrics, originally a branch of psychology, has developed an increasingly important place within the health sciences for measuring complex constructs such as quality of life and the assessment of clinical performance. This report is primarily focused on evaluating the assessment of surgical performance. The validity and reliability of different assessment methods are carefully measured. We will briefly outline the conceptual basis of these measurements, relating them closely to the context of performance assessment.

Validity is the extent to which a result reflects the construct it intends to measure and not something else.[58] To establish the validity of assessments intended to measure surgical performance requires evidence to support whether or not the assessment actually measures surgical performance. There are many sources of validity evidence that can be drawn upon in making judgements of assessment validity to allow a meaningful interpretation of assessment scores.[59] As a bare minimum, a valid assessment should appear to be measuring what is intended (*face validity*) and must include the relevant performance criteria and elements of the skill or behaviour being tested (*content validity*). There should also be agreement with other assessments intended to measure the same construct (*criterion validity*). For example, tool A and tool B both assess non-technical surgical skills and, despite being different assessment methods that are completed separately, could be expected to agree as they measure the same construct of surgical performance. *Construct validity* concerns the extent to which assessment scores correspond to an assessment *construct*, which could be an attribute, ability or skill, as predicted by some rationale or theory. It is measured by testing hypotheses about the construct of interest (for example, surgical skill) and evaluating whether these are confirmed or refuted by the assessment scores. A surgeon's surgical skill could be predicted to improve with years of training and previous surgical experience. If one correctly hypothesises that more senior and experienced surgeons will obtain higher scores for surgical skill (because theory shows that the acquisition of surgical expertise requires many hours of surgical experience), the assessment may have construct validity. Construct validity should be demonstrated by an accumulation of evidence. There will be more

confidence in the assessment score when more strategies are used to demonstrate its construct validity, provided the evidence is convincing.

Predictive and consequential validity both provide more stable evidence of the validity of assessments. *Predictive (outcome) validity* is the extent to which an assessment score predicts expected scores on some criterion measure. For example, the validity of an assessment of surgical performance would be demonstrated by a significant correlation of scores with surgical outcomes or clinical supervisor performance ratings. Large sample sizes and follow-up data collection are required to demonstrate predictive validity. *Consequential validity* considers the educational impact of assessments, in particular what trainees learn and how they learn it, to evaluate whether assessments encourage good or poor learning behaviours. For example, WBAs of surgical skill are designed to aid learning through supervised operating and feedback from clinical supervisors, but if they failed to support trainees to gain operating experience and develop their surgical skills, their consequential validity would be judged as low.

Reliability is the extent to which an assessment score reflects all possible measurements of the same construct.[58] It reflects the reproducibility of assessment scores. A reliable test should give the same result if repeated or if a different assessor is used.[60] An evaluation of reliability should involve a clear statement of the circumstances that the results are meant to represent.[34] For example, this particular assessment has demonstrated good reliability for assessing the surgical skill of all levels of vascular specialist registrars, using no fewer than four assessors on four cases, within a typical teaching hospital context.

To understand how big an impact the case, the judge and other important contextual factors have on the assessment score requires that all these sources of error (termed variability) are quantified. Estimating reliability then depends on comparing the effect of assessor-to-assessor and case-to-case variability in scores with overall trainee-to-trainee variability in scores. Assessor and case variability represent the greatest threats to the reliability of WBA.[61] Generalisability theory provides the most robust and meaningful reliability estimates by simultaneously examining assessment scores for these different types of variability (see *Glossary* for a detailed description). Generalisability theory is the statistical approach used for the analysis of assessment scores within this study.

The observation of real-time performance in the workplace is essential for achieving the most authentic assessment, as this method approximates to the 'real world' as closely as possible. The more frequently WBA is integrated into routine practice, the better the validity of the assessment.[27] Trainers and trainees are being encouraged to use every clinical encounter and surgical case as an opportunity for WBA, moving away from the mini-exam mentality that is produced by infrequent assessment. WBA frequently lacks evidence of reliability because of the difficulties involved in producing accurate estimates of reliability, the methods themselves are new, and the evaluations have yet to be done. The main threats to reliability include case specificity, variations in case complexity, assessor subjectivity, differences in rating scales and methods (see *Performance-based assessment* above). The PMETB has acknowledged the difficulties involved in demonstrating reliability for WBA and encouraged the use of van der Vleuten's utility index[56] (see *Figure 3*) and the triangulation of evidence from different assessment methods.[62]

Approaches to assessing surgical skills

Our intention is to provide an overview for some of the approaches to surgical skill assessment to illustrate the transition to more performance-focused assessment methods. This overview

includes formalised assessment processes, whether simulation-based or in the workplace, as well as surrogate measures of surgical skill. The rigour (reliability and validity) of an assessment method is an essential component of its overall utility, which we draw on as the focus for discussing the relative advantages and disadvantages of the various approaches for surgical skill assessment. The literature search for this section was updated in July 2009.

Speed, quality of product and patient outcomes

These three outcomes, while not formalised assessment processes, have been explored for use as surrogate measures of surgical skill.

Operative speed may provide a measure of technical skill, although there is a paucity of literature on this subject. Robert Liston, a London surgeon working before the introduction of anaesthesia, was proud of his operative speed and once challenged observers, 'Now gentlemen, time me, 28 seconds before placing an amputated limb in the sawdust'.[63] However, measuring competence merely by setting time targets for a certain procedure is crude and probably unacceptable, as a fast surgeon is not necessarily a good one. Lord Lister, in comparison, was observed to have 'none of the dramatic dash and haste of the surgeon of previous times. He proceeded calmly, deliberately and carefully and, as he told his students, anaesthetics have abolished the need for operative speed and they allow time for careful procedure'.[63]

The Toronto group undertook a small study to examine whether time and operative product could serve as measures of technical skill using bench model simulations.[64] Twenty general surgery residents participated in a six-station bench model examination, in which 'time to completion' was recorded and 'quality of the final product' was assessed using a global five-point rating scale by two assessors per station. The mean inter-rater reliability was 0.59 for product quality. Interstation reliability (Cronbach's α) was 0.59 for analysis of product quality and 0.72 for time to completion. Both measures demonstrated construct validity, with more senior trainees producing better products in less time, although there was poor agreement with previous OSATS examination scores. These measures offer a time- and cost-efficient, if less reliable, alternative to direct observation for the assessment of technical skill. However, the quality of the final product may be relatively easy to measure on a simulation, whereas it may be more difficult to assess in the operating theatre and is likely to depend on the specific procedure. Similarly, time to completion is a straightforward measure for simulations but in the operating theatre is affected by many case-specific variables as well as the performance of the whole surgical team.

Expert surgeons appreciate that operative speed is important. For instance, surgeons performing cardiac and vascular procedures seek to minimise cardiac bypass and vessel cross-clamping time to reduce operative complications. Some evidence with respect to time and surgical performance comes from the Medical Research Council (MRC) European Carotid Surgery trial, which reported a relationship between procedure time and adverse outcomes.[65]

Efficiency in operating time is important to maximise service delivery, which is increasingly pressured under the current policies for patient management pathways and waiting list times. However, such efficiency requires large numbers of procedures, which trainees could not usually expect to obtain, and is therefore a more appropriate aspiration for newly appointed consultants.

Although attractive, measurement of the performance of surgeons based upon patient outcomes is fraught with difficulty owing to variation in case mix and the large numbers required for reliability.[66] Errors made by trainees are often corrected, and therefore masked, by their supervising consultant. In addition, patient outcomes reflect the performance of the whole surgical team, both within the operating theatre and during the postoperative period, and therefore do not provide a reliable assessment of an individual surgeon. It may be a good

screening method for consultant surgical skill, but tests of competence will be required for those consultants for whom there is cause for concern.

OpComp

Logbooks form a useful record of procedural experience[67] but do not reflect the performance level achieved by trainees, lacking both validity and reliability as a method of assessing trainees' surgical skill.[68] The Operative Competency (OpComp) form was introduced by the Specialty Advisory Committee (SAC) in General Surgery in 2003 to complement the information provided by logbooks on trainees' surgical skills.[69] The OpComp form asked educational supervisors to assess the ability of a trainee to perform the specific index procedures (relevant to their specialty) against defined criteria at the end of a clinical placement. The following rating scale was used to make these summary judgements, derived from the then current 'Training the Trainers' Course[70] and modified on the basis of the pilot studies:[69]

U = Unknown or insufficient evidence to support a judgement.

D = Unable to perform the procedure, or part observed, under supervision.

C = Able to perform the procedure under supervision.

B = Able to perform the procedure with minimum supervision but needs occasional help.

A = Competent to perform the procedure unsupervised and can deal with most complications.

A checklist of technical skills was provided on the reverse of the OpComp form to assist educational supervisors in making judgements. This checklist's content was systematically derived using a Delphi survey of surgeons in Scotland to establish the 'essential' technical skills required of trainees.[71] The OpComp form was shown to have good construct validity, and trainers found it simple to complete.[68] Although an advance, there were aspects of this method that undermined its reliability. The form suffers from retrospective recall of numerous procedures over the course of a placement, risking loss of important training information and cross-contamination between procedures. Almost half of the trainees surveyed during its pilot indicated that a trainer had rated their ability to perform a procedure unseen.[68] In addition, the reliance on a single assessor per clinical placement opens up this method to various types of assessor bias, including the halo effect, whereby an assessor provides a final opinion rather than providing a discriminatory rating for each item,[72] and expectation bias, in which the knowledge of a trainee's seniority influences assessors' ratings,[73] as well as bias arising from the primary influence of a trainee's interpersonal skills, rather than their technical skills, on supervisor ratings.[74] The use of a collection of surgical skill assessments at the end of a clinical placement did not promote the opportunity for 'on-the-job' training and feedback. It can be appreciated that the move to current WBAs, using immediate assessment and feedback after single procedures/operations using multiple assessors during a clinical placement, moves the validity and reliability of assessment methodology forward, beyond that offered by OpComp.

Simulators

The controlled environment of a skills centre permits reliable assessment of technical skills on simulations such as bench models, animals and computer models. However, their validity depends upon the fidelity of the simulation,[75] with high-fidelity simulations possessing a high level of realism compared with a living human patient. Characteristics of high-fidelity simulations include visual and tactile cues, feedback capabilities and interaction with the trainee, with the opportunity for trainees to complete surgical procedures rather than isolated tasks.

Although low-fidelity bench models and video box methods show less realism, these are often used to assess trainees because of lower cost, portability and the potential for repetitive use.

Parallel assessments of surgical skill, using live animals and bench simulations, have been compared by the Toronto group.[52] Performance was graded for both assessment formats using task-specific checklists, global ratings and pass/fail judgements (i.e. OSATS). Using 20 surgical residents, the correlations between live and bench scores were high (0.69–0.72), and the mean inter-rater reliability across stations ranged from 0.64 to 0.72. This study showed that using simulations with OSATS is a valid and reliable method of assessing surgical skill, with bench models giving equivalent results to live animal model simulations. Lentz *et al.*[76] repeated this comparison of bench and live animal assessments within a surgical laboratory curriculum and found that skills improved over time for individual trainees and as a cohort by year of training, suggesting that simulations can also be successfully used to assess progress during training.

Whenever possible, the assessment method used in the workplace should be used for the relevant simulation because this will aid validity and transferability.[77] Studies of the transferability of simulation-based assessments to actual operating theatre performance are of fundamental importance in establishing their validity as an assessment method. Several authors have demonstrated that assessments of technical skill on low-fidelity simulations predict performance in the operating theatre.[52,78–80] It does appear that complete procedures can be deconstructed into tasks suitable for trainees to rehearse and be assessed for competence, before moving on to surgical training in patients.

The more recent development of high-fidelity simulations for surgical assessment offers the potential to reduce assessor time while providing automatic 'objective' output data for feedback and assessment purposes. Output metrics for trainees performing virtual simulations include economy of motion, length of path movements and instrument errors.[81] Several studies have shown that both the minimally invasive surgical trainer–virtual reality (MIST-VR) (Mentice, Gothenburg, Sweden) and LapSim® Gyn VR systems (Surgical Science, Gothenburg, Sweden) are valid and reliable methods of assessing psychomotor skills for various laparoscopic procedures.[81–85] Furthermore, randomised controlled studies have demonstrated the benefits of high-fidelity simulation training for performing laparoscopic cholecystectomy,[86,87] endoscopy[88] and catheter-based interventions.[89]

The main advantage of simulations is that they allow unlimited practice for trainees to learn a new procedure before moving to patients, thus providing a new opportunity to learn safely from mistakes.[90] Assessment using simulations provides access to surgical skill training and assessment in situations where training in patients is unavailable or very limited[91] and/or too high risk.[92] Simulations can also be used to assess the performance of individuals within emergency teams[93] and surgical teams.[94] It appears that simulators can take on a well-defined assessment role that complements the assessment of real-time operating theatre performance. However, practice and assessment on simulators are no substitute for real operating experience, although they offer trainees the opportunity to progress their surgical skills before training in the complex operating theatre environment.

Direct observation in the operating theatre

It seems axiomatic that direct observation of surgical performance in the operating theatre represents the 'gold standard' in terms of both content and construct validity. However, unstructured directly observed assessments suffer from halo error[95] and other types of assessor bias.[73] In fact, the main issue with unstructured assessments using expert/consultant assessors is that the direct observation itself is not wholly successful, i.e. assessments are completed based on indirect observation or incomplete direct observation, which limits its reliability as an assessment method.[96]

Standardising direct observation depends upon the use of a structured assessment, such as OSATS or PBA. Although the validity and reliability of structured direct observation has been comprehensively researched for simulations, the evidence for assessing surgical performance in the operating theatre is limited. This study seeks to directly fill this gap in the research.

There are two main types of rating scales used for structuring assessment of surgical skill: task analysis checklists and global ratings. These rating scales are compared in *Table 2*. Dual assessments using separate task analysis and global ratings may be time consuming to perform, but each method may have different roles.[53] Global ratings seem useful when assessing more complex operations, especially when there is more than one method of performing the task correctly, or when assessing experts for the purposes of certification or revalidation. Task analysis checklists provide a trainee with detailed instructions and feedback of how to undertake the operation in an approved way. The PBA and OSATS tools use both types of rating scales, in combination and separately respectively.

Assessment tools used in this study

There is a plethora of assessment tools that have been developed worldwide to assess surgical performance. The background presented here is limited to the assessment tools considered within this study, all of which are used to directly observe and assess dimensions of surgical performance. These WBA tools are not research-only tools but are either directly or indirectly relevant to current UK surgical training practice.

Two of the tools (PBA and OSATS) are primarily concerned with the assessment of technical skill, although they both include some non-technical skills. Both tools are in current use in UK postgraduate training programmes. These conform to the assessment principles laid down by the PMETB in 2005 and are designed to measure all the domains of *Good Medical Practice*.[50] PBAs were introduced for orthopaedic trainees by the OCAP in 2005 and for all other surgical trainees by the ISCP in 2007. The OSATS tool has been adopted as the method of assessing technical skills within O&G and ophthalmology specialties since 2007.

One concern about PBA, OSATS and other similar assessments is that they may not reflect 'higher-order' skills that underpin technical proficiency, such as situation awareness, decision-making, team-working and leadership. The Non-technical Skills for Surgeons (NOTSS) tool is designed to assess and debrief trainee surgeons on their non-technical skills, although it is not currently used within UK training programmes.

The formal training of surgeons predominantly focuses on developing knowledge, clinical expertise and technical skills, with the focus of assessment on the observation of technical skills and surgical performance. Non-technical skills have been defined as the critical cognitive (e.g. decision-making) and interpersonal (e.g. teamwork) skills that complement surgeons' technical skills.[97] There is increasing recognition of the need for explicit training and assessment in non-technical skills because of the importance of these skills for patient safety. Case reviews and

TABLE 2 Task analysis checklists vs global ratings

Task analysis	Global ratings
Procedure specific	Items common to any procedure (e.g. handling of instruments)
Checklist of steps that represent one safe way to perform a procedure	Useful in assessing complex operations
Good for assessing trainees and for feedback	Good for assessing experts as there is no 'right' way

studies of operating theatre behaviour have consistently shown that failures in non-technical skills are implicated in surgical adverse events and errors,[98,99] Therefore, technical skills appear to be a prerequisite but are insufficient to ensure patient safety in the operating theatre. Fostering non-technical skills within training is likely to support surgeons in maintaining high levels of performance over time.

Examples of the three assessment tools are found in *Appendices 1–3*. The full guidance notes for each assessment tool are available on the relevant websites: www.iscp.ac.uk for PBA, www.rcog.org.uk for OSATS and www.abdn.ac.uk/iprc/notss/ for NOTSS.

Procedure-based assessment

Procedure-based assessment is a method for assessing surgical skills in the operating theatre during interventional procedures. It is designed to be used in conjunction with the surgical logbook for all the index procedures for a particular surgical specialty. PBA was originally developed by the OCAP for trauma and orthopaedic surgery[100] and PBAs have now been written for all other surgical specialties by the relevant specialty associations and SACs within the ISCP. Trainees already in training when PBA was introduced have been encouraged to use it, both as an aid to learning and to complement logbook experience, but the use of PBA has been made compulsory only for those entering surgical training since 2005 for orthopaedics and 2007 for the other surgical specialties. PBA has not yet been adopted by surgical training organisations outside the UK.

The assessment form itself has two principal parts. The first consists of a series of competencies within six core domains covering 'consent', 'preoperative planning', 'preoperative preparation', 'exposure and closure', 'intraoperative technique' and 'postoperative management'. The consent and preoperative planning domains address perioperative competencies, whereas the remaining domains encompass intraoperative competencies. It is not expected that all PBA domains will be completed at any one time as consent and preoperative planning are often undertaken at a different time and place. While many of the competencies are common to all procedures (global items), others are specific to the particular procedure (task specific), particularly within the intraoperative technique domain. Each competency is assessed as satisfactory (S), unsatisfactory/development required (U/D) or not assessed (N). The assessment form is supported by a worksheet, originally used as part of the validation process, which gives examples of desirable and undesirable behaviours for each competency. Therefore, the first part of the PBA uses a combination of task and global items which are rated with a single binary rating scale. The second part of the assessment form consists of a four-level summary judgement in which the assessor rates the ability of the trainee to perform the observed elements of the procedure on that occasion with or without supervision (see *Appendix 2*). It uses similar levels to those used on the OpComp form. The content and construct validity of PBAs has been validated for index procedures in both general and orthopaedic surgery.[52,100]

It is assumed that the assessor (clinical supervisor) will normally be scrubbed and supervising the trainee. Trainees carry out the procedure, or part of it, explaining what they intend to do throughout. The assessor will provide verbal prompts to remind the trainee to make explanations, if required, and will intervene if patient safety is at risk or the quality of treatment may be compromised. The form has been designed to allow the assessor to score items at the end of the procedure by using a simple binary rating scale. In addition, the completed form is intended to structure the provision of immediate constructive feedback (e.g. in the coffee room between cases). A PBA may be undertaken every time an index procedure is undertaken, as the primary aim is to aid learning.

The satisfactory standard for each competency is the level required for the CCT. At the end of a placement, a collection of PBAs, together with the logbook, will enable the educational supervisor or programme director to make a summary judgement about the competence of a trainee to perform an index procedure to the required standard.

Objective Structured Assessment of Technical Skills

The OSATS system was introduced by the RCOG (www.rcog.org.uk/education-and-exams/curriculum) as a formal requirement of their new training and education programme for all grades of trainees since August 2007. Prior to this, OSATS was used informally within training, particularly for junior O&G trainees.

The development of OSATS by Reznick *et al.*[52,101] at the University of Toronto in the 1990s initiated the current trend towards structured observational assessment. OSATS was originally developed for use on bench model simulations and was designed to be completed in real time by assessors, as with objective structured clinical examinations, rather than at the end of the procedure. OSATS has been shown to possess good inter-rater reliability and construct validity for assessing general surgical trainees performing common operations on both cadaver and live animal simulations.[52] Goff *et al.*[102] have also demonstrated that OSATS possesses construct validity and good inter-rater reliability for blinded and unblinded assessment of O&G trainees performing common procedures on lifelike models in a multiple station exam format. Direct observation or videoing of real surgical procedures in the operating theatre using structured checklists based on OSATS can demonstrate high inter-rater reliability and construct validity for simple operations such as varicose veins surgery.[53,103]

The original OSATS was developed by Winckel *et al.*[101] and consisted of a technical checklist (rated on a numerical scale using 0 for 'not performed', 1 for 'performed poorly' and 2 for 'performed well') that was specific for the procedure. The second part was a generic assessment of 10 global items (using a five-point numerical rating scale from 0 'poorly, or never' to 4 'excellent or always') that were common to all procedures. The global rating assessment was modified by Martin *et al.*,[52] reducing the number of global items to seven, changing the five-point numerical rating scale to a behaviourally anchored scale using descriptors (e.g. 'makes many unnecessary moves' to 'fluid moves with instruments and no awkwardness' for global item time and motion) and including a pass/fail judgement at the end of the global assessment. Therefore, the OSATS form uses separate task and global assessments which are rated using two different rating scales, with a summary pass/fail judgement.

The OSATS adopted by the RCOG for the assessment of 10 O&G index procedures uses a similar form to the modified Martin *et al.*'s OSATS version.[52] Part 1 is a technical checklist in which task items, specific to the index procedure, are rated as 'done independently' or 'needs help'. Part 2 provides the generic assessment, which has been reduced to a three-point behaviourally anchored rating scale of seven global items (see *Appendix 1*). These seven global items have been further modified by the RCOG to combine some global items (e.g. 'instrument handling' and 'knowledge of instruments' have been combined as 'knowledge and handling of instruments'), whereas other global items are new additions (e.g. 'suturing and knotting skills', 'relations with patient and the surgical team', 'insight/attitude' and 'documentation of procedures'). This is combined with a summary pass/fail judgement (this version was in use at the time of this study). However, the pass/fail terminology has since been revised to 'competent in all areas included in this OSATS' and 'working towards competence' to enforce the formative nature of the OSATS.[104] For OSATS to count as a pass, an assessment algorithm is used: all items on the technical checklist must be ticked as 'performed independently', and the majority of global items ringed in the middle to

right of the rating scale, while insight/attitude must be consistently ringed 'fully understands areas of weakness'.[105] Trainees are required to achieve a set number of passes for each index procedure in order to 'sign off' their logbook competency for that procedure. Currently, the requirement is for three completed OSATSs by at least two trainers in order for the relevant logbook competency to be signed off as competent for independent practice. Validity and reliability studies have not been undertaken for the application of this revised tool for assessing trainees in the operating theatre environment.

From our literature review and from email correspondence with surgical training organisations in North America, Continental Europe and Australasia, it appears that OSATS is not routinely used within any surgical curriculum in the workplace.

Non-technical Skills for Surgeons

The NOTSS system is a behavioural rating system developed using a multidisciplinary group of surgeons, psychologists and an anaesthetist from the Royal College of Surgeons of Edinburgh, in collaboration with the University of Aberdeen.[106] The development of NOTSS built on work from a similar project that developed a behaviour rating system for anaesthetists called Anaesthetist's Non-technical Skills (ANTS).[107] While ANTS is not used formally in UK anaesthetic training, it has been piloted for use within the Australian and New Zealand College of Anaesthetists training programme. NOTSS is not currently part of the curricula for surgeons in training in the UK. However, there are a number of ongoing process trials worldwide that are considering the adoption of NOTSS. Furthermore, the Royal Australian College of Surgeons has adapted and expanded NOTSS to establish a surgical performance framework for assessment purposes.[108]

Behavioural rating systems are already used to structure training and evaluation of non-technical skills in anaesthesia, civil aviation and nuclear power, to improve safety and efficiency. They are rating scales based on skills taxonomies, with examples of good and poor behavioural markers, and are used to identify observable, non-technical behaviours that contribute to superior or substandard performance. Behavioural rating systems are context specific and should be developed in the domain in which they are to be used.

Non-technical Skills for Surgeons describes the main observable non-technical skills associated with good surgical practice. It has been designed to provide surgeons with explicit ratings and feedback on their non-technical skills, either within the operating theatre or operating theatre simulator. The system comprises only behaviours that are directly observable or can be inferred through communication during the intraoperative ('gloves on, scrubbed up') phase of surgery.

The NOTSS system comprises a three-level hierarchy consisting of categories (at the highest level), elements, and behaviours. Four skill categories and 12 elements make up the skills taxonomy (see *Appendix 3*). Each category and element is defined with examples of good and poor behaviours for each element. These exemplar behaviours were generated by consultant surgeons and are intended to be indicative rather than comprehensive. The aim is to provide a common terminology and assessment framework for surgical trainees and consultants to structure their training needs, in order to develop their non-technical skills in the workplace.[106]

The development and design of the NOTSS system has been a rigorous and structured process, from initial task analysis through to system evaluation. For more details on NOTSS development see Yule *et al.*[109] The psychometric evaluation of NOTSS to date has been carried out by the NOTSS development team in the operating theatre simulator.[110] Six video scenarios were designed and filmed by practising surgeons, anaesthetists and nurses with experience

in non-technical skills, to illustrate a range of surgeons' non-technical skills. A group of 44 consultant surgeons, trained in the use of NOTSS, observed and rated these videos using NOTSS assessment forms. Their ratings were compared with expert ratings for 'accuracy' and assessed for inter-rater reliability. In this study, the NOTSS system demonstrated a consistent internal structure, and internal reliability was high for all four categories (overall mean difference of 0.25 scale points between categories and elements). The sensitivity ('accuracy') of the system was moderate (mean sensitivity across all categories was 0.67 scale points difference from expert ratings) with the 'decision-making' category most sensitive and the 'situation awareness' category least sensitive. The inter-rater reliability (mean within-group agreement of ratings across NOTSS categories) was acceptable (> 0.7) for the categories 'communication and 'teamwork' and 'leadership', although below acceptable for categories 'situation awareness' and 'decision-making'.

The NOTSS development team has analysed the level of agreement between ratings of expert versus novice raters in more detail.[111] The mode ratings of NOTSS category ratings for each video scenario showed 50% agreement. Where there was disagreement, novice raters were more likely to give harsher ratings, with the widest differences between rater groups within 'communication and teamwork' and 'leadership' domains. Of note, 23% of 'situation awareness' ratings were scored not applicable by novice raters, compared with 0% of experts, indicating that experts in non-technical skills are more equipped to rate these behaviours. This highlights the role of training in human factors training and the rehearsal of non-technical skill ratings for NOTSS assessors.

The NOTSS development team has also considered the usability of the NOTSS system[112] within the operating theatre for observing common surgical procedures. Questionnaire responses from consultant surgeons participating in the usability trial have indicated that NOTSS provides a structure and language to rate and provide trainee surgeons with feedback on their non-technical behaviours. The main concerns with using the NOTSS system were reported as: difficulty in understanding the behavioural descriptors; difficulty in rating behaviours when trainees do not verbalise adequately; assessments being more suitable for senior than for junior surgical trainees; the use of routine cases raising insufficient decisions for rating behaviours in the decision-making category; and the 'negative' impact of a scrubbed consultant upon trainee-led leadership. These user-satisfaction responses from NOTTS assessors provide great insight into the challenges of rating behaviours in the operating theatre.

This study moves forward the psychometric evaluation of the NOTSS system into the operating theatre environment. The tool will be used for observing and rating surgeons' non-technical skills in vivo. In addition, this study provides further work on the usability of the NOTSS system in the operating theatre.

Working with the pace of change

Workplace-based assessment is a relatively recent development within postgraduate assessment and therefore this research field is fast evolving. Since this study began in April 2007, there have been significant changes to the policy of WBA implementation and also the use of assessment tools, particularly with respect to RCOG. We will fully elaborate upon these more recent changes in *Chapter 4*. It was envisaged that WBA would evolve during the timescale of this study. For this reason, our study design was not solely focused on the evaluation of validity and reliability but also the wider issues of acceptability and feasibility.

Aims of the study

The primary aims of this study were to compare the user satisfaction and acceptability, and reliability and validity, of three different methods of assessing the surgical skills of trainees by direct observation in the operating theatre across a range of different surgical specialties and procedures. 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 O&G, upper gastrointestinal (GI) surgery, colorectal surgery, cardiac surgery, vascular surgery and orthopaedic surgery. Two to four index procedures were chosen in each specialty.

Information on user satisfaction and acceptability of each assessment method from both assessor and trainee perspectives were obtained from structured questionnaires. The reliability of each method was measured using generalisability theory. Aspects of validity included the internal structure of the structured tools and correlation between tools, construct validity, predictive validity, interprocedural differences, the effect of assessor designation and the effect of assessment on performance. User satisfaction/acceptability, reliability and validity are all important because they equally affect the utility of an assessment method.

A secondary aim was to study the feasibility and fidelity of video recording in the operating theatre with a view to evaluating the reliability of subsequent blinded assessment.

We anticipate that the information provided by this study will be of value to the following organisations:

- the ISCP
- the OCAP
- the RCOG
- the Academy of Medical Royal Colleges
- the Academy of Medical Educators
- the Conference of Postgraduate Medical Deaneries
- the PMETB
- the GMC (Revalidation and Performance Procedures)
- the National Clinical Assessment Authority.

Chapter 2

Methods

Ethics

Ethical approval for the study was obtained from Trent Main Research Ethics Committee on 15 March 2007. The original study proposal can be found in *Appendix 6*. An ethical amendment was approved on 1 June 2007, primarily to include the specialty of O&G to enable evaluation of the OSATS tool within a surgical specialty in which it is used for training but also to include additional index procedures. In May 2009, when recruitment was complete, the ethics committee approved the collection of anonymised trainee data from the deanery regarding any formal or informal training concerns identified, including ARCP2 [formerly Record of In-Training Assessment (RITA) D] and ARCP3 (formerly RITA E) statements.

Participants

Participants were a large, heterogeneous group consisting of patients, surgical trainees, consultant surgeons, anaesthetists and theatre practitioners. Ethical approval had originally been obtained for the study to assess consultant surgeons as well as surgical trainees. However, once the study began in the clinical setting there was concern raised by some consultants that the study might constitute an attempt at 'revalidation by the back door'. It was essential that the consultants were supportive of the study as their assessments of trainees would provide the data necessary to answer the primary research question. The initial reason for inclusion of consultant assessments had been to provide a reference group against which to compare the trainee assessments. However, the study statistician advised that the scores given by consultants to their peers would not provide valid reference scores. The study was therefore confined to the assessment of trainees only from its outset. The aim was to concentrate assessments on trainees within specialty training (ST), i.e. specialist training (ST3–7), although there was no exclusion of those in core training (ST1–2) or doctors in non-training posts on the basis that all doctors require regular assessment. Hereafter in the report these participants are referred to collectively as trainees.

Setting

Although ethical approval was granted for the study to be undertaken as a multicentre study in Sheffield, Nottingham and Leeds, the study proceeded as a single-centre study. The multicentre approach was seen initially to offer the greatest potential for recruitment and to strengthen the generalisability of the results. However, there was only sufficient funding from the grant to employ a single study co-ordinator, who was based in Sheffield. Once recruitment began, it became apparent that it would not be feasible to undertake the study in more than one location without a co-ordinator in each centre. Our study statistician also recommended that the limited resources would be better dedicated to obtaining multiple assessments on each trainee participant rather than single assessments on many participants. These views were supported by the Steering Committee. Therefore we focused recruitment within a single city at three teaching hospitals: Royal Hallamshire Hospital, Northern General Hospital and Jessop Wing for Women.

Timescale and schedule

The study ran from April 2007 to June 2009. Initial funding and ethical approval was given until March 2009. However, a 3-month funded and ethically approved extension was granted in recognition of the impact that the Medical Training Application Service (MTAS) had had upon recruitment during the initial 3 months of the study. During this period, many clinical supervisors and trainees were preoccupied with specialty applications and selection.

A Gantt chart was used to provide a framework for the study schedule. It was amended to take into consideration the delayed start of recruitment (see *Appendix 4*). Recruitment took place from June 2007 to May 2009. Interim reports were provided to the Health Technology Assessment programme at 6, 12 and 18 months along the study's timescale. An interim analysis was provided to the Steering Committee at 12 months.

Study design and methodology

The design was a prospective, observational study within the operating theatres of the three teaching hospitals in Sheffield. Trainees were directly observed performing named index surgical procedures in six specialties.

The methodology of the assessments was direct observation of trainee's surgical performance (encompassing technical and non-technical skills) followed by the provision of structured assessment ratings by trained assessors according to the criteria, standards and rating scales of each individual tool. We considered the role of the assessors in this study to be *observer-as-participant,* part-way along the continuum from complete observer to complete participant.[113] This takes account of the context of performing assessments, in which we were not purely observers but part of a working surgical team. For example, clinical supervisors acted as scrubbed-in first assistants for the assessments.

Specialties and index procedures

The index procedures were selected by the research team in collaboration with each surgical specialty and after subsequent approval by the Steering Committee (*Table 3*).

The index procedures represented typical procedures for each specialty that were performed on a regular basis, allowing for frequent assessments over a spread of trainee grades. They were also chosen to reflect a range of procedural difficulty/complexity and the breadth of surgery in each specialty, e.g. open and laparoscopic procedures.

Assessment tools and questionnaires

Procedure-based assessment

Procedure-based assessments were available for all the non-O&G index procedures, having been written by the relevant SACs for the OCAP and ISCP. Specifically for the purpose of this study, PBAs were developed for the O&G index procedures. These were drafted by the research team by combining the generic PBA template from the ISCP with the relevant task-specific OSATS checklist. They were circulated for review and consensus of agreement by five O&G clinical supervisors, including the programme director. Additional task items were included from suggestions proposed by the clinical supervisors. The final PBAs were the result of three iterations before final consensus was achieved. These PBAs were then approved by the Chairperson of the Steering Committee.

TABLE 3 Index procedures selected for each specialty

Specialty	Index surgical procedure
Cardiac	Coronary artery bypass grafts
	Aortic valve replacement
Colorectal	Right hemicolectomy
	Anterior resection
Upper GI	Laparoscopic cholecystecomy
	Open inguinal hernia repair
Orthopaedic	Primary hip replacement
	Primary knee replacement
Vascular	Varicose vein surgery
	Aortic aneurysm repair
	Carotid endarterectomy
O&G	Elective caesarean section
	Diagnostic laparoscopy
	Evacuation of uterus
	Urgent caesarean section

Outwith this study, it was never intended that all sections of a PBA must be completed for any given assessment, although trainees must demonstrate competence in all domains of surgical practice using a collection of PBA assessments. The remit of this study concerns the assessment of surgical skills in the operating theatre. Therefore, PBA domains concerning consent (Part I) and preoperative planning (Part II) were excluded as they are completed in other hospital settings (wards, clinics) in advance of the operative case. Parts III–V of the PBA assessment form were completed, corresponding to the preoperative preparation, exposure and closure, and intraoperative technique domains of the PBA, all directly attending to the assessment of intraoperative skills. The postoperative management domain (Part VI) was excluded for logistical reasons, as the independent assessor required this time to co-ordinate the completion and collection of assessment forms and observe feedback.

The PBA was used to assess trainees in all specialties by the clinical supervisor and one or more independent assessor. Within non-O&G specialties a PBA was completed for every case. Within O&G, where the PBA and OSATS were under comparison, either a PBA or OSATS was completed.

Objective Structured Assessment of Technical Skills

These were available for all the O&G index procedures, having been adapted from Reznick's original tool[52,101] for use by the RCOG. OSATSs are not used as a surgical skills' assessment tool in any specialty other than O&G. Therefore, we used OSATS solely within this specialty. All parts of the OSATS form were completed. OSATSs were completed by the clinical supervisor and one or more independent assessors. As previously highlighted, it was used interchangeably with PBA to allow a comparison between both tools in O&G.

Non-technical Skills for Surgeons

Use of the NOTSS tool within the study was approved by the NOTSS development team from the University of Aberdeen,[106] while being freely available for use within surgical practice. All parts of the NOTSS form were used for observing and rating non-technical behaviours.

The NOTSS form was used to assess trainees in all specialties by one or more independent assessors, anaesthetists, scrub nurses and surgical care practitioners. This was in keeping with its intended design for use as a non-specialty-specific assessment tool. The intention was that NOTSS assessments were completed for every case.

User-satisfaction and acceptability questionnaires

The study questionnaires (see *Appendix 5*) were drafted by the research team following the review of published guidance on questionnaire design[114–116] and examples of education questionnaire-based papers.[117–119] The questionnaires were reviewed by the Chairperson of the Steering Committee and these suggestions were included in the final versions. The NOTSS team was consulted on the design of the NOTSS questionnaire, and some of the questions used were adapted with permission from their published questionnaire-based research.[112] The addition of the O&G specialty in June 2007 (following the ethical amendment) required additional questionnaire development to evaluate the OSATS tool, as well as question modifications because the PBA tool is not used for training within this specialty. The format of the O&G questionnaires was revised significantly to account for these O&G differences with advice from a lecturer in social sciences with an expertise in questionnaire design.

Sampling

The sampling aim was that at least two assessors would assess each surgical trainee undertaking each index procedure in their specialty on at least two occasions, equating to a minimum of eight assessments per trainee in those specialties with two index procedures and more assessments per trainee in those specialties with additional index procedures. 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.

This sampling grid guided the sampling plan but, during implementation, a pragmatic approach to recruitment had to be taken, depending on which surgical trainees and clinical supervisors were available for assessments of suitable index procedures.

Videos of cases were recorded where patient consent was given to do so and when it was practically possible to film the case. Two cameras were used to record an operating theatre view and an operative field view. During laparoscopic procedures, the operative field view was swapped to the laparoscope camera view. The two images were recorded screen-in-screen on to a hard-disk recorder together with dual sound recordings of the clinical supervisor and trainee, who wore radio microphones.

Sample size

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 reliable estimates it is essential to sample each relevant factor (trainees, case, assessors, etc.) as widely and representatively as possible.

The proposal at the start of the study was to recruit 50–60 surgical cases for each index procedure, which gave a total of 450–540 cases. In view of initial difficulties with recruitment and the addition of further index procedures, the Steering Committee recommended a total of at least 300 cases with a good spread across specialties and index procedures. Although this number was seen as a minimum requirement for the overall study, it was increased in light of the need to

compare PBA and OSATS within O&G. An overall target of 450 cases was subsequently set, of which 150 were intended to be within O&G.

Informing, recruitment, consent and training of participants

Ethical approval required written informed consent from patient participants. Patients were sent a study information leaflet by post prior to admission for surgery or it was given on the ward if they were already an inpatient. They were given at least 24 hours to consider the information. A member of the research team then discussed the study with the patient who was given an opportunity to ask questions and raise any concerns regarding the study. The patient was made aware that they could withdraw at any time without it affecting their medical care or legal rights. He or she was then asked to sign a study consent form if willing to participate. This included an agreement for the procedure to be videoed. Patients were also given the option to agree to participate but decline to be videoed.

Written consent for participation by any of the staff groups was not stipulated in the ethical approval. The surgical trainees, however, as those being assessed using the study tools, were seen by the research team as additional study participants. Therefore, a letter of invitation was sent to all of them explaining the study and stating that their participation in the study was entirely voluntary. Trainees were informed that the study was designed to assess the reliability of several assessment methods over a large number of cases/assessors/trainees and not to assess an individual's performance. If trainees asked for guidance from the research team in the operating theatre, they were simply told to perform their surgical case as they would usually do, i.e. within their own limits of competence and asking for guidance/assistance from their clinical supervisor as they usually would.

Information regarding the study and training resources for use of the tools were disseminated to all staff assessor groups. Each individual was given a study overview, the relevant assessment tool for that staff group, and training guidance in the use of the tool. This information was sent to everyone in electronic format via e-mail and by post as a paper copy. It was followed up by face-to-face discussion/training, where accepted, with each individual in advance of the first case that they were involved in. Clinical supervisors were asked to be scrubbed in to be able to directly observe the case. There was an expectation that the cases were observed with an emphasis on assessment, rather than training, and that clinical supervisors would allow trainees to lead within their limits of competence, only prompting or intervening in the interests of good patient care.

Training of the research team

The individual roles and responsibilities of the research team members are outlined clearly in the *Contribution of authors* section in the acknowledgements. All study personnel had up-to-date training in good clinical practice in the conduct of research.

It was essential that the independent assessors of the research team were surgically credible, had an appropriately high level of knowledge about concepts of current WBA methods and were trained in the use of the study tools. All independent assessors continued to practice in surgery on a regular basis throughout the study period. They also attended the Royal College of Surgeons' 'Training the Trainers' course. Two training sessions were provided by the expert NOTSS team from Aberdeen University/Royal College of Surgeons of Edinburgh in human factors and the use of the NOTSS tool in the clinical environment.

Study team members attended and presented at a number of conferences and workshops which were relevant to the field being studied, some of which are outlined below:

- six presentations at specialty conferences
- PBA workshops at specialty workshops
- international clinical skills conferences – presentation and attendance
- three international workshops on behavioural science applied to surgery – presentation and attendance
- Royal College of Surgeons' education conference – presentation and attendance
- RCOG – two presentations to the assessment subcommittee
- Trent Regional Anaesthetics meeting – presentation on non-technical skills
- cardiac simulation centre training day – presentation on non-technical skills.

The aims of attendance were to increase the team's knowledge of WBA, to ensure that the team kept up to date with the latest developments and to publicise our study as widely as possible.

Study implementation

The implementation of the study within a single surgical specialty is illustrated by the flowchart in *Figure 4*.

The study co-ordinator and research fellow checked the diaries of the relevant surgical departments every few days to obtain the details of potentially suitable index procedures. A screening/recruitment log was completed, indicating which patients had been sent a letter with the study information sheet.

The clinical supervisor and surgical trainee were informed once suitable lists and operations were identified. If both the clinical supervisor and trainee consented to participate, the research team proceeded with approaching the patient for consent. The anaesthetist and theatre practitioner were also informed in advance, but often it was possible to do so only on the morning of the procedure because of staff allocation to specific lists.

All trainee and assessor participants were asked to complete a demographic questionnaire in advance of their initial assessment. The trainee's questionnaires included years of UK and non-UK surgical training as well as previous total and recent (in the last 6 months) experience and confidence in performing the specialty-specific index procedures.

The PBA, OSATS and NOTSS assessment forms were shown to the relevant staff assessors prior to each procedure to refamiliarise them with the tools. Assessors were asked to observe the trainee's performance and to provide assessment ratings to reflect their observations on this occasion. It was made explicit to assessors that ratings should not be based upon previous experience or knowledge of the trainee's performance.

The independent assessor had the full guidance notes of each tool available for the reference of assessors if required. Laminated copies of the NOTSS behavioural markers were provided in theatre to enable each NOTSS assessor to complete the assessments.

At the first suitable point after the operation, the assessment tools were completed independently by the relevant staff participants (including independent assessors themselves). At this stage, the research team did not provide further guidance on completion of the assessment forms. Formal

training with refamiliarisation before the start of the case had already been provided and the research team felt that ratings could be biased by discussion between assessors and independent assessors. If assessors asked for advice from the research team, they were told to provide a PBA/OSATS assessment as they would normally conduct one, based upon their previous experience and/or training in use of the tool. In the case of NOTSS assessors, reference to the laminated NOTSS behavioural markers charts was advised, while acknowledging that their ratings needed to provide a personal judgement of non-technical skills, for which there were no 'right' or 'wrong' answers. However, where sections of the forms were noted to be incomplete, the independent assessor did prompt full completion by the assessor.

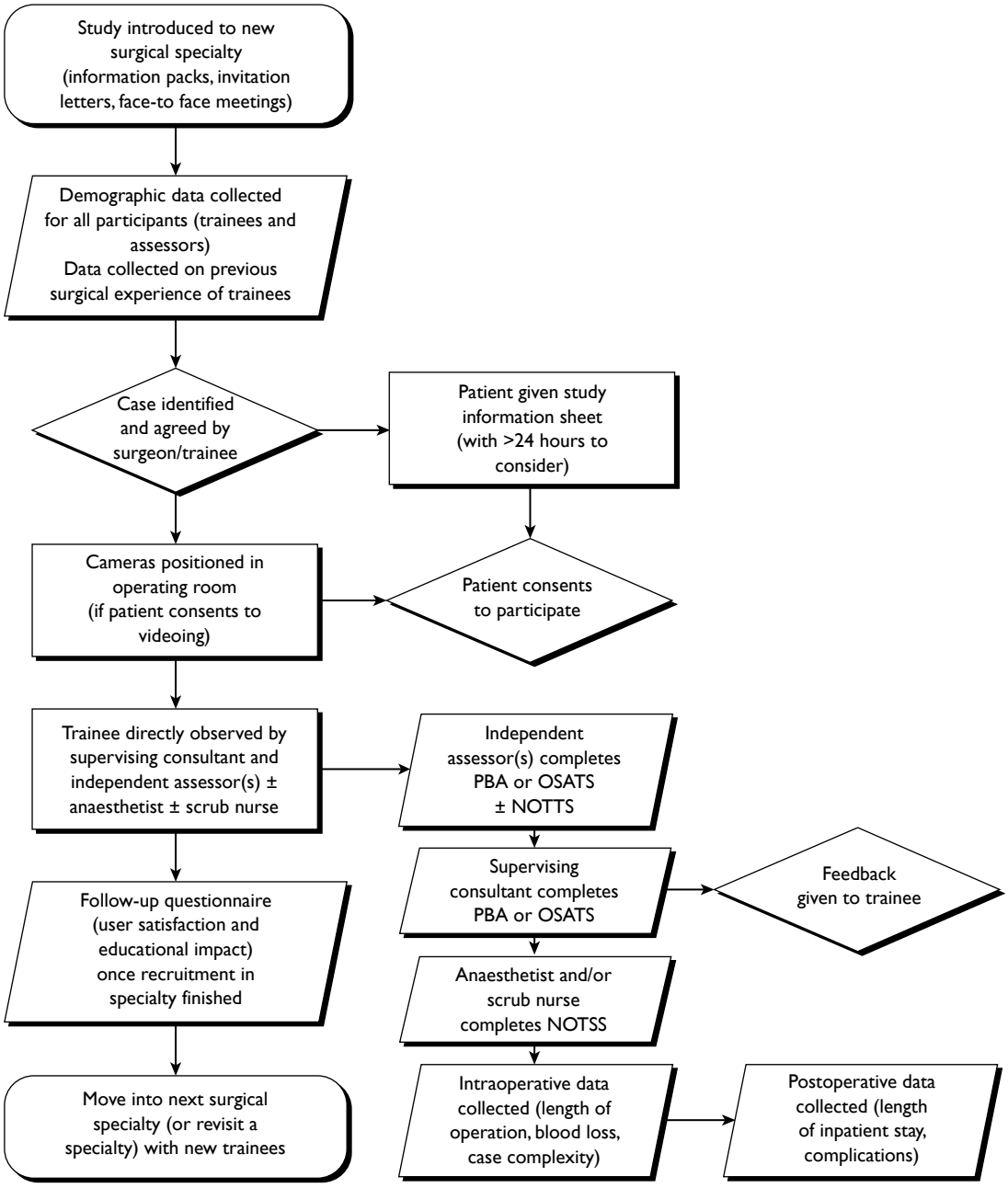


FIGURE 4 Flowchart of the study implementation. Adapted from Marriott *et al*.[120]

All assessment forms were collated by the independent assessor. Photocopies of the PBA and OSATS forms that had been completed by their clinical supervisor were provided to the trainees for their training portfolios on request.

Formative feedback to the trainee by the clinical supervisor was given in the usual way following completion of the PBA or OSATS tool. The independent assessor observed and timed the verbal feedback given to the surgical trainee on his or her surgical performance. In some cases there was no feedback provided. The NOTSS tool was originally designed to be used by clinical supervisors to give trainees feedback on their non-technical skills. However, as our study design used non-surgeon NOTSS assessors, feedback using NOTSS was not included within this study.

Where consent for videoing had been given by the patient, the video-recording equipment was assembled in the operating theatre by a medical photography expert prior to the beginning of the procedure. The case was filmed from entry of the patient into the operating theatre until application of surgical dressing at the end of the procedure. This was recorded screen-in-screen, together with sound, on to DVD.

For each case, the independent assessor also recorded the American Society of Anesthesiologists (ASA) status of the patient, the duration of the operation, the difficulty of the operation, blood loss and any intraoperative complications.

Following each case, the independent assessor provided the trainee and clinical supervisor separately with a written slip asking whether surgical performance had been affected by the assessment conditions. This question was added to the study method at an early stage in recruitment (from case 95 onwards), following the suggestion by the Steering Committee. It was intended to provide both trainee and trainer perspectives on whether they judged that the assessment itself had had any impact on their assessment scores for that particular procedure.

The study co-ordinator collected further outcome data relating to inpatient stay (e.g. intensive care unit (ICU)/high-dependency unit (HDU) admission, duration of inpatient stay) and postoperative complications (e.g. returns to theatre). These were retrieved, where possible, directly from the patient's medical records as the primary data source. In the few cases where patient records could not be located, information was sourced from the hospital electronic systems.

Following their involvement in the study, all staff participants were sent a follow-up questionnaire to complete. This was concerned with previous use of the study tools, satisfaction with their use and educational impact.

In line with local trust policy and professional guidelines on accountability to share concerns regarding serious failings in practice or conduct, the research team needed an operational guideline to follow if such an event was observed while undertaking the study. The agreed action was that concerns raised by any of the independent assessors would be reported to the principal investigator, who would notify the relevant programme training director.

Statistical analysis

All analyses were conducted using the Statistics Package for the Social Sciences (SPSS; SPSS Inc., Chicago, IL, USA) version 14.

Recruitment, study context, and user satisfaction and acceptability

Numerical data in the form of total numbers, medians and averages are presented using frequency tables. They were calculated directly from responses on structured questionnaires and are presented as proportions of those who responded. No statistical tests were applied as the sample sizes were too small. Therefore, comparative statements are used purely to explain observed differences in our questionnaire response data. Box-and-whisker plots are used to illustrate any differences between clinical supervisor and trainee perspectives. Qualitative data from the same questionnaires have been collected and stored but not yet analysed.

Reliability

Reliability is a measure of precision and discrimination. Broadly, it reflects the reproducibility of assessment rankings. In the context of this study, reliability represents how well a trainee's score based on x assessors observing y cases would predict that same trainee's score if a different x assessors observed a different y cases. More exactly, it represents how two different trainees would score relative to one another if they were both dual assessed according to that pattern.

If we call the stable differences between trainees 'true variance', then the impact of case variation, assessor variation and other sources of unwanted variation might be called 'error variance'. Reliability (R) is given by the equation:

$$R = \text{true variance}/(\text{true variance} + \text{error variance})$$

The coefficient R will be a fraction between 0 and 1, where 0 is the worst possible reliability (all error) and 1 is the best possible reliability (all true). A reliability coefficient of ≥ 0.7 is generally accepted as sufficient for 'low-stakes' assessment situations; ≥ 0.8 is generally accepted as sufficient for 'high-stakes' assessment situations.

We calculated reliability using generalisability theory.[121] Generalisability studies apply variance component analysis to estimate the impact on an assessment score of every relevant factor (G-study). They then combine the sources of variance using equations derived by Cronbach to model the reliability coefficient in a given assessment situation with a particular number of assessors and cases (D-study). A modelled reliability coefficient is thus called a generalisability coefficient (G).

Our G-study used the VARCOMP procedure. We selected the MINQUE method because of its superior handling of unbalanced data.[122] Our regression model assumes that each factor contains a random sample from an infinite universe and estimates the factors in *Table 4*.

Our D-studies are reported in tabular form to show how G varies with increasing numbers of cases and assessors. To mimic real assessment formats the axes of the table are 'cases' and 'assessors per case'. This means that (for example) cell 2,2 represents the reliability of the scores from four different assessors, two observing one case and two observing another. The tables assume that judges are drawn from a similar mix of designations in every trainee's assessment.

We have modelled reliability for two situations. Firstly, we have estimated the reliability of a format where trainees all perform the same index procedure. The D-studies for comparing trainees within a procedure use the formula:

$$G = V_p/(V_p + (V_i/N_i) + ((V_j + V_{\text{des}})/N_j) + (V_{i*j}/(N_j \times N_i)) + (V_{\text{proc}*j}/N_j))$$

where N means 'the number of'.

TABLE 4 Factors estimated by the regression model

GT label	Component label	Meaning	Short description
V_p	Var(trainee)	Consistent differences between trainees after correcting for the different samples of procedures done and assessors assessing	Trainee ability
V_i	Var(case)	Case-to-case variation (nested within trainee)	Trainee case-to-case variation
V^{proc}	Var(proccod)	Consistent differences between index procedures after correcting for the different samples of trainees doing, and assessors assessing, each index procedure	Procedure difficulty
V_j	Var(assessor)	Consistent differences between assessors after correcting for the different samples of trainees and procedures assessed	Assessor stringency
V_{des}	Var(designation)	That part of Var(assessor) that can be explained by designation on the basis that it is consistent within a group of assessors of the same designation	Assessor designation stringency
V_{p*proc}	Var(trainee*proccod)	The consistent tendency for a particular trainee to score more highly (or poorly) on one index procedure than their general performance	Trainee procedure aptitude
V_{i*j}	Var(case*assessor)	The tendency for assessors to give different scores to a particular case that is not explained by their baseline differences in stringency	Assessor subjectivity over case
V_{proc*j}	Var(proccod*assessor)	That part of Var(case*assessor) that is explained by the index procedure being assessed on the basis that the assessor differences are consistent within an index procedure	Assessor subjectivity over procedure
V_e	Var(error)	Score variation that is not already explained by one of the factors above	Residual variation

GT, generalisability theory.

Secondly, we have estimated the reliability of a format whereby trainees all perform an identical mix of two procedures. The *D*-studies for these tables use the formula:

$$G = V_p/(V_p + (V_i/N_i) + ((V_j + V_{des})/N_j) + (V_{p*proc}/2) + (V_{i*j}/(N_j \times N_i)) + (V_{proc*j}/(N_j \times 2)))$$

Validity

Validity indicates how well the score reflects the intended construct of 'surgical performance'. The study provides many sources of information about validity and these will all be presented in evidence for or against the validity of the three assessment methods. If valid, the following hypotheses will be fulfilled:

1. Scores obtained by each assessment will correlate with the other assessments that set out to measure the same aspect of performance. These correlations will operate within instruments (internal structure) and between instruments.
2. Scores will increase with duration of surgical training and number of similar procedures performed (experience).
3. Higher-scoring operations will result in less operative time and blood loss, fewer perioperative and postoperative complications and a shorter length of hospital stay.
4. Mean scores, and scores for each element, will not be significantly different across the fifteen different index procedures.
5. Assessor designation (clinical supervisor versus independent assessor for PBA and OSATS, anaesthetists and scrub nurses versus independent assessor for NOTSS) will not affect assessment stringency.
6. Assessment (plus or minus video-recording equipment) will not affect the performance of trainees.

Each of these hypotheses will be tested as follows: Pearson's method will be used for hypotheses 1, 2 and 3. In addition, for hypothesis 1, factor analysis will test internal structure (principal

axis factoring with varimax rotation). For hypothesis 2, categorical regression (CATREG procedure) will regress to correct for intercorrelations between predictor variables. Regarding hypothesis 4, the raw mean scores for each procedure are confounded because they are based on very different samples of trainees and assessors in each case. Thus the best test of whether the different procedures themselves produce different scores is to examine the value of Var(proc) in the *G*-study tables for each instrument. Var(proc) represents the 'consistent differences between index procedures after correcting for the different samples of trainees doing, and assessors assessing, each index procedure'.

Exactly the same considerations apply to hypothesis 5. The raw mean scores for each assessor designation are confounded by the uneven sampling of individual assessors and of trainees and procedures. The value of Var(designation) in the *G*-study tables, however, corrects for this and represents 'that part of Var(assessor) that can be explained by designation on the basis that it is consistent within a group of assessors of the same designation'.

For hypothesis 6, we do not have the data to compare the performance of 'not assessed' and 'assessed' cases. We therefore present the questionnaire-based perceptions of trainees and supervising surgeons as a proxy indicator. In addition, we compare the scores given to those cases in which either party felt that assessment affected performance with the scores given with those cases in which neither party perceived an effect. We also compare the scores given to video-recorded and non-video-recorded cases. Both comparisons use unpaired *t* tests.

Where appropriate, the Bonferroni correction was used to raise the threshold for significance where several outcomes were used to evaluate the same hypothesis.

An interim analysis was performed at 1 year and submitted to the Steering Committee. After discussion with the Steering Committee, a post hoc analysis was performed to explore the reasons for differences in the reliability of the PBA and OSATS tools.

Chapter 3

Results

Recruitment

Patients

All patients had to be provided with written information in advance of their informed consent prior to their recruitment within the study. Information was sent to 832 patients whose surgery was identified as a suitable training case. A total of 274 (33%) of these patients were not consented for the study. A small proportion declined consent ($n = 14$), but the vast majority ($n = 260$) were not approached for consent because of lack of availability of an inpatient bed, alteration or cancellation of the operating list, or known non-availability of any trainee or research team member for the operating list. Of the 558 patients who were consented for participation in the study, 121 (22%) were not included. The flowchart in *Figure 5* illustrates the percentages of consented and included patient cases.

Table 5 displays the number of cases consented, included and not included for each specialty. The greatest numbers of patients included were from O&G ($n = 183$) and vascular ($n = 91$) specialties, while the lowest numbers were from colorectal ($n = 25$) and orthopaedic specialties ($n = 36$).

The percentages of included versus non-included cases (i.e. lost to assessment) for each specialty are shown in *Figure 6*. Orthopaedic (31%) and cardiac (27%) surgery had the largest proportion of lost cases.

We recorded the reason for non-inclusion in each case. Fourteen cases were not included because no member of the research team was available to co-ordinate the assessments in theatre. Within O&G surgery, 29 cases were lost because a provisionally planned caesarean section

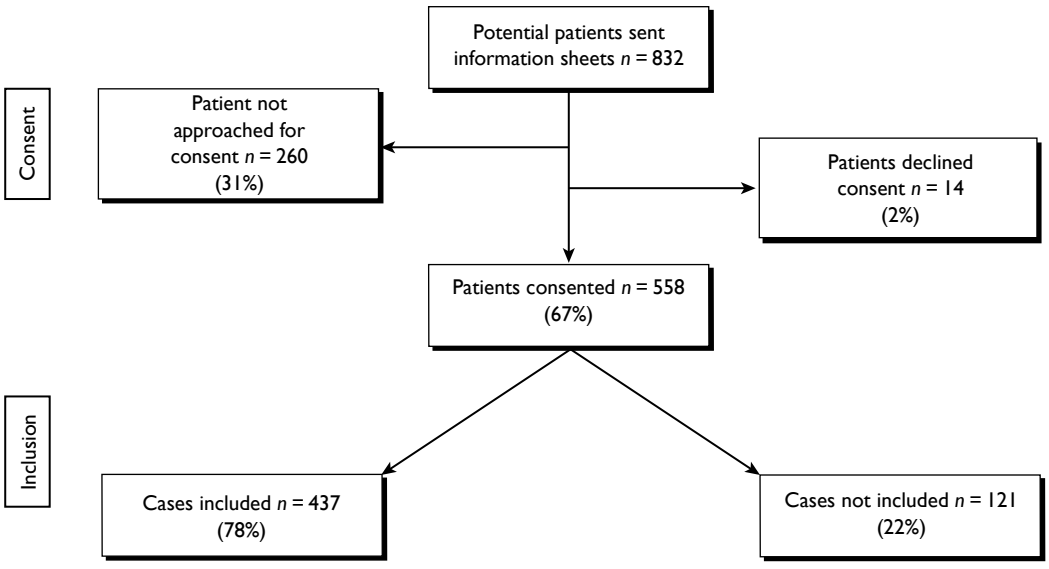


FIGURE 5 Flowchart of cases consented and included.

TABLE 5 Numbers of patients consented and included per specialty

	Cardiac	Colorectal	Upper GI	Orthopaedic	O&G	Vascular	All specialties
Patients consented	55	28	78	52	243	102	558
Patients included	40	25	62	36	183	91	**437**
Patients not included	15	3	16	16	60	11	121

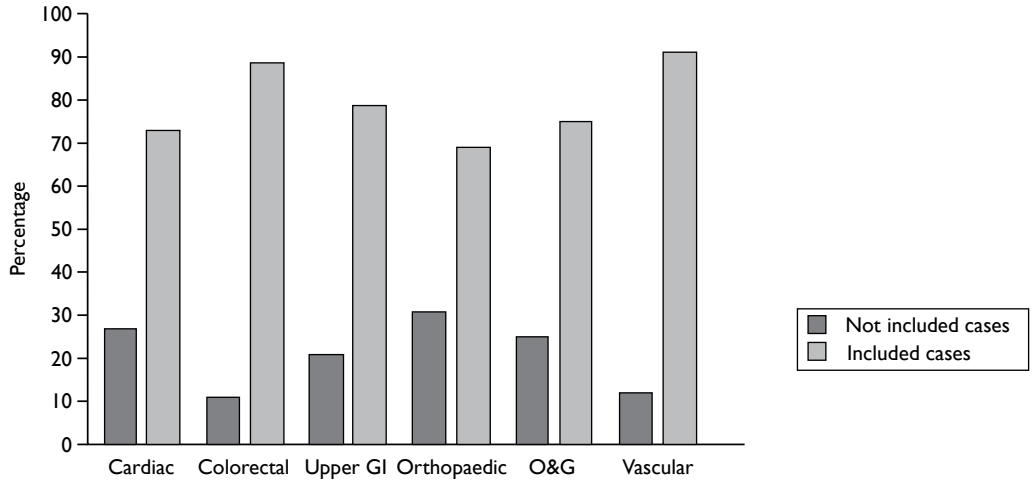


FIGURE 6 Percentage of included versus non-included cases for each surgical specialty.

did not proceed. These cases were excluded from further analysis as we wanted to accurately reflect the clinical obstacles to WBA for trainees. The most common reasons for non-inclusion across all surgical specialties were no trainee available to perform the case (27%), the clinical supervisor performing the case, either because he or she felt the case was unsuitable for training the trainee available or because he or she chose to perform the case personally (22%), and no list time available for surgical training (12%). The contribution of each factor to the loss of cases is illustrated in *Figure 7*.

Assessor/trainee consent and participation

Having received information about the study and an invitation to participate, all clinical supervisors and all but two trainees within the six surgical specialties, while the study was active in those specialties, gave verbal consent to participate in the study. The two trainees who declined to consent were within O&G and were near achieving their CCT. Verbal consent from these groups was sought before each round of recruitment in that specialty or when an individual entered the specialty if this occurred part-way through the recruitment round. Four clinical supervisors (two in orthopaedics, one in cardiac surgery and one in colorectal surgery) did not go on to undertake any assessments because of logistical reasons which have been identified above. Six trainees (two in orthopaedics and one in colorectal, one in cardiac and two in upper GI surgery) did not go on to be assessed, also for logistical reasons. All anaesthetists, nurses and surgical-care practitioners (SCPs) working within the operating theatres were given prior information about the study and invited to participate. Individual assessors were approached for verbal consent as far in advance of their first case as possible. However, these professional groups sometimes rotated through the operating theatres in an unpredictable manner and it was not always possible to seek consent until the morning of the list. Owing to the large numbers

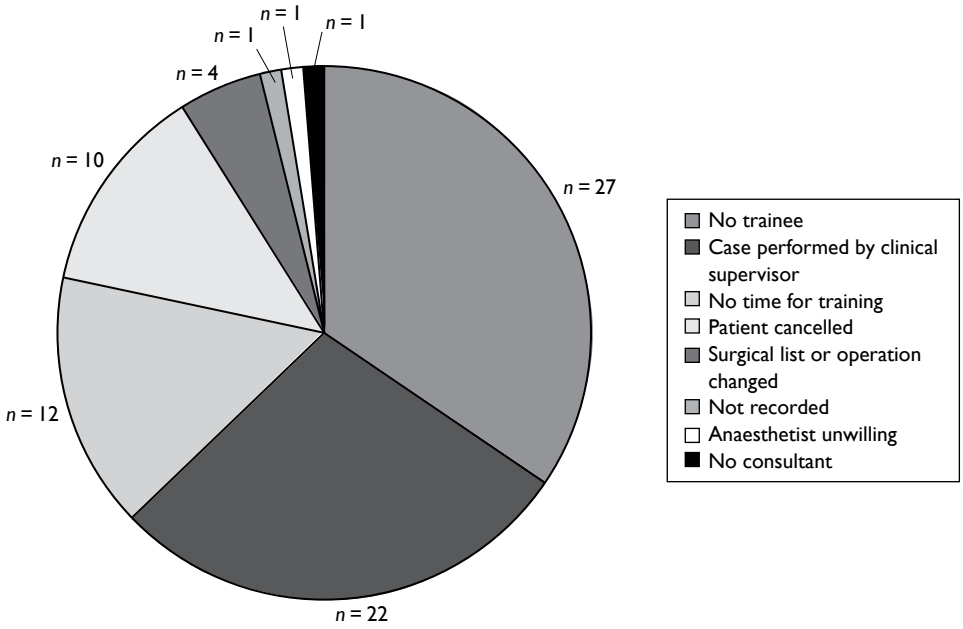


FIGURE 7 Reasons for lost assessment cases.

of personnel within these groups, verbal consent was sought only from those individuals attached to the relevant operating lists and not from all potential recruits as was the case with clinical supervisors and trainees. No nurses and only one anaesthetist who were approached to participate declined to do so. Numbers of non-recruited anaesthetists and nurses were not recorded. Demographic information was not collected for any non-participant assessors or trainees as they had not given consent for us to do so.

Surgical cases

Fifty-one clinical supervisors, 56 anaesthetists, 39 scrub nurses, two SCPs and four independent assessors provided 1635 assessments on 85 trainees undertaking 437 cases. The distribution of assessments for these surgical cases using the three assessment tools is shown in *Table 6*. A total of 749 PBAs, 695 NOTSS assessments and 191 OSATS were performed. There are many more assessments than cases owing to the multiple assessments provided by different assessor groups. In addition, the independent assessor completed an OSATS/PBA *and* NOTSS assessment on many occasions. Independent assessor (IA)1 completed more than one assessment tool in 75% of cases, IA2 in 73% of cases and IA3 in 86% of cases.

Study context

Assessor and trainee demographics

Surgical trainee demographics

Of 85 surgical trainees, 82 provided almost complete demographic data (98% response rate) and there are some data for all trainees. The descriptive statistics are presented as a proportion of those who responded (this is also the case for presentation of the assessor demographic data). The greatest share of trainees were from O&G (38%), while the remaining trainees were fairly evenly split between the other five surgical specialties (*Figure 8*).

The trainees' demographics are summarised in *Table 7* and illustrated in *Figures 9–13*. The majority of all trainees were male (65%). Male gender predominated in all specialties

TABLE 6 Overall study recruitment by surgical specialty and procedure

	Cardiac		Colorectal		Upper GI		Orthopaedic		O&G				Vascular			
	AVR	CABG	Anterior resection	Right hemicolectomy	Hernia	Laparoscopic cholecystectomy	Hip replacement	Knee replacement	Diagnostic laparoscopy	Elective caesarean	Evacuation of uterus	Urgent caesarean	Aortic aneurysm	Carotid endarterectomy	Varicose veins	Totals
Trainees	3	10	8	6	11	12	7	7	23	22	15	4	4	7	9	
Cases	5	35	14	11	19	43	18	18	73	60	45	5	15	25	51	**437**
PBA forms																
Surgeon	4	34	14	10	9	35	15	16	40	33	15	3	15	25	51	**319**
IA1	5	35	14	11	10	28	15	18	15	11	1	0	13	20	42	**238**
IA2	0	5	2	1	4	4	1	2	4	4	2	0	3	2	10	**44**
IA3	0	0	0	0	15	23	4	0	38	34	17	4	0	1	3	**139**
IA4	0	0	0	0	3	6	0	0	0	0	0	0	0	0	0	**9**
NOTSS forms																
Anaesthetist	3	26	6	8	10	19	1	0	24	26	8	5	12	11	24	**183**
IA1	4	29	11	10	8	22	14	13	14	6	1	0	9	16	30	**187**
IA2	0	2	2	1	0	0	1	3	2	4	3	0	3	6	9	**36**
IA3	0	0	0	0	11	10	4	0	61	57	42	5	0	1	3	**194**
Nurse	0	3	6	6	1	3	0	1	33	20	11	3	4	1	0	**92**
SCP	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	**3**
OSATS forms																
Surgeon	0	0	0	0	0	0	0	0	33	27	28	1	0	0	0	**89**
IA1	0	0	0	0	0	0	0	0	6	4	0	0	0	0	0	**10**
IA2	0	0	0	0	0	0	0	0	0	3	2	0	0	0	0	**5**
IA3	0	0	0	0	0	0	0	0	32	26	28	1	0	0	0	**87**

AVR, aortic valve replacement; CAGB, coronary artery bypass grafting.

TABLE 7 Trainee demographics by surgical specialty

		Cardiac	Colorectal	Upper GI	Orthopaedic	O&G	Vascular
Total number	Frequency	10	9	13	9	33	11
	%	12	10.5	15	10.5	39	13
Number males	Frequency	10	6	12	9	9	9
	%	100	67	92	100	28	90
Age (years)	Median	37	33	33	35	32	34
	Range	28–45	28–39	27–43	34–40	25–40	27–46
UK graduate	Frequency	1	7	9	7	13	6
	%	10	78	69	78	41	60
ST level	Median	8	3	3	8	3	3
	Range	3–8	3–8	1–7	3–8	0–7	2–7
Total years' surgical training	Median	11	6	9	9	5.5	6
	Range	4–17	2–13	0–14	4–18	0–15	2–19
Years' UK surgical training	Median	11	6	4	9	3	5
	Range	1–13	2–12	0.2–12	4–12	0.5–9	2–10
Years' non-UK surgical training	Median	3	2.5	0	0	0	0
	Range	0–6	1–4	0–7	0–7	0–6	0–9
Total index procedures	Median	57.5	7	30	116	46	3.5
	Range	0–450	0–54	0–200	5–350	0–400	0–500
Recent index procedures	Median	13.5	4	17.5	15	10	1.5
	Range	0–60	1–10	0–58	3–60	0–45	0–20

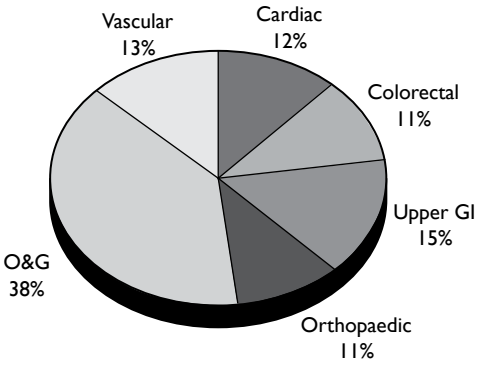


FIGURE 8 Proportion of trainees in each surgical specialty.

(67%–100%), with the exception of O&G, where 78% were female; 100% of all trainees in cardiac and orthopaedic specialties were male.

Half of all trainees graduated in the UK (51%), although the proportion of UK and non-UK graduates in each specialty differed substantially. Colorectal and orthopaedic specialties had a large majority of UK graduates (both 78%), whereas cardiac trainees were predominantly non-UK graduates (89%) with the highest median for number of years of non-UK surgical training.

The most experienced trainees, in terms of age, total years of surgical training and total number of index procedures previously performed, were from the cardiac and orthopaedic specialties. Trainees in vascular surgery had the greatest range in years of total and non-UK surgical training

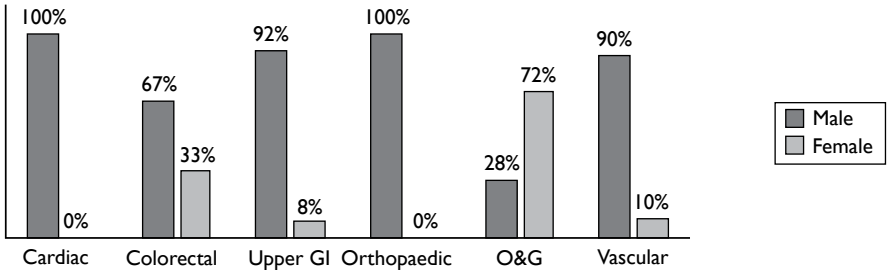


FIGURE 9 Percentage of male and female trainees by surgical specialty.

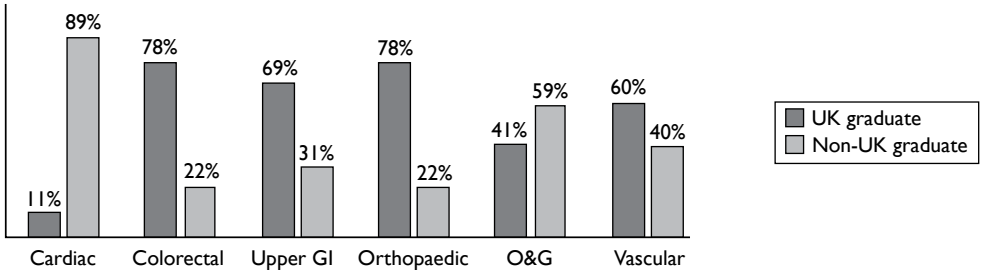


FIGURE 10 Percentage of UK and non-UK graduates by surgical specialty.

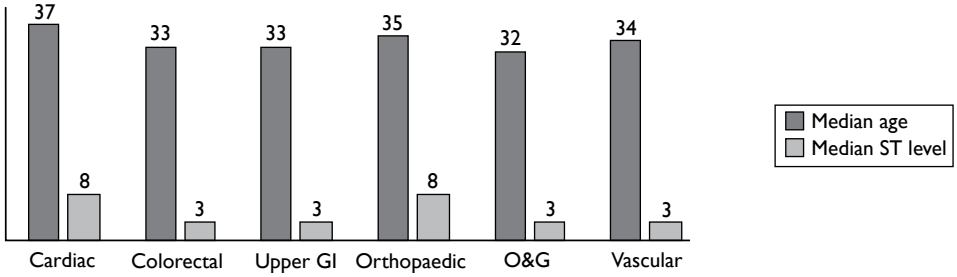


FIGURE 11 Median age (years) and ST level of trainees by surgical specialty.

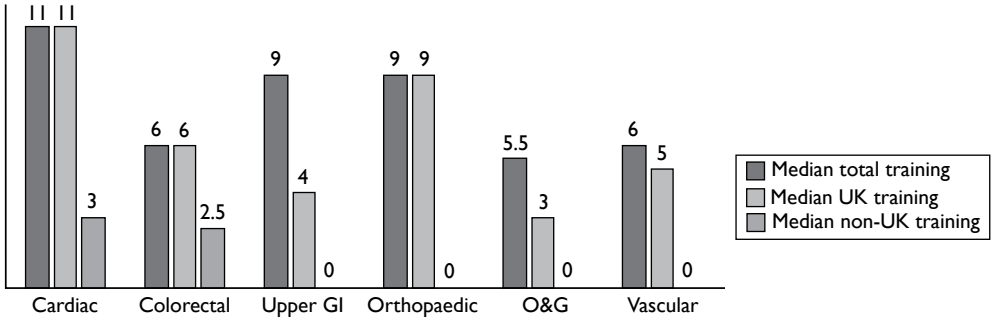


FIGURE 12 Medians for total years of surgical training, UK surgical training and non-UK surgical training by surgical specialty.

(2–19 years and 0–9 years respectively) and the greatest range of previous total number of index procedures performed (0–500).

All levels of ST were represented within the trainees recruited across the study. The highest median ST level was within cardiac and orthopaedic specialties.

The spread of trainees according to their year of ST within each of the surgical specialties is presented in *Figure 14*. The percentage of trainees at each ST level is presented in *Figure 15*. The greatest percentage of trainees was at either a junior level (ST2 or ST3) or a senior level (ST7 and ST8), reflecting the tertiary referral centre/teaching hospital setting for the study.

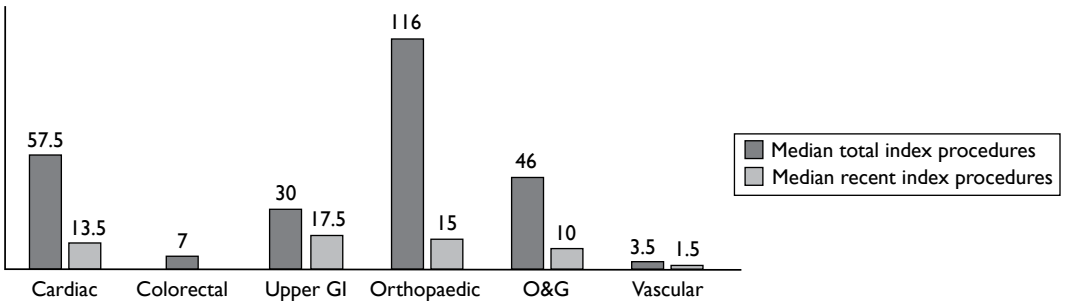


FIGURE 13 Medians for total number and number of recently performed index procedures prior to taking part in the study by surgical specialty.

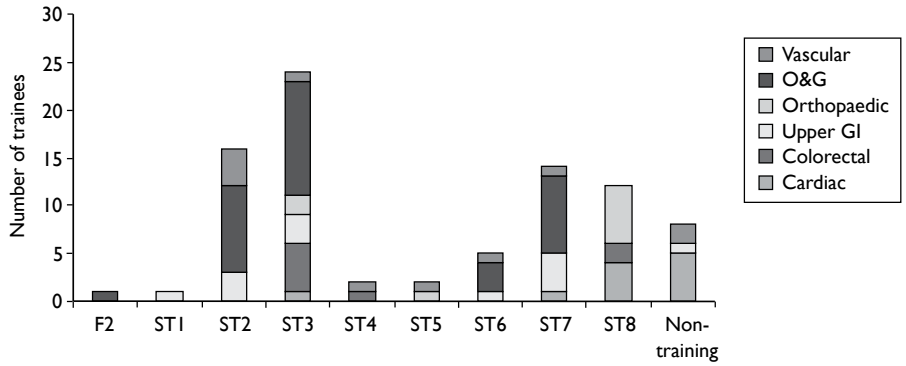


FIGURE 14 Trainees by specialty and year of training.

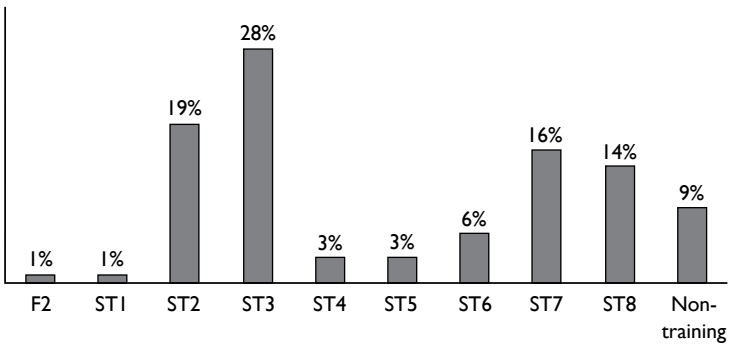


FIGURE 15 Percentage of trainees at each ST level.

Clinical supervisor demographics

Of 51 clinical supervisors, 50 provided almost complete demographic data (98% response rate) and there are some data for all clinical supervisors. The greatest share of clinical supervisors were from O&G (37%), while the remainder were fairly evenly split between the other five surgical specialties (*Figure 16*).

The majority of clinical supervisors were male (90%) and were UK graduates (78%). There was a broad age range across the specialties (37–61 years), although the median age was similar within each specialty (45–48 years). There was also a broad range of experience (1–26 years) across all specialties. The median years of experience was highest in O&G (13 years) and cardiac (10.5 years) and lowest in upper GI (5 years). Six senior trainees acted as assessors in a small number of cases ($n = 15$, 3% of total). Their demographic data are not included in *Table 8* but rather in the demographic data for trainees.

Independent assessor demographics

The demographic information for the four independent assessors was:

- IA1 an SCP of 4 years' experience who was a 37-year-old female UK graduate
- IA2 a consultant vascular surgeon of 17 years' experience who was a 51-year-old male UK graduate
- IA3 an ST4 trainee in O&G who was a 29-year-old female UK graduate

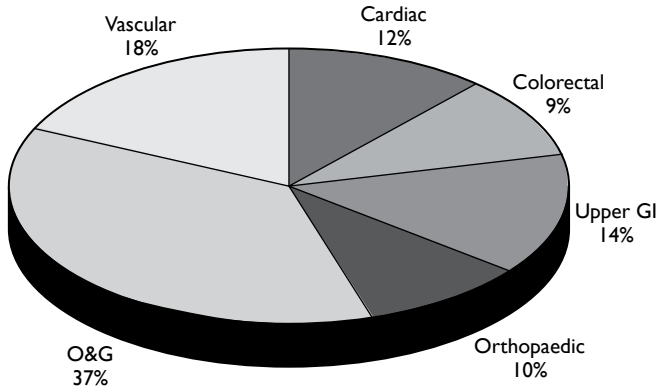


FIGURE 16 Percentage of clinical supervisors in each surgical specialty.

TABLE 8 Clinical supervisor demographics by surgical specialty

		Cardiac	Colorectal	Upper GI	Orthopaedic	O&G	Vascular
Total number of clinical supervisors	Frequency	6	5	7	5	19	9
	%	12	9.5	14	9.5	37	18
Age (years)	Median	47	48	45	45	47	45
	Range	43–53	43–61	38–60	39–51	37–60	40–52
Number males	Frequency	6	3	7	5	13	9
	%	100	75	100	100	76	100
UK graduate	Frequency	5	3	6	5	12	7
	%	83	75	100	100	63	78
Experience (years as a consultant)	Median	10.5	6.5	5	9	13	8
	Range	8–18	2–8	1–26	3–16	1–25	1–17

- IA4 an ST2 trainee in vascular surgery who was a 33-year-old male non-UK graduate.

Anaesthetist demographics

Of 56 anaesthetists, 48 provided almost complete demographic data (86% response rate). They had a median age of 41 years (range 27–57 years) and 73% were male. Eighty-eight per cent were UK trained and they had a median of 8 years' experience as a consultant (range 0–27 years).

Scrub nurse and SCP demographics

Of 39 scrub nurses, 33 provided almost complete demographic data (85% response rate). They had a median age of 39 years (range 26–58 years) and 18% were male. Eighty-two per cent were UK trained and they had a median of 10 years' experience as a scrub nurse (range 1–38 years).

The two SCPs who provided NOTSS assessments were both UK-trained males. They were 33 and 51 years old and had 4 and 8 years' SCP experience respectively.

Cases and outcomes

Table 9 displays the ASA grade and some important intra- and postoperative outcome data for the surgical cases included within the study. The intra- and postoperative data are almost complete for all of the 437 cases. Blood loss was not available for aortic valve replacement (AVR) procedures because of the difficulties of estimating blood loss for some cardiac procedures.

The data show the anticipated patterns. The different surgical specialties have different profiles with respect to the comorbidity of patients and the complexity of surgery. These more complex operative cases on more dependent patients typify the index cases within cardiac, vascular and colorectal specialties, while upper GI and O&G specialties include fewer high-risk index cases on fitter patients. Coronary artery bypass grafting had the longest median operating time and evacuation of the uterus the shortest. Aortic aneurysm repair had the highest median blood loss while hernia repairs had the lowest.

There is a large range in the length of operations, which is influenced by case complexity as well as surgical skill/aptitude. Longer and higher blood loss operations on more dependent patients are associated with greater lengths of hospital stay and a greater likelihood of HDU/ICU stay and postoperative complications. As expected, cardiothoracic and vascular cases were associated with the greatest number of HDU/ICU stays and postoperative complications. Patients in O&G, who are often young and fit, tolerate surgery well and have a short length of hospital stay. There are some notable exceptions with respect to postoperative stay: for orthopaedic patients, postoperative rehabilitation increased postoperative stay in excess of the surgical recovery time.

Experience and training in use of the study tools

All participants were asked to report their *prior* experience with the relevant assessment tool, either for the purpose of assessing and providing feedback in the case of the study assessors or for being assessed and receiving feedback for the surgical trainees. All assessors received face-to-face training supported by written and/or e-mail information packs from the research team before their involvement in the study. Participants' perspectives on the adequacy of the training they received, either through their involvement in the study or from other training avenues, were addressed within the questionnaires.

The descriptive statistics for experience and training are all presented as a proportion of those who responded.

TABLE 9 Surgical case outcomes by surgical specialty

		Cardiac		Colorectal		GI		Orthopaedic		O&G				Vascular		
		AVR	CABG	Anterior resection	Right hemicolectomy	Hernia	Laparoscopic cholecystectomy	Hip replacement	Knee replacement	Diagnostic laparoscopy	Elective caesarean	Evacuation of uterus	Urgent caesarean	Aortic aneurysm	Carotid endarterectomy	Varicose veins
Number of cases		5	35	14	11	19	43	18	18	73	60	45	5	15	25	51
ASA grade (1–4)	Median	4	3	2	2	2	2	2	2	1	1	1	1	3	3	1
	Minimum	2	3	1	2	1	1	1	1	1	1	1	1	2	2	1
	Maximum	4	4	4	4	3	3	4	3	3	2	2	2	4	3	2
Length of operation (minutes)	Median	205	210	190	155	60	70	87.5	97.5	32	45	10	44	195	143	45
	Minimum	175	140	120	55	35	42	70	60	17	25	6	39	135	115	20
	Maximum	215	315	420	315	80	130	120	135	185[a]	71	21	74	390	185	75
Blood loss (ml)	Median		250	500	400	0	0	500	100	0	500	200	500	1500	200	100
	Minimum		0	250	100	0	0	200	0	0	300	50	400	600	100	0
	Maximum		500	3500	700	0	400	1200	500	400	1000	450	900	2000	500	500
Length of postoperative stay (days)	Median	6	6	9	15	1	1	7	7	0	3	0	3	9	3	0
	Minimum	5	4	4	5	0	0	3	3	0	3	0	2	6	1	0
	Maximum	18	19	55	55	4	4	38	15	4	23	1	7	26	27	1
HDU stay (days)	Median	1	1	1	1	0	0	0	0	0	0	0	0	2.5	1	0
	Minimum	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
	Maximum	5	11	2	5	0	0	1	2	0	2	0	0	7	3	0
ICU stay (days)	Median	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
	Minimum	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Maximum	1	11	0	0	0	0	0	0	0	0	0	0	0	0	0
No. intraoperative complications	Sum	0	1	2	0	0	2	0	0	2	3	1	0	2	0	0
No. postoperative complications	Sum	1	14	4	5	1	2	2	2	4	4	1	0	6	2	1
Returns to theatre	Sum	0	1	0	1	0	0	0	0	0	1	0	0	1	0	0
In-hospital deaths	Sum	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0

CABG, coronary artery bypass grafting.
a Includes operative laparoscopic time.

Procedure-based assessment – prior experience

Procedure-based assessment questionnaires were given to all clinical supervisors and trainees in the six surgical specialties. Questions relating specifically to previous experience with the PBA tool were excluded from questionnaires for O&G clinical supervisors and trainees. PBA was introduced to O&G for the purposes of our study. Therefore, they would not have used the tool before.

Twenty-eight of the 38 non-O&G clinical supervisors (response rate 74%) provided almost complete responses about their prior experience and training in using the PBA tool for assessment and feedback. They reported a range of prior experience and a range of training (*Table 10*). Overall, 39% of them reported that they had not assessed trainees using PBA prior to their involvement in the study. However, where PBA had been used in WBA, 44% them had given feedback to trainees as part of the assessment process.

Forty-one of 53 non-O&G trainees (response rate 77%) provided responses about their experience of being assessed and receiving feedback using PBA. They reported a range of prior experience (*Table 11*). Forty-nine per cent had never previously been assessed with PBA.

Procedure-based assessment – training

Forty-four of 52 clinical supervisors (response rate 85%) provided almost complete responses about their PBA training. The questionnaires asked them about the types of PBA training they had received and the overall adequacy of this training for using PBA (*Table 12*). Eighty-nine per cent of them reported that they had undergone some form of training and 70% agreed that the training had been adequate. Clinical supervisors were asked about the different types of training received, and multiple responses were possible. Face-to-face and written information (60% of clinical supervisors) were the most common types of training, reflecting the training approach adopted by the research team, with 20% of clinical supervisors having received a combination of two or more types of training.

TABLE 10 Non-O&G clinical supervisors' prior PBA experience

		n	%
Assessing with PBA	>15 times	2	7
	6–15 times	8	29
	1–5 times	7	25
	Never	11	39
Giving feedback with PBA	Always	12	44
	Sometimes	4	15
	Never	0	0
	Not applicable	11	41

TABLE 11 Non-O&G trainees' prior PBA experience

		n	%
Being assessed with PBA	>15 times	3	7
	6–15 times	6	15
	1–5 times	12	29
	Never	20	49
Receiving feedback with PBA	Always	13	32
	Sometimes	9	22
	Never	0	0
	Not applicable	19	46

Sixty-six out of 85 trainees (response rate 78%) provided almost complete responses about their PBA training (*Table 13*). Sixty-five per cent of them reported that they had received some form of training and 41% agreed that they had had adequate training. A greater proportion of trainees accessed the ISCP web guidance for training, although face-to-face and written information remained the most common types of training.

Objective Structured Assessment of Technical Skills – prior experience

Objective Structured Assessment of Technical Skills questionnaires were given to all clinical supervisors and trainees in O&G. As OSATS had been used informally in clinical practice for 2 years before the study began, prior experience with OSATS was assessed according to length of time used.

Seventeen of 21 O&G clinical supervisors (response rate 81%) provided responses about OSATS (*Table 14*). There was high previous experience of using the OSATS tool. Only 6% of them had never previously assessed a trainee using OSATS prior to their involvement in the study.

TABLE 12 Clinical supervisors' PBA training

	n	%
Face to face	27	60
Video	1	2
ISCP web guidance	10	22
Written information	27	60
Combination of methods	9	20
None	5	11

TABLE 13 Trainees' PBA training

	n	%
Face to face	19	31
Video	0	0
ISCP web guidance	16	25
Written information	15	24
Combination of methods	10	15
None	23	35
Missing/blank	19	23

TABLE 14 O&G clinical supervisors' prior OSATS experience

		n	%
Assessing	> 2 years	2	12
	1–2 years	6	35
	Since 1 August 2007	8	47
	Never	1	6
Giving feedback	Always	5	29
	Mostly	7	41
	Sometimes	4	24
	Never	0	0
	Not applicable	1	6

Twenty-eight of 33 O&G trainees using OSATS (response rate 85%) provided responses about the method (*Table 15*). They reported similarly substantial experience in previous use of OSATS to their O&G clinical supervisors, with only 4% never having been assessed with OSATS before.

Objective Structured Assessment of Technical Skills – training

Sixty-four per cent of O&G clinical supervisors who responded were satisfied with their training in the use of the OSATS tool (*Table 16*). This is similar to the proportion of the total clinical supervisor cohort who were satisfied with their training in the use of PBA (70%).

The O&G trainees were less satisfied than their O&G clinical supervisors with training for the use of OSATS (43% vs 64%) (*Table 17*). This is also similar to the proportion of the total trainee cohort who were satisfied with their training in the use of PBA (41%). Both groups had accessed similar training methods, and the same proportion of both groups (29%) had used a combination of training methods.

Non-technical Skills for Surgeons

Non-technical Skills for Surgeons questionnaires were given to all anaesthetist and scrub nurse assessors. NOTSS assessments are not in current training use and therefore questions regarding prior experience with this tool were not relevant. The training methods outlined reflect the training received for the study's purpose alone.

Thirty of 56 anaesthetists (response rate 54%) provided responses about the NOTSS tool. The anaesthetists reported a range of perceived training (*Table 18*). Responses on whether the adequacy of training was sufficient were evenly divided between agreement and disagreement.

TABLE 15 O&G trainees' prior OSATS experience

		n	%
Being assessed	>2 years	6	21
	1–2 years	10	36
	Since 1 August 2007	11	39
	Never	1	4
Getting feedback	Always	1	4
	Mostly	19	68
	Sometimes	7	25
	Never	0	0
	Not applicable	1	4

TABLE 16 O&G clinical supervisors' OSATS training

		n	%
Reported training activity	Face to face	4	24
	Training workshop	6	35
	RCOG web guidance	2	12
	A combination	5	29
Response to question: 'I have had sufficient training'	Strongly agree	5	29
	Agree	6	35
	Neither agree or disagree	4	24
	Disagree	2	12
	Strongly disagree	0	0

TABLE 17 O&G trainees' OSATS training

		n	%
Reported training activity	Face to face	8	29
	Training workshop	4	14
	RCOG web guidance	6	21
	A combination	8	29
	None/other	2	7
Response to question: 'I have had sufficient training'	Strongly agree	0	0
	Agree	12	43
	Neither agree or disagree	7	25
	Disagree	4	14
	Strongly disagree	5	18

TABLE 18 Anaesthetists' NOTSS training

		n	%
Reported training activity	Face to face	12	40
	NOTSS booklet	7	23
	A combination	7	23
	None	4	13
Response to question: 'I have had sufficient training'	Strongly agree	2	7
	Agree	12	40
	Neither agree or disagree	2	7
	Disagree	13	43
	Strongly disagree	1	3

Twenty-six of 39 nurses (response rate 67%) provided almost complete responses about using NOTSS. They reported a range of perceived training (*Table 19*). Eighty-four per cent of them reported having received face-to-face training compared with 63% of anaesthetists. The majority of them (82%) felt that the NOTSS training received was adequate compared with 47% of anaesthetists.

User satisfaction and acceptability

All participants were asked to report their satisfaction with and acceptability of the tools that they used as part of the study. A postrecruitment structured questionnaire was used for this purpose. The descriptive statistics for user satisfaction and acceptability are all presented as a proportion of those who responded.

Procedure-based assessment user satisfaction and acceptability

Forty-four of 52 clinical supervisors (response rate 85%) provided almost complete responses about their satisfaction with the PBA assessment and feedback process. They provided mixed, but predominantly positive, responses about the use of the PBA tool (*Table 20*). The majority of them agreed or strongly agreed that PBA was valuable as a tool for providing feedback (77%), for formative assessment (72%), for summative assessment (68.5%), and to support reflective practice (63%). Of those clinical supervisors who strongly disagreed or disagreed that PBA was a valuable tool for assessment, more disagreed that it was useful in summative assessment than formative assessment (15.5% vs 6%).

TABLE 19 Scrub nurses' NOTSS training

		n	%
Reported training activity	Face to face	18	69
	NOTSS booklet	2	8
	A combination	4	15
	None	2	8
Response to question: 'I have had sufficient training'	Strongly agree	0	0
	Agree	18	82
	Neither agree or disagree	0	0
	Disagree	4	18
	Strongly disagree	0	0

TABLE 20 Clinical supervisors' opinions on the value of the PBA tool (*n*, %)

Response to questionnaire statements	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
PBAs are a useful tool for providing feedback after an operation	0	3	8	29	4
	0%	7%	18%	66%	9%
PBAs are a valuable formative assessment tool	0	2	7	21	2
	0%	6%	22%	66%	6%
PBAs are a valuable summative assessment tool	1	4	5	18	4
	3%	12.5%	16%	56%	12.5%
PBAs are a useful tool to support reflective practice or to provide insight	0	3	12	23	5
	0%	7%	27%	52%	11%

Clinical supervisors were asked to report their opinions of PBA based upon personal use of the tool. A categorical scale from 0 to 10 was used for these questions where 0 was 'not at all', 5 was 'moderately' and 10 was 'very much' (*Table 21* – median scores are shaded). The clinical supervisors' overall satisfaction with PBA was moderately good. They reported that they thought trainees found their feedback moderately useful (median score 6, interquartile range 5–7). The extent to which PBA enhanced their ability to assess was reported as moderate (median score 5, interquartile range 4–7), although 7% reported that it did not help them at all to assess trainees (score 0). Most clinical supervisors felt comfortable reporting their concerns about a trainee's surgical skills on a PBA form (median score 7, interquartile range 4.5–8). Two questions were asked only of non-O&G clinical supervisors, for whom PBA is the current WBA tool within the ISCP. These questions related to the role of PBA in surgical education and whether the clinical supervisor would use PBA in the future if given the choice. The majority of non-O&G clinical supervisors felt that PBAs were moderately important in surgical education (median score 6, interquartile range 5–7). However, there was a wide range of opinion as to whether, if given the choice, they would use PBA in the future (median score 6, interquartile range 3.5–8). The majority (68%) felt that they were more likely than not to choose to use PBA in the future (scores ≥ 5).

Sixty-six of 85 trainees (response rate 78%) provided almost complete responses about their satisfaction with PBA and its feedback process. They provided mixed, but predominantly positive, responses about the use of the PBA tool (*Table 22*). The majority of trainees agreed that PBA was useful for providing feedback (88%), for formative assessment (72%), for summative assessment (64%), and to support reflective practice (74%). In contrast to clinical supervisor responses, there was slightly more agreement/strong agreement among trainees for the use of PBA as a formative assessment than for its use as a summative assessment (72% vs 64%). In keeping with

TABLE 21 Clinical supervisors' opinions on their use of PBA (medians shaded) (*n*, %)

	Not at all →					Moderately		→		Very much
Response to questions	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
How useful do you think your trainees found the feedback you gave in these sessions?	0	0	1	2	0	11	11	9	5	4	1
	0%	0%	2%	5%	0%	25%	25%	20%	11%	9%	2%
To what extent do you think PBA enhances your ability to assess your trainees?	3	0	3	4	6	8	5	9	6	0	0
	7%	0%	7%	9%	14%	18%	11%	20%	14%	0%	0%
How comfortable do you feel recording on a PBA that you have concerns about a trainee?	0	1	2	3	5	6	0	6	12	4	5
	0%	2%	5%	7%	11%	14%	0%	14%	27%	9%	11%
How important do you think PBAs are in surgical education?	1	0	0	3	2	7	6	5	2	1	0
	4%	0%	0%	11%	7%	26%	22%	19%	7%	4%	0%
How likely is it that you would use PBAs in the future if given the choice?	3	0	1	3	2	3	4	2	5	0	5
	11%	0%	4%	11%	7%	11%	14%	7%	18%	0%	18%

TABLE 22 Trainees' opinions on the value of the PBA tool (*n*, %)

Response to questionnaire statements	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
PBAs are a useful tool for providing feedback after an operation	0	5	3	45	12
	0%	8%	5%	69%	19%
PBAs are a valuable formative assessment tool	1	3	10	31	5
	2%	6%	20%	62%	10%
PBAs are a valuable summative assessment tool	2	6	10	30	2
	4%	12%	20%	60%	4%
PBAs are a useful tool to support reflective practice or to provide insight	2	4	11	39	9
	3%	6%	17%	60%	14%

clinical supervisors' responses, more trainees disagreed or strongly disagreed with the use of PBA for summative assessment than for formative assessment (16% vs 8%). A greater proportion of trainees than clinical supervisors felt that PBA was useful for providing feedback (88% vs 77%) and to support reflective practice (74% vs 63%).

Trainees were also asked to report their opinions of PBA, based upon personal use of the tool (*Table 23* – median scores are shaded). Trainees reported that they found the feedback provided by the clinical supervisors moderately to very useful (median score 7, interquartile range 6–8). The extent to which trainees felt that PBA had enhanced the ability of their clinical supervisors to assess them was reported as moderate (median score 6, interquartile range 4.5–7), although 6% of trainees felt that it did not help their clinical supervisors at all to assess them (score 0). Most trainees, like the clinical supervisors, felt comfortable with clinical supervisors reporting their concerns about a trainee's surgical skills on a PBA form (median score 7, interquartile range 5–8). Non-O&G trainees were asked the same two additional questions as non-O&G clinical supervisors. The majority of non-O&G trainees felt that PBA was moderately to very important in surgical education (median score 7, interquartile range 5–8). The vast majority of trainees (84% with scores ≥ 5) reported that, if given the choice, they would be likely to choose to use PBA in the future (median score 8, interquartile range 6–9).

Key differences in the responses of clinical supervisors and trainees regarding their experiences in using PBA are illustrated by box-plots (*Figures 17–20*), with median scores displayed by a

TABLE 23 Trainees' opinions on their use of PBA (medians shaded) (*n*, %)

	Not at all →					Moderately		→		Very much
Response to questions	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
How useful did you find the feedback given to you in these sessions?	0	0	0	2	3	5	7	21	18	7	2
	0%	0%	0%	3%	5%	8%	11%	32%	28%	11%	3%
To what extent do you think PBA enhances your trainer's ability to assess you?	4	1	5	4	2	7	12	16	8	4	1
	6%	2%	8%	6%	3%	11%	19%	25%	13%	6%	2%
How comfortable do you feel about a trainer recording on a PBA that they have concerns about a trainee?	0	3	0	6	4	13	4	12	10	6	6
	0%	5%	0%	9%	6%	20%	6%	19%	16%	9%	9%
How important do you think PBAs are in surgical education?	2	1	1	5	2	13	7	15	13	3	3
	3%	2%	2%	8%	3%	20%	11%	23%	20%	5%	5%
How likely is it that you would use PBA in the future if given the choice?	3	1	0	3	9	3	4	6	8	4	9
	7%	2%	0%	7%	9%	7%	10%	15%	20%	10%	22%

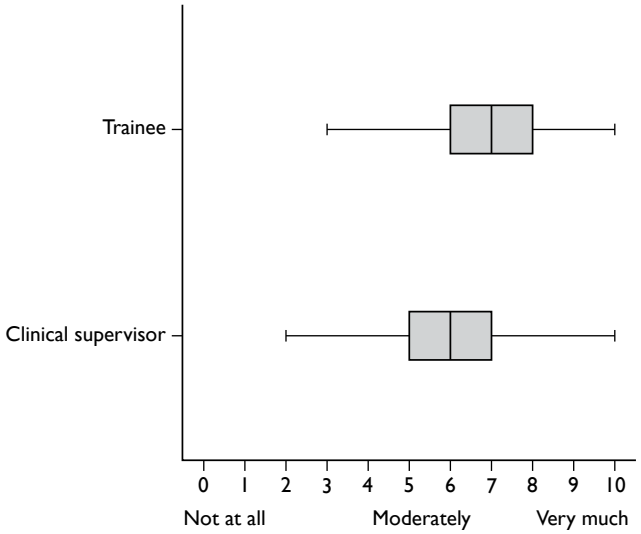


FIGURE 17 Comparing clinical supervisors' and trainees' opinion of the usefulness of feedback with PBA.

vertical black line, the interquartile range of scores by the grey box and the range of scores by the horizontal whisker lines, with outliers identified as small numbered circles. There were differences of opinion regarding the value of the PBA feedback process, with trainees finding the feedback provided more useful than the clinical supervisors themselves felt it was useful to trainees. Trainees also perceived that PBA enhanced the process of assessing surgical skills more than clinical supervisors did. Trainees placed more value upon the importance of PBA within surgical education than did clinical supervisors. The most striking difference of opinion was regarding the likelihood of choosing to continue using PBA in the future as a method of assessing surgical skills.

Objective Structured Assessment of Technical Skills user satisfaction and acceptability

Seventeen of 20 O&G clinical supervisors using OSATS (response rate 85%) provided almost complete responses about their satisfaction with OSATS and its feedback process. They provided predominantly positive responses about the use of the OSATS tool, but the greatest number of

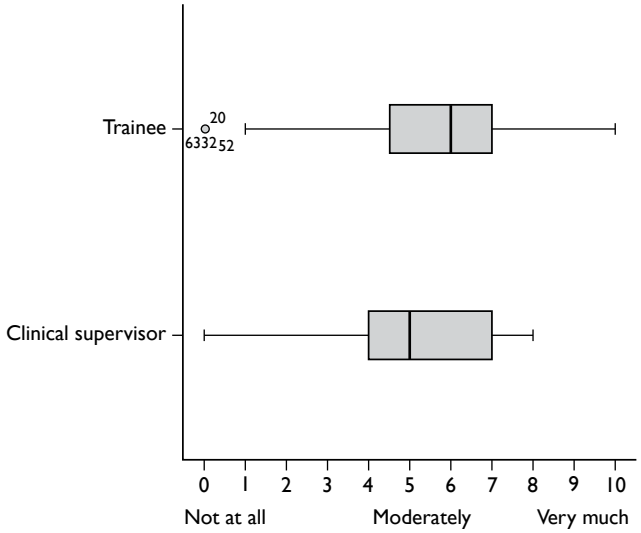


FIGURE 18 Comparing clinical supervisors' and trainees' opinion of enhancement of the assessment process with PBA.

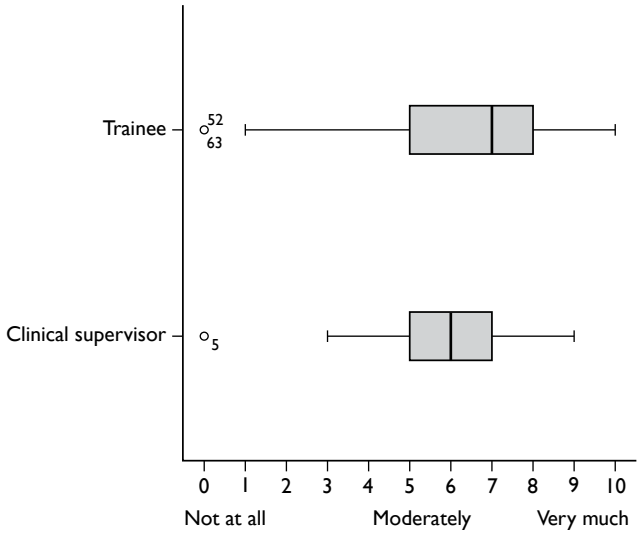


FIGURE 19 Comparing clinical supervisors' and trainees' opinion of the importance of PBA in surgical education.

negative responses were over its summative use (*Table 24*). The majority of them felt that OSATS was valuable as a tool for providing feedback (88%), for formative assessment (76%), for summative assessment (59%), and to support reflective practice (71%). There was a clear difference in agreement or strong agreement for the use of the OSATS tool as either a formative or a summative assessment (76% vs 59%), with more trainers disagreeing or strongly disagreeing with the use of OSATS as a summative assessment than as a formative assessment (24% vs 6%). In addition, the majority who responded felt that OSATS did not add time to their operating list (66%).

Overall, O&G clinical supervisors' satisfaction with OSATS was moderately good (*Table 25*). Clinical supervisors reported the extent to which OSATS enhanced their ability to assess as moderately good (median score 7, interquartile range 5–7). Their overall satisfaction with OSATS as an assessment tool was moderately good (median score 7, interquartile range 4.5–7.5). However, the 15 O&G clinical supervisors who had used both OSATS and PBA expressed a slightly greater overall satisfaction with PBA as an assessment tool (median score 7, interquartile

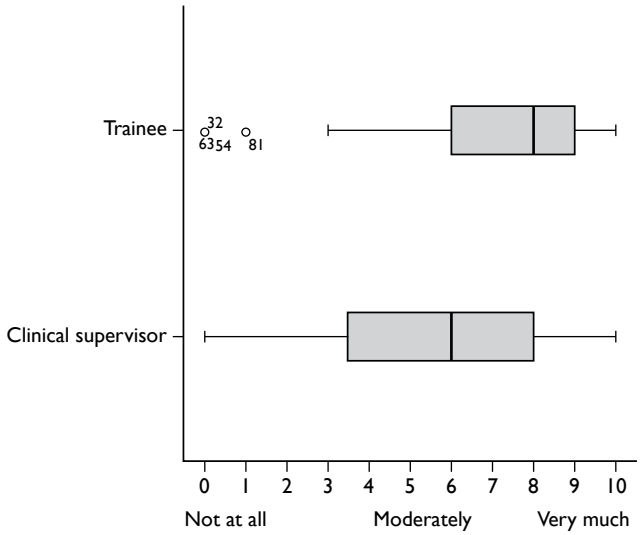


FIGURE 20 Comparing clinical supervisors' and trainees' opinion of the likelihood of using PBA in the future if given the choice.

TABLE 24 O&G clinical supervisors' opinions on the value of the OSATS tool (*n*, %)

Response to questionnaire statements	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
OSATS is a useful tool for providing feedback after an operation	0	0	2	13	2
	0%	0%	12%	76%	12%
OSATS is a valuable formative assessment tool	1	0	3	12	1
	6%	0%	18%	70%	6%
OSATS is a valuable summative assessment tool	2	2	3	10	0
	12%	12%	18%	59%	0%
OSATS is a useful tool to support reflective practice or to provide insight	0	1	4	11	1
	0%	6%	24%	65%	6%

TABLE 25 O&G clinical supervisors' opinions on their use of OSATS (medians shaded) (*n*, %)

	Not at all →					Moderately		→		Very much
Response to questions	0	1	2	3	4	5	6	7	8	9	10
To what extent do you think OSATS enhances your ability to assess your trainees?	0	1	0	1	0	4	1	7	3	0	0
	0%	6%	0%	6%	0%	24%	6%	41%	18%	0%	0%
I am satisfied with OSATS as an assessment tool	0	0	1	1	2	2	1	6	4	0	0
	0%	0%	6%	6%	12%	12%	6%	35%	24%	0%	0%
I am satisfied with PBA as an assessment tool	0	0	0	0	0	4	2	2	3	4	0
	0%	0%	0%	0%	0%	27%	13%	13%	20%	27%	0%

range 5–9). In response to the direct question 'Do you prefer OSATS or PBA?', five (29%) preferred PBA, two (12%) preferred OSATS and 10 (59%) expressed no preference.

Twenty-eight of 33 O&G trainees using OSATS (response rate 85%) provided almost complete responses about their satisfaction with OSATS and its feedback process. They provided mixed

responses about the use of OSATS (*Table 26*), which were less positive than those of their clinical supervisors. Similarly to the clinical supervisors' responses, the greatest number of negative responses related to its summative use. The only response that was answered positively by the majority of trainees (83%) was regarding the value of OSATS in providing feedback after an operation. Fifty per cent were of the opinion that OSATS is valuable as a formative assessment tool and 36% as a summative assessment tool. Almost half (46%) felt OSATS was a useful tool to support reflective practice.

The overall satisfaction of O&G trainees with OSATS was not particularly high (*Table 27*). The extent to which trainees felt that OSATS had enhanced the ability of their clinical supervisor to assess them was reported as moderate (median score 6, interquartile range 5–7). Their overall satisfaction with OSATS as an assessment tool was just moderate (median score 5, interquartile range 3.25–6.75). Directly comparing the overall satisfaction of trainees with clinical supervisors, 43% versus 65%, respectively, were satisfied with OSATS as an assessment tool (scores ≥ 6). However, the 24 trainees who had used both OSATS and PBA expressed a greater overall satisfaction with PBA as an assessment tool (median score 6.5, interquartile range 4–7.75). In response to the direct question 'Do you prefer OSATS or PBA?', 11 (39%) preferred PBA, five (18%) preferred OSATS and 12 (43%) expressed no preference.

Key differences in the responses of clinical supervisors and trainees regarding their experiences in using OSATS are illustrated in box-plots (*Figures 21–23*). Clinical supervisors perceived that OSATS enhanced the process of assessing surgical skills more than trainees. In addition, clinical

TABLE 26 O&G trainees' opinions on the value of the OSATS tool (*n*, %)

Response to questionnaire statements	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
OSATS is a useful tool for providing feedback after an operation	0	1	4	22	1
	0%	4%	14%	79%	4%
OSATS is a valuable formative assessment tool	0	6	8	14	0
	0%	21%	29%	50%	0%
OSATS is a valuable summative assessment tool	2	8	8	10	0
	7%	29%	29%	36%	0%
OSATS is a useful tool to support reflective practice or to provide insight	1	4	10	13	0
	4%	14%	36%	46%	0%

TABLE 27 O&G trainees' opinions on their use of OSATS (medians shaded) (*n*, %)

	Not at all →					Moderately		→		Very much
Response to questions	0	1	2	3	4	5	6	7	8	9	10
To what extent do you think OSATS enhances the ability of your clinical supervisor to assess you?	0	0	2	2	1	8	5	6	4	0	0
	0%	0%	7%	7%	4%	29%	18%	21%	14%	0%	0%
I am satisfied with OSATS as an assessment tool	0	1	2	4	1	8	5	4	3	0	0
	0%	4%	7%	14%	4%	29%	18%	14%	11%	0%	0%
I am satisfied with PBA as an assessment tool	0	1	1	2	3	1	4[a]	6[a]	5	1	0
	0%	4%	4%	8%	13%	4%	17%	25%	21%	4%	0%

a Median score 6.5.

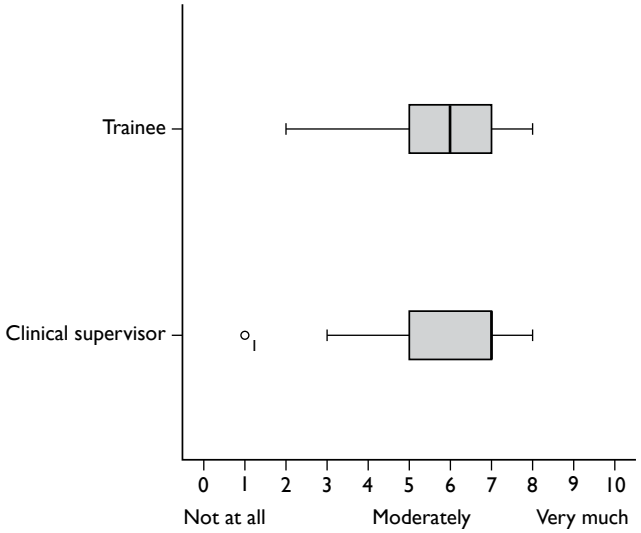


FIGURE 21 Comparing O&G clinical supervisors' and trainees' opinion of enhancement of the assessment process with OSATS.

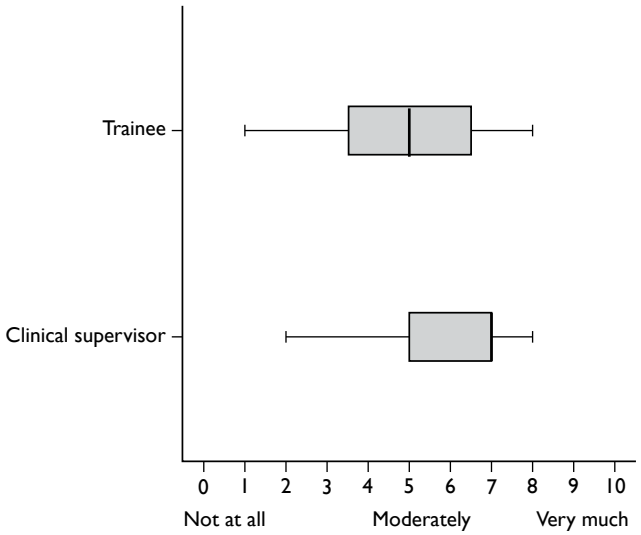


FIGURE 22 Comparing O&G clinical supervisors' and trainees' opinion of overall satisfaction with OSATS as an assessment tool.

supervisors expressed greater overall satisfaction than trainees with both OSATS and PBA as assessment tools. The O&G pattern of responses for OSATS, with O&G clinical supervisors' user satisfaction greater than that of O&G trainees, contrasts with the PBA user satisfaction results, in which non-O&G trainees reported greater satisfaction with PBA than their clinical supervisors.

Non-technical Skills for Surgeons user satisfaction and acceptability

Thirty of 56 anaesthetists (response rate 54%) provided almost complete responses about the NOTSS tool (*Table 28*). More agreed that they were able to easily assess interpersonal skills than cognitive skills using NOTSS (60% vs 27%). A majority agreed that the tool was useful for reflection (73%) and that it added value to the use of surgical skill assessments (60%). All of the 46% of anaesthetists who responded to this question felt NOTSS was useful for providing feedback. However, they were evenly split on whether or not it would enhance patient safety.

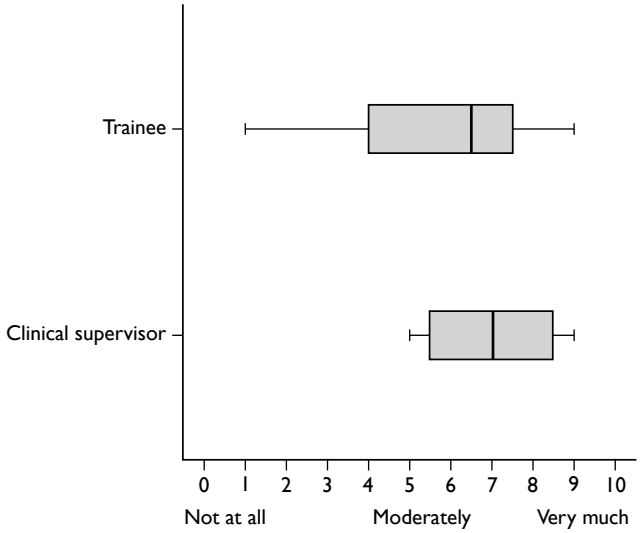


FIGURE 23 Comparing O&G clinical supervisors' and trainees' opinion of overall satisfaction with PBA as an assessment tool.

TABLE 28 Anaesthetists' opinions on their use of NOTSS (*n*, %)

Response to questionnaire statements	No response	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
NOTSS provides a common language to discuss non-technical skills	0	0	1	8	17	4
	0%	0%	3%	27%	57%	13%
It was easy to rate cognitive skills (situation awareness, decision-making)	0	0	12	10	8	0
	0%	0%	40%	33%	27%	0%
It was easy to rate interpersonal skills	0	0	4	8	18	0
	0%	0%	13%	27%	60%	0%
Using NOTSS added too much time to my list	0	5	16	8	0	1
	0%	17%	53%	27%	0%	3%
NOTSS is a useful tool to support reflective practice or to provide insight	0	0	0	8	19	3
	0%	0%	0%	27%	63%	10%
NOTSS is a valuable adjunct to tools that assess surgical skills, e.g. PBA, OSATS	2	0	0	10	17	1
	7%	0%	0%	33%	57%	3%
Routine use of the NOTSS system will enhance safety in the operating theatre	1	1	8	12	8	0
	3%	3%	27%	40%	27%	0%
NOTSS provides useful feedback for the trainee	16	0	0	0	10	4
	53%	0%	0%	0%	33%	13%

Only one anaesthetist felt that using NOTSS added too much time to the list, but anaesthetists declined to complete a NOTSS form on some occasions because they said they were too occupied with their clinical duties at the beginning of the procedure.

Twenty-six of 39 nurses (response rate 67%) provided complete responses about using NOTSS (*Table 29*). Almost all of them agreed that NOTSS is a valuable tool for reflective practice (96%) and a large majority saw its value as an adjunct to surgical skill assessments (81%). A large majority of nurses perceived that they were able to easily assess interpersonal and cognitive

TABLE 29 Scrub nurses' opinions on their use of NOTSS (*n*, %)

Response to questionnaire statements	No response	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
NOTSS provides a common language to discuss non-technical skills	0	0	0	4	22	0
	0%	0%	0%	15%	85%	0%
It was easy to rate cognitive skills (situation awareness, decision making)	0	0	2	2	20	2
	0%	0%	8%	8%	77%	8%
It was easy to rate interpersonal skills (communication and teamwork, leadership)	0	0	1	1	22	2
	0%	0%	4%	4%	85%	8%
Using NOTSS added too much time to my list	0	4	9	9	4	0
	0%	15%	35%	35%	15%	0%
NOTSS is a useful tool to support reflective practice or to provide insight	0	0	0	1	19	6
	0%	0%	0%	4%	73%	23%
NOTSS is a valuable adjunct to tools that assess surgical skills, e.g. PBA/OSATS	0	0	0	5	19	2
	0%	0%	0%	19%	73%	8%
Routine use of the NOTSS system will enhance safety in the operating theatre	0	0	1	8	12	5
	0%	0%	4%	31%	46%	19%
NOTSS provides useful feedback for the trainee	1	0	0	0	21	4
	4%	0%	0%	0%	81%	15%

skills (92% and 85% respectively). This was noticeably different from the anaesthetists where the proportion was 60% and 27% respectively for being easily able to assess interpersonal and cognitive skills. Sixty-five per cent of nurses felt that NOTSS would enhance patient safety. However, four scrub nurses felt that NOTSS assessments added too much time to their lists.

Reliability

Assessment ratings for assessment tools

In order to make it possible to appreciate the results, we have summarised the structure and rating profile of each tool.

Procedure-based assessment rating profile

The PBA instrument consists of a checklist of up to 62 items, each of which may be scored as 'performed to a satisfactory standard for CCT', 'development required', or 'not observed/not appropriate'. The items are clustered into the domains of 'consent', 'preoperative planning', 'preoperative preparation', 'exposure and closure', 'intraoperative technique' and 'postoperative management'. Each of the 15 index procedures has a different number and variety of task-specific items, so comparison at the item level is not possible. The structured checklist is therefore presented as the adjusted total item score (ATIS). This is a proportion of the perfect score of 1, based on the mean of the completed/appropriate items by converting satisfactory to 1 and development required to 0. In addition, assessors make a summary global assessment of progression towards independent practice. This is presented as the level, whereby level 1 is unable to perform the procedure, or unable to perform that part of the procedure that was observed, under supervision; level 2 is able to perform the procedure under supervision; level 3 is able to perform the procedure with minimum supervision (needed occasional help) and level 4 is competent to perform the procedure unsupervised (could deal with any complications that arose).

Objective Structured Assessment of Technical Skills rating profile

The OSATS instrument consists of three parts. Part 1 contains a checklist of up to 17 tasks unique to each index procedure. Each task may be scored as 'performed independently' (1), 'needs help' (0) or 'not applicable'. The checklist is presented as the adjusted total task score (ATTS) – calculated in exactly the same way as the ATIS for PBA as a proportion of 1. Part 2 is a generic assessment with eight global dimensions of performance:

1. 'respect for tissue'
2. 'time, motion and flow of operation and forward planning'
3. 'knowledge and handling of instruments'
4. 'suturing and knotting skills as appropriate for the procedure'
5. 'technical use of assistants'
6. 'relations with patient and the surgical team'
7. 'insight/attitude'
8. 'documentation of procedures'.

Each generic skill may be assessed at one of three levels (0, 1, 2) based on behaviourally anchored descriptors of performance. The generic section is presented as the adjusted total generic score (ATGS), based on the mean of the completed elements out of a perfect score of 2. In Part 3 the assessor is required to provide a 'pass/fail' judgement by which the trainee has 'achieved' (1) or 'failed to achieve' (0) the OSATS competency. According to the guidance material for OSATSs, this judgement is to be based on a clear assessment algorithm that requires all checklist items (Part 1) to be scored 'performed independently' plus the general skill for 'insight' scored at the highest level of 2.

Non-technical Skills for Surgeons rating profile

The NOTSS instrument consists of four domains of performance, which are described in the NOTSS tool as categories. These constitute the skill *categories*: 'situation awareness', 'decision-making', 'communication/teamwork' and 'leadership'. Within each category there are three *element scores*, which reflect behaviours relevant to the category. There are clear positive and negative behavioural examples provided for each category as these are designed to assess separate performance domains. There is also one overall *category score* which is an overall expert judgement based upon balancing the three element scores. Each category and element may be scored as 'poor' (1), 'marginal' (2), 'acceptable' (3) or 'good' (4). As the instrument is clearly intending to measure four separate domains, each NOTSS assessment is presented as four separate category scores: 'situation awareness domain score' (SDS), 'decision-making domain score' (DDS), 'communication/teamwork domain score' (CDS), and 'leadership domain score' (LDS). For the purposes of this analysis a 'global score' (GS) was also calculated as the mean of the four category scores to provide a *G* coefficient for the tool as a whole.

These 10 possible assessment ratings across the three tools are presented as a function of procedure and assessor type in *Table 30*.

Reliability of procedure-based assessment

Tables 31 and *32* display the variance component analyses for the ATIS and the level score respectively. For both scores, the 'overall' ability of the trainee (as judged across all procedures, cases and assessors) makes the largest contribution to any given score (variance 33% and 36% respectively).

Of the two ratings, the level score appears to provide a more reliable indicator of the trainee's performance than the ATIS because it highlights the trainee's aptitude for a particular procedure

DOI: 10.3310/hta15010

TABLE 30 Scores for study tools by assessor type for all specialties and index procedures

		Cardiac		Colorectal		Upper GI		Orthopaedic		O&G				Vascular			
		AVR	CABG	Anterior resection	Right hemicolectomy	Hernia	Laparoscopic cholecystectomy	Hip replacement	Knee replacement	Diagnostic laparoscopy	Elective caesarean	Evacuation of uterus	Urgent caesarean	Aortic aneurysm	Carotid endarterectomy	Varicose veins	Mean
PBA form																	
ATIS	Surgeon	1.0	0.9	0.9	0.9	0.7	0.8	1.0	1.0	0.8	0.9	0.9	0.9	0.8	0.8	0.8	**0.9**
	IA1	0.9	0.8	0.8	0.8	0.8	0.8	0.9	0.9	0.7	0.9	0.9	–	0.8	0.8	0.7	**0.8**
	IA2	–	0.9	0.8	0.6	0.7	0.8	1.0	0.8	0.8	0.8	0.9	–	0.9	0.6	0.7	**0.8**
	IA3	–	–	–	–	0.8	0.8	1.0	–	0.7	0.9	0.8	0.8	–	0.9	0.9	**0.8**
	IA4	–	–	–	–	0.9	0.8	–	–	–	–	–	–	–	–	–	**0.9**
	Mean	**1.0**	**0.9**	**0.8**	**0.8**	**0.8**	**0.8**	**1.0**	**0.9**	**0.8**	**0.9**	**0.9**	**0.9**	**0.8**	**0.8**	**0.8**	
Level	Surgeon	3.5	3.1	2.7	3.0	2.6	2.9	3.7	3.3	2.8	3.4	3.3	3.3	2.6	2.5	2.5	**3.0**
	IA1	3.6	3.1	2.6	3.0	3.5	3.0	3.7	3.2	2.3	3.4	3.0	–	2.7	2.6	2.3	**3.0**
	IA2	–	3.0	2.5	2.0	3.5	3.0	4.0	3.0	2.8	2.5	3.5	–	3.0	2.0	2.5	**2.9**
	IA3	–	–	–	–	3.4	3.1	4.0	–	2.7	3.3	3.1	3.5	–	3.0	3.0	**3.2**
	IA4	–	–	–	–	3.3	3.3	–	–	–	–	–	–	–	–	–	**3.3**
	Mean	**3.6**	**3.1**	**2.6**	**2.7**	**3.3**	**3.1**	**3.9**	**3.2**	**2.7**	**3.2**	**3.2**	**3.4**	**2.8**	**2.5**	**2.6**	
NOTSS form																	
SDS	Anaesthetist	3.6	3.3	3.4	3.3	3.1	3.5	4.0	–	2.9	3.0	2.9	2.7	3.2	3.0	3.2	**3.2**
	IA1	3.4	2.8	2.9	3.0	2.9	3.1	3.2	3.4	2.0	2.9	2.3	–	3.1	3.2	2.3	**2.9**
	IA2	–	3.4	2.4	2.0	–	–	3.0	3.0	2.5	2.7	2.4	–	3.0	2.6	2.7	**2.7**
	IA3	–	–	–	–	3.0	2.7	3.5	–	2.4	2.8	2.7	2.5	–	3.2	2.5	**2.8**
	Nurse	–	2.6	3.5	3.0	2.8	3.0	–	3.3	3.1	3.1	3.3	2.3	3.3	4.0	–	**3.1**
	SCP	4.0	3.4	–	–	–	–	–	–	–	–	–	–	–	–	–	**3.7**
	Mean	**3.7**	**3.1**	**3.1**	**2.8**	**3.0**	**3.1**	**3.4**	**3.2**	**2.6**	**2.9**	**2.7**	**2.5**	**3.2**	**3.2**	**2.7**	

continued

TABLE 30 Scores for study tools by assessor type for all specialties and index procedures *(continued)*

		Cardiac		Colorectal		Upper GI		Orthopaedic		O&G				Vascular			
		AVR	CABG	Anterior resection	Right hemicolectomy	Hernia	Laparoscopic cholecystectomy	Hip replacement	Knee replacement	Diagnostic laparoscopy	Elective caesarean	Evacuation of uterus	Urgent caesarean	Aortic aneurysm	Carotid endarterectomy	Varicose veins	Mean
DDS	Anaesthetist	3.3	3.2	3.2	3.2	3.2	3.5	–	–	2.9	3.0	2.9	2.8	3.3	2.9	3.3	**3.1**
	IA1	3.7	3.0	3.0	3.1	3.2	3.2	3.4	3.5	2.5	2.9	3.0	–	3.1	3.2	2.5	**3.1**
	IA2	–	3.0	3.0	3.0	–	–	3.2	3.0	3.0	3.0	3.0	–	2.8	2.8	3.1	**3.0**
	IA3	–	–	–	–	3.0	2.9	4.0	–	2.5	2.7	2.7	2.6	–	4.0	2.9	**3.0**
	Nurse	–	3.0	3.4	3.1	2.2	2.7	–	4.0	2.9	3.1	3.3	2.3	3.4	3.8	–	**3.1**
	SCP	3.8	3.5	–	–	–	–	–	–	–	–	–	–	–	–	–	**3.7**
	Mean	**3.6**	**3.1**	**3.2**	**3.1**	**2.9**	**3.1**	**3.5**	**3.5**	**2.8**	**2.9**	**3.0**	**2.6**	**3.2**	**3.3**	**3.0**	
CDS	Anaesthetist	3.6	2.9	3.8	3.2	3.1	3.5	3.8	–	2.8	3.0	3.0	2.4	3.2	3.1	3.2	**3.2**
	IA1	3.3	2.8	3.1	2.9	2.4	3.0	2.9	3.0	2.2	2.6	2.0	–	2.9	3.2	2.3	**2.8**
	IA2	–	3.2	2.9	2.3	–	–	3.0	3.0	3.1	3.0	2.8	–	2.7	2.7	2.7	**2.9**
	IA3	–	–	–	–	2.8	2.9	3.1	–	2.4	2.7	2.7	2.6	–	3.7	2.7	**2.8**
	Nurse	–	3.1	3.6	3.0	2.8	3.0	–	4.0	3.0	3.0	3.2	2.3	3.5	3.8	–	**3.2**
	SCP	4.0	2.9	–	–	–	–	–	–	–	–	–	–	–	–	–	**3.5**
	Mean	**3.6**	**3.0**	**3.4**	**2.9**	**2.8**	**3.1**	**3.2**	**3.3**	**2.7**	**2.9**	**2.7**	**2.4**	**3.1**	**3.3**	**2.8**	
LDS	Anaesthetist	3.7	3.2	3.6	3.2	3.4	3.7	4.0	–	3.2	3.1	2.9	2.6	3.4	3.3	3.2	**3.3**
	IA1	3.7	2.8	3.0	2.9	3.3	3.2	3.2	3.3	2.2	2.6	3.0	–	3.1	3.3	2.3	**3.0**
	IA2	–	3.0	3.0	3.0	–	–	3.0	2.8	3.2	2.9	3.0	–	2.9	2.9	2.7	**2.9**
	IA3	–	–	–	–	3.2	3.2	3.3	–	2.6	2.8	2.8	2.5	–	3.0	3.0	**2.9**
	Nurse	–	3.5	3.8	3.2	2.8	3.1	–	4.0	3.0	3.1	3.1	2.4	3.4	3.3	–	**3.2**
	SCP	4.0	3.1	–	–	–	–	–	–	–	–	–	–	–	–	–	**3.6**
	Mean	**3.8**	**3.1**	**3.4**	**3.1**	**3.2**	**3.3**	**3.4**	**3.4**	**2.8**	**2.9**	**3.0**	**2.5**	**3.2**	**3.2**	**2.8**	

DOI: 10.3310/hta15010

		Cardiac		Colorectal		Upper GI		Orthopaedic		O&G				Vascular			
		AVR	CABG	Anterior resection	Right hemicolectomy	Hernia	Laparoscopic cholecystectomy	Hip replacement	Knee replacement	Diagnostic laparoscopy	Elective caesarean	Evacuation of uterus	Urgent caesarean	Aortic aneurysm	Carotid endarterectomy	Varicose veins	Mean
GS	Anaesthetist	3.5	3.1	3.5	3.2	3.2	3.6	3.9	–	2.9	3.0	2.9	2.6	3.3	3.1	3.2	**3.2**
	IA1	3.5	2.9	3.0	3.0	2.9	3.1	3.2	3.3	2.2	2.7	2.6	–	3.0	3.2	2.4	**2.9**
	IA2	–	3.1	2.8	2.6	–	–	3.0	2.9	2.9	2.9	2.8	–	2.9	2.7	2.8	**2.9**
	IA3	–	–	–	–	3.0	2.9	3.4	–	2.5	2.7	2.7	2.6	–	3.5	2.8	**2.9**
	Nurse	–	3.0	3.6	3.1	2.7	2.9	–	3.8	3.0	3.1	3.2	2.3	3.4	3.8	–	**3.2**
	SCP	4.0	3.2	–	–	–	–	–	–	–	–	–	–	–	–	–	**3.6**
	Mean	**3.7**	**3.1**	**3.2**	**3.0**	**3.0**	**3.1**	**3.4**	**3.3**	**2.7**	**2.9**	**2.8**	**2.5**	**3.2**	**3.3**	**2.8**	
OSATS form																	
ATTS	Surgeon	–	–	–	–	–	–	–	–	0.9	0.9	1.0	0.8	–	–	–	**0.9**
	IA1	–	–	–	–	–	–	–	–	0.7	0.9	–	–	–	–	–	**0.8**
	IA2	–	–	–	–	–	–	–	–	–	0.9	0.9	–	–	–	–	**0.9**
	IA3	–	–	–	–	–	–	–	–	0.8	0.9	0.9	0.5	–	–	–	**0.8**
	Mean	–	–	–	–	–	–	–	–	**0.8**	**0.9**	**0.9**	**0.7**	–	–	–	
ATGS	Surgeon	–	–	–	–	–	–	–	–	1.5	1.6	1.8	1.5	–	–	–	**1.6**
	IA1	–	–	–	–	–	–	–	–	0.9	1.4	–	–	–	–	–	**1.2**
	IA2	–	–	–	–	–	–	–	–	–	1.7	1.6	–	–	–	–	**1.7**
	IA3	–	–	–	–	–	–	–	–	1.4	1.4	1.7	0.8	–	–	–	**1.3**
	Mean	–	–	–	–	–	–	–	–	**1.3**	**1.5**	**1.7**	**1.1**	–	–	–	
Pass/fail	Surgeon	–	–	–	–	–	–	–	–	0.6	0.8	0.8	0.0	–	–	–	**0.6**

CABG, coronary artery bypass grafting.

TABLE 31 Variance component analysis for PBA (ATIS)

Component	Estimate	%	Meaning
Var(trainee)	0.011	33	Trainee ability
Var(case)	0.003	9	Trainee case-to-case variation
Var(proccod)	0.002	7	Procedure difficulty
Var(assessor)	0.005	15	Assessor stringency
Var(designation)	0.000	0	Assessor designation stringency
Var(trainee*proccod)	0.004	13	Trainee procedure aptitude
Var(case*assessor)	0.007	21	Assessor subjectivity over case
Var(proccod*assessor)	0.001	2	Assessor subjectivity over procedure
Var(error)	0.000	0	Residual variation

TABLE 32 Variance component analysis for PBA (level score)

Component	Estimate	%	Meaning
Var(trainee)	0.267	36	Trainee ability
Var(case)	0.082	11	Trainee case-to-case variation
Var(proccod)	0.079	11	Procedure difficulty
Var(assessor)	0.024	3	Assessor stringency
Var(designation)	0.000	0	Assessor designation stringency
Var(trainee*proccod)	0.158	21	Trainee procedure aptitude
Var(case*assessor)	0.117	16	Assessor subjectivity over case
Var(proccod*assessor)	0.010	1	Assessor subjectivity over procedure
Var(error)	0.000	0	Residual variation

(variance 21% vs 13%), brings assessors of different stringencies into line (variance 3% vs 15%), and reduces assessor subjectivity (variance 16% vs 21%). Using either rating, trainees vary in their performance from case to case (variance 11% vs 9%) and the index procedures appear to differ in their difficulty (variance 11% vs 7%).

Table 33 shows the *D*-study table for the ATIS, comparing trainees on the same procedure and on a mix of procedures. Reliability for the same procedure ($G > 0.8$) can be achieved using four cases with one assessor per case, which equates to four individual expert judgements. Reliability on a mix of procedures can be achieved only with large numbers of cases and assessors.

Table 34 shows the *D*-study table for the level score, comparing trainees on the same procedure and on a mix of procedures. Reliability for the same procedure (using $G > 0.8$) can be achieved using three cases with one assessor per case, which equates to three individual expert judgements. Reliability on a mix of procedures can be achieved only with large numbers of cases and assessors.

Reliability of Objective Structured Assessment of Technical Skills

Tables 35 and *36* display the relative contributions that a number of key factors make to score variation for the ATTS and ATGS respectively. The pass/fail judgement was completed too infrequently by the supervising consultant (50% of cases) to allow useful variance component analysis. For both the ATTS and ATGS, the largest contribution to any given score across all procedures, cases and assessors is the overall ability of the trainee (variance 34% and 25% respectively). However, the influence of the other factors is very different for the two ratings.

TABLE 33a *D*-study for PBA (ATIS): comparing trainees on the same procedure

	Assessors per case		
Cases	1	2	3
1	0.42	0.54	0.60
2	0.65	0.75	0.79
3	0.76	0.83	0.85
4	0.82	0.87	0.89
5	0.85	0.90	0.91
6	0.88	0.91	0.93
7	0.89	0.93	0.94
8	0.91	0.94	0.95

TABLE 33b *D*-study for PBA (ATIS): comparing trainees on a mix of procedures

	Assessors per case		
Cases	1	2	3
1	0.38	0.49	0.54
2	0.58	0.65	0.68
3	0.66	0.71	0.73
4	0.70	0.74	0.76
5	0.73	0.76	0.78
6	0.75	0.78	0.79
7	0.76	0.79	0.79
8	0.77	0.79	0.80

TABLE 34a *D*-study for PBA (level score): comparing trainees on the same procedure

	Assessors per case		
Cases	1	2	3
1	0.55	0.64	0.67
2	0.76	0.81	0.83
3	0.85	0.88	0.89
4	0.89	0.91	0.91
5	0.91	0.93	0.93
6	0.93	0.94	0.94
7	0.94	0.95	0.95
8	0.95	0.95	0.96

TABLE 34b *D*-study for PBA (level score): comparing trainees on a mix of procedures

	Assessors per case		
Cases	1	2	3
1	0.47	0.53	0.56
2	0.62	0.65	0.66
3	0.67	0.69	0.70
4	0.70	0.71	0.72
5	0.72	0.73	0.73
6	0.73	0.73	0.74
7	0.73	0.74	0.74
8	0.74	0.74	0.75

Of the two, the ATGS appears to be a more reliable assessment because it highlights the trainee's aptitude for a particular procedure (variance 7% vs 0%) and reduces the extent to which case-to-case variation influences the score (variance 8% vs 26%). However, assessor stringency has a greater impact on the ATGS than on the ATTS (variance 14% vs 9%), and both are much higher than the 3% for the PBA level score. Both the ATTS and ATGS attract a significant score variance, which is attributable to assessor case subjectivity (variance 23% and 21%), although this is similar to the PBA ATIS and the level score (variance 21% and 16%).

The striking difference compared with the PBA scores is the large element of assessor stringency variation that appears to be attributable to the assessor's designation within OSATS scores (variance 4% and 16% compared with 0%). This reflects the differences in scores that are attributable to our two main assessor groups (clinical supervisors and independent assessors), whereby these assessor groups differ more in their stringency when using OSATS and agree more when using PBAs.

The ATTSs show 0% variance owing to both trainee procedure aptitude and assessor subjectivity over procedure. These are the two additional procedure-specific components used in the generalisability theory model for mix of procedures analysis. Therefore, the reliability of the ATTS, whether modelled for the same procedure or for a mix of procedures, generates the same *D*-study table estimates. For the ATTS, reliability ($G > 0.8$) can be achieved using five cases with one assessor per case (*Table 37*).

TABLE 35 Variance component analysis for OSATS (ATTS)

Component	Estimate	%	Meaning
Var(trainee)	0.0092	34	Trainee ability
Var(case)	0.0072	26	Trainee case-to-case variation
Var(proccod)	0.0010	4	Procedure difficulty
Var(assessor)	0.0023	9	Assessor stringency
Var(designation)	0.0012	4	Assessor designation stringency
Var(trainee*proccod)	0.0000	0	Trainee procedure aptitude
Var(case*assessor)	0.0064	23	Assessor subjectivity over case
Var(proccod*assessor)	0.0000	0	Assessor subjectivity over procedure
Var(error)	0.0000	0	Residual variation

TABLE 36 Variance component analysis for OSATS (ATGS)

Component	Estimate	%	Meaning
Var(trainee)	0.041	25	Trainee ability
Var(case)	0.014	8	Trainee case-to-case variation
Var(proccod)	0.012	7	Procedure difficulty
Var(assessor)	0.022	14	Assessor stringency
Var(designation)	0.026	16	Assessor designation stringency
Var(trainee*proccod)	0.012	7	Trainee procedure aptitude
Var(case*assessor)	0.034	21	Assessor subjectivity over case
Var(proccod*assessor)	0.004	3	Assessor subjectivity over procedure
Var(error)	0.000	0	Residual variation

TABLE 37 *D*-study for OSATS (ATTS): comparing trainees on the same procedure or on a mix of procedures

	Assessors per case		
Cases	1	2	3
1	0.37	0.44	0.48
2	0.59	0.65	0.67
3	0.70	0.75	0.76
4	0.77	0.80	0.81
5	0.81	0.84	0.85
6	0.84	0.86	0.87
7	0.86	0.88	0.89
8	0.88	0.89	0.90

Table 38 shows the *D*-study table for the ATGS, comparing trainees on the same procedure and on a mix of procedures. Reliability for the same procedure ($G > 0.8$) can be achieved using five cases with one assessor per case. Reliability on a mix of procedures can be achieved only with large numbers of cases and assessors.

The ATTS and ATGS for the same procedure require the same number of cases and assessors per case (equivalent to five expert judgements) to achieve $G > 0.8$. The ATGS shows better overall reliability for assessing within each procedure, considering both the variance component analysis

TABLE 38a *D*-study for OSATS (ATGS): comparing trainees on the same procedure

	Assessors per case		
Cases	1	2	3
1	0.37	0.50	0.56
2	0.61	0.71	0.76
3	0.72	0.80	0.83
4	0.79	0.85	0.87
5	0.83	0.88	0.90
6	0.86	0.90	0.92
7	0.88	0.91	0.93
8	0.89	0.92	0.94

TABLE 38b *D*-study for OSATS (ATGS): comparing trainees on a mix of procedures

	Assessors per case		
Cases	1	2	3
1	0.35	0.46	0.51
2	0.55	0.64	0.68
3	0.65	0.72	0.74
4	0.70	0.75	0.77
5	0.73	0.78	0.79
6	0.76	0.79	0.81
7	0.77	0.81	0.82
8	0.79	0.81	0.82

results and the slightly higher *G*-values within the *D*-study tables (0.83 vs 0.81). However, as the ATTS fails to reflect procedural variance, it would appear to be more reliable for making comparisons across procedures.

Reliability of Non-technical Skills for Surgeons

Table 39 displays the relative contributions that a number of key factors make to score variation for the NOTSS GS. The 'overall' ability of the trainee (as judged across all procedures, cases and assessors) makes the biggest contribution (variance 31%), but only by a small margin. The reason for the small margin is the fact that assessor stringency variation and assessor subjectivity strongly influence the GS (variance 27% and 20% respectively). NOTSS GSs show more variance owing to assessor stringency than either PBA or OSATS scores, although variance arising from assessor subjectivity is similar across all the tools (variance range 16%–23%).

The specific index procedure being observed has a much smaller influence on the score in the context of NOTSS-based non-technical skills assessment than it does in PBA-based surgical skills assessment (variance 3% compared with 11% for procedure difficulty and 5% compared with 21% for procedure aptitude). Baseline trainee case-to-case variation is similar to PBA ratings (variance 10% compared with 9% and 11%).

Table 40 summarises the factors that influence the scores across the four category ratings. Differences exist between the category ratings, some demonstrating greater or lesser proportions of true variance and case-to-case variation. The most striking observation is that, as with the GS, they are all highly influenced by assessor stringency and subjectivity, while demonstrating relatively procedure-independent scores.

Table 41 shows the *D*-study tables for the NOTSS GS, comparing trainees on the same procedure and on a mix of procedures. Because of the strong influence of assessor stringency and subjectivity, the reliability of NOTSS is lower than that for PBA. Reliability for the same procedure ($G > 0.8$) can be achieved using six cases and one assessor per case, i.e. six individual assessor judgements.

As procedure-specific factors exert a lesser influence on NOTSS scores than on either PBA or OSATS scores, the reliability of NOTSS scores on a mix of procedures is more similar to the reliability for the same procedure. Reliability between procedures ($G > 0.8$) can be achieved using eight cases and one assessor per case, i.e. eight assessor judgements.

TABLE 39 Variance component analysis for NOTSS (GS)

Component	Estimate	%	Meaning
Var(trainee)	0.111	31	Trainee ability
Var(case)	0.035	10	Trainee case-to-case variation
Var(proccod)	0.012	3	Procedure difficulty
Var(assessor)	0.097	27	Assessor stringency
Var(designation)	0.000	0	Assessor designation stringency
Var(trainee*proccod)	0.019	5	Trainee procedure aptitude
Var(case*assessor)	0.072	20	Assessor subjectivity over case
Var(proccod*assessor)	0.013	4	Assessor subjectivity over procedure
Var(error)	0.000	0	Residual variation

TABLE 40 Variance component analysis for NOTSS (separate category scores)

	Component score (% variance)			
Component	SDS	DDS	CDS	LDS
Trainee ability	25	29	20	29
Trainee case-to-case variation	12	5	14	6
Procedure difficulty	3	0	3	5
Assessor stringency	21	22	28	21
Assessor designation stringency	3	0	0	0
Trainee procedure aptitude	3	8	9	2
Assessor subjectivity over case	25	32	27	33
Assessor subjectivity over procedure	8	5	0	4
Residual variation	0	0	0	0

TABLE 41a *D*-study for NOTSS (GS): comparing trainees on the same procedure

	Assessors per case		
Cases	1	2	3
1	0.35	0.48	0.55
2	0.57	0.69	0.74
3	0.68	0.78	0.82
4	0.75	0.83	0.86
5	0.79	0.86	0.89
6	0.82	0.88	0.90
7	0.85	0.90	0.92
8	0.86	0.91	0.93

TABLE 41b *D*-study for NOTSS (GS): comparing trainees on a mix of procedures

	Assessors per case		
Cases	1	2	3
1	0.34	0.46	0.52
2	0.53	0.64	0.69
3	0.63	0.72	0.76
4	0.69	0.77	0.80
5	0.73	0.80	0.82
6	0.76	0.82	0.84
7	0.78	0.83	0.85
8	0.80	0.84	0.86

Validity

The evidence for the validity of the three assessment methods is presented by using various hypotheses to confirm or refute our interpretation of assessment scores. Content validity has not been directly evaluated within this study as the items for each assessment method have been systematically derived to reflect the key competencies of surgical performance.

Hypothesis 1: Correlation within and between assessment methods

Scores obtained by each assessment will correlate with the other assessments that set out to measure the same aspect of performance. These correlations will operate within instruments (internal structure) and between instruments.

Where an assessment instrument or part of an instrument is setting out to measure a particular construct, all the items addressing that construct should correlate more highly with each other than they do with other items. This intercorrelation within domains gives the instrument a 'factor structure' which can be tested by factor analysis. If the expected factor structure is observed, then it demonstrates that the items are being used in a rational and non-random fashion and provides one source of evidence that those items are being used to reflect the intended construct. Most of the items on PBA and OSATS are not designed to measure a stable construct; they are simply checklists of the actions or behaviours that should be observed during a particular surgical procedure. However, the generic section of the OSATS instrument is intended to reflect globally relevant behaviours that can be linked to constructs. The NOTSS instrument sets out explicitly to measure four constructs ('situation awareness', 'decision-making', 'communication/teamwork' and 'leadership'), which should be reflected in the factor structure of the scores.

There is no prior theory on the factor structure of the OSATS generic items. *Table 42* displays the observed factor structure. All the items are explained by a single factor (factor 1) with the exception of two (the factor with the highest loading is highlighted in bold). The items intended to measure 'insight/attitude' and 'documentation of procedures' each load on to a separate factor (factors 2 and 3 respectively). These are the only two items that are not designed to reflect the operative process, so the observed structure seems rational.

Table 43 displays the factor structure of the NOTSS scores (factor loadings < 0.5 are not displayed for clarity). The observed factors match the intended constructs almost perfectly, with only one item (setting and maintaining standards) falling outside the pattern by loading almost equally on two factors ('situation awareness' and 'leadership'). This suggests that assessors are rating similarly the separate elements within each of the four categories as separate domains of performance.

Table 44 displays the correlations across cases between the eight ratings available from the three instruments. There were insufficient data to correlate PBA with OSATS, as to carry out both a PBA and an OSATS during an O&G case would have required the presence of two independent assessors, which was rarely possible. All the ratings correlate strongly and statistically significantly. The strongest correlations are within each instrument as might be expected (0.73 between ATIS and level on PBA, 0.74–0.76 between SDS, DDS, CDS and LDS on NOTSS, and 0.84 between ATTS and ATGS on OSATS). The correlations between ratings across instruments

TABLE 42 OSATS factor structure

		Factor		
Item	Meaning	1	2	3
1	Respect for tissue	**0.381**	0.370	0.356
2	Time, motion and flow of operation and forward planning	**0.712**	0.323	
3	Knowledge and handling of instruments	**0.631**		
4	Suturing and knotting skills as appropriate for the procedure	**0.407**	0.404	
5	Technical use of assistants	**0.835**		
6	Relations with patient and the surgical team	**0.580**		0.317
7	Insight/attitude			**0.835**
8	Documentation of procedures		**0.774**	

TABLE 43 NOTSS factor structure

		Factor			
Item	Meaning	1	2	3	4
1	SITUATION AWARENESS			0.76	
2	Gathering information			0.83	
3	Understanding information			0.59	
4	Projecting and anticipating future state	0.50		0.60	
5	DECISION-MAKING	0.82			
6	Considering options	0.72			
7	Selecting and communicating options	0.76			
8	Implementing and reviewing decisions	0.76			
9	COMMUNICATION AND TEAMWORK		0.83		
10	Exchanging information		0.73		
11	Establishing shared understanding		0.69		
12	Co-ordinating team activities		0.67		
13	LEADERSHIP				0.76
14	Setting and maintaining standards			0.54	0.54
15	Supporting others				0.73
16	Coping with pressure				0.68

are all between 0.40 and 0.67 and remain statistically significant even if correcting for multiple correlation tests using the Bonferroni method by raising the threshold for p to 0.002 (0.05/24). Because the instruments are designed to measure different aspects of surgical performance, we should not expect complete correlation – but the hypothesis is fulfilled. It is interesting to note that the strongest correlations between the non-technical instrument NOTSS and the 'technical' instruments PBA and OSATS are in the 'decision-making' domain in both cases.

Hypothesis 2: Construct validity

Scores will increase with duration of surgical training and number of similar procedures performed (experience).

Table 45 displays the correlations across trainees between age, year of ST, total years of surgical experience (including further division into years of surgical experience in the UK and overseas), with their mean score on all eight ratings. Correlations that are statistically significant are highlighted in bold. Year of ST is the strongest predictor of every outcome where there is a positive association. Total years of surgical training is also a strong predictor, but years of overseas training make no contribution at all. The strongest and most statistically significant relationships are observed with the PBA ratings. In contrast, none of the factors appears to be a significant predictor of OSATS ratings. All predict 'decision-making' and 'situation awareness' ratings in the NOTSS system, but only year of ST and UK years of surgical training reach significance as a predictor of 'communication/teamwork' or 'leadership' after p-value after correcting for multiple correlation tests using the Bonferroni method by raising the threshold for p to 0.01 (0.05/5).

Table 46 displays the correlations between the number of cases of an index procedure that a trainee has performed ever and recently (last 6 months) on recruitment to the study, and his or her ratings on that particular index procedure. All the index procedures have been pooled for this analysis because, as *Table 7* shows, the total number of cases within many index procedures is quite small. Correlations that are statistically significant are highlighted in bold. The strongest and most statistically significant relationships are observed with the PBA ratings. After correcting

TABLE 44 Rating correlations between the assessment instruments

		PBA		NOTSS				OSATS	
		Level	ATIS	SDS	DDS	CDS	LDS	ATTS	ATGS
PBA									
ATIS	Pearson correlation			0.54	0.59	0.51	0.53		
	Significance (two-tailed test)			0.000	0.000	0.000	0.000		
	n			315	308	317	315		
Level	Pearson correlation		0.73	0.48	0.55	0.43	0.49		
	Significance (two-tailed test)		0.000	0.000	0.000	0.000	0.000		
	n		351	315	308	317	315		
NOTSS									
SDS	Pearson correlation				0.76	0.74	0.74	0.65	0.58
	Significance (two-tailed test)				0.000	0.000	0.000	0.000	0.000
	n				391	401	399	90	90
DDS	Pearson correlation					0.75	0.75	0.67	0.57
	Significance (two-tailed test)					0.000	0.000	0.000	0.000
	n					392	390	88	88
CDS	Pearson correlation						0.74	0.48	0.40
	Significance (two-tailed test)						0.000	0.000	0.000
	n						401	90	90
LDS	Pearson correlation							0.56	0.50
	Significance (two-tailed test)							0.000	0.000
	n							90	90
OSATS									
ATTS	Pearson correlation								0.84
	Significance (two-tailed test)								0.000
	n								90
ATGS	Pearson correlation								
	Significance (two-tailed test)								
	n								

for multiple correlation tests using the Bonferroni method by raising the threshold for p to 0.025 (0.05/2), procedural experience does not appear to be a significant predictor of OSATS ratings. Total experience only predicts 'decision-making' and 'situation awareness' ratings in the NOTSS system, but recent experience also predicts 'communication/teamwork' and 'leadership'.

Many of the predictors examined above are themselves correlated: for example, older trainees are likely to be in higher training years, have more years of surgical experience, and have performed more of the index cases. Regression analysis leaves year of ST and the number of recent index procedures performed as significant independent predictors of performance using PBA level. Hypothesis 2 is confirmed for PBA and NOTSS but not for OSATS.

Hypothesis 3: Relationship of ratings with case outcomes

Higher-scoring operations will result in less operative time and blood loss, fewer perioperative and postoperative complications and a shorter length of hospital stay.

As *Table 9* showed, the different index procedures are associated with very different case outcomes. Five of these outcomes (operating time, blood loss, length of hospital stay, HDU and

TABLE 45 Correlation between scores and trainees' age and training

		Age	ST level	Total years' surgical training	UK years' surgical training	Non-UK years' surgical training
PBA						
Number of trainees		73	73	77	75	75
ATIS	Pearson correlation	**0.36**	**0.53**	**0.34**	**0.41**	0.03
	Significance (two-tailed test)	0.00	0.00	0.00	0.00	0.81
Level	Pearson correlation	**0.51**	**0.71**	**0.51**	**0.56**	0.12
	Significance (two-tailed test)	0.00	0.00	0.00	0.00	0.32
NOTSS						
Number of trainees		75	75	79	77	77
SDS	Pearson correlation	**0.29**	**0.57**	**0.30**	**0.49**	–0.15
	Significance (two-tailed test)	0.01	0.00	0.01	0.00	0.19
DDS	Pearson correlation	**0.31**	**0.57**	**0.34**	**0.47**	–0.04
	Significance (two-tailed test)	0.01	0.00	0.00	0.00	0.73
CDS	Pearson correlation	0.22	**0.40**	0.20	**0.36**	–0.14
	Significance (two-tailed test)	0.06	0.00	0.07	0.00	0.21
LDS	Pearson correlation	0.18	**0.46**	0.23	**0.40**	–0.14
	Significance (two-tailed test)	0.11	0.00	0.04	0.00	0.22
OSATS						
Number of trainees		28	30	29	28	28
ATTS	Pearson correlation	0.16	0.15	0.16	0.27	0.01
	Significance (two-tailed test)	0.42	0.42	0.39	0.16	0.96
ATGS	Pearson correlation	–0.01	–0.01	0.08	0.11	0.09
	Significance (two-tailed test)	0.97	0.96	0.67	0.57	0.64

ICU admission) are examined in *Table 47* in relation to the assessment scores for each procedure. A regression analysis shows that the procedure is the predominant determinant of all these outcome measures. This means that it is not possible to pool the different procedures to examine the relationship between ratings and outcomes. Outcomes must be compared with ratings on a procedure-by-procedure basis.

Using the ratings 'level' for PBA, 'GS' for NOTSS, and 'ATGS' for OSATS, there are 118 possible associations to test. *Table 47* shows which of these relationships reach statistical significance at the $p = 0.05$ level (those highlighted in bold). Chance alone would result in six associations in each direction. Because the ratings are positive but the outcomes are negative (worse with

TABLE 46 Correlation between scores and trainees' procedural experience

		Number of procedures done	
		Total	Recent
PBA			
Number of trainee procedure combinations		110	113
ATIS	Pearson correlation	**0.31**	**0.34**
	Significance (two-tailed test)	0.00	0.00
Level	Pearson correlation	**0.50**	**0.49**
	Significance (two-tailed test)	0.00	0.00
NOTSS			
Number of trainee procedure combinations		118	122
SDS	Pearson correlation	**0.22**	**0.36**
	Significance (two-tailed test)	0.02	0.00
DDS	Pearson correlation	**0.28**	**0.35**
	Significance (two-tailed test)	0.00	0.00
CDS	Pearson correlation	0.06	**0.25**
	Significance (two-tailed test)	0.50	0.01
LDS	Pearson correlation	0.13	**0.33**
	Significance (two-tailed test)	0.16	0.00
OSATS			
Number of trainee procedure combinations		38	40
ATTS	Pearson correlation	0.28	0.12
	Significance (two-tailed test)	0.08	0.45
ATGS	Pearson correlation	0.31	0.06
	Significance (two-tailed test)	0.06	0.71

increasing size), the hypothesis anticipates negative associations. In this analysis, 13 negative associations and four counterintuitive positive associations were observed. This cannot be said to provide statistical confirmation for hypothesis 3 because there is no appropriate single statistical test.

Hypothesis 4: Interprocedural differences

Mean scores, and scores for each element, will not be significantly different across the nine different procedures.

Table 30 displays the scores across procedures as well as across different rating types and rater designations (clinical supervisors, independent assessors, anaesthetists and scrub nurses). Some procedures received lower mean scores then others. However, examining the *G*-study tables shows that the differences are related to the uneven distribution of rater stringency and trainee experience across the procedures rather than to a validity problem. Var(proc) accounts

TABLE 47 Correlation between scores and outcomes by specialty and index procedure

		Cardiac		Colorectal		Upper GI		Orthopaedic		O&G				Vascular		
		AVR	CABG	Anterior resection	Right hemicolectomy	Hernia	Laparoscopic cholecystectomy	Hip replacement	Knee replacement	Diagnostic laparoscopy	Elective caesarean	Evacuation of uterus	Urgent caesarean	Aortic aneurysm	Carotid endarterectomy	Varicose veins
Records	5	35	14	11	19	43	18	18	73	60	45	5	15	25	51	
Length of operation (minutes)	Level	**0.92 (*p*=0.03)**	–0.05 (*p*=0.76)	0.18 (*p*=0.52)	–0.39 (*p*=0.24)	–0.42 (*p*=0.07)	–0.24 (*p*=0.12)	**–0.47 (*p*=0.05)**	**–0.76 (*p*=0.00)**	–0.08 (*p*=0.62)	–0.23 (*p*=0.18)	0.20 (*p*=0.44)	**–0.95 (*p*=0.05)**	–0.22 (*p*=0.43)	–0.11 (*p*=0.58)	**–0.45 (*p*=0.00)**
	GS	0.12 (*p*=0.88)	–0.03 (*p*=0.84)	0.35 (*p*=0.23)	**–0.71 (*p*=0.01)**	–0.09 (*p*=0.80)	–0.21 (*p*=0.22)	–0.29 (*p*=0.27)	**–0.52 (*p*=0.04)**	0.07 (*p*=0.56)	–0.09 (*p*=0.50)	–0.20 (*p*=0.19)	–0.73 (*p*=0.16)	0.14 (*p*=0.61)	–0.02 (*p*=0.94)	–0.24 (*p*=0.12)
	ATGS									–0.24 (*p*=0.18)	–0.31 (*p*=0.11)	**–0.72 (*p*=0.00)**				
Blood loss (ml)	Level			0.10 (*p*=0.73)	–0.42 (*p*=0.20)		0.14 (*p*=0.40)	–0.44 (*p*=0.08)	–0.28 (*p*=0.30)	0.02 (*p*=0.90)	–0.19 (*p*=0.27)	–0.11 (*p*=0.67)	0.10 (*p*=0.89)	–0.12 (*p*=0.68)	–0.04 (*p*=0.86)	–0.30 (*p*=0.06)
	GS			0.39 (*p*=0.18)	–0.36 (*p*=0.27)		0.02 (*p*=0.91)	–0.19 (*p*=0.48)	–0.24 (*p*=0.43)	0.08 (*p*=0.52)	–0.06 (*p*=0.62)	**–0.35 (*p*=0.02)**	0.18 (*p*=0.77)	**–0.81 (*p*=0.00)**	–0.20 (*p*=0.41)	–0.08 (*p*=0.67)
	ATGS									–0.18 (*p*=0.31)	–0.08 (*p*=0.65)	**–0.64 (*p*=0.00)**				
Length of stay (days)	Level	0.48 (*p*=0.41)	0.04 (*p*=0.83)	–0.30 (*p*=0.29)	–0.10 (*p*=0.76)	0.08 (*p*=0.74)	–0.01 (*p*=0.96)	0.10 (*p*=0.69)	**0.56 (*p*=0.02)**	–0.11 (*p*=0.51)	0.11 (*p*=0.53)		0.47 (*p*=0.53)	0.08 (*p*=0.76)	**–0.43 (*p*=0.03)**	–0.10 (*p*=0.47)
	GS	0.03 (*p*=0.97)	0.28 (*p*=0.12)	–0.07 (*p*=0.80)	0.20 (*p*=0.56)	0.20 (*p*=0.46)	–0.07 (*p*=0.68)	0.29 (*p*=0.25)	0.48 (*p*=0.06)	–0.01 (*p*=0.94)	**0.30 (*p*=0.02)**	0.07 (*p*=0.62)	0.55 (*p*=0.34)	–0.09 (*p*=0.75)	–0.37 (*p*=0.10)	–0.20 (*p*=0.19)
	ATGS									–0.07 (*p*=0.71)	0.23 (*p*=0.23)	0.10 (*p*=0.58)				
HDU stay (days)	Level	0.54 (*p*=0.34)	–0.21 (*p*=0.24)	–0.12 (*p*=0.68)	–0.37 (*p*=0.25)			0.18 (*p*=0.49)	–0.14 (*p*=0.59)		–0.15 (*p*=0.40)			0.32 (*p*=0.27)	–0.36 (*p*=0.08)	
	GS	0.87 (*p*=0.12)	0.06 (*p*=0.76)	0.51 (*p*=0.07)	**–0.64 (*p*=0.03)**			0.12 (*p*=0.64)	0.08 (*p*=0.76)		**0.30 (*p*=0.02)**			0.08 (*p*=0.79)	**–0.49 (*p*=0.02)**	
	ATGS										0.15 (*p*=0.44)					
ICU stay (days)	Level		0.09 (*p*=0.59)													
	GS	–0.35 (*p*=0.65)	0.18 (*p*=0.34)													
	ATGS															

for 3%–11% of score variance, which shows that, after controlling for trainee and assessor differences, very little score variation is due to differences between the index procedures.

Five index procedures received a mean PBA level score < 2.8 (anterior resection, right hemicolectomy, diagnostic laparoscopy, carotid endartectomy and varicose veins). The highest scoring procedures were AVR and hip replacement (mean PBA level 3.5 and 3.8 respectively). *Table 48* examines the relationship between PBA level scores for each index procedure and the mean level of ST for the subgroup of trainees performing them. The clinical supervisor's judgement as to whether the case was 'more difficult than usual' is also included. This illustrates that the index procedures with the lowest PBA scores either are performed by more junior trainees (mean ST level < 5.5) or have a higher proportion (> 50%) of cases judged to be more difficult than usual. Conversely, the highest PBA-scored procedures were performed by more senior trainees, with a lower proportion of cases judged to be more difficult than usual.

Hypothesis 5: Stringency of assessors by designation

Assessor designation will not affect assessment stringency.

Table 30 displays the scores across assessor designations as well as across different rating types and procedures. Some designations gave lower mean scores than others. However, the *G*-study tables show that the differences, in all but one case, are related to the uneven distribution of individual rater stringency (not designation), trainee experience and procedure mix across the designations rather than to a validity problem. Var(designation) accounts for 0%–4% of score variance, with the exception of the OSATS ATGS. ATGS was significantly influenced by assessor designation, which contributed 16% to score variance. In summary, assessor designation does not affect assessment stringency after controlling for other variables, except in the case of the OSATS ATGS. The hypothesis is confirmed.

Hypothesis 6: Effect of assessment on performance

Assessment (plus or minus video-recording equipment) will not affect the performance of trainees.

Trainees' and clinical supervisors' perspectives were sought on whether surgical performance was affected by the assessment conditions from case 96 onwards, representing a total of 342 cases. Both trainees and clinical supervisors were asked for their judgement ('yes' or 'no') as to whether the trainee performed differently for each separate case that was assessed within the study. Some clinical supervisors were unable to comment ($n = 50$) if they had not previously supervised a trainee either operating or performing a particular index procedure. Similarly, some trainees felt unable to comment ($n = 14$) for corresponding reasons. The anaesthetists, nurses and SCPs were not asked, therefore only PBAs and OSATSs are considered.

Table 49 displays the responses. Trainees felt that their performance was affected in 70 (20%) of the cases where they were able to give a response. Clinical supervisors felt that performance was affected in 41 (12%) of cases where they were able to give a response. In total, there were 92 cases in which either trainees or supervisors judged that performance was affected. This represents 27% of cases where the question was asked.

Owing to the complex hierarchical nature of the data (interdependent variables), there is no feasible statistical approach for testing this hypothesis robustly on the basis of actual score differences. However, we do provide the following post hoc raw comparison of scores: (a) concerned that assessment has affected performance versus unconcerned that assessment has affected performance (subjective); and (b) video-recorded versus not video-recorded (objective).

TABLE 48 Relationship between PBA level scores, year of training and case difficulty by index procedure

	AVR	CAGB	Anterior resection	Right hemicolectomy	Hernia	Laparoscopic cholecystectomy	Hip replacement	Knee replacement	Diagnostic laparoscopy	Elective caesarean	Evacuation of uterus	Urgent caesarean	Aortic aneurysm	Carotid endarterectomy	Varicose veins
PBA level score															
Mean	3.5	3.1	2.6	2.7	3.3	3.1	3.8	3.2	2.6	3.1	3.2	3.4	2.8	2.5	2.6
Maximum	4	4	4	4	4	4	4	4	4	4	4	4	3	3	4
Minimum	2	2	2	2	1	1	2	2	2	2	1	2	2	1	1
ST level															
Mean	8	6.7	5.5	4.2	3.9	5.0	7.2	6.2	3.5	4.1	2.6	2.6	6	5.0	2.7
Maximum	8	8	8	8	7	7	8	8	7	7	7	3	7	7	4
Minimum	8	3	3	3	1	2	3	3	2	2	0	2	3	2	2
% Case difficulty	0	12	62	60	44	46	27	37	32	47	7	50	60	38	16

CABG, coronary artery bypass grafting.

We do not offer these comparisons to answer hypothesis 6, rather to raise new hypotheses for further work.

Table 50 compares the mean scores given on those cases where either the trainee or clinical supervisor responded 'yes' with the mean score of the remaining cases. Cases in which either the trainee or supervisor responded 'yes' received lower scores on every measure. These differences reach statistical significance for all but the ATTS and ATGS, even after applying the Bonferroni correction [$p < 0.006$ (0.05/8)].

A total of 120 cases were video recorded during the study, representing 27% of all recruited cases ($n = 437$). Considering the cases where trainee and trainer opinions were sought on the effect of assessment on performance ($n = 342$), 91 (26.6%) of these cases were video recorded.

Table 51 compares the mean scores given on those cases that were video recorded with the mean score of those that were not. Cases that were video recorded received similar scores to cases that were not. None of the small differences in the table reached statistical significance after applying the Bonferroni correction [$p < 0.006$ (0.05/8)].

TABLE 49 Trainees' and clinical supervisors' perspectives on the impact of assessment

		Clinical supervisor					
		No	Yes	Not relevant	Not done	Total trainee	% of those asked
Trainee	No	208	22	28	0	258	75
	Yes	41	19	10	0	70	20
	Not relevant	2	0	12	0	14	4
	Not done	0	0	0	95	95	–
Total clinical supervisor		251	41	50	95	437	
% of those asked		73	12	15	–		

TABLE 50 Relationship between scores and perspectives on assessment

	Did assessment affect performance?							
	No			Yes			Independent samples *t*-test	
	Mean	*n*	SD	Mean	*n*	SD	*t*	Significance
PBA								
Level	2.98	187	0.75	2.63	72	0.66	–3.64	0.000
ATIS	0.84	187	0.14	0.74	72	0.17	–4.22	0.000
NOTSS								
SDS	2.94	231	0.56	2.63	87	0.52	–4.71	0.000
DDS	2.97	223	0.55	2.74	85	0.47	–3.64	0.000
CDS	2.87	231	0.57	2.68	87	0.52	–2.83	0.005
LDS	2.99	231	0.54	2.72	87	0.50	–4.17	0.000
OSATS								
ATTS	0.90	64	0.13	1.41	23	0.17	–1.79	0.083
ATGS	1.63	64	0.30	1.41	23	0.36	–2.58	0.014

SD, standard deviation.

TABLE 51 Relationship between scores and perspectives on video assessment

	Was the case video recorded?							
	No			Yes			Independent samples *t*-test	
	Mean	*n*	SD	Mean	*n*	SD	*t*	Significance
PBA								
Level	2.94	264	0.81	2.99	87	0.77	0.43	0.669
ATIS	0.83	264	0.15	0.82	87	0.15	–0.54	0.590
NOTSS								
SDS	2.94	288	0.59	2.78	113	0.54	–2.60	0.010
DDS	3.00	279	0.56	2.83	113	0.54	–2.68	0.008
CDS	2.87	290	0.57	2.76	113	0.59	–1.70	0.091
LDS	3.00	288	0.57	2.86	113	0.50	–2.54	0.012
OSATS								
ATTS	0.87	56	0.15	0.88	34	0.13	0.35	0.727
ATGS	1.54	56	0.36	1.64	34	0.25	1.51	0.135

SD, standard deviation.

A significant minority (20% of trainees and 12% of supervisors) perceived that assessment affected their performance. Interestingly, trainees who felt that observation affected their performance (subjective) performed less well than trainees who did not feel affected. However, trainees who were video recorded (objective) did not perform less well than trainees who were not video recorded. This post hoc analysis cannot answer hypothesis 6, but it does raise the interesting hypothesis that observation or video recording do not affect actual performance, rather a poor performance may be attributed to assessment or video recording.

Video recordings

A total of 120 cases were recorded during the study, representing 27% of all recruited cases. Some cases were lost because the patient consented to the observational part of the study but not to video recording. Other cases were not filmed because of the lack of availability of the audiovisual technician or owing to logistical difficulties posed by setting up video equipment for cases at different hospital sites or for cases that were changed at short notice. *Table 52* summarises the number of video recordings for each index procedure.

Recordings were attempted for all index procedures but proved too difficult within orthopaedic theatres because camera access to the operative field was prevented by the high vertical laminar flow system of the Charnley tent used to limit contamination. In cardiac surgery, the amount of equipment surrounding the operating table, including the bypass pump and ultrasound machine, also prevented adequate views of the operative field.

With regard to the other specialty procedures, the operative field was frequently obscured by the surgeons operating during anterior resection, right hemicolectomy, hernia repair and aortic aneurysm repair. Therefore, video assessment of these procedures was not pursued, with efforts to video alternative procedures maximised. Good views of the operative field were obtained during caesarean section, carotid endarterectomy and varicose vein operations (saphenofemoral ligation). Varicose vein operations were not routinely recorded as we have previously shown

TABLE 52 Distribution of video-recorded index procedures

	Cardiac		Colorectal		Upper GI	Orthopaedic		O&G			Vascular		
	AVR	CABG	Anterior resection	Right hemicolectomy	Hernia	Laparoscopic cholecystectomy	Hip replacement	Knee replacement	Diagnostic laparoscopy	Elective caesarean	Evacuation of uterus	Urgent caesarean	Aortic aneurysm	Carotid endarterectomy	Varicose veins
Number of cases	0	0	1	3	7	11	2	1	34	26	16	4	4	6	5
% of cases	0	0	7	27	37	26	11	6	47	43	36	80	27	24	10
% of total cases	0	0	0.8	2.5	5.8	9.2	1.6	0.8	28	21.6	13.3	3.3	3	5	4.2

CABG, coronary artery bypass grafting.

good reliability of video recordings for this particular procedure.[53] The best operative views were obtained during laparoscopic cholecystectomy and diagnostic laparoscopy, but even for these procedures we were unable to record sufficient numbers of individual index procedures for a dependable estimate of the reliability of blinded video assessment.

Chapter 4

Discussion

Reliability of assessment methods

Procedure-based assessment

The excellent reliability of $G > 0.8$ for three PBA level scores, using one assessor per case for the same index procedure, is exceptional for a WBA tool. The reliability of the ATIS, which informs the level score, approaches the same reliability. The reason for the high reliability of PBA scores is because trainee ability is the largest contributor to variance for both the level score and the ATIS (36% and 33% respectively). Unsurprisingly, the reliability of PBA on a mix of procedures is much lower. This is because trainee procedure aptitude contributes significantly to the variance within PBA scores, which reflects the procedure-related specificity of the assessment and the lack of transferability of competence between different procedures. The implication of these results is that a trainee needs to be adequately assessed on every index procedure to establish his or her competence, not just on a sample of different procedures.

Procedure-based assessment is primarily intended to be an assessment for learning. Assessments that are primarily intended for a low-stakes formative purpose (assessment *for* learning) can have lower reliability than those for a summative purpose (assessment *of* learning), as they are intended to be performed more frequently with the aim of providing constructive feedback to drive learning rather than determining high-stakes decisions about progression of training or certification. However, our results indicate that PBA reliability is sufficient for use within summative assessments.

Objective Structured Assessment of Technical Skills

Both the ATTS and ATGS require a larger number of cases and assessors than PBA (equivalent to five assessor judgements on one case each) to achieve $G > 0.8$ for the same procedure. This is because all assessors and the different designated assessor groups vary more in the stringency of their ratings when using OSATS and agree more when using PBA. The reasons for this are explored in the *Post hoc analysis* section below. However, trainee ability remains the largest contributor to variance for both the ATTS and ATGS of the OSATS tool.

The ATGS has better overall reliability than the ATTS for assessing trainees on the same procedure. This agrees with previous research by Martin *et al.*,[52] who showed better reliability for global than task-specific ratings of surgical skill. As expected, the ATGS reliability is better for assessments of the same procedure than for a mix of procedures. However, the ATTS shows no procedure-specific variance (0% trainee procedure aptitude and 0% assessor subjectivity over case) and therefore its reliability is identical for the same procedure or a mix of procedures.

Non-technical Skills for Surgeons

The reliability of the NOTTS GS is lower than either PBA or OSATS (six assessor judgements on one case each to achieve $G > 0.8$ for the same procedure). However, scores are relatively procedure-independent, so the reliability of NOTSS GS scores on a mix of procedures approaches the reliability for the same procedure (eight assessor judgements on one case each to achieve $G > 0.8$). This is an achievable number, particularly for assessments over a mix of cases, as members of the operating team, including scrub nurses and anaesthetists, could act as assessors

in addition to the clinical supervisor. Interestingly, the number of judgements required to achieve reliability is similar to multisource feedback tools, such as the Mini-Peer Assessment Tool (Mini-PAT), which also use non-medical coworkers as assessors.

Trainee ability remains the largest contributor to variance for the NOTSS GS, although only by a small margin as scores show more variance from assessor stringency than either PBA or OSATS. This suggests that the NOTSS instrument may require more assessor training to achieve a better reliability. It may also be that NOTSS assesses an inherently more subjective domain than the instruments designed primarily to assess surgical skill.

Post hoc analysis

Interim analysis of the results at 1 year highlighted significant differences in the reliability between the PBA and OSATS tools. There were several possible reasons why OSATS might perform differently from PBA:

1. The instrument is inherently less reliable by virtue of its design.
2. The nature of the speciality of O&G, or the nature of the selected index cases within O&G, is less easy to assess reliably.
3. The particular cohort of O&G trainees or assessors recruited to this study is atypical in a way that negatively affects assessment precision or discrimination (e.g. highly homogeneous trainees or unusually subjective assessors).

One way to see if the reliability difference is instrument specific is to compare the reliability of PBAs in O&G with PBAs in the other five surgical specialties. The relatively large proportion of PBA assessments within O&G (30% of total PBA assessments) permitted this comparison. The other surgical specialties did not individually recruit sufficient cases to permit additional comparative reliability analyses.

Comparing the variance component analyses for PBA ATISs for O&G and non-O&G procedures (*Tables 53* and *54*), the ratings obtained within O&G show lower reliability. In particular, 'true' variance contributes less to O&G than non-O&G procedural scores (variance 22% vs 37%), scores are more subject to case-to case variation (variance 16% vs 8%), with greater variance owing to assessor designation subjectivity and stringency (variance 11% vs 0%). Nevertheless, O&G ATISs anchor assessor stringency better (variance 5% vs 15%).

Comparing the PBA level score for O&G and non-O&G procedures (*Tables 55* and *56*), the scores obtained within O&G are markedly less reliable than the level scores obtained within non-O&G specialties. Whereas 'true' variance is extremely high within non-O&G procedural scores

TABLE 53 Variance component analysis for O&G PBA (ATIS)

Component	Estimate	%	Meaning
Var(trainee)	0.006	22	Trainee ability
Var(case)	0.004	16	Trainee case-to-case variation
Var(proccod)	0.002	6	Procedure difficulty
Var(assessor)	0.001	5	Assessor stringency
Var(designation)	0.003	11	Assessor designation stringency
Var(trainee*proccod)	0.002	9	Trainee procedure aptitude
Var(case*assessor)	0.007	27	Assessor subjectivity over case
Var(proccod*assessor)	0.001	3	Assessor subjectivity over procedure
Var(error)	0.000	0	Residual variation

TABLE 54 Variance component analysis non-O&G PBA (ATIS)

Component	Estimate	%	Meaning
Var(trainee)	0.014	37	Trainee ability
Var(case)	0.003	8	Trainee case-to-case variation
Var(proccod)	0.003	7	Procedure difficulty
Var(assessor)	0.005	15	Assessor stringency
Var(designation)	0.000	0	Assessor designation stringency
Var(trainee*proccod)	0.005	12	Trainee procedure aptitude
Var(case*assessor)	0.007	19	Assessor subjectivity over case
Var(proccod*assessor)	0.001	2	Assessor subjectivity over procedure
Var(error)	0.000	0	Residual variation

TABLE 55 Variance component analysis O&G PBA (level score)

Component	Estimate	%	Meaning
Var(trainee)	0.029	5	Trainee ability
Var(case)	0.093	16	Trainee case-to-case variation
Var(proccod)	0.103	18	Procedure difficulty
Var(assessor)	0.033	6	Assessor stringency
Var(designation)	0.000	0	Assessor designation stringency
Var(trainee*proccod)	0.141	25	Trainee procedure aptitude
Var(case*assessor)	0.169	30	Assessor subjectivity over case
Var(proccod*assessor)	0.000	0	Assessor subjectivity over procedure
Var(error)	0.000	0	Residual variation

TABLE 56 Variance component analysis non-O&G PBA (level score)

Component	Estimate	%	Meaning
Var(trainee)	0.395	48	Trainee ability
Var(case)	0.084	10	Trainee case-to-case variation
Var(proccod)	0.068	8	Procedure difficulty
Var(assessor)	0.004	1	Assessor stringency
Var(designation)	0.005	1	Assessor designation stringency
Var(trainee*proccod)	0.153	19	Trainee procedure aptitude
Var(case*assessor)	0.093	11	Assessor subjectivity over case
Var(proccod*assessor)	0.020	2	Assessor subjectivity over procedure
Var(error)	0.000	0	Residual variation

(accounting for 48% of total variance), it is extremely low (5%) for O&G scores. Furthermore, O&G level scores demonstrate higher case-to-case variation in trainee performance (variance 16% vs 10%) with a greater impact on scores from assessor subjectivity (variance 30% vs 11%).

As the reliability of both PBA and OSATS tools across all specialties has already been shown to be greater for the same procedure rather than a mix of procedures, the *D*-study tables presented here only compare trainees for the same procedure. *Table 57* demonstrates the reliability of PBA ATIS ratings within O&G and non-O&G cases. Whereas $G > 0.8$ can be achieved using five cases with

TABLE 57a *D*-study for O&G versus non-O&G PBA ATIS ratings: cases assessed within O&G specialty

	Assessors per case		
Cases	1	2	3
1	0.32	0.41	0.46
2	0.56	0.64	0.67
3	0.69	0.74	0.76
4	0.76	0.80	0.82
5	0.81	0.84	0.85
6	0.84	0.87	0.87
7	0.86	0.88	0.89
8	0.88	0.90	0.90

TABLE 57b *D*-study for O&G versus non-O&G PBA ATIS ratings: cases assessed in non-O&G specialties

	Assessors per case		
Cases	1	2	3
1	0.47	0.60	0.66
2	0.70	0.79	0.82
3	0.79	0.86	0.88
4	0.84	0.89	0.91
5	0.88	0.92	0.93
6	0.90	0.93	0.94
7	0.91	0.94	0.95
8	0.92	0.95	0.96

one assessor per case (five assessor judgements) within O&G, this level of reliability is achieved using four cases with one assessor per case (four assessor judgements) for non-O&G cases.

A direct comparison of PBA level scores across O&G and non-O&G cases shows a striking difference in reliability (*Table 58*). Within O&G, a good level of reliability ($G > 0.8$) is not achieved using feasible numbers of cases and assessors per case. Conversely, the reliability within non-O&G cases is exceptionally high, with only two assessor judgements (two cases with one assessor per case) required to achieve good reliability.

This comparison shows that the lower reliability of OSATS in O&G is not instrument specific. Even the highly reliable PBA instrument does not produce reliable assessment results in O&G. This reliability gap is greatest when the 'level of independence' judgement is made. This judgement is the most reliable form of assessment outside the specialty of O&G, and the least reliable form of assessment inside the specialty of O&G.

To see if the reliability problem in O&G might be related to the cohort of trainees, we calculated the proportion of trainees within the study cohort for whom there were any training concerns. This was not part of our original demographic data set and required information from the deanery and the O&G programme director, after obtaining ethics approval. Formal training concerns are documented as ARCP2 (RITA D) and ARCP3 (RITA E) at annual review. An ARCP outcome 2 is recommended by the ARCP panel for focused training to acquire specific competencies using a timescale agreed with the trainee. An ARCP outcome 3 is recommended by the ARCP panel if additional remedial training is required, with the ARCP panel responsible for judging the intended outcome and timescale. Other informal training concerns are simply noted in the trainee's file. The proportion of trainees who had informal or formal training concerns relative to the overall number of trainees recruited from each specialty is shown in *Table 59*. During the study period, there was a markedly higher proportion of O&G trainees with training issues (42%, $n = 14$) than in all other specialties (4%, $n = 2$).

Table 60 displays the number of trainees with identified training concerns by level of training for each specialty. The differences at ST6 and ST7 levels were striking, with seven senior O&G trainees having identified training concerns, compared with only one trainee from all other specialties combined. It is highly likely, therefore, that the high proportion of trainees in O&G with identified training concerns, especially at a senior level, contributed to the lower reliability of PBA and OSATS in O&G (as well as to the poor construct validity of OSATS).

TABLE 58a *D*-study for O&G versus non-O&G PBA level score: cases assessed within O&G specialty

	Assessors per case		
Cases	1	2	3
1	0.09	0.13	0.15
2	0.22	0.28	0.31
3	0.32	0.39	0.42
4	0.41	0.47	0.50
5	0.48	0.54	0.56
6	0.53	0.59	0.61
7	0.58	0.63	0.65
8	0.61	0.66	0.68

TABLE 58b *D*-study for O&G versus non-O&G PBA level score: cases assessed in non-O&G specialties

	Assessors per case		
Cases	1	2	3
1	0.69	0.75	0.77
2	0.85	0.88	0.89
3	0.91	0.92	0.93
4	0.93	0.94	0.94
5	0.95	0.95	0.96
6	0.96	0.96	0.96
7	0.96	0.97	0.97
8	0.97	0.97	0.97

TABLE 59 Proportion of trainees with training concerns per specialty

	Cardiac	Colorectal and upper GI	Orthopaedic	O&G	Vascular	Total
Trainees in study cohort	10	23	9	33	11	**85**
Number with training concerns	0	1	0	14	1	**16**
Percentage with training concerns	0	5	0	42	9	**19**

Validity of assessment methods

Validity has many different aspects, as defined within *Chapter 1* of this report. All assessments require validity, and evidence for validity requires multiple sources. WBAs hold high face validity as assessments of day-to-day performance, as trainees are being assessed on direct observation of their real clinical practice. All assessments included in this study were carried out using a direct observation methodology.

Content validity is assured because the content of the study assessment tools was systematically derived from an iterative process involving many consultant surgeons and surgical educators during their development. The strong correlations between the ATIS and the level score for PBA (0.73), the ATTS and ATGS for OSATS (0.84), and the four categories within NOTSS (range 0.74–0.76) are an indication of the good internal content validity of each tool. The confirmed factor structure of the NOTSS tool also indicates that each NOTSS category is measuring a different competency domain, as intended by the content design of the tool. The good correlation between PBA or OSATS and NOTSS is encouraging, although perfect correlation should not be expected as they are intended to measure different competencies. The strongest correlations between NOTSS and the 'technical' instruments (PBA and OSATS) are in the 'decision-making' domain. This is to be expected, as a number of the items in both PBA and OSATS relate to decision-making, and provides evidence for criterion validity, in which instruments measuring the same construct should correlate.

The evidence for construct validity for PBA is demonstrated by the good correlation of scores with all demographic measures of age and experience we considered (age, ST level, total years and UK years of surgical training, total and recent experience of relevant index procedure) except

TABLE 60 Training concerns by specialty and year of training

	Cardiac			Colorectal and upper GI			Orthopaedic			O&G			Vascular		
	Concerns	ARCP2	ARCP3	Concerns	ARCP2	ARCP3	Concerns	ARCP2	ARCP3	Concerns	ARCP2	ARCP3	Concerns	ARCP2	ARCP3
F2										0	0	0			
ST1				0	0	0									
ST2				0	0	0				0	3	0	0	0	0
ST3	0	0	0	0	0	0	0	0	0	2	1	1	0	0	0
ST4					0	0							0	1	0
ST5							0	0	0				0	0	0
ST6				0	0	0				1	0	1	0	0	0
ST7				0	0	1				1	1	3	0	0	0
ST8	0	0	0	0	0	0	0	0	0						
Total	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**4**	**5**	**5**	**0**	**1**	**0**

F2, foundation year 2.

years of non-UK surgical training. All of the NOTSS categories demonstrated construct validity for many of these measures. For the 'situation awareness' and 'decision-making' categories, there was good correlation of scores with age, ST level, total years and UK years of surgical training, and recent procedural experience. For the 'communication' and 'leadership' categories, there was correlation with ST level, UK surgical training and recent procedural experience.

We would hypothesise that previous procedural experience would have a greater effect on the 'technical' scores (PBA and OSATS) than on the 'non-technical' scores (NOTTS) and that, within NOTTS, 'situation awareness' and 'decision-making' would depend more on procedure-specific experience than would 'communication' or 'leadership'. Our data showed this to be the case for PBA and NOTSS.

The OSATS tool did not demonstrate any evidence for construct validity, as there was no correlation of scores with any of the measures for age and experience. This may be explained by the factors previously addressed within the post hoc analysis discussion section. The high proportion of senior level O&G specialist trainees with training concerns, whose surgical competence was appraised as being below the expected standard for their level of training, undermines the expected hypothesis relating surgical skill with training/experience.

We have demonstrated a correlation between the PBA and NOTSS scores with years of UK surgical training but not with years of non-UK surgical training. While this may raise a question regarding the consistency of overseas training, the numbers are small and there may be other confounders, therefore no direct inferences can be made. Other factors that might affect assessment, such as language skills and cultural background, could also be associated with overseas training. This area warrants further research.

Predictive (outcome) validity provides the most important long-term evidence to support the validity of assessment methods. However, it is often the hardest to demonstrate, particularly without large sample sizes. We believe that there is a 'hint' of outcome validity in this study, albeit without reaching statistical significance. There were more than twice as many significant associations between surgical case outcomes and assessment scores than could be expected by chance alone. This is the first time that such an effect has been demonstrated for trainee assessments. Measuring outcome validity for procedures performed by trainees has previously been discounted because of the chance that outcomes could be influenced by consultant/team intervention and other confounders, such as patient-related risk factors. Many specialties are now providing risk-stratified comparative outcomes for consultants for the purpose of revalidation. There is an argument for providing the same facility for senior trainees, especially those who are approaching consultant status. Outcome data could be linked to both logbook and WBA data for commonly performed index procedures to triangulate the evidence on a trainee's surgical performance.

Consequential validity considers the educational impact or effectiveness of an assessment method. Evaluating the educational impact of the study's three WBA methods was not one of our primary research aims. However, the study's user-satisfaction questionnaires provide some evidence of good educational impact. Kirkpatrick described four levels on which to focus educational evaluation, which have been adapted for use in health education by Freeth *et al.*[17,18] The Kirkpatrick model can be used to evaluate the process and outcomes of assessment, or to evaluate any other aspect of education. Level 1 concerns learners' views on the learning experience, and our user-satisfaction questionnaires provide a wealth of data on trainees' views towards the assessment methods, which are discussed below. Level 2a concerns a change in attitudes among participants. Our user-satisfaction questionnaires asked trainees and clinical supervisors to consider their attitudes towards future use of the PBA, which we consider

is a surrogate marker to indicate a change in attitude. Encouragingly, a majority of clinical supervisors and a large majority of trainees reported that they were likely to use PBA in the future if given the choice. Levels 3 and 4 concern a change in behaviour and change in patient outcomes respectively, which were beyond the scope of this study.

Although it would be expected that trainees are accustomed to WBA undertaken by their clinical supervisor, the more formal nature of the assessments within this study, which included the presence of an independent assessor (in some cases also video recording), may have affected the performance of some trainees. The independent assessor made every attempt to remain unobtrusive to ensure that the operating and assessment conditions for the trainee were as authentic as possible. However, this was sometimes hard to achieve while ensuring good views of the procedure to allow direct observation and rating of a trainee's performance. Although either the trainee or clinical supervisor felt that performance had been affected in a quarter of cases, there was little agreement between their judgements for individual surgical cases. The cases judged to have been negatively affected by direct observational assessment (subjective) had slightly but significantly lower scores, but the video-recorded cases (objective) did not. This raises the possibility that trainees and assessors might have attributed a poor performance to the assessment conditions when the real reason lay elsewhere.

User satisfaction and acceptability

Procedure-based assessment

Clinical supervisors and trainees in all surgical specialties provided mixed, but predominantly positive, responses about the use of PBA for assessment and feedback. The majority of clinical supervisors and trainees were positive about its value as both an 'assessment *for* learning' (formative purpose) and an 'assessment *of* learning' (summative purpose), indicating that they are comfortable with the dual purpose of PBA for on-the-job training as well as for informing the ARCP. However, more clinical supervisors and trainees were negative about its use for a summative purpose. The high acceptability of the tool for formative and summative purposes indicates that successful implementation of the tool should be feasible. Scepticism towards the use of PBA as a summative assessment method might be reduced once evidence for its validity and reliability is within the public domain.

Comparing the perspectives of clinical supervisors with trainees in their responses to the applications of the PBA tool, trainees were relatively more positive about using PBA in the future if given the choice (were it not mandatory), the value of PBA in surgical education, the value of PBA in enhancing the assessment process and its usefulness for giving feedback. This positive perspective towards the value of feedback supports the evidence that feedback aids learning.[123,124] It also reinforces the need for feedback to be acknowledged as an integral aspect of WBA that requires clinical supervisors to invest in giving trainees timely feedback on their surgical performance.

Information from the ISCP indicates that engagement of registered trainees and clinical supervisors with WBA in the Yorkshire and Humber Deanery is 90%, which is similar to other deaneries (unpublished data). This suggests that our data for user satisfaction and usability are generalisable within the UK.

Objective Structured Assessment of Technical Skills

There is a marked difference between O&G clinical supervisor and trainee perspectives for the use of OSATS, with trainees reporting less overall satisfaction and acceptability than clinical supervisors. This contrasts with the PBA clinical supervisor and trainee perspectives, in

which trainees reported greater satisfaction than clinical supervisors. The majority of clinical supervisors agreed that OSATS was useful as both a formative and a summative assessment method, although more were positive about its formative use. Trainees were markedly less positive about the use of OSATS as a formative method and especially as a summative assessment method, with these perspectives falling equal to or below a majority view respectively. The proportion of clinical supervisors and trainees who disagreed with its use as a summative method outweighed the proportion of those who disagreed with its formative use. This finding was more marked than for PBA. These responses show trainees' concern about the summative design and purpose of OSATS. This could reflect the OSATS tool design at the time of the study, in which a pass/fail summary judgement was assigned by the clinical supervisor. This reinforced the summative purpose of OSATS and may have been viewed by trainees as a practical 'mini-exam'.

The clinical supervisors and trainees in O&G who had used both OSATS and PBA expressed greater overall satisfaction with PBA. The reasons for this preference are being explored within subsequent focus groups with trainees (see *Further research* section) and include the use of a single assessment form, detailed items that provide a structure for feedback, a simpler binary rating system that uses an explicit standard (CCT level) and the use of a four-level outcome with clear descriptors concerning readiness for independent practice, rather than a pass/fail judgement of their performance.

Non-technical Skills for Surgeons

The scrub nurses provided more positive responses than the anaesthetists to every question regarding the NOTSS tool and its uses. A majority of both assessor groups agreed that the NOTSS system provided a common language for discussing non-technical skills, which provides encouraging evidence for the face validity of the tool. In addition, the majority of both groups agreed that NOTSS was useful for reflective practice and providing feedback and a valuable adjunct to the use of surgical skill assessments. Anaesthetists were evenly split on whether or not NOTSS would enhance patient safety, whereas scrub nurses indicated a strong agreement. These results would be encouraging were NOTSS to be considered for use in routine surgical practice, as both assessor groups see a definite role for NOTSS within surgical training.

The majority of anaesthetists felt that they were able to more easily assess interpersonal skills (e.g. communication and leadership) than cognitive skills (e.g. situation awareness and decision-making) using NOTSS. This contrasts with the scrub nurse perspectives, who found it equally easy to assess both skill domains, and with the usability trial conducted by Yule *et al.*,[112] in which clinical supervisors found interpersonal skills more difficult to rate than cognitive skills. In our study, more scrub nurses than anaesthetists perceived that they were able to easily assess both cognitive and interpersonal skills. This may be because they possess more technical knowledge of the procedure, which may inform their understanding and interpretation of surgeons' non-technical behaviours.

The differing assessor groups' perspectives on their ability to rate non-technical skills may provide an argument for the use of NOTSS by a multidisciplinary group of assessors, similar to multisource feedback tools, e.g. Mini-PAT. While the overall NOTSS reliability is stable, it may be that each disciplinary group offers a unique perspective on non-technical performance that may enhance the assessment process.

Time is a precious commodity in the operating theatre and our NOTSS assessors were asked to voluntarily complete NOTSS assessments in the knowledge that these would be used only for research purposes. The overwhelming majority of participating NOTSS assessors indicated that it was feasible to complete the assessments within the time constraints of the list. However, some scrub nurses and anaesthetists declined to complete an assessment at the start of a case because

of other constraints (staff shortages, need to supervise a trainee, complex case, list running over time), which was entirely justified to prioritise patient safety. Our results agree with those of Yule *et al.*,[112] who found that only 9% of clinical supervisors in their usability trial thought NOTSS added too much time to their list. These results provide further weight to the proposal that NOTSS should be a feasible tool to implement within the operating theatre.

Implications for assessment and training

Study recruitment and opportunities lost for workplace-based assessment

In this section, we make use of our recruitment results to illustrate the opportunities lost for WBA within surgical training. We feel that this reflects implications for practice, given the expectation that current surgical assessment programmes advocate using WBA as frequently as possible and making use of every possible surgical case for training. Shorter training schemes combined with the reduction in hours required by the EWTD mean that training must necessarily become more efficient. 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.

Successful recruitment to the study relied upon the timing and alignment of many factors to enable a WBA to be completed, e.g. a suitable case, an available clinical supervisor, an appropriate trainee and sufficient training time. The practical difficulties that we encountered shed light on why training opportunities are lost in practice. There will always be pressures on training time within the NHS, given the dual and often conflicting concerns of service provision with training. There is a need for trainees and clinical supervisors to be fully aware of the obstacles to training and to focus their joint responsibility on securing suitable training opportunities. We offer some possible solutions to the obstacles we have identified.

Some training obstacles are not directly within the control of clinical supervisors and trainees, being determined at an organisational level, although successful lobbying may influence organisational decision-making. We found that there was a large proportion of consented cases that could not be assessed within the study because there was insufficient operating theatre time for supervised training. It is possible that, with more allocation of operating theatre time per case, a better balance could be achieved between provision of training opportunities and service provision. There was a similar proportion of cases that were lost to study recruitment owing to lack of availability of an inpatient bed, cancellation or alteration of the case/operating list and staff leave or shortages. Two solutions could be to provide ring-fenced beds for elective inpatient admissions in tertiary referral centres and to establish training opportunities within local diagnosis and treatment centres.

Other training obstacles, although they may be perceived as organisational, are determined at a professional level, being subject to the guiding principles of the medical professionals themselves. These might be suitably addressed through forward planning and reorganisation of clinical workload by clinical supervisors and trainees. The most common reason for non-recruitment in our study was the lack of a trainee at short notice, usually owing to the emergency rota or other clinical commitments. With appropriate rota design and reorganisation to allocate trainees to daytime training lists, trainee availability for daytime supervised surgical training could be better prioritised, e.g. larger (hospital-at-night) emergency rotas or taking trainees off being on call at night. This might be difficult within the constraints of the EWTD. However, trainees might wish to consider using the voluntary additional 8 hours of opt-out above the 48-hour EWTD maximum for additional training rather than service.

Further training obstacles are immediately suitable for change as they are under the direct control of trainees and clinical supervisors at an individual level. The second most common reason for cases being lost to assessment within the study was that the clinical supervisor performed the case as he or she deemed it unsuitable for the trainee available. However, it appeared that some clinical supervisors stated that the case was unsuitable for training when it seemed no more difficult than usual. Some trainees were not given the opportunity to be involved in certain elements of an operation appropriate to their level of training, including opening and closure of the operative site. There also appeared to be conflicts of interest as some clinical supervisors wanted to make the most of the surgical opportunity themselves while others appeared reluctant or uneasy supervising the trainees. Clinical supervisors also blamed time pressure on the operating list: 'It's quicker to do it myself'. Therefore, it is imperative that a surgical training culture is engendered in which clinical supervisors are encouraged to regularly supervise trainees to perform suitable cases in part or whole and in which trainees are empowered to identify and ask for opportunities for WBA. In addition, improved collaboration between clinical supervisors and trainees would enable better matching of suitable trainees to appropriate cases.

Some issues were more common in certain specialties. For example, orthopaedic surgery had a high proportion of complex cases that were deemed unsuitable for junior trainees. Cardiac, vascular and O&G surgery suffered more from lack of availability of trainees owing to other clinical commitments. O&G trainers stated more often that the case was unsuitable for training, which may reflect the factors discussed within the post hoc analysis.

Frequency and timing of workplace-based assessment

Although our study shows that the number of cases required to achieve a reliable assessment using the tools is quite low, it would be a mistake to equate this with the number of assessments actually required. The main purpose of WBA is assessment *for* learning, in which repeated supervised surgical practice with timely feedback aids surgical skill acquisition. Furthermore, the more frequently assessments take place and are integrated into training, the better the validity of the assessment in terms of relating to actual performance.

Defining a minimum or set number of assessments may encourage trainees and trainers to view them as 'mini-exams'. This is certainly the experience of the Foundation Programme and Core Surgical Training Programme, where it has been acknowledged that trainees practise procedures informally without asking for an assessment and only ask for assessment later on when they feel confident of achieving a good score.[125] Viewing WBA as 'mini-exams' may also increase the pressure on an assessor to give a good rating, particularly where there is the option only to pass or fail a trainee for a given assessment. This can undermine the overall reliability of an assessment tool. This raises the question of whether the lower reliability of OSATS than PBA was, in part, related to its use of a pass/fail summative judgement. In our study, only half of O&G assessors completed the pass/fail summary judgement within OSATSs. This culture of putting emphasis on assessment as summative rather than formative defeats the principal aim of WBA, which is to aid learning.

At present, some surgical specialties set target numbers of index procedures for trainees to achieve in order to progress with their training. These numbers are often based on what is achievable or on a consensus of opinion, rather than guidance from an evidence base. The ISCP will be in a position to correlate PBA levels with logbook experience from the large database of PBAs completed and submitted electronically. This will provide information about learning curves and the calculation of confidence intervals for the number of procedures required to achieve competence. One system for monitoring a trainee's competency progression over time is with cumulative sum (CUSUM) charts. CUSUM charts were first used in industry in the 1950s as a quality control tool,[126] but have been applied to medicine to monitor progress in trainees'

surgical,[127] anaesthetic[128] and clinical skills.[129] Both the OCAP and ICSP plan to collate the PBA level scores and logbook data for each trainee so that progression of experience and performance can be mapped.

A CUSUM chart provides an objective graph of performance for a number of consecutive procedures, with the CUSUM score representing a running total of successful and unsuccessful attempts. The upper and lower limits of performance can be set for the procedure in question, which may be further adjusted for different ST levels.[130] For example, if the accepted standard is that ST4 trainees should achieve a PBA level 3 or 4 for an inguinal hernia repair on 9 out of 10 occasions, the accepted 'failure rate' is 10%. Each unsuccessful procedure (PBA level < 3) will have an incremental value of 0.9, and each successful procedure (PBA level 3 or above) a decrement of 0.1. When the graph shows a consistent downward sloping line, an acceptable performance has been achieved for the relevant procedure. CUSUM charts can readily identify poorly performing trainees or weaknesses in training. They also enable the calculation of the numbers of procedures required for the majority of trainees to acquire competence for a particular procedure. This has a direct bearing on both the average number of assessments required to demonstrate a trainee's competence as well as the number of training opportunities required by trainees. The addition of CUSUM charts to trainees' e-portfolio may be a useful aid to learning, as well as an audit of actual assessment and training practice, particularly within the run-through training structure whereby large numbers of novice surgeons enter training at the same time.

Another important question is whether WBA should be undertaken every time a trainee performs an index procedure or less frequently (i.e. triggered assessment). Evidence for how often WBA can be feasibly undertaken is emerging. A study in Bath suggested that weekly PBA is feasible and acceptable.[131] Although triggered assessments may be more acceptable to trainees and assessors as they involve less time and work, the philosophy of assessment *for* learning should be 'the more the better' because of the constructive feedback provided. Furthermore, early and frequent assessment must be encouraged in order to demonstrate a progression in surgical skills that should precede achievement of competence. Our results demonstrate reliability for the tools using feasible numbers of assessments, but the emphasis for an assessment *for* learning should remain on the feedback element of the assessment, which these tools structure and support. Clinical supervisors and trainees require reassurance that 'development required' or 'needs help' simply means that more practice is required.

The other factor in this 'numbers game' is the frequency of complications specific to a particular procedure. A trainee undertaking a complex procedure with multiple possible complications will need to be assessed more frequently to ensure that he or she can deal with them all. Uncommon and serious adverse events are probably better rehearsed on simulators, wherever possible.

Returning to our study results, we have shown that the tools that primarily assess technical skills (PBA/OSATS) require adequate assessment of each individual index procedure for reliability. Therefore, in any decision-making regarding the frequency and timing of WBA, the interests of ensuring adequate sampling of cases for reliability and providing the maximum educational benefit from ongoing feedback need to be balanced.

Purpose and design of workplace-based assessment

The purpose of WBA needs to be clear to both clinical supervisors and trainees, as it exerts a powerful effect on the type of teaching and learning that the assessment method supports. As a research team we gained insight into the impact of the assessment purpose upon training through the recruitment from different surgical specialties. The ISCP, which applies to all specialties studied except O&G, states that the main purpose of WBA is formative, i.e. assessment *for* learning. PBA has been made mandatory only for those trainees entering the ISCP surgical

curriculum at ST3 level from August 2007, using a competency-based training approach, for whom this purpose is explicit and understood from their foundation-level training experience. However, use of PBA has not been made mandatory for more senior trainees who have trained within the former apprenticeship system, in recognition that the expectations and training under this system were fundamentally different. Conversely, in O&G, OSATS has been a requirement for trainees at all levels, with a set minimum number of assessments that a trainee must pass to achieve logbook competencies. This may increase pressure on both the trainee and the clinical supervisor to produce good OSATS scores. It may explain the lower reliability of OSATS found in our results if there is reluctance on the part of clinical supervisors to use OSATS in making a pass/fail summative judgement. The low overall trainee user satisfaction/acceptability of OSATS in our study, with the most negative responses relating to its summative use, are likely to reflect the trainees' dissatisfaction with this issue. Encouragingly, the pass/fail terminology has since been revised to 'competent in all areas included in this OSATS' and 'working towards competence' by the RCOG Assessment and Examination Committee in order to reinforce its formative purpose.[104] Our experience resonates with recent policy documents aimed at improving the implementation of assessment one of which states, 'A period of re-education is required to change the mindset of mini-exam towards WBA'.[27]

Both the PMETB and the Academy of Medical Royal Colleges have recently issued guidance to stress the importance of clarifying and communicating the dual purpose of WBA.[16,27] Ongoing WBA, as assessment *for* learning, provides part of the body of evidence for the assessment *of* learning at ARCP. Rather than being at 'crossed purposes', both assessment purposes should be complementary in reinforcing feedback and self-directed learning. It is evident that clearly defining the purpose of WBA has direct implications for the utility of the assessment tool, particularly its user satisfaction and reliability.

The dual purpose of PBA was made explicit from the beginning, embedded in the ISCP framework, web-based and available to all stakeholders. In our study, PBA stands out as the most reliable tool with the greatest user satisfaction and acceptability. The majority of clinical supervisors and trainees valued its dual assessment purpose as assessment *for* learning and assessment *of* learning.

There is emerging evidence that the types of rating scales used for rating performance of doctors have a strong influence on assessors' judgements.[125,132] The latest PMEMB recommendation is that assessors should make judgements against word descriptors rather than numerical rating scales,[16] as performance-based descriptions of what is being judged, and at what standard, help assessors achieve accuracy and consistency in their ratings. For example, rating scales that use either numerical (e.g. Likert) or relative scales (e.g. poor/average/excellent performance) are outperformed by scales using descriptive, behaviourally anchored rating scales (e.g. able or not able to perform independently). Furthermore, expressing concerns about a trainee's competence is more appropriate using clear performance descriptors. Our results add to this emerging body of evidence. The exceptional reliability of the PBA level rating scale leads to the possible conclusion that medical assessors are more able to be consistent and discriminatory when they are asked to make judgements that they are familiar with (such as readiness for independent practice).

Experience in using workplace-based assessment

The ISCP and RCOG implemented their new competency-based curricula concurrently in August 2007, adopting PBA and OSATS respectively. However, OSATS had been in informal use for 2 years before that. Around half of the O&G clinical supervisors and trainees in our study had previous experience of using OSATS during that period. Encouragingly, this reflects a positive uptake of OSATS before its mandatory implementation in the new curriculum. In contrast, half

of non-O&G clinical supervisors and trainees had never used PBA before their involvement in the study. As there was no prior informal use of PBA, it could have been anticipated that the initial cohorts of non-O&G clinical supervisors and trainees had yet to gain direct experience of using the tool. Interestingly, following the introduction of the new curriculum, a further half of O&G assessors and one-third of O&G trainees reported that they had begun to use OSATS. This suggests that a change in educational policy promotes a change in educational practice.

Assessor training for workplace-based assessment

The general principle of WBA is that assessors should possess the relevant clinical expertise with regard to the task being undertaken and have been trained in that assessment method and in giving feedback. One aim of training is to improve the reliability of assessment through a thorough understanding of the design and purpose of the assessment method being used, as well as the standard required. Training may help assessors to make consistent and defensible (i.e. reliable) judgements.[16] Another aim is to improve the quality of teaching during the procedure and the feedback afterwards, through the use of constructive methods such as Pendleton's rules.[133] Our results suggest that, for PBA in particular, achieving good assessment reliability may not require rigorous training of clinical supervisors. However, it is required to help trainees and clinical supervisors understand the process and purpose of assessment. Assessor training is probably required to improve the quality of teaching and feedback, although this was outside the scope of this study.

Assessors for PBA and OSATS are normally expected to be clinical supervisors in the relevant specialty, who are competent to perform the procedure being assessed. Although written guidance and web-based training is available via the ISCP, OCAP and RCOG, all organisations advocate that assessor training is best carried out through face-to-face workshops. It has since been acknowledged that the hurried implementation of ST assessment systems to meet the PMETB approval in 2007, which coincided with problems caused by the MTAS, resulted in inconsistent assessor training.[27] The pattern of our results for clinical supervisor training reflects this inconsistency, as the coverage of training across different methods was patchy. Some clinical supervisors reported that they had received training using a combination of methods, whereas others reported training with one method. This suggests that training may occur ad hoc and could benefit from a more systematic approach. Similar proportions of both non-O&G and O&G clinical supervisors had utilised web-based training, indicating that the provision of online training is an essential resource to support other training methods.

Clinical supervisors' perceptions of the provision and adequacy of training appear to be very subjective. Our results demonstrate that not all clinical supervisors access training where it is made available to them. We provided all clinical supervisors with face-to-face training in the use of PBA/OSATS, supported by written and/or e-mail packs, yet a large proportion reported not having received training. Some clinical supervisors found it acceptable that they had undertaken very little training. Conversely, other clinical supervisors who had reported receiving training using more than one method did not perceive this as adequate. These findings reflect the real-life surgical training culture, in which there are wide variations in clinical supervisors' engagement and attitudes towards their personal training.

The PBA tool demonstrated excellent reliability for a WBA tool, despite patchy assessor training. This suggests that intensive assessor training may not be required for the use of PBA. This may be because the self-explanatory design of the form is intuitive to use, and the descriptive anchor statements for establishing a standard (based upon the ability of a trainee to perform a procedure with or without supervision) are well understood without the need for training.

Despite offering training in the appropriate use of the tools, our team observed some recurring inconsistencies in the way the tools were used:

- prompting trainees too readily during the procedure
- inability to allow trainees to lead the case *within* their level of competence by taking over decision-making or the surgical instruments
- directing trainees to operate using the supervisor's preferred surgical sequence and/or technique, even if the trainee's sequence/technique was acceptable
- reluctance to score competencies negatively (and/or give difficult feedback), particularly for senior trainees.

We recognise that some of these training styles may have influenced the ratings of trainees' skills and behaviours in our study. For example, if a trainee is directed to operate using a different technique, he or she may not be as smooth in its delivery, which could affect WBA scores. This highlights the difficulties that trainees experience when a clinical supervisor combines training with assessment during a surgical case. Although we have analysed quantitatively the proportion of trainees who judged that their performance was affected, we still have descriptive data to analyse, which may shed more light on this issue.

The most successful training opportunities and complete WBA were observed to be during those cases in which the clinical supervisor permitted the trainee to operate within his or her limits of competence and granted him or her the leadership to carry this out, prompting or intervening only when required or requested. Clinical supervisor training to this level was beyond the scope of our study, but it is likely to be required of clinical supervisors in the future if surgical training is to become more effective. As a response to our team observations, we have begun to draw upon these experiences through facilitating clinical supervisor training workshops, using videos to explore their surgical training techniques.

Face-to-face training may also be required to improve the quality of feedback provided by clinical supervisors. We observed and timed the feedback given by clinical supervisors to trainees after all cases, and in some cases we recorded the characteristics and quality of the feedback using structured observation charts. These feedback data, which are yet to be analysed using quantitative and qualitative methods, constitute ongoing team research work.

One of the acknowledged obstacles to achieving good assessor training is the difficulty in prioritising time and resources to prepare clinical supervisors for their educational role and responsibilities, within the constraints of service provision. The 2007 PMETB trainers' survey found that over half of training programme directors had not undergone appraisal for their educational duties.[134] This sheds doubt on whether clinical supervisors' training is systematically directed and may undermine confidence in the consistency and standards of clinical supervisors' training within surgical training programmes. However, it is also imperative that clinical supervisors themselves take ownership and responsibility for their educational responsibilities in a bottom-up approach. In our study we noted that some clinical supervisors did not engage with the face-to-face WBA training being offered in preparation for their inclusion in the study. Sometimes insufficient time was given as a reason, while in other cases it appeared not to be a personal priority.

It has been openly recognised by WBA stakeholders that achieving full compliance with the PMETB's 2008 'standards for trainers'[135] may be difficult owing to organisational issues and time constraints on training all clinical supervisors. This was reflected in a deadline for full compliance deferred to January 2010. In rolling out WBA, it may be considered most credible to use clinical

supervisors as assessors. However, our results suggest that the assessment burden could be eased by using non-surgeon assessors. The independent assessor ratings using PBA were as reliable as those provided by clinical supervisors, with assessor designation not affecting the stringency of either the ATIS or level PBA ratings. One independent was an SCP who completed assessments in all specialties, and the three other independent assessors were 'relative non-expert' surgeons, as they completed assessments outside, as well as within, their surgical specialty. It may be that, with additional PBA training, SCPs could assess the surgical skills of trainees for the purposes of WBA. However, the use of SCPs as trained assessors requires further evaluation. A standard-setting exercise conducted by the Vascular Society of Great Britain and Ireland suggested that scrub nurses were able to discriminate between different levels of operative performance from videos of operations.[103] We did not use theatre nurses or SCPs to complete PBA (except our SCP independent assessor who was trained in WBA) but this could be evaluated in a further study. However, it may be that it is the experience of SCPs performing surgery themselves, experience that scrub nurses do not have, that makes them suitable as assessors of technical skills.

Non-technical Skills for Surgeons was also originally developed and designed for use by trained surgeons. After discussion with the NOTSS development team, we chose to address in our study whether NOTSS could be reliably used by non-surgeon assessors. We used both anaesthetists and scrub nurses, along with the independent assessor, as our NOTSS assessors. Although the reasons for this were primarily pragmatic, as it was not desirable for clinical supervisors to complete NOTSS in addition to PBA/OSATS, we sought to extend the validation and application of NOTSS for WBA purposes. We hypothesised that members of the wider surgical team could reliably assess non-technical behaviours, in a similar way to members of the medical team using multisource feedback assessments to assess professional behaviours. Both anaesthetists and scrub nurses were believed to offer excellent assessors' perspective. They are familiar with the operating theatre environment, with a good understanding of the steps of surgery within their own practising subspecialty, while being present throughout the operation to directly observe trainee behaviours, particularly to capture the preoperative behaviours that the clinical supervisor was often not present to observe. In addition, we recognised that there was some familiarity among anaesthetists with non-technical skills and human factor concepts that are inherent to their clinical role and that scrub nurses regularly articulated surgical safety issues.

All of our NOTSS assessors received training, largely by provision of a NOTSS booklet together with face-to-face training from an independent assessor. None of our assessors had any prior experience of using NOTSS, although some of our anaesthetists had a working knowledge of ANTS and human factors training. NOTSS training, as advocated by the developers of the tool, is comprehensive: background knowledge on human performance, error management and non-technical skills, an understanding of the principles of using psychometric tools for rating performance, familiarity with the NOTSS tool, and practice in observing non-technical skills and rating behaviours using NOTSS. At the time of the study, the NOTSS development team were delivering these elements within a 1-day training workshop. Attendance on such a course for all NOTSS assessors was not feasible for our study, in terms of both study resources and staff time/availability. Furthermore, we wanted to assess the reliability of NOTSS 'in the real world' where limited training opportunities may exist. In practice, both the evaluation and usability trials conducted by the NOTSS development team have used a shorter 3-hour training session to train clinical supervisors,[110,112] which suggests that more limited training may be sufficient. The vast majority of clinical supervisors involved in these trials reported that they had received adequate training. With the increasing body of evidence that relates non-technical performance to patient surgical outcomes and safety, the need for non-technical training becomes axiomatic. This is supported by the surgeons' perspective, with a recent study of human factors training finding that views on the impact of human factors upon surgical performance had changed among surgeon participants.[136]

All independent assessors in our study provided 'relative expert' NOTSS ratings, having completed a training course led by the NOTSS development team as well as attending 'Behavioural Science Applied to Surgery' conferences annually. Despite the more limited training received by anaesthetist and scrub nurse assessors, our results show that they did not differ significantly in the stringency of their scores compared with independent assessors, with no variance owing to assessor designation for NOTSS ratings. This provides evidence that less intensive NOTSS training does not compromise the tool's reliability.

Implementation of workplace-based assessment and research

As a research team, we became interested in the issues and challenges we faced over the course of the study in implementing both WBA and research in the operating theatre. Sharing these experiences and the lessons learnt could be a useful resource for those involved in implementing WBA into surgical training and researchers working in this field. The main themes relating to our experiences of implementing WBA are summarised here, but appear in full in Marriott *et al.*[120] Since our publication, issues of WBA implementation have come to the forefront, and recent educational policy is now seeking to address this gap between assessment theory and practice.[27]

Relating the study design to the research aim

The surgical workplace is a complex and unpredictable assessment context. Additionally, performance is a complex assessment construct. To find order in this complexity demands a structured design and overarching theoretical framework. Our commitment to answering clear research questions drove the study's design, e.g. for evaluating the question of tool validity, the study's design included trainee demographics for training and experience (construct validity), questionnaire data (consequential validity) and surgical outcomes (predictive validity), which provided evidence to confirm or refute our validity hypotheses. In addition, the study design was heavily influenced by two theoretical frameworks: van der Vleuten's utility model (see *Figure 3*) and the Cambridge model of performance assessment (see *Figure 2*), which are illustrated and discussed within *Chapter 1*. However, even when the design is structured by clear research questions and informed by educational theory, developments to the design may be required during implementation, e.g. collecting data of the trainees'/clinical supervisors' perspectives for the effect of assessment and/or video recording on performance.

Key lessons:

- Use structured questions and theory to inform the design of assessment research.
- Flexibility and foresight are required to manage developments to the study design during implementation.

Matching the research team to the study design

A multidisciplinary research team with expertise and confidence in surgery, operating theatre etiquette, principles and practice of education, educational research and research governance is required to evaluate WBA. All four independent assessors within our team were practising in surgery, trained in assessment and had a surgical education research interest. Our psychometrician had established statistical experience in WBA.

Although it is impossible to outline all the essential skills and attributes required from a research team working in this field, we consider the following to be essential to the process of implementation:

- expertise in surgical knowledge, skills, attributes and competence

- familiarity and confidence with working in the operating theatre environment
- firm research governance knowledge and 'good clinical practice' training
- statistical expertise, independent from the grass-roots researchers
- diplomacy in negotiating sociopolitical surgical frameworks
- tenacity towards recruitment of cases and engagement of trainees/assessors.

Engaging surgical teams

The start of the study in 2007 coincided with major shifts in surgical training and assessment, including the MTAS and the new competency-based training curricula. The engagement of staff was initially difficult as there was resistance to further change. However, we noticed a change in attitude over the timeframe of the study, with increasing acceptance and value placed upon WBA. Our approach was to familiarise surgical teams with the study's aim and purpose in advance to ensure that the research wasn't seen to be imposed upon the theatre environment, which may have been viewed as threatening and/or unnecessary. Engagement was best achieved by e-mail and written information packs supported by face-to-face discussion in the workplace. Unsurprisingly, we encountered cynicism among some trainees and clinical supervisors, so giving time for discussion was a key part of the process.

It is very encouraging that we successfully engaged all clinical supervisors and all but two trainees across the six surgical specialties to participate in the study. It is not known whether the decision to exclude assessments of consultant surgeons prior to implementing our study affected their engagement positively, but it appeared to be a useful step in separating the study aims from the controversial issue of 'revalidation'. Individual trainees became increasingly engaged with the study as it became clear that it was providing ring-fenced opportunities for them to undertake training cases that contributed to their training portfolio. The engagement of trainees and clinical supervisors with the study shows that implementation of WBA is achievable even where scepticism may exist.

Key lessons:

- Explain the aims of the study in advance within the context of WBA and surgical training.
- Engage staff using face-to-face discussion, supported by written and/or e-mail information.
- Revise your approach to overcome barriers in the workplace.

Ethical considerations for participants: patient and trainee consent

The ethics committee viewed the patients as study participants primarily on the basis of the use of video assessment. However, normal UK surgical practice does not routinely seek informed patient consent for the involvement of surgical trainees in their operations under supervision. It was noteworthy that patient consent was not a major limiting factor for case recruitment. For the few patients who declined to consent, their decision was usually surrounded by misconceptions about trainee involvement in performing supervised elective surgery. We provided an open discussion of the role of supervised operating in surgical training, and some patients were then happy to consent, while others wished for consultant-led care.

During implementation of the study, we considered trainees as additional participants and viewed their informed consent to be an important ethical consideration. The added requirements of this study beyond normal WBA training requirements included video assessments, NOTSS assessments and the presence of independent assessors. We used trainee invitation letters in advance of study recruitment within a specialty and verbal consent to ensure the voluntary involvement of trainees. An important part of this consent process was highlighting that the study's purpose was to evaluate the tools themselves across different trainees, cases and specialties and not to evaluate an individual's level of surgical skill. Once this purpose was clearly

communicated, the very small proportion of hesitant trainees felt comfortable with participating. There were only two trainees, who were approaching CCT level, who declined to participate. While we are entirely comfortable with our ethical approach to consent, the reality that some trainees declined to participate raises the possibility of a shortcoming in the ethics process. There remains a literature gap on the ethics of trainee involvement in educational research where there is an overlap between training and research requirements.

Key lessons:

- Consent of participants within educational research is complex.
- Use the patient's perspective towards surgical care systems and surgical training to inform your consent process.
- Consider the ethics of trainees' involvement within educational research.

Research versus training agenda

Given that our study sought to validate WBA tools that were in current use within surgical training, there were opportunities for the research to form collaborations or conflicts with the training agenda. Some key examples from both camps are summarised here.

Examples of collaboration:

- provision of valuable, timely training on the tools for trainees and clinical supervisors
- ring-fenced opportunities for WBA
- encouraging appropriate use of tools for formative assessment, e.g. using parts of PBA for junior surgical trainees
- practical demonstration of the feasibility of workplace learning and assessment
- 'field testing' prompting tool modification, e.g. separating cystic duct and artery ligation tasks for appropriate assessment within laparoscopic cholecystectomy PBA.

Examples of dichotomy:

- Conflicts may exist between providing assessors with sufficient tool information without prejudicing usage. This may be limited by advance training and ongoing guidance.
- Upholding the research agenda of assessing the utility of WBA in the real-life setting may compromise timely guidance on 'correct' use of the tool.
- If multiple assessors are used, care is needed to avoid contamination of ratings, limiting necessary discussion until after ratings are assigned.
- The role of WBA assessors may be more aligned to observer-as-participant than complete observer if the assessor is part of the working surgical team.

Limitations of the study

The original study plan was to include teaching hospital trusts in three cities, and site-specific ethics approval was obtained for all three. During the initial phase of the study, it became apparent that the difficulty and workload involved in identifying, consenting and assessing suitable surgical cases would have required the appointment of a trial co-ordinator/independent assessor at each centre. Funding was not available for this within the finite resources of the grant. It is therefore possible that the results would not be not generalisable to other centres in the UK. Our sampling strategy of assessing as many cases per trainee as feasible, providing the most dependable reliability data, was achieved by focusing our efforts on gaining multiple assessments for trainees at the single hospital trust. We also revisited most surgical specialties on two separate

occasions during the course of the study, providing the opportunity for the assessment of trainees who had rotated in from other peripheral hospitals. Within the trust there were three hospital sites that we observed to have a range of working and training cultures. Furthermore, the willingness of all potential clinical supervisors and all bar two potential trainees to participate in the study increases our confidence in the generalisability of our results to the rest of the UK.

The overall recruitment figures for the study fell slightly short of the target we set ourselves, although they well exceeded those proposed by the Steering Committee. Within recruitment as a whole the number of non-O&G cases fell short of our target although, once again, it was above the target set by the Steering Committee. It is disappointing that the recruitment figures for orthopaedic and colorectal surgery were noticeably lower than intended, despite concerted efforts to optimise numbers in the second rounds of recruitment. However, this does not negatively affect the primary outcome of reliability evaluation, although it does limit the extent to which we can explore specialty differences in reliability, validity and feasibility.

Inevitably, the study could not include all surgical specialties or all index procedures. Specialties were selected on the basis of having an adequate number of trainees with a sufficient workload of the selected index procedures to allow an evaluation of WBA tools within a real-life training environment. Various surgical subspecialties including breast, endocrine, plastic, ENT (ear, nose and throat), maxillofacial, ophthalmic, neurosurgery and urology did not satisfy these selection criteria. We did not include specialties for which PBA had not been developed (e.g. endovascular), except O&G, within which new PBAs were developed to allow a comparison of PBA utility with OSATS. We excluded any specialties that require operating microscopes for their index procedures because of the difficulty of direct observation by independent assessors, e.g. ophthalmic surgery. However, having included 15 index procedures across six specialties, we believe that the tools have been adequately evaluated for interspecialty and interprocedural differences. PBA did not show interprocedural variations in scores that could be attributed to index procedures per se, but rather were due to the confounding effect of assessor stringency and procedural difficulty across all index procedures. We have no reason to suspect that PBA would be less valid or reliable in other surgical specialties; however, future work may be useful.

Only domains 3–5 of PBA were used within the context of the study. The main reason for this was that these sections covered performance in the operating theatre, as do OSATS and NOTSS. Furthermore, in practice it is not anticipated that an entire PBA is necessarily completed for any one case. The fact that we have shown excellent reliability for domains 3–5 means that there should be less concern about the validity and reliability of the tool in its entirety. However, it would be helpful to do further work on the other PBA domains, especially preoperative planning and consent.

Although all ST levels were represented, the greatest proportion of surgical trainees in this study were at junior (ST2 or ST3) and senior (ST7 and ST8) levels. This reflects the organisation of the surgical training programmes. Trainees are often placed centrally at the main teaching hospital sites at the beginning of their training. They then move to clinical placements in the surrounding district general hospitals for intermediate-level training, before moving back centrally as senior trainees for advanced subspecialty training. To include the surrounding district general hospitals in this study would have required additional resources. Despite these restrictions on trainee sampling, we did achieve a good spread of WBA scores, indicating that a full range of surgical performance was represented within our cohort of trainees.

Reliability is a measure of how consistently an assessment method can discriminate between candidates. This requires a precise instrument and a heterogeneous population of candidates with stable differences. Indeed, reliability is the ratio of precision to performance spread. Therefore,

reliability is reduced where precision is poor or where the population is very homogeneous. It is important that the sample of trainees is reasonably large and representative if the reliability results of the evaluation are to be extrapolated to the general population of trainees. Our interim analysis demonstrated that OSATS reliability was significantly different from PBA, raising the possibility of a cohort-specific effect. A high proportion of senior trainees with training concerns was found in our O&G cohort, which made this population more homogeneous. To optimise conditions for evaluating the reliability of PBA/OSATS within O&G, we sustained our attempt to capture as wide a sample of O&G trainees and clinical supervisors as possible within the timescale of the study. We sampled another entire trainee cohort after the August 2008 changeover. Overall, the recruitment of trainees within O&G approached 100% (two trainees declined to consent).

The response rates for some of the user satisfaction and acceptability follow-up questionnaires were suboptimal. While the response rates of clinical supervisors and trainees were very good and well above accepted questionnaire response rate thresholds, those for the anaesthetists and nurses were disappointing. This may affect the generalisability of the NOTSS questionnaire results. Initially the response rates were poor for all participant groups, and our efforts at resending e-mails and hard-copy questionnaires proved unsuccessful in increasing response rates. It became clear that the most successful method of retrieving questionnaires was by approaching personnel face to face with a hard copy at a time that was convenient for them to complete it. We had not anticipated quite how time consuming this approach would be. We made a decision to prioritise retrieving trainees' and clinical supervisors' questionnaires. In so doing, we lost responses from the other professional groups. We would have needed to allocate much more time to this aspect of data collection for NOTSS questionnaire response rates to have been significantly improved.

As OSATS and PBA appear to be alternative tools for use in WBA, it would have been ideal to evaluate their validity by estimating the correlation between them when used together across cases. However, the estimation of reliability requires multiple assessors to use the same tool as they observe a given case. Unless there are three or more simultaneous 'technical' assessors, it is not possible to gather good reliability data for two 'technical' instruments in parallel. In taking a pragmatic approach, we prioritised obtaining reliability evidence for the tools.

At an early stage of the study, the Steering Committee suggested that each independent assessor should minimise the number of cases for which they used a 'technical' PBA or OSATS with the non-technical NOTSS simultaneously, owing to concern about cross-contamination of the ratings. This would have required two independent assessors to be present for every case. However, this was not often possible because of the priority we gave to sampling as many cases as possible for each trainee. Therefore, for these cases, it is possible that the completion of more than one tool informed the independent assessor ratings, particularly with respect to the non-technical skills that form part of the content of the 'technical' tools. In addition, a further potential confounder to assessment ratings could stem from the recognised different methodological approaches. For the 'technical' tools there is an expectation that trainees should verbalise their intentions throughout the procedure. This is made explicit on the PBA form itself. However, the NOTSS tool is designed for use without prompting verbalisation of behaviours. The independent assessor's approach was not to prompt the trainees to fulfil the methodological expectations for a given tool, but the supervising clinical supervisor often had an expectation that trainees would verbalise their intentions and actions throughout the cases. However, except for the OSATS ATGS, the independent assessors did not differ in the stringency of their ratings compared with other relevant assessor groups. Given the overall similarity between assessor designation stringency, it strongly suggests that the independent assessors' ratings were not unduly influenced by the completion of more than one tool.

The proportion of index procedures that could be video recorded was disappointing. The reasons contributing to this are detailed in *Chapter 3*. As we were unable to record sufficient numbers of individual index procedures, we were unable to complete work on the reliability of the study tools using video assessment. Furthermore, the quality of the recordings was suboptimal for some index procedures. Better recordings were obtained with cameras built into the theatre lights than with the study's filming equipment. However, the majority of our recordings relied upon using study equipment because of the limited availability of trust operating theatres with purpose-built recording equipment. The study team's ongoing work with video assessment is discussed within the further research section.

Reflections and lessons learned

Reflecting on the study's design, implementation and results offers the opportunity to explore, in hindsight, how we might have approached fulfilling the study's primary outcomes differently. This is a complex argument to present as the implementation of the study was developmental, its design evolving in response to the clinical setting in which it took place. While we are satisfied with the majority of the primary outcomes, particularly the validity and reliability results, the user satisfaction and acceptability data were less robust. It may be that these outcomes could have been better fulfilled by allocating more of the team's resources to optimising questionnaire response rates. Also, the use of additional methodology, such as focus groups or interviews, could have been employed in the design. During the course of the study's implementation, we explored the advantage of using focus groups to support the questionnaire data under collection. However, there were insufficient research team resources to take this forward. On balance, the questionnaire methodology allowed the widest sampling of participant opinions.

Inevitably, discussing how resources could have been better focused on a specific research outcome demands a decision on how finite resources could have been reallocated. There was a significant, and ongoing, investment of the research team's resources involved in the video recording of cases. In fact, the validity (fidelity) of video recording as an indicator of directly observed performance was a secondary research outcome. This, together with the shortcomings in achieving sufficient numbers for reliability analysis, leads us to reflect that this aspect of the study was probably allocated a disproportionate amount of time and resources.

Working within the confines of the grant resources proved difficult for the research team. In practice, the process of identifying cases and undertaking assessments was extremely time consuming. This had very real implications for the study, which quickly became apparent during implementation. The study protocol had ambitiously aimed to recruit from three different cities. However, it was unfeasible for the single study co-ordinator to travel to other centres when a full-time presence on site was essential for successful recruitment. Effectively these constraints limited the study to one centre. There may be lessons to be learned from this for researchers wishing to undertake similar workplace-based studies when making grant funding applications.

Further research

Ongoing work

Some of the ongoing research projects that we as a research team are continuing to develop have already been highlighted within the relevant discussion sections. These can be summarised as follows:

1. Does WBA capture authentic performance? The narrative comments from trainees and clinical supervisors on the effect of assessment on performance will be subjected to qualitative analysis to supplement our quantitative data.
2. What is the quality of the feedback that clinical supervisors provide when completing WBA? We have collected structured observations of 56 feedback sessions and these will be analysed qualitatively to explore the characteristics and quality of verbal feedback given to trainees. It is anticipated that the qualitative analysis will identify key themes that can then be explored in subsequent focus groups with trainees and clinical supervisors.
3. Can our NOTSS data be used to provide further validation and development of the NOTSS system? Our NOTTS assessors completed the tools by providing examples of behaviours to justify their ratings. These behavioural markers will be subjected to qualitative analysis that may identify new markers for the NOTSS system. The collaboration with the NOTSS team is ongoing and we obtained ethical approval to share our DVD material with this expert group. The proposal is to use the DVDs to identify more observable non-technical behaviours, particularly in leadership. This research forms part of an ongoing process for improving the sensitivity and validity of the NOTSS system as well as extending the NOTSS framework for use in surgical training assessment programmes.
4. What influences opinions on WBA acceptability and user satisfaction? Although our questionnaires have been subjected to quantitative analysis for this report, there remains a body of narrative evidence to be analysed qualitatively to fully explore the trainees' and trainers' attitudes to these WBA tools. This is especially relevant for our O&G trainees and clinical supervisors for whom user satisfaction of OSATS and PBA can be compared.
5. Can DVDs be used to provide reliable performance assessments of trainees? Good inter-rater reliability between direct and video assessment of saphenofemoral ligation has been demonstrated.[53] However, Scott *et al.*[137] found that assessment of edited videotapes of laparoscopic cholecystectomies did not correlate well with direct observation. A study conducted on behalf of the Vascular Society also found that silent video recordings of trainees performing carotid endarterectomies could not be scored reliably because of difficulty in gauging how much help was provided by the clinical supervisor (unpublished data). The reliability of video assessment would be anticipated to be improved by dual recordings of the operative field and the operating theatre, combined with voice recordings, as used in our study. Unfortunately, we were unable to video sufficient cases of individual index procedures to allow dependable reliability estimates. However, video assessment by several assessors per DVD may be used to compensate for our small sample sizes. The excellent reliability of direct observation using PBA means that testing the reliability of video assessment may become less important. However, video recordings of operations may have other important educational roles for training trainees and assessors, which are discussed within the *New research* section.
6. Can using DVDs of supervised surgical cases improve training for assessors? The use of videos within training workshops or educational courses could help assessors understand how to use WBA tools in making judgements of trainee surgical performance. The evidence from the occupational psychology literature shows that the most effective method for training assessors to provide accurate performance ratings is through frame-of-reference training.[138] This involves providing assessors with examples of performances at different levels of competence with clear standards for assessments, with opportunities to practise giving scores and feedback. We suggest that DVDs of operations used in these settings would provide ideal frame-of-reference material for training clinical supervisors. Therefore, we have begun to use DVDs within workshops at various Royal Colleges of Surgeons with good feedback from participants. We plan to use the data collected from the workshops to study the effect of such training on the reliability of WBA tools.

New research

In addition, there are a number of new research areas that we have identified as being worthy of further pursuit:

1. Does WBA demonstrate educational effectiveness, i.e. does it enable trainees to achieve the required surgical training competencies and improve patient outcomes? This study provided some evidence of educational effectiveness at level 1 and level 2 of the Kirkpatrick model as adapted by Freeth *et al.*,[18] including user satisfaction perspectives and changes in attitudes, although it was not one of our study's aims. Evidence for changes in behaviours and changes in patient outcomes at levels 3 and 4, respectively, require research of assessment programmes rather than of individual WBA tools using longitudinal and integrated research methodologies.
2. Is there a relationship between surgical experience, performance and outcomes? Can surgical outcome data be used to assess trainees? This study shows some promising evidence of outcome validity. Further large-scale studies are required to evaluate this as a potential method of assessing the surgical performance of more senior trainees.
3. Can we establish learning curves and minimum number requirements for index procedures? Longitudinal collection of PBA forms and logbook data nationally should permit calculation of the numbers of procedures required for the 'average' trainee to achieve competence.
4. Are the PBA domains that were not part of this study (e.g. preoperative planning and consent) also valid and reliable?
5. Is WBA affected by factors such as language skills and cultural background?
6. Can non-surgeons assess the surgical performance of trainees using WBA? We have shown that non-surgeons, e.g. anaesthetists, scrub nurses and SCPs, can reliably assess the non-technical skills of trainees using NOTSS. Our results also indicated that SCPs could produce as reliable PBA ratings as those by clinical supervisors, as one independent assessor was an SCP. Further research is needed to establish whether SCPs are reliable as an assessor group.
7. Can DVDs be used to provide trainees with additional feedback on their surgical performance? Videos can be used by trainees to review the key stages of a procedure before entering the operating theatre and to review the management of adverse events. They may also have a role in reinforcing the feedback provided after direct observation by a clinical supervisor. Feedback on performance using videos is well established within general practice for patient consultation skills.[139–141] There is also evidence that giving trainees feedback on their surgical performance improves their surgical skill.[142] Trainees could be provided with DVDs of their operations to review, with the PBA/OSATS assessment provided by their clinical supervisor, for additional feedback on their surgical performance.

Chapter 5

Conclusions

Procedure-based assessment possesses high reliability ($G > 0.8$ using three assessors for the same index procedure), excellent construct validity and positive user satisfaction and acceptability perspectives from trainees and clinical supervisors. Given its high validity, reliability and acceptability, PBA demonstrates good evidence of overall assessment utility. These results indicate that PBA is highly suitable as an assessment *for* learning and as part of an assessment *of* learning. Therefore, 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 of PBA is procedure specific and requires that trainees are adequately assessed for each individual index procedure. We have no reason to believe that PBA would be less valid or reliable in other surgical specialties; however, further evaluation within other specialties may be useful.

Objective Structured Assessment of Technical Skills is a less reliable method than PBA as a tool for assessing predominately 'technical' skills. However, good reliability for assessing the same procedure remains achievable using feasible numbers of surgical cases ($G > 0.8$ using five assessors for the same index procedure). OSATS failed to show construct validity for all demographic measures of age and experience. However, the context of our OSATS evaluation within the specialty of O&G had fundamental cohort differences from the other surgical specialties in which PBA was evaluated. The high proportion of senior O&G trainees with training concerns made the population more homogeneous, resulting in reduced estimated reliability and undermining the construct validity evidence for OSATS.

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, with evidence of a valid internal structure and good construct validity. Good reliability ($G > 0.8$) can be achieved using eight assessors for a mix of procedures, without intensive assessor training. Given the prerequisite that reliable assessment using PBA/OSATS demands adequate assessment of individual procedures, there would be no foreseeable difficulty in obtaining an adequate sample of a mix of procedures to permit a reliable assessment of non-technical skills using NOTSS. NOTSS may complement the predominantly 'technical' WBA tools especially for trainees who have mastered the technical aspects of a procedure. The 'technical' tools showed concurrent validity with NOTSS, demonstrated by score correlations between OSATS/PBA and NOTSS. This suggests that NOTSS is valid for providing a supplementary assessment of surgical skill, as part of the overarching assessment construct intending to measure surgical performance. Surgical training programmes may wish to consider the inclusion of NOTSS into their assessment framework and/or considering integrating elements of the NOTSS into their 'technical' WBA tools.

Taking into account all the study assessment tools, there is some evidence of predictive (outcome) validity for WBA. We found twice as many significant correlations between case outcomes and scores than could be expected by chance alone.

The variance component analyses used to estimate reliability reveal that assessor designation (i.e. the different assessors involved) does not affect their scoring stringency using PBA and NOTSS. 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. These results suggest that PBA could be completed by SCPs and NOTSS by anaesthetists, scrub nurses and SCPs.

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.

Regarding the reported impact of assessment on surgical performance, there was little agreement between clinical supervisor and trainee perspectives. Although the cases judged to have been affected by direct observational assessment had lower scores, the video-recorded cases did not, suggesting neither affect performance a priori but rather that a poor performance may be attributed to assessment or video recording. This supports the intention of WBA in assessing authentic surgical performance.

Our difficulties with study recruitment 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 may 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. While we acknowledge that conflicts between training and service provision are inherent within the context of the NHS, we have identified three levels of obstacles to achieving systematic supervised training in the operating theatre with corresponding solutions presented:

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

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 investigate the reliability, validity and user acceptability of, and satisfaction with, PBA, OSATS and NOTSS – were achieved.

Acknowledgements

Steering Committee

We gratefully acknowledge the many useful comments and suggestions from experts in WBA around the world. We are particularly grateful for the advice provided by our Steering Committee:

- Professor David Rowley (Chairperson), Director of Education, Royal College of Surgeons of Edinburgh
- Professor Richard Reznick, Head of Surgery, University of Toronto, Canada
- Professor Brian Jolly, Head of Department, Centre for Medical and Health Sciences Education, Monash University, Australia
- Mr William Thomas, Vice President, Royal College of Surgeons of England.

Sadly, our colleague Dr Helena Davies was forced to retire as Chairperson of the Steering Committee at an early stage owing to ill health. We wish her well and thank Professor Rowley for taking over the Chair.

Contribution of authors

Jonathan Beard (Professor of Surgical Education) was the Principal Investigator. He wrote the project grant application, obtained ethics approval, supervised the study, acted as an independent assessor and edited the draft and revised report.

Helen Purdie (Senior Research Sister) was the study co-ordinator and acted as an independent assessor. She was responsible for day-to-day study management including compliance with research governance policies and guidelines and communication with external bodies and committees, which included the writing of interim reports. She created, inputted and maintained the study databases. She co-authored the drafts of *Chapters 2* and *4* and codrafted the revised report.

Joy Marriott (Research Fellow) was appointed as a further independent assessor. She took responsibility for the recruitment and co-ordination of assessments within O&G, including ethical amendments for inclusion of this specialty. She was also responsible for developing PBA forms and follow-up questionnaires in O&G. She undertook some of the statistical analysis, drafted *Chapters 1–4* of the report and compiled the reference database. She also co-drafted the revised report.

Jim Crossley (Senior Research Fellow in Medical Education) advised on study design, analysed the data, assisted with the interpretation of findings and contributed in full to the final report.

Contribution of others

We also wish to thank the following for their help, advice and support:

All of the surgical trainees and consultant clinical supervisors who participated in the study.

The NOTSS team from Aberdeen and Edinburgh: Steven Yule, Rhona Flin, Nicola Maran, David Rowley and Simon Paterson-Brown.

Members of the ISCP, the OCAP and the Departments of Education at the Royal Colleges of Surgeons of England and Edinburgh: Maria Bussey, Adrian Woodthorpe, Ruth McKee, David Rowley and David Pitts.

William Ledger, Tom Farrell, Peter Stewart and Diana Fothergill from the Jessop Wing, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield.

Jean Russell, Statistician in Computing Services at the University of Sheffield.

Georgina Jones, Senior Lecturer in Social Sciences at the University of Sheffield.

The Education Committee of the Royal College of Obstetricians and Gynaecologists.

David Equeall from the Medical Photography Department, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield.

Jane Eyre, Personal Assistant to Professor Beard.

References

1. Thayer WS. Osler, the teacher. *Johns Hopkins Hospital Bulletin* 1919;**30**:198–200.
2. Halsted WS. The training of the surgeon. *Bull Johns Hopkins Hospital* 1904;**15**:267–76.
3. Calman KC, Temple JG, Naysmith R, Cairncross RG, Bennett SJ. Reforming higher specialist training in the United Kingdom – a step along the continuum of medical education. *Med Educ* 1999;**33**:28–33.
4. Donaldson L. *Unfinished business – proposals for reform of the Senior House Officer grade*. London: Department of Health; 2002.
5. Postgraduate Medical Education and Training Board. *Standards for curricula and assessment systems*. London: PMETB; 2008. Available from: www.pmetb.org.uk/fileadmin/user/Standards_Requirements/PMETB_Scas_July2008_Final.pdf (accessed 29 March 2010).
6. Bloom B. *Taxonomy of educational objectives: the classification of educational goals: Handbook 1: Cognitive domain*. New York, NY: David MacKay; 1971.
7. Otter S. *Learning outcomes in higher education*. London: Further Education Unit/Unit for the Development of Adult Continuing Education; 1992.
8. Wolf A. *Competence based assessment*. Milton Keynes: Open University Press; 1995.
9. Department of Education. *Working together: education and training*. London: HMSO; 1986.
10. Fordham AJ. Using a competency based approach in nurse education. *Nurs Standard* 2005;**19**:41–8.
11. Harden RM, Crosby JR, Davis MH. AMEE guide no. 14: Outcome-based education: Part 1 An introduction to outcome education. *Med Teacher* 1999;**21**:7–14.
12. Newble D, Stark P, Bax N, Lawson M. Developing an outcome-focused core curriculum. *Med Educ* 2005;**39**:680–7.
13. Beard J, Strachan A, Davies H, Patterson F, Stark P, Ball S, *et al.* Developing an education and assessment framework for the Foundation Programme. *Med Educ* 2005;**39**:841–51.
14. MMC. *Modernising Medical Careers*. 2009. Available from: www.mmc.nhs.uk (accessed 10 June 2009).
15. Foundation Programme. *Training and assessment*. 2009. Available from: www.foundationprogramme.nhs.uk/pages/home/training-and-assessment (accessed 10 June 2009).
16. Postgraduate Medical Education and Training Board and Academy of Medical Royal Colleges. *Workplace based assessment (WPBA): a guide for implementation*. London: PMETB and AOMRC; 2009.
17. Kirkpatrick DL. *Evaluating training programs: the four levels.* San Francisco, CA: Berrett-Koehler Publishers, Inc.;1994.
18. Freeth D, Hammick M, Reeves S, Barr H. *Critical review of evaluations of interprofessional education*. London: Higher Education Academy Learning and Teaching Support Network for Health Sciences and Practice; 2002.
19. Overeem K, Faber MJ, Arah OA, Elwyn G, Lombarts KMJMH, Wollerssheim HC, *et al.* Doctor performance assessment in daily practice: does it helps doctors or not? A systematic review. *Med Educ* 2007;**41**:1039–49.

20. Fox AT, Palmer RD, Crossley JGM, Sekaran D, Trewavas ES, Davies HA. Improving the quality of outpatient clinic letters using the Sheffield Assessment Instrument for Letters (SAIL). *Med Educ* 2004;**38**:852–8.

21. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;**39**:309–17.

22. Martin M, Vashisht B, Frezza E, Ferone T, Lopez B, Pahuja M, *et al.* Competency-based instruction in critical invasive skills improves both resident performance and patient safety. *Surgery* 1998;**124**:313–17.

23. Porte MC, Xeroulis G, Reznick RK, Dubrowski A. Verbal feedback from an expert is more effective than self-assessed feedback about motion efficiency in learning new surgical skills. *Am J Surg* 2007;**193**:105–10.

24. van Sickle KR, Gallagher AG, Smith CD. The effect of escalating feedback on the acquisition of psychomotor skills for laparoscopy. *Surg Endosc* 2007;**21**:220–4.

25. Katory M, Singh S, Beard JB. Twenty Trent trainees: a comparison of operative competence after BST. *Ann R Coll Surg Engl* 2001;**83**(Suppl.):328–30.

26. Department of Health. *HSC 2003/001 – Protecting staff, delivering services: implementing the European Working Time Directive for doctors in training*. London: DoH; 2003. Available from: www.dh.gov.uk/en/publicationsandstatistics/lettersandcirculars/healthservicecirculars/DH_4003588 (accessed 10 June 2009).

27. Academy of Medical Royal Colleges. *Improving assessment*. London: Academy of Medical Royal Colleges; 2009. Available from: www.aomrc.org.uk/aomrc/admin/reports/docs/IMPROVING_ASSESSMENT_EMAIL.pdf (accessed 15 July 2009).

28. Healthcare Assessment and Training. *Directly observed procedural skills*. 2009. Available from: www.hcat.nhs.uk/foundation/ (accessed 22 June 2008).

29. Intercollegiate Surgical Curriculum Programme. *Procedure based assessment*. 2009. Available from: www.iscp.co.uk/Assessment/WBA/PBA.aspx (accessed 2 June 2009).

30. Royal College of Obstetricians and Gynaecologists. *Core OSATS*. 2009. Available from: www.rcog.org.uk/files/rcog-corp/uploaded-files/Ed-Core-OSATS.pdf (accessed 22 June 2009).

31. Halpin G, Halpin G. Experimental investigation of the effects of study and testing on student learning, retention, and ratings of instruction. *J Educ Psychol* 1982;**74**:32–8.

32. Newble DI, Jaeger K. The effects of assessments and examinations on the learning of medical students. *Med Educ* 1983;**17**:165–71.

33. Stillman PL, Haley HL, Regan MB, Philbin MM. Positive effects of a clinical performance assessment program. *Acad Med* 1991;**66**:481–3.

34. Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ* 2002;**36**:800–4.

35. Epstein RM, Hundert EM. Defining and assessing professional competence (review). *JAMA* 2002;**287**:226–35.

36. Hays R, Wellard R. In-training assessment in postgraduate training for general practice. *Med Educ* 1998;**32**:507–13.

37. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;**65**(Suppl.):S63–7.

38. Rethans J, Norcini J, Baron-Maldonado M, Blackmore D, Jolly BC, LaDuca T, *et al.* The relationship between competence and performance: implications for assessing practice performance. *Med Educ* 2002;**36**:901–9.

39. Reznick RK. Teaching and testing technical skills. *Am J Surg* 1993;**165**:358–61.

40. Rogers CR. *Freedom to learn*. Columbus, OH: Merrill; 1969.

41. Ram P, van der Vleuten C, Rethans JJ, Schouten B, Hobma S, Grol R. Assessment in general practice: the predictive value of written-knowledge tests and a multiple-station examination for actual medical performance in daily practice. *Med Educ* 1999;**33**:197–203.

42. Rethans JJ, Sturmans F, Drop R, van der Vleuten C, Hobus P. Does competence of General Practitioners predict their performance? Comparison between examination setting and actual practice. *BMJ* 1991;**303**:1377–80.

43. Ramsey PG, Wenrich MD, Carline J. Use of peer ratings to evaluate physician performance. *JAMA* 1993;**269**:1655–60.

44. Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons from the health professions. *Educ Researcher* 1995;**24**:5–11.

45. Kopelow ML, Schnabl GK, Hassard TH, Tamblyn RM, Klass DJ, Beazley G, *et al.* Assessment of performance in the office setting with standardised patients: assessing practicing physicians in two settings using standardised patients. *Acad Med* 1992;**67**:S19–21.

46. Norman G. Editorial: The long case versus objective structured clinical examinations. *Med Educ* 2002;**324**:748–9.

47. Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL. Knowledge and clinical problem solving. *Med Educ* 1985;**19**:344–56.

48. Elstein AS, Shulman LS, Sprafka SS. *Medical problem solving*. Cambridge, MA: Harvard University Press; 1978.

49. Streiner DL. Global rating scales. In: Neufield VR, Norman GR (eds). *Assessing clinical competence*. New York, NY: Springer Publishing Company; 1985. pp. 114–41.

50. General Medical Council. *Good Medical Practice*. London: GMC; 2006.

51. Regehr G, MacRae H, Reznick R, Szalay D. Comparing the psychometric properties of checklists and global ratings for assessing performance on an OSCE-format examination. *Acad Med* 1998;**73**:993–7.

52. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, *et al.* Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;**84**:273–8.

53. Beard JD, Jolly BC, Newble DI, Thomas WEG, Donnelly J, Southgate LJ. Assessing the technical skills of surgical trainees. *Br J Surg* 2005;**92**:778–82.

54. Schuwirth LWT, Southgate LH, Page GG, Paget NS, Lescop JMJ, Lew SR, *et al.* When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ* 2002;**36**:925–30.

55. Southgate L, Cox J, David T, Hatch D, Howes A, Johnson N, *et al.* The assessment of poorly performing doctors: the development of the assessment programmes for the General Medical Council's performance procedures. *Med Educ* 2001;**35**(Suppl.1):2–8.

56. van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1996;**1**:47–67.

57. Postgraduate Medical Education and Training Board. *Principles for an assessment system for postgraduate medical training*. London: PMETB; 2004.

58. Streiner DL, Norman GR. *Health measurement scales*, 2nd edn. New York, NY: Oxford University Press; 1995.

59. Jolly B, Grant J. *The good assessment guide: a practical guide to assessment and appraisal for higher specialist training.* London: Joint Centre for Education in Medicine; 1997.

60. Beard JD. Assessment of surgical skills of trainees in the UK. *Ann R Coll Surg Engl* 2008;**90**:282–5.

61. Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ* 2004;**38**:1006–12.

62. Holsgrove G. *Reliability issues in the assessment of small cohorts*. London: Postgraduate Medical Education and Training Board, 2009.

63. Ellis H. *Famous operations.* Malvern, PA: Harwal Medical Publications; 1984.

64. Szalay D, MacRae H, Regehr G, Reznick RK. Using operative outcome to assess technical skill. *Am J Surg* 2000;**180**:234–7.

65. European Trialists' Collaborative Group. MRC European Carotid Surgery Trial: interim results for symptomatic patients with severe (70–99%) or with mild (0–29%) carotid stenosis. *Lancet* 1991;**337**:1235–43.

66. Prytherch DR, Ridler BMF, Beard JB, Earnshaw JJ, on behalf of the Audit and Research Committee of the Vascular Society of Great Britain and Ireland. A model for national outcome audit in vascular surgery. *Eur J Vasc Endovasc Surg* 2001;**21**:477–83.

67. Galasko C, MacKay C. Unsupervised surgical training – logbooks are essential for assessing progress. *BMJ* 1997;**315**:1306–7.

68. Thornton M, Donlon M, Beard JB. The operative skill of higher surgical trainees: measuring competence acheived rather than experience undertaken. *Ann R Coll Surg Engl* 2003;**85**(Suppl.):190–3.

69. Burt CG, Chambers E, Maxtad M, Grant JR, Markham N, Watts H, *et al.* The evaluation of a new method of operative competence assessment for surgical trainees. *Bull R Coll Surg* 2003;**85**(Suppl.):152–5.

70. Royal College of Surgeons. *Training the trainers: learning and teaching.* London: RCS;1996.

71. Baldwin PJ, Paisley AM, Paterson-Brown S. Consultant surgeons' opinion of the skills required of basic surgical trainees. *Br J Surg* 1999;**86**:1078–82.

72. Thorndike EL. A constant error in psychological ratings. *J Appl Psychol* 1920;**4**:25–9.

73. Kent RN, Foster SL. Direct observational procedures: methodological issues in naturalistic settings. In: Ciminero AR, Calhoun KS, Adams HE (eds). *Handbook of behavioural assessment.* New York, NY; 1977.

74. Risucci DA, Tortolani AJT, Ward RJ. Ratings of surgical residents by self, supervisors and peers. *Surg Gynecol Obstet* 1989;**169**:519–26.

75. Gould DA, Reekers JA. The role of simulation in training endovascular interventions. *Eur J Vasc Endovasc Surg* 2008;**35**:633–6.

76. Lentz GM, Mandel LS, Goff BA. A six-year study of surgical teaching and skills evaluation for obstetric/gynaecologic residents in porcine and inanimate surgical models. *Am J Obstet Gynecol* 2005;**193**:2056–61.

77. Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teacher* 2005;**27**:10–28.

78. Scott DJ, Bergen PC, Rege RV, Laycock R, Tesfay ST, Valentine RJ, *et al.* Laparoscopic training on bench models: better and more cost effective than operating room experience? *J Am Coll Surg* 2000;**191**:272–83.

79. Datta V, Bann S, Beard J, Mandalia M, Darzi A. Comparison of bench test evaluations of surgical skill with live operating performance assessments. *J Am Coll Surg* 2004;**199**:603–6.

80. Coleman RL, Muller CV. Effects of a laboratory-based skills curriculum on laparoscopic proficiency: a randomised trial. *Am J Obstet Gynecol* 2002;**186**:836–42.

81. Moorthy K, Munz Y, Sarker SK, Darzi A. Objective assessment of technical skills in surgery. *BMJ* 2003;**327**:1032–7.

82. Larsen CR, Grantcharov T, Aggarwal R, Tully A, Sorensen JL, Dalsgaard T, *et al.* Objective assessment of gynecologic laparoscopic skills using the LapSimGyn virtual reality simulator. *Surg Endosc* 2006;**20**:1460–6.

83. Gallagher AG, Satava RM. Virtual reality as a metric for the assessment of laparoscopic psychomotor skill. *Surg Endosc* 2002;**16**:1746–52.

84. Gallagher AG, Smith CD, Bowers SP, Seymour NE, Pearson A, McNatt S, *et al.* Psychomotor skills assessment in practicing surgeons experienced in performing advanced laparoscopic skills. *J Am Coll Surg* 2003;**197**:479–88.

85. Grantcharov T, Bardram L, Funch-Jensen P, Rosenberg J. Learning curves and impact of previous operative experience on performance on a virtual reality simulator to test laparoscopic skills. *Am J Surg* 2003;**185**:146–9.

86. Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Bansal VK, Andersen DK, *et al.* Virtual reality training improves operating room performance: results of a randomised, double-blinded study. *Ann Surg* 2002;**236**:458–63.

87. Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jenson P. Randomised clinical trial of virtual reality simulation for laparoscopic skills training. *Br J Surg* 2004;**91**:146–50.

88. Haycock AV, Youd P, Bassett P, Saunders BP, Tekkis P, Thomas-Gibson S. Simulator training improves practical skills in therapeutic GI endoscopy: results from a randomised, blinded, controlled study. *Gastrointest Endosc* 2009;**70**:835–45.

89. Chaer RA, DeRubertis BG, Lin SC, Bush HL, Karowski JK, Birk D, *et al.* Simulation improves resident performance in catheter based intervention. *Ann Surg* 2006;**244**:343–9.

90. Dawson S. Procedural simulation: a primer. *J Vasc Intervent Radiol* 2006;**17**:205–13.

91. Tubbs RJ, Murphy B, Mainiero MB, Shapiro M, Kobayashi L, Lindquist D, *et al.* High-fidelity medical simulation as an assessment tool for radiology residents' acute contrast reaction management skills. *J Am Coll Radiol* 2009;**6**:582–7.

92. Overly FL, Sudikoff SN, Shapiro MJ. High-fidelity medical simulation as an assessment tool for pediatric residents' airway management skills. *Pediatr Emerg Skills* 2007;**23**:11–15.

93. Wayne DB, Didwania A, Feinglass J, Fudala MJ, Barsuk JH, McGaghie WC. Simulation-based education improves quality of care during cardiac arrest team responses at an academic teaching hopsital. *Chest* 2008;**133**:56–61.

94. Paige JT, Kozmenko V, Yang T, Gururaja RP, Hilton CW, Cohn I, *et al.* High-fidelity, simulation-based, interdisciplinary operating room team training at the point of care. *Surgery* 2009;**145**:138–46.

95. Cooper WH. Ubiquitous halo. *Psychol Bull* 1984;**90**:218–44.

96. Paisley AM, Baldwin PJ, Patterson-Brown S. Accuracy of medical staff assessment of trainees' operative performance. *Med Teacher* 2005;**27**:634–8.

97. Yule S, Flin R, Paterson-Brown S, Maran N. Non-technical skills for surgeons in the operating room: a review of the literature. *Surgery* 2006;**139**:140–9.

98. Gawande AA, Zinner MJ, Studdert DM, Brennan TA. Analysis of error reported by surgeons at three teaching hospitals. *Surgery* 2003;**133**:614–21.

99. Christian C, Gustafson M, Roth E, Sheridan T, Gandhi T, Dwyer K, *et al.* A prospective study of patient safety in the operating room. *Surgery* 2006;**139**:159–73.

100. Pitts D, Rowley DI, Sher JL. Assessment of performance in orthopaedic training. *J Bone Joint Surg* 2005;**87**:1187–91.

101. Winckel CP, Reznick RK, Cohen R, Taylor B. Reliability and construct validity of a structured technical skills assessment form. *Am J Surg* 1994;**167**:423–7.

102. Goff B, Nielsen PE, Lentz G, Chow GE, Chalmers RW, Fenner D. Surgical skills assessment: a blinded examination of obstetrics and gynaecology residents. *Am J Obstet Gynecol* 2002;**186**:613–17.

103. Beard JB, on behalf of the Education and Training Commitee of the Vascular Society of Great Britain and Ireland. Setting standards for the assessment of operative competence. *Eur J Vasc Endovasc Surg* 2005;**30**:215–18.

104. RCOG. *Training and curriculum amendments*. 2009. Available from: www.rcog.org.uk/files/rcog-corp/uploaded-files/ED-Curriculum-Amends-March2009.pdf#page=1 (accessed 9 April 2009).

105. RCOG. *Trainees' frequently asked questions.* 2008. Available from: www.rcog.org.uk/index.asp?PageID=2326 (accessed 2 February 2009).

106. Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons' non-technical skills. *Med Educ* 2006;**40**:1098–104.

107. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *Br J Anaesth* 2003;**90**:580–8.

108. Dickinson I, Watters D, Graham I, Montgomery P, Collins J. Guide to the assessment of competence and performance in practicing surgeons. *Aust N Z J Surg* 2009;**79**:198–204.

109. Yule S, Flin R, Maran N, Rowley D, Youngson G, Duncan J, *et al.* Development and evaluation of the NOTSS behaviour rating system for intraoperative surgery. In: Flin R, Mitchell L (eds). *Safer surgery: Analysing behaviour in the operating theatre*. Farnham: Ashgate; 2009. pp. 7–25.

110. Yule S, Flin R, Maran N, Rowley D, Youngson G, Paterson-Brown S. Surgeons' Non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. *World J Surg* 2008;**32**:548–56.

111. Yule S, Rowley D, Flin R, Maran N, Youngson G, Duncan J, *et al.* Experience matters: comparing novice and expert ratings of non-technical skills using the NOTSS system. *Aust N Z J Surg* 2009;**79**:1–7.

112. Yule S, Flin R, Maran N, Youngson G, Mitchell A, Rowley D, *et al.* Debriefing surgeons on non-technical skills. *Cognition Technol Work* 2008;**10**:265–74.

113. Gold RL. Roles in sociological field observations. *Social Forces* 1958;**36**:217–23.

114. Cohen L, Manion L, Morrison K. *Research methods in education*, 5th edn. London: Routledge Falmer; 2000.

115. Morrison J. ABC of learning and teaching in medicine: evaluation. *BMJ* 2003;**326**:385–7.

116. Woodward CA. Questionnaire construction and question writing for research in medical education. *Med Educ* 1988;**22**:345–63.

117. Cassar K. Development of an instrument to measure the surgical operating theatre learning environment as perceived by basic surgical trainees. *Med Teacher* 2004;**26**:260–4.

118. Roff S, McAleer S, Skinner A. Development and validation of an instrument to measure the postgraduate clinical learning and teaching educational environment for hospital-based junior doctors in the UK. *Med Teacher* 2005;**27**:326–31.

119. Grant J, Kilminster S, Jolly B, Cottrell D. Clinical supervision of SpRs: where does it happen, when does it happen and is it effective? *Med Educ* 2003;**37**:140–8.

120. Marriott JC, Purdie H, Crossley J, Beard JD. Implementing the assessment of surgical skills and non-technical behaviours in the operating room. In: Flin R, Mitchell L (eds). *Safer surgery: Analysing behaviour in the operating theatre*. Farnham: Ashgate; 2009. pp. 47–66.

121. Crossley J, Davies H, Humphris G, Jolly BC. Generalisability: a key to unlock professional assessment. *Med Educ* 2002;**36**:972–8.

122. Crossley J, Russell J, Jolly BC, Ricketts C, Roberts C, Schuwirth LWT, *et al.* I'm pickin' up good regressions: the governance of generalisability analyses. *Med Educ* 2007;**41**:926–34.

123. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No.31. *Med Teacher* 2007;**29**:855–71.

124. Hattie J, Timperley H. The power of feedback. *Rev Educ Res* 2007;**77**:81–112.

125. Beard JB, Rowley D, Woodthorpe A, Foulkes J. Workplace-based assessment: an evalution of the use of surgical DOPS in the intercollegiate surgical curriculum project. *Br J Surg* 2009;**96**(Suppl. 4):76.

126. Page ES. Continuous inspection schemes. *Biometrika* 1954;**41**:100–14.

127. Van Rij AM, McDonald JR, Pettigrew RA, Putterill MJ, Reddy CK, Wright JJ. Cusum as an aid to early assessment of the surgical trainee. *Br J Surg* 1995;**82**:1500–3.

128. Fradkin D, Tolhurst-Cleaver S, Palmer J. A learning curve for all: CUSUM curves in initial assessment of competency. *R Coll Anaesth Bull* 2009;**54**:13–15.

129. Williams SM, Parry BR, Schlup MMT. Quality control: an application of the cusum. *BMJ* 1992;**304**:1359–61.

130. Lanigan C, Blanco R. CUSUM scoring: theory and practice. *R Coll Anaesth Bull* 2009;**54**:16–19.

131. James K, Cross K, Lucarotti ME, Fowler AL, Cook TA. Undertaking procedure-based assessment is feasible in clinical practice. *Ann R Coll Surg Engl* 2009;**91**:110–12.

132. Postgraduate Medical Education and Training Board. *Developing and maintaining an assessment system – a PMETB guide to good practice*. London: PMETB; 2007.

133. Hewson MG, Little ML. Giving feedback in medical education: verification of recommended techniques. *J Gen Intern Med* 1998;**13**:111–16.

134. Riley S, Smith D, Le Rolland P. *National Survey of Trainers 2007 Summary Report*. London: Postgraduate Medical Education and Training Board; 2007.

135. Postgraduate Medical Education and Training Board. *Generic standards for training Version 1.1. September* 2009. Available from: www.pmetb.org.uk/fileadmin/user/Standards_Requirements/PMETB_Gst_Sept2009.pdf (assessed 29 March 2010).

136. Mason V, Balloo S, Upton D, Heer K, Higton P, Shiralkar U. Surgeons' experience of learning psychological skills: a preliminary evaluation of a psychological skills training course. *Ann R Coll Surg Engl* 2009;**91**:321–5.

137. Scott DJ, Rege RV, Bergen PC, Guo WA, Laycock R, Tesfay ST, *et al.* Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. *J Laparoendosc Adv Surg Techniques* 2000;**10**:183–90.

138. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: a quantitative review. *J Occupat Organizat Psychol* 1994;**67**:189–205.

139. Royal College of General Practitioners. Brief guide to workplace based asessment in the nMRCGP. London: RCGP; 2008. Available from: www.rcgp-curriculum.org.uk/nmrcgp/wpba.aspx (accessed 10 July 2009).

140. Campbell LM, Howie JGR, Murray TS. Use of videotaped consultations in summative assessment of trainees in general practice. *Br J Gen Pract* 1995;**45**:137–41.

141. Ram P, Grol R, Rethans JJ, Schouten B, van der Vleuten C, Kester A. Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. *Med Educ* 1999;**33**:447–54.

142. Grantcharov TP, Schulze S, Kristiansen VB. The impact of objective assessment and feedback on improvement of laparoscopic performance in the operating room. *Surg Endosc* 2007;**21**:2240–3.

Appendix 1

OSATS forms for caesarean section

Objective structured assessment of technical skill

CAESAREAN SECTION

Trainee Name:		Assessor Name:		Date:
Level of training: Grade/Year		Post:		

Clinical details of complexity/ difficulty of case	

Item under observation	**Done independently**	**Needs help**
Appropriate skin incision e.g. length, position		
Safe entry of peritoneal cavity		
Careful management of bladder		
Appropriate uterine incision e.g. length, position		
Safe and systematic delivery of baby		
Appropriate delivery of placenta		
Check uterine cavity e.g. intact, empty, configuration		
Safe securing of uterine angles		
Check for ovarian pathology		
Appropriate closure of rectus sheath		
Attention to haemostasis		
Neatness of skin closure		
Comments		

Examples of minimum levels of complexity for each stage of training:

ST1	First or second Caesarean section with longitudinal lie
Core Training	Twins/ transverse lie Preterm greater than 28 weeks
CCT	Preterm less than 28 weeks/ Grade 4 Placenta praevia Fibroids in lower uterine segment

Signed.....

Objective structured assessment of technical skill

GENERIC TECHNICAL SKILLS ASSESSMENT

Assessor, please ring the candidate's performance for each of the following factors:

Respect for tissue	Frequently used unnecessary force on tissue or caused damage by inappropriate use of instruments.	Careful handling of tissue but occasionally causes inadvertent damage	Consistently handled tissues appropriately with minimal damage.
Time, motion and flow of operation and forward planning	Many unnecessary moves. Frequently stopped operating or needed to discuss next move.	Makes reasonable progress but some unnecessary moves Sound knowledge of operation but slightly disjointed at times	Economy of movement and maximum efficiency. Obviously planned course of operation with effortless flow from one move to the next.
Knowledge and handling of instruments	Lack of knowledge of instruments.	Competent use of instruments but occasionally awkward or tentative	Obvious familiarity with instruments.
Suturing & knotting skills	Placed sutures inaccurately or tied knots insecurely, and lacked attention to safety.	Knotting and suturing usually reliable but sometimes awkward	Consistently placed sutures accurately with appropriate and secure knots, and with proper attention to safety.
Technical use of assistants Relations with patient and the surgical team	Consistently placed assistants poorly or failed to use assistants. Communicated poorly or frequently showed lack of awareness of the needs of the patient and/or the professional Team	Appropriate use of assistant most of the time Reasonable communication and awareness of the needs of the patient and/or of the professional team	Strategically used assistants to the best advantage at all times. Consistently communicated and acted with awareness of the needs of the patient and/or of the professional team
Insight/Attitude	Poor understanding of areas of weakness	Some understanding of areas of weakness	Fully understands areas of weakness
Documentation of Procedures	Limited documentation Poorly written	Adequate documentation, but with some omissions, or areas that need elaborating	Comprehensive legible documentation, indicating findings, procedure and postoperative management

Based on the checklist and the Generic Technical Skills Assessment, Dr ……..has achieved/failed* to achieve the OSAT competency

Needs further help with:	Competent to perform the entire procedure without the need for supervision
* * *	
Date	Date
Signed	Signed

* Delete where applicable, and date and sign the relevant box

RCOG Resources for Education

Appendix 2

PBA form for caesarean section

Obstetrics & Gynaecology PBA: Caesarean Section

Trainee:	Assessor:	Date:
StR Year :	Start time:	End time:

Elective / Emergency? (circle one)

Operation more difficult than usual? Yes / No If yes, state reason (e.g. high BMI)

Levels of complexity for each stage of training *(circle one)*

Basic Training First or second caesarean with longitudinal lie

Intermediate Training Twins or transverse lie, preterm >28 weeks

Advanced Training Preterm <28 weeks/major placenta praevia/fibroids in lower segment

The Trainee should explain what he/she intends to do throughout the procedure
The Assessor should provide verbal prompts, if required, and intervene if patient safety is at risk.

Rating : N = Not observed or not appropriate D = Development required
S = Satisfactory standard for CCT (no prompting or intervention required)

	Competencies and Definitions	Rating N/D/S	Comments
I.	**Consent**		
C1	Demonstrates sound knowledge of indications and contraindications including alternatives to surgery		
C2	Demonstrates awareness of sequelae of operative or non operative management		
C3	Demonstrates sound knowledge of complications of surgery		
C4	Explains the perioperative process to the patient and/or relatives or carers and checks understanding		
C5	Explains likely outcome and time to recovery and checks understanding		
II.	**Pre operative planning**		
PL1	Demonstrates recognition of anatomical and pathological abnormalities (and relevant co-morbidities) and selects appropriate operative strategies/techniques to deal with these e.g. nutritional status		
PL2	Demonstrates ability to make reasoned choice of appropriate equipment, materials or devices (if any) taking into account appropriate investigations e.g. ultrasound, MRI		
PL3	Checks materials, equipment and device requirements with theatre staff and blood bank if major haemorrhage is anticipated		
PL4	Ensures the operation site is marked where applicable		Not applicable
PL5	Checks patient records, personally reviews investigations		
III.	**Pre operative preparation**		
PR1	Checks in theatre that consent has been obtained		
PR2	Gives effective briefing to theatre team, including paediatrician		
PR3	Ensures proper and safe positioning of the patient on the operating table		
PR4	Demonstrates careful skin preparation		
PR5	Demonstrates careful draping of the patient's operative field		
PR6	Ensures general equipment and materials are deployed safely (e.g. urinary catheter, diathermy)		
PR7	Ensures appropriate drugs administered (e.g. prophylactic antibiotics, oxytocin)		
PR8	Arranges for and deploys specialist supporting equipment (e.g. cell salvager) effectively		

	Competencies and Definitions	Rating N/D/S	Comments
IV.	**Exposure and closure**		
E1	Demonstrates knowledge of optimum skin incision/access		
E2	Achieves an adequate exposure through purposeful dissection in correct tissue planes and identifies all structures correctly		
E3	Completes a sound wound repair (rectus sheath)		
E4	Protects the wound with dressings and inserts drains safely where appropriate		
V.	**Intra operative Technique: global (G) and task-specific (T) items**		
IT1(G)	Follows an agreed, logical sequence or protocol for the procedure		
IT2(G)	Consistently handles tissue well with minimal damage		
IT3(G)	Controls bleeding promptly by an appropriate method		
IT4(G)	Demonstrates a sound technique of knots and sutures/staples		
IT5(G)	Uses instruments appropriately and safely		
IT6(G)	Proceeds at appropriate pace with economy of movement		
IT7(G)	Anticipates and responds appropriately to variation e.g. anatomy		
IT8(G)	Deals calmly and effectively with unexpected events/complications if they occur		
IT9(G)	Uses assistant(s) to the best advantage at all times		
IT10(G)	Communicates clearly and consistently with the scrub team		
IT11(G)	Communicates clearly and consistently with the anaesthetist		
IT12 (T)	Safely enters peritoneal cavity		
IT13 (T)	Carefully mobilises bladder		
IT14 (T)	Performs appropriate uterine incision		
IT15 (T)	Safely and systematically delivers baby		
IT16 (T)	Completes delivery of placenta and membranes safely		
IT17(T)	Checks uterine cavity (empty, configuration)		
IT18(T)	Identifies and manages uterine atony appropriately		
IT19(T)	Safely secures both uterine angles		
IT20(T)	Identifies and safely secures any uterine extensions		
IT21(T)	Checks for pelvic pathology (uterus, tubes, ovaries)		
IT22(T)	Completes swabbing out of paracolic gutters		
IT23(T)	Ensures contracted uterus and complete haemostasis before closing abdomen		
IT24(T)	Performs vaginal toilet and checks vaginal loss		
IT25(T)	Checks urine colour at end of procedure		
IT26(T)	Closes the skin with attention to cosmesis and healing (e.g. excision of old scar)		
VI.	**Post operative management**		
PM1	Ensures the patient is transferred safely from the operating table to bed		
PM2	Constructs a clear operation note		
PM3	Records clear and appropriate post operative instructions		
PM4	Deals with specimens. Labels and orientates specimens appropriately		Not applicable

Global summary

Level at which completed elements of the PBA were performed on this occasion		Tick as appropriate
Level 0	Insufficient evidence observed to support a summary judgement	
Level 1	Unable to perform the procedure, or part observed, under supervision	
Level 2	Able to perform the procedure, or part observed, under supervision	
Level 3	Able to perform the procedure with minimum supervision (needed occasional help)	
Level 4	Competent to perform the procedure unsupervised (could deal with any complications that arose)	

Comments by Assessor (including strengths and areas for development):

Comments by Trainee:

Trainee Signature:	**Assessor Signature:**

Appendix 3

NOTSS form, rating scale and descriptors

Hospital ..... **Trainer name** ..... **Date** .....

Trainee name ..... **Operation** .....

Category	Category rating*	Element	Element rating*	Feedback on performance and debriefing notes
Situation Awareness		Gathering information		
		Understanding information		
		Projecting and anticipating future state		
Decision Making		Considering options		
		Selecting and communicating option		
		Implementing and reviewing decisions		
Communication and Teamwork		Exchanging information		
		Establishing a shared understanding		
		Co-ordinating team activities		
Leadership		Setting and maintaining standards		
		Supporting others		
		Coping with pressure		

* **1** Poor; **2** Marginal; **3** Acceptable; **4** Good; **N/A** Not Applicable

1 Poor	Performance endangered or potentially endangered patient safety, serious remediation is required
2 Marginal	Performance indicated cause for concern, considerable improvement is needed
3 Acceptable	Performance was of a satisfactory standard but could be improved
4 Good	Performance was of a consistently high standard, enhancing patient safety; it could be used as a positive example for others
N/A	Not Applicable

The NOTSS rating scale

The scale below is used to rate non-technical skills based on observed behaviour. The same scale is used to rate category and element-level skills. If a skill is not required or not relevant in the particular case being observed then 'N/A' should be used. If a skill should be displayed but is lacking, then '1 – poor' should be used.

NOTSS System Rating Options

Rating Label	Description
4 – Good	Performance was of a consistently high standard, enhancing patient safety; it could be used as a positive example for others
3 – Acceptable	Performance was of a satisfactory standard but could be improved
2 – Marginal	Performance indicated cause for concern, considerable improvement is needed
1 – Poor	Performance endangered or potentially endangered patient safety, serious remediation is required
N/A – Not Applicable	Skill was not required or relevant in this case

Not all skill elements may be required or desirable in any given clinical encounter.

You should expect to see behaviours in order to provide ratings 2 (marginal), 3 (acceptable), or 4 (good). You should expect to see poor behaviours or the absence of required behaviours to rate 1 (poor). Rating N/A means that you did not see behaviours to rate because they were not required or not relevant for the clinical encounter being rated.

Situation Awareness: Developing and maintaining a dynamic awareness of the situation in theatre based on assembling data from the environment (patient, team, time, displays, equipment); understanding what they mean, and thinking ahead about what may happen next.

Gathering information – Seeking information in the operating theatre from the operative findings, theatre environment, equipment, and people.

Good behaviours:
- Carries out pre-operative checks of patient notes, including investigations and consent
- Ensures that all relevant investigations (e.g. imaging) have been reviewed and are available
- Liaises with anaesthetist regarding anaesthetic plan for patient
- Optimises operating conditions before starting e.g. moves table, lights, AV equipment
- Identifies anatomy/ pathology clearly
- Monitors ongoing blood loss
- Asks anaesthetist for update

Poor behaviours:
- Arrives in theatre late or has to be repeatedly called
- Does not ask for results until the last minute or not at all
- Does not consider the views of operating room staff
- Fails to listen to anaesthetist
- Fails to review information collected by team
- Asks for information to be read from patient notes during procedure because has not been read before operation started

Understanding information – Updating one's mental picture by interpreting the information gathered, and comparing it with existing knowledge to identify the match or mismatch between the situation and the expected state.

Good behaviours:
- Acts according to information gathered from previous investigation and operative findings
- Looks at CT scan and points out relevant area
- Reflects and discusses significance of information

Poor behaviours:
- Overlooks or ignores important results
- Misses clear sign (e.g. on CT scan)
- Asks questions which demonstrate lack of understanding
- Discards results that don't 'fit the picture'

Projecting and anticipating future state – Predicting what may happen in the near future as a result of possible actions, interventions or non-intervention.

Good behaviours:
- Plans operating list taking into account potential delays due to surgical or anaesthetic challenges
- Verbalises what equipment may be required later in operation
- Shows evidence of having a contingency plan ('plan B') (e.g. by asking scrub nurse for potentially required equipment to be available in theatre)
- Cites contemporary literature on anticipated clinical event

Poor behaviours:
- Overconfident manoeuvres with no regard for what may go wrong
- Does not discuss potential problems
- Gets into predictable blood loss, then tells anaesthetist
- Waits for a predicted problem to arise before responding
- Operates beyond level of experience

NON-TECHNICAL SKILLS FOR SURGEONS

Decision Making: Skills for diagnosing the situation and reaching a judgement in order to choose an appropriate course of action.

Considering options – Generating alternative possibilities or courses of action to solve a problem. Assessing the hazards and weighing up the threats and benefits of potential options.

Good behaviours:
- Recognises and articulates problems
- Initiates balanced discussion of options, pros and cons with relevant team members
- Asks for opinion of other colleagues
- Discusses published guidelines

Poor behaviours:
- No discussion of options
- Does not solicit views of other team members
- Ignores published guidelines

Selecting and communicating option – Choosing a solution to a problem and letting all relevant personnel know the chosen option.

Good behaviours:
- Reaches a decision and clearly communicates it
- Makes provision for and communicates 'plan B'
- Explains why contingency plan has been adopted

Poor behaviours:
- Fails to inform team of surgical plan
- Is aggressive/ unresponsive if plan questioned
- Shuts down discussion on other treatment options
- Only does what she/he thinks is best or abandons operation
- Selects inappropriate manoeuvre that leads to complication

Implementing and reviewing decisions – Undertaking the chosen course of action and continually reviewing its suitability in light of changes in the patient's condition. Showing flexibility and changing plans if required to cope with changing circumstances to ensure that goals are met.

Good behaviours:
- Implements decision
- Updates team on progress
- Reconsiders plan in light of changes in patient condition or when problem occurs
- Realises 'plan A' is not working and changes to 'plan B'
- Calls for assistance if required

Poor behaviours:
- Fails to implement decisions
- Makes same error repeatedly
- Does not review the impact of actions
- Continues with 'plan A' in face of predictably poor outcome or when there is evidence of a better alternative
- Becomes hasty or rushed due to perceived time constraints

Communication and Teamwork: Skills for working in a team context to ensure that the team has an acceptable shared picture of the situation and can complete tasks effectively.

Exchanging information – Giving and receiving knowledge and information in a timely manner to aid establishment of a shared understanding among team members.

Good behaviours:
- Talks about the progress of the operation
- Listens to concerns of team members
- Communicates that operation is not going to plan

Poor behaviours:
- Fails to communicate concerns with others
- Attempts to resolve problems alone
- Does not listen to team members
- Needs help from assistant but does not make it clear what assistant is expected to do

Establishing a shared understanding – Ensuring that the team not only has necessary and relevant information to carry out the operation, but that they understand it and that an acceptable shared 'big picture' of the case is held by team members.

Good behaviours:
- Provides briefing and clarifies objectives and goals before commencing operation
- Ensures team understand the operative plan before starting
- Encourages input from all members of the team
- Ensures relevant members of team are comfortable with decisions
- Checks that assistant knows what they are expected to do
- Debriefs relevant team members after operation, discussing what went well and problems that occurred

Poor behaviours:
- Does not articulate operative plan to team
- Does not make time for collective discussion and review of progress
- Fails to discuss the case beforehand with unfamiliar team members
- Makes no attempt to discuss problems and successes at end of operation
- Fails to keep anaesthetist informed about procedure (e.g. to expect bleeding)
- Appears uncomfortable discussing the operative plan if challenged

Co-ordinating team activities – Working together with other team members to carry out cognitive and physical activities in a simultaneous, collaborative manner.

Good behaviours:
- Checks that other team members are ready to start operation
- Stops operating when asked to by anaesthetist or scrub nurse
- Ensures that team works efficiently by organising activities in a timely manner

Poor behaviours:
- Does not ask anaesthetist if it is OK to start operation
- Proceeds with operation without ensuring that equipment is ready

NON-TECHNICAL SKILLS FOR SURGEONS

Leadership: Leading the team and providing direction, demonstrating high standards of clinical practice and care, and being considerate about the needs of individual team members.

Setting and maintaining standards – Supporting safety and quality by adhering to acceptable principles of surgery, following codes of good clinical practice, and following theatre protocols.

Good behaviours:
- Introduces self to new or unfamiliar members of theatre team
- Clearly follows theatre protocol
- Requires all team members to observe standards (e.g. sterile field)

Poor behaviours:
- Fails to observe standards (e.g. continues even though equipment may be contaminated or inadequate)
- Breaks theatre protocol
- Shows disrespect to the patient

Supporting others – Providing cognitive and emotional help to team members. Judging different team members' abilities and tailoring one's style of leadership accordingly.

Good behaviours:
- Modifies behaviour according to trainee needs
- Provides constructive criticism to team members
- Ensures delegation of tasks is appropriate
- Establishes rapport with team members
- Gives credit for tasks performed well

Poor behaviours:
- Does not provide recognition for tasks performed well
- Fails to recognise needs of others
- Engages in 'tunnel vision' approach to technical aspects of operation
- Shows hostility to other team members (e.g. makes sarcastic comments to nurses)

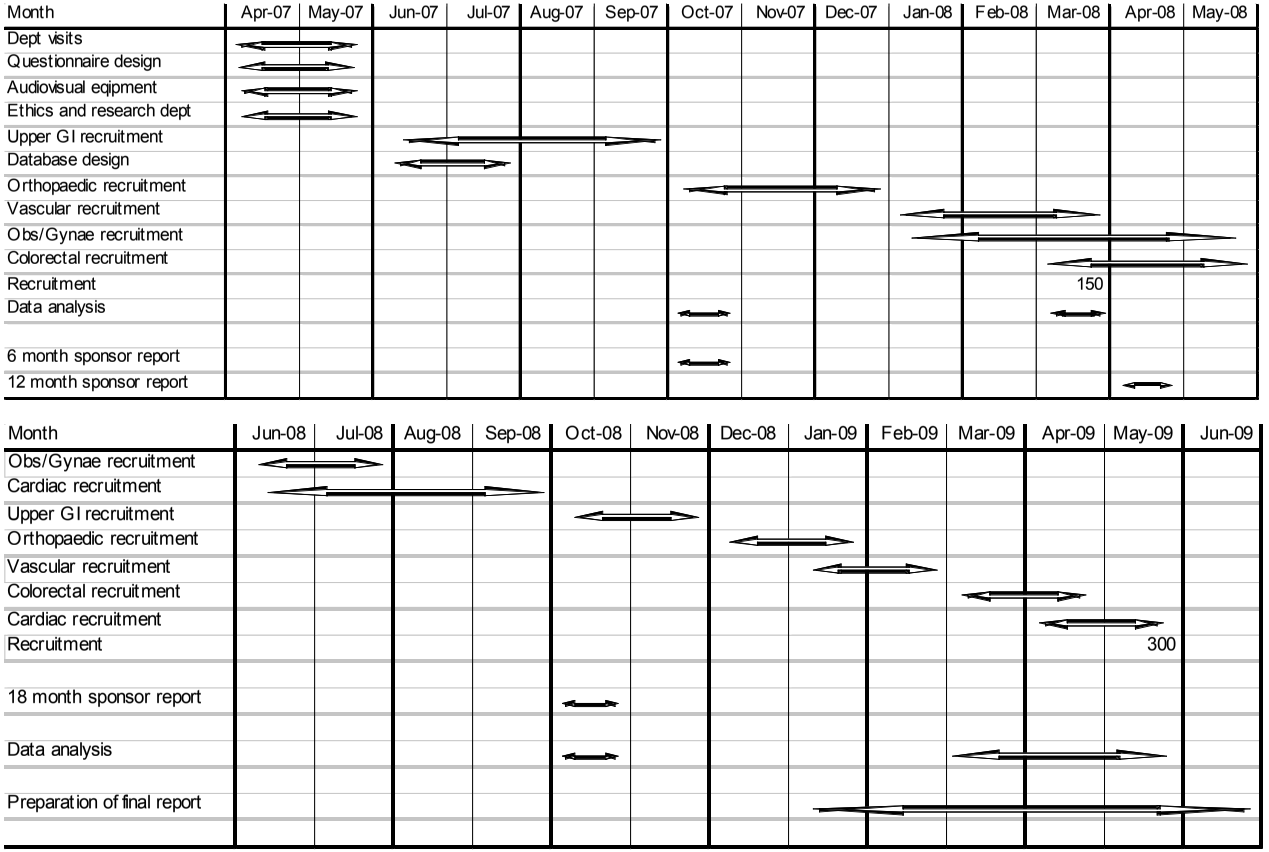
Coping with pressure – Retaining a calm demeanour when under pressure and emphasising to the team that one is under control of a high-pressure situation. Adopting a suitably forceful manner if appropriate without undermining the role of other team members.

Good behaviours:
- Remains calm under pressure
- Emphasises urgency of situation (i.e. by occasionally raising voice)
- Takes responsibility for the patient in emergency/ crisis situation
- Makes appropriate decision under pressure
- Delegates tasks in order to achieve goals
- Continues to lead team through emergency

Poor behaviours:
- Suppresses concern over clinical problem
- 'Freezes' and displays inability to make decisions under pressure
- Fails to pass leadership of case when technical challenge requires full attention
- Blames everyone else for errors and does not take personal responsibility
- Loses temper

Appendix 4

Gantt chart of study progress


Month
Apr-07
May-07
Jun-07
Jul-07
Aug-07
Sep-07
Oct-07
Nov-07
Dec-07
Jan-08
Feb-08
Mar-08
Apr-08
May-08
Dept visits
Questionnaire design
Audiovisual eqipment
Ethics and research dept
Upper GI recruitment
Database design
Orthopaedic recruitment
Vascular recruitment
Obs/Gynae recruitment
Colorectal recruitment
Recruitment
150
Data analysis
6 month sponsor report
12 month sponsor report
Month
Jun-08
Jul-08
Aug-08
Sep-08
Oct-08
Nov-08
Dec-08
Jan-09
Feb-09
Mar-09
Apr-09
May-09
Jun-09
Obs/Gynae recruitment
Cardiac recruitment
Upper GI recruitment
Orthopaedic recruitment
Vascular recruitment
Colorectal recruitment
Cardiac recruitment
Recruitment
300
18 month sponsor report
Data analysis
Preparation of final report

Appendix 5

Trainee and assessor questionnaires

A. PBA questionnaire for clinical supervisors

Thank you for using PBAs in your surgical sessions as part of the Surgical Skills Study. Please can you now complete this short survey to give us feedback on your experience of using this assessment form.

1. How many times had you assessed trainees with PBAs prior to this study? (please circle)

Never	1–5 times	6–15 times	> 15 times

2. How often did you give feedback from PBAs at the time? (please circle)

Not applicable	Never	Sometimes	Always

3. Please indicate the extent to which you disagree or agree with the following statements about PBAs: (please tick appropriate box)

		1	2	3	4	5
		Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
a.	PBAs are a useful tool for providing debriefing (feedback) after an operation					
b.	PBAs are a valuable tool in formative assessment ie. to show progress in training					
c.	PBAs are a valuable tool in summative assessment ie. to show a level of competency has been achieved					
d.	PBAs are a useful tool to support reflective practice or to provide insight					

4. The amount of training I have undertaken to use PBAs is (please circle):

Too little	Just right	Too much

5. What training have you undertaken? (please circle more than one as appropriate)

ISCP web guidance	Training workshop	Written information from research team	Face-to-face discussion with reseach team	None

6. Where do you normally conduct PBA debrief sessions? (please circle more than one as appropriate)

Operating theatre	Coffee room	Office	Other

If other, please specify location: (please enter text)

---

7. How useful do you think your trainees found the feedback given by you in these sessions? (please circle a number on the scale)

0 1 2 3 4 5 6 7 8 9 10
|---|---|---|---|---|---|---|---|---|---|
Not at all useful Very useful

Please give details if you wish: (please enter text)

---

8. To what extent do you think PBAs enhanced your ability to assess your trainees? (please circle)

0 1 2 3 4 5 6 7 8 9 10
|---|---|---|---|---|---|---|---|---|---|
Not at all Very much

Please give details if you wish (please add text)

---

9. Do you think trainees performed differently as a result of being assessed? (please circle):

Yes	No

Please give details and/or examples (please add text)

---

10. How comfortable do you feel about formally recording on the PBA that you have concerns about a trainee's surgical skills?

0 1 2 3 4 5 6 7 8 9 10
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
Very uncomfortable Very comfortable

Please give details if you wish: (please add text)

---

11. How important do you think PBAs are in surgical education? (please circle a number on the scale)

0 1 2 3 4 5 6 7 8 9 10
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
Irrelevant Very important

Please give details if you wish: (please add text)

---

12. Have you used PBAs outside of the study since the study began? (please circle)

Yes	No

13. If yes, in what context have you used PBAs outside the study? (please enter text)

-

-

14. If no to Q.12, what was the reason for not using PBAs? (Please circle more than one as appropriate)

No trainee to assess	No index procedures performed	Lack of time	Other

If other, please give reason: (please enter text)

---

15. PBAs are now part of the new surgical curriculum, but if given the choice, how likely would you be to use PBAs in the future? (please circle a number on the scale)

0 1 2 3 4 5 6 7 8 9 10
|---|---|---|---|---|---|---|---|---|---|
Very unlikely Very likely

16. Please make any further comments regarding PBAs or this study: (please enter text)

B. PBA questionnaire for surgical trainees

Thank you for using PBAs in your surgical sessions as part of the Surgical Skills Study. Please can you now complete this short survey to give us feedback on your experience of using this assessment tool.

2. How many times had you been assessed with PBAs prior to this study? (please circle)

Never	1–5 times	6–15 times	> 15 times

2. How often did you receive feedback from PBAs at the time? (please circle)

Not applicable	Never	Sometimes	Always

4. Please indicate the extent to which you disagree or agree with the following statements about PBAs: (please tick appropriate box) :

		1	2	3	4	5
		Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
a.	PBAs are a useful tool for providing debriefing after an operation					
b.	PBAs are a valuable tool in formative assessment ie. to show progress in training					
c.	PBAs are a valuable tool in summative assessment ie. to show a level of competency has been achieved					
d.	PBAs are a useful tool to support reflective practice or to provide insight					

4. The amount of training I have undertaken to use PBAs is (please circle):

Too little	Just right	Too much

5. What training have you undertaken? (please circle more than one as appropriate)

ISCP web guidance	Training workshop	Written information from research team	Face-to-face discussion with research team	None

6. Where are your PBA debrief sessions normally conducted? (please circle more than one as appropriate)

Operating theatre	Coffee room	Office	Other

If other, please specify location: (please enter text)

---

7. How useful have you found the feedback given by your supervising consultant in these sessions? (please circle a number on the scale)

0 1 2 3 4 5 6 7 8 9 10
|---|---|---|---|---|---|---|---|---|---|
Not at all useful Very useful

Please give details if you wish: (please enter text)

---

8. To what extent do you think PBAs enhanced your trainer's ability to assess you? (please circle)

0 1 2 3 4 5 6 7 8 9 10
|---|---|---|---|---|---|---|---|---|---|
Not at all Very much

Please give details if you wish: (please add text)

---

9. Do you think you performed differently as a result of being assessed? (please circle):

Yes	No

Please give details and/or examples (please add text)

---

10. How comfortable do you feel about a consultant formally recording on the PBA that they have concerns about a trainee's surgical skills?

0 1 2 3 4 5 6 7 8 9 10

Very uncomfortable Very comfortable

Please give details if you wish: (please add text)

---

11. How important do you think PBAs are in surgical education? (please circle a number on the scale)

0 1 2 3 4 5 6 7 8 9 10

Irrelevant Very important

Please give details if you wish: (please add text)

---

12. Have you used PBAs outside of the study since the study began? (please circle)

Yes	No

13. If yes, in what context have you used PBAs outside the study? (please enter text)

-

-

14. If no to Q.12, what was the reason for not using PBAs? (Please circle more than one as appropriate)

No index procedures performed	Consultant unwilling	Lack of time	Other

If other, please give reason: (please enter text)

---

15. PBAs are now part of the new surgical curriculum, but if given the choice, how likely would you be to use PBAs in the future? (please circle a number on the scale)

0 1 2 3 4 5 6 7 8 9 10
|---|---|---|---|---|---|---|---|---|---|
Very unlikely Very likely

16. Please make any further comments regarding PBAs or this study: (please enter text)

C. OSATS questionnaire for clinical supervisors

Thank you for using OSATS and PBA (procedure-based assessment) forms in your surgical sessions as part of the Surgical Skills Study. Please could you complete this short survey to give us feedback on your experience of using these assessment forms.

PART 1: OSATS Assessments

1. How long have you been using OSATS forms to assess trainees surgical skills?

Not before this study	Since 01/08/07 (launch of new curriculum)	1–2 years	2 years
☐	☐	☐	☐

2. When using OSATS forms *outside of this study*, how frequently have you given trainees feedback on their surgical skills?

Not applicable	Never	Sometimes	Mostly	Always
☐	☐	☐	☐	☐

3. Please indicate the extent to which you disagree or agree with the following statements about OSATS: (please select appropriate box)

		1	2	3	4	5
		Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
a.	OSATS are a useful tool for providing feedback after an operation/procedure	☐	☐	☐	☐	☐
b.	OSATS are a valuable formative assessment method, i.e. an aid to learning	☐	☐	☐	☐	☐
c.	OSATS are a valuable summative assessment method, i.e. to show a satisfactory level of competency has been acheived	☐	☐	☐	☐	☐
d.	OSATS are a useful tool to support reflective practice or to provide insight	☐	☐	☐	☐	☐

4. I have had sufficient training in using the OSATS tool (please select appropriate box to indicate the extent of your agreement or disagreement with this statement)

☐----------☐----------☐----------☐----------☐

Strongly disagree | Disagree | Neither agree or disagree | Agree | Strongly agree

Give details if you wish (enter text):

5. What training have you undertaken? (please select more than one as appropriate)

None	RCOG web guidance	Training workshop	Written information from research team	Face-to-face discussion with research team	Other
☐	☐	☐	☐	☐	☐

If you selected 'other' box, please give details:

6. Using OSATS adds too much time to my operating list (please select appropriate box to indicate the extent of your agreement or disagreement with this statement)

☐----------☐----------☐----------☐----------☐

Strongly disagree | Disagree | Neither agree or disagree | Agree | Strongly agree

Give details if you wish:

7. Do you think OSATS have helped you to assess the surgical skills of trainees? i.e. helped you identify trainees' surgical strengths and areas for development (please select a number on the scale)

1 2 3 4 5 6 7 8 9 10

☐--☐--☐--☐--☐--☐--☐--☐--☐--☐

Not at all | Moderately | Very much

Give details if you wish:

8. What is your level of *overall satisfaction* with the OSATS tool for assessing trainees surgical skill and providing feedback?

1 2 3 4 5 6 7 8 9 10

☐--☐--☐--☐--☐--☐--☐--☐--☐--☐

No satisfaction | Moderately satisfied | Highly satisfied

Please give details if you wish:

9. OSATS are now part of the new specialty curriculum, but if given the choice, would you continue to use the OSATS tool to assess trainees' surgical skills?

Never	Sometimes	Always	Not sure
☐	☐	☐	☐

Please give details if you wish:

10. Have you any suggestions to improve the current OSATS tool?

Yes	No
☐	☐

If yes, please give details:

11. Do you think there have been any differences with the OSATS you completed within this study, compared to the training OSATS assessments you usually complete?

Yes	No
☐	☐

If yes, please give details:

PART 2: PBA Assessments

1. Please indicate the extent to which you disagree or agree with the following statements about PBAs: (please select appropriate box)

		1	2	3	4	5
		Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
a.	PBAs are a useful tool for providing feedback after an operation/procedure	☐	☐	☐	☐	☐
b.	PBAs are a valuable formative assessment method, i.e. an aid to learning	☐	☐	☐	☐	☐
c.	PBAs are a valuable summative assessment method, i.e. to show a satisfactory level of competency has been achieved	☐	☐	☐	☐	☐
d.	PBAs are a useful tool to support reflective practice or to provide insight	☐	☐	☐	☐	☐

2. I have had sufficient training in using the PBA tool (please select appropriate box)

☐----------☐----------☐----------☐----------☐

Strongly disagree Disagree Neither agree or disagree Agree Strongly agree

Give details if you wish (enter text):

3. What training have you undertaken? (please select more than one as appropriate)

None	ISCP web guidance	Training workshop	Written information from research team	Face-to-face discussion with research team	Other
☐	☐	☐	☐	☐	☐

If you selected 'other' box, please give details:

4. Using PBAs added too much time to my list (please select appropriate box to indicate the extent of your agreement or disagreement with this statement)

☐----------☐----------☐----------☐----------☐

Strongly disagree Disagree Neither agree or disagree Agree Strongly agree

Give details if you wish:

5. Do you think PBAs have helped you to assess the surgical skills of trainees? (please select a number on the scale)

1 2 3 4 5 6 7 8 9 10

☐--------☐--------☐--------☐--------☐--------☐--------☐--------☐--------☐----☐

Not at all Moderately Very much

Give details if you wish:

6. What is your level of *overall satisfaction* with the PBA tool for assessing trainees' surgical skills and providing feedback?

1 2 3 4 5 6 7 8 9 10

☐--------☐--------☐--------☐--------☐-------☐--------☐--------☐-------☐-------☐

No satisfaction Moderately satisfied Highly satisfied

Please give details if you wish:

7 (a). Comparing your experience of using the OSATS and PBA forms to assess trainees' surgical skills and provide feedback, which tool do you prefer?

OSATS	PBA	No preference
☐	☐	☐

(b) If you do have a preference for one of the tools, please explain why:

8. Have you any suggestions to improve the PBA tool?

Yes	No
☐	☐

If yes, please give details:

PART 3: General questions

1. Where do you usually give trainees feedback on their OSATS/PBA assessment?
(Please select all those appropriate)

Operating theatre	Coffee room	Office	Other
☐	☐	☐	☐

If other, please specify location (please enter text):

2. I usually give OSATS/PBA feedback in a suitable (i.e. confidential) place.
(Please select appropriate box to indicate the extent of your agreement or disagreement with this statement)

☐----------☐----------☐----------☐----------☐

Strongly disagree Disagree Neither agree or disagree Agree Strongly agree

Please give details if you wish:

3. How helpful do you think trainees find the feedback you give in these sessions?
(please select a number on the scale)

1 2 3 4 5 6 7 8 9 10

☐--------☐--------☐--------☐--------☐--------☐--------☐--------☐--------☐-----☐

No help Some help Very helpful

Please give details if you wish:

4. How comfortable do you feel about formally recording on the OSATS/PBA form that you have concerns about a trainee's surgical competence?

1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10 ☐

Uncomfortable Neutral Comfortable

Please give details if you wish:

5. Has the completion of an OSATS/PBA form resulted in you raising concerns about a trainee's surgical competence?

Yes ☐	No ☐

If yes, please provide general comments only to protect trainees' confidentiality:

6. Do you have any suggestions for the development of a new assessment tool, or for alternative methods of assessing the surgical competence of trainees?

Yes ☐	No ☐

If yes, please give details:

7. Have you any further comments or concerns regarding OSATS/PBA assessments or this research study?

Yes ☐	No ☐

If yes, please give details:

D. OSATS questionnaire for trainees

Thank you for using OSATS in your surgical sessions. Please can you now complete this short survey to give us feedback on OSATS.

1. How long have you been using OSATS forms in the assessment of your surgical skills? (please select appropriate box)

Not before this study ☐	Since 01/08/07 (launch of new curriculum) ☐	1–2 years ☐	> 2 years ☐

2. How many times had you been assessed with OSATS *prior to this study*?

Never ☐	1–5 times ☐	6–15 times ☐	>15 times ☐

3. How often did you receive feedback from OSATS (*outside this study*) at the time?

Not applicable ☐	Never ☐	Sometimes ☐	Mostly ☐	Always ☐

4. I have had sufficient training in using the OSATS tool (please select appropriate box)

☐----------☐----------☐----------☐----------☐

Strongly disagree Disagree Neither agree or disagree Agree Strongly agree

Please give details if you wish (enter text):

5. What training have you undertaken? (please select more than one as appropriate)

None ☐	RCOG web guidance ☐	Training workshop ☐	Written information ☐	Face-to-face discussion ☐	Other ☐

If you selected 'other' box, please give details:

6. My trainers appear to have had sufficient training in using the OSATS tool:

☐----------☐----------☐----------☐----------☐

Strongly disagree Disagree Neither agree or disagree Agree Strongly agree

Please give details if you wish:

7. Please indicate the extent to which you disagree or agree with the following statements about OSATS: (please tick appropriate box)

		1	2	3	4	5
		Strongly disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
a.	OSATS are a useful tool for providing debriefing (feedback) after an operation	☐	☐	☐	☐	☐
b.	OSATS are a valuable tool in formative assessment, i.e. to help learning and show progress in training	☐	☐	☐	☐	☐
c.	OSATS are a valuable tool in summative assessment, i.e. to show a level of competency has been achieved	☐	☐	☐	☐	☐
d.	OSATS are a useful tool to support reflective practice or to provide insight	☐	☐	☐	☐	☐

8. Where do you usually receive your OSATS feedback? (Please select all those appropriate)

Operating theatre	Coffee room	Office	Other
☐	☐	☐	☐

If other, please specify location (please enter text):

9. I usually receive OSATS feedback in a suitable (i.e. confidential) place.

☐----------☐----------☐----------☐----------☐

Strongly disagree Disagree Neither agree or disagree Agree Strongly agree

Please give details if you wish:

10. How helpful have you found the feedback given by your supervising consultant in these sessions? (please select a number on the scale)

1 2 3 4 5 6 7 8 9 10

☐--------☐--------☐--------☐--------☐--------☐--------☐--------☐--------☐-----☐

No help Some help Very helpful

Please give details if you wish:

11. The helpfulness of feedback I have received following OSATS assessment depends on the trainer who gives it.

☐----------☐----------☐----------☐----------☐

Strongly disagree | Disagree | Neither agree or disagree | Agree | Strongly agree

Please give details if you wish:

12. To what extent do you think OSATS enhanced your trainer's ability to assess your surgical skill?

1 2 3 4 5 6 7 8 9 10

☐--------☐--------☐--------☐--------☐--------☐--------☐--------☐--------☐-----☐

Not at all — Moderately — Very much

Please give details if you wish:

13. The scores in past OSATS assessments have been an accurate representation of my surgical/technical skills

☐----------☐----------☐----------☐----------☐

Strongly disagree | Disagree | Neither agree or disagree | Agree | Strongly agree

Please give details if you wish:

14. How comfortable do you feel about a consultant formally recording on the OSATS form that they have concerns about a trainee's surgical skills?

1 2 3 4 5 6 7 8 9 10

☐--------☐--------☐--------☐--------☐--------☐--------☐--------☐--------☐----☐

Uncomfortable — Neutral — Comfortable

Please give details if you wish:

15. How important do you think OSATS are in surgical education?

1 2 3 4 5 6 7 8 9 10

☐--------☐--------☐--------☐--------☐--------☐--------☐--------☐--------☐-----☐

Not at all — Moderately — Very much

Please give details if you wish:

16. Do you think OSATS have improved your surgical training? i.e. helped you build on strengths and improve areas for development that were identified during feedback

1 2 3 4 5 6 7 8 9 10

Not at all Moderately Very much

Please give details if you wish:

17. OSATS are part of the new specialty curriculum, but, if given the choice, would you continue to use OSATS as an assessment of your surgical skills?

Never	Sometimes	Always	Not sure
☐	☐	☐	☐

Please give details if you wish:

18. What is your level of *overall satisfaction* with the OSATS tool for assessment and feedback on your surgical skills?

1 2 3 4 5 6 7 8 9 10

No satisfaction Moderately satisfied Highly satisfied

Please give details if you wish:

19. Comparing your experience of assessment and feedback using the OSATS and PBA, which tool do you prefer?

OSATS	PBA	No preference
☐	☐	☐

If you have a preference, please explain why:

20. Have you any suggestions to improve the current OSATS tool?

Yes	No
☐	☐

If yes, please give details:

21. Do you have any suggestions for the development of a new assessment tool for assessing surgical competency?

Yes ☐	No ☐

If yes, please give details:

22. Do you feel your surgical performance *in this study* was affected by assessment?

Yes ☐	No ☐

If yes, please give details:

23. Do you think there has been any difference in the OSATS performed within this study compared to the usual OSATS assessments in your training?

Not applicable ☐	Yes ☐	No ☐

If yes, please give details:

24. Have you any further comments regarding OSATS assessment or this study?

Yes ☐	No ☐

If yes, please give details:

E. NOTSS questionnaire for anaesthetists and scrub nurses

Thank you for using NOTSS forms in your theatre sessions. Please could you now complete this short survey to give us feedback on your experience of using these forms.

1. Please indicate the extent to which you disagree or agree with the following statements about NOTSS: (please select appropriate box)

		1	2	3	4	5
		Strongly Disagree	Disagree	Neither agree or disagree	Agree	Strongly agree
a.	NOTSS provides a common language to discuss non-technical skills	☐	☐	☐	☐	☐
b.	It was easy to rate cognitive skills (e.g. decision making) situation awareness,	☐	☐	☐	☐	☐
c.	It was easy to rate interpersonal skills (e.g.communication and teamwork, leadership)	☐	☐	☐	☐	☐
d.	Using NOTSS forms added too much time to my operating list	☐	☐	☐	☐	☐
e.	NOTSS may be a useful tool to support reflective practice or provide insight for surgeons in training	☐	☐	☐	☐	☐
f.	NOTTS is a valuable adjunct to the available assessment tools for surgeons in training	☐	☐	☐	☐	☐
g.	Routine use of the NOTSS system will enhance safety in the operating theatre	☐	☐	☐	☐	☐
h.	NOTSS may be a useful tool for providing a trainee with feedback after an operation	☐	☐	☐	☐	☐

2. I have had sufficient training in using the NOTSS tool (please select appropriate box to indicate the extent to which you agree or disagree with this statement)

☐----------☐----------☐----------☐----------☐

Strongly disagree Disagree Neither agree or disagree Agree Strongly agree

Give details if you wish (enter text):

3. What training have you undertaken? (please select more than one as appropriate)

None	NOTSS Booklet	NOTSS video	Training workshop	Face-to-face discussion with research team	Other
☐	☐	☐	☐	☐	☐

If you selected 'other' box, please give details:

4. Did you work at the category level, element level, or both? (please select appropriate box)

Category	Element	Both
☐	☐	☐

5. Which of the four NOTSS categories did you focus on? (Please select more than one box as appropriate)

Situation awareness	Decision making	Communication	Teamwork/leadership
☐	☐	☐	☐

6. If you did not use all the categories, could you explain why (please enter text):

7. How easy did you find it to rate the non-technical behaviours of surgeons using NOTSS forms? (please select a number on the scale)

1 2 3 4 5 6 7 8 9 10

☐---------☐---------☐---------☐---------☐---------☐---------☐---------☐---------☐--------☐

Very difficult Very easy

If you had difficulties rating behaviours using NOTSS please explain why:

8. How important do you think it is *within surgical education* to rate the non-technical skills of surgeons in training? (please select a number on the scale)

1 2 3 4 5 6 7 8 9 10

No importance — Of moderate importance — Very important

Give details if you wish:

9. What is your level of *overall satisfaction* with the NOTSS tool for assessing the non-technical skills of surgeons in training? (please select a number on the scale)

1 2 3 4 5 6 7 8 9 10

No satisfaction — Moderately satisfied — Highly satisfied

Please give details if you wish:

10. Would you be prepared to use NOTSS in your theatre sessions again? (please select appropriate box)

Never	Sometimes	Always	Not sure
☐	☐	☐	☐

Please give details if you wish:

11. Have you any comments or suggestions to improve the NOTSS tool *or* for alternative methods of rating non technical skills?

Yes	No
☐	☐

If yes, please give details:

12. Have you any further comments or concerns regarding NOTSS or this research study?

Yes	No
☐	☐

If yes, please give details:

Appendix 6

Original study proposal

Background

Surgical training in the UK has traditionally been based upon an apprenticeship and examination model. Trainees must complete a set number of years of training and pass the Intercollegiate Examination of the Royal Colleges of Surgeons, in order to achieve their Certificate of Completion of Specialist Training. Technical skills are not formally assessed. Logbooks form a useful record of experience (Galasko and Mackay, 1997) but this does not guarantee competence, as we have shown (Thornton *et al.*, 2003). Competence can be defined as 'what a person does in a controlled representation of professional practice', e.g. when a trainee performs an operation under supervision (Rethans *et al.*, 2002). Competence usually comes from experience (practice) combined with positive feedback (Reznick, 1993), and positive feedback has been defined as 'an informed, non-evaluative, objective appraisal that is intended to improve clinical skills' (Rogers, 1969). Performance can be defined as 'what a person does in actual professional practice' (Rethans *et al.*, 2002). The opportunity to gain experience in the operating room is also decreasing. We, and others, have shown a reduction in the numbers of operations undertaken and the level of competence achieved by surgical trainees (Katory *et al.*, 2001). The reasons for this reduction include shorter training time following the Calman Report (Calman *et al.*, 1999), the EWTD (Department of Health, 2003) and new working practices which mean that more operations are performed by consultants. Thus, the traditional apprenticeship model, where technical competence was usually achieved through many years and long hours, seems no longer appropriate.

Although attractive, measurement of the performance of consultant surgeons based on outcomes is fraught with difficulty due to variation in case-mix, and the large numbers required for reliability (Prytherch *et al.,* 2001). It is probably a good screening method, but tests of competence will be required for those consultants in whom there is cause for concern. Measurement of performance using outcomes of trainee surgeons may be even more difficult, because errors made by trainees are often corrected (masked) by their supervising consultant (Szalay *et al.*, 2000). For this reason, the skills assessments developed by the GMC Performance Procedures (Beard *et al.*, 2005a) and by the ISCP (www.iscp.uk) have been competence-based. The ISCP is a collaborative venture between all the Specialty Surgical Associations and the Royal Colleges of Surgeons in the UK and Ireland. Trainees' progress through the new Intercollegiate Surgical Curriculum will be measured by an integrated framework of WBAs, annual reviews (RITAs) and examinations. The various assessment instruments are designed to provide a mixture of formative feedback to trainees, and summative assessments which must be cleared in order to progress. The overall assessment strategy and the individual assessment tools conform to the assessment principles laid down by the PMETB (PMETB, 2005), and the assessment tools are designed to measure all the domains of *Good Medical Practice* (GMC, 1998).

It seems axiomatic that direct observation of technical skill in the operating theatre represents the 'gold standard' in terms of content and construct validity. The technical skills of trainees in the operating theatre was first assessed objectively, using a two-part structured technical skills assessment form, by Reznick's group in Toronto (Winckel *et al.*, 1994). Part 1 consists of the essential components of the procedure (task-specific checklist). Part 2 is a global rating form

which consists of more non-specific items, e.g. handling of instruments and communication with the theatre staff. The group used the same assessment methodology, renamed OSATS, on surgical simulations in the skills laboratory with similar results (Martin *et al.*, 1997). They also showed that global ratings possessed slightly better construct validity when comparing a mixed group of trainees and consultants (Regehr *et al.*, 1998). We have recently confirmed this finding but interestingly found that checklists were more discriminatory for trainees (Beard, 2007). The assessment method used depends upon the purpose of assessment. One purpose is to provide feedback to aid learning (formative assessment), e.g. during training. Another is to check that a level of competence has been achieved or maintained (summative assessment), e.g. for certification or revalidation. These two purposes are not mutually exclusive – there is no reason why a 'summative' assessment should not provide feedback, and a collection of formative assessments can also be viewed summatively.

Dual assessments are time consuming to perform and each method may have different roles. Global ratings seem useful when assessing more complex operations, especially when there is more than one method of performing the task correctly, or when assessing experts for the purposes of certification or revalidation. Task-specific checklists provide a trainee with detailed instructions and feedback on how to undertake the operation in an approved way. We have developed a simpler assessment tool for saphenofemoral ligation which combines task-specific and global items (Beard *et al.*, 2005b). This has been validated against the standard global rating method and seems a good test of technical skills for intermediate trainees. PBA, adopted by the ISCP as the main WBA for surgical trainees, is a similar combination of task-based and global items, together with a summary judgement about the competence of the trainees to perform that operation. However, little validation of PBA has been done, especially regarding its transferability to a wide range of specialties and procedures.

One concern about PBA and other such technical assessments is that they may not reflect 'higher-order' skills that underpin technical proficiency, such as situation awareness, decision-making, team working and leadership. The NOTSS tool has been developed by the Department of Psychology at the University of Aberdeen, in collaboration with the Royal College of Surgeons of Edinburgh, to address these areas of NOTSS (www.abdn.ac.uk/iprc/notss). The NOTSS system comprises a three-level hierarchy consisting of categories (at the highest level), elements and behaviours: four skills categories (situation awareness, decision-making, communication and teamwork, leadership) and 12 elements make up the skills taxonomy with examples of good and poor behaviours provided for each element. The aim is to provide a common terminology that allows all those working in this area to understand each other, and a framework for trainee and consultant surgeons to develop their abilities in the workplace (Yule *et al.*, 2006).

Another question is whether such assessments can be reliably performed by other health-care professionals, e.g. theatre nurses, as this could ease the assessment burden for consultants. A standard-setting exercise conducted by the Vascular Society suggested that theatre nurses were able to reliably discriminate between different levels of operative competence (Beard *et al.,* 2005b). Multi professional assessment has been shown to possess good reliability for the multi-source feedback tool (Mini-PAT) which has been adopted by the Foundation Programme and by the ISCP to assess aspects of professional behaviour (Archer *et al.*, 2005). Self-assessment is another important component of Mini-PAT as this provides valuable information about insight, which seems vital for the development of competence (Hays and Jolly, 2002). There have been few studies of the reliability of self-assessment and none in surgery (Fitzgerald *et al.*, 2003).

Video-recording of operations for subsequent analysis may prove useful when external assessment is required. A portfolio of recorded consultations forms part of the requirement for the Membership of the Royal College of General Practitioners (Joint Committee on Postgrad

Training for General Practice, 2004). Many operating theatres are now equipped with camera lights and video monitors. We have shown good inter-rater reliability between direct and video assessment of saphenofemoral ligation (Beard *et al.*, 2005c). However, Scott *et al.* found that global assessment of edited videotapes of laparoscopic choleystectomies did not correlate well with direct observation (Scott *et al.,* 2000). A study conducted on behalf of the Vascular Society also found that video recordings of trainees performing carotid endarterectomies could not be scored reliably without information on the amount of help provided by the trainer who was assisting (unpublished data). Reliability for more complex operations may be improved by dual recordings of the operative field and the operating room, combined with voice recordings. Video recordings, combined with structured assessment forms may provide a powerful feedback tool (Backstein *et al.*, 2004).

Purpose of research

The aim of this study is to compare the validity, reliability and user satisfaction of three different methods of assessing surgical skills in the operating theatre. Content validity (whether it contains all the components required), construct validity (whether it measures what it is supposed to), predictive validity (correlation with outcome) and educational validity (impact on learning) will be studied. The reliability of various assessors and video recordings (inter-rater reliability) and inter-specialty differences will also be studied as will insight, acceptability and educational impact. The information provided by this study will be of great value to the ISCP, the GMC Revalidation and Performance Procedures and the National Clinical Assessment Authority. It will also inform the selection of performance objectives and metrics for subsequent simulation design.

Subject group, location and sample size

Consultant and trainee surgeons in upper GI, colorectal, vascular, orthopaedic and cardiac surgery at three teaching hospitals (Sheffield, Leeds and Nottingham) will be assessed. The advantage of using three hospitals is that a larger number of assessments can be obtained in the time available. Fifty to one hundred assessments over 16 months for each of the nine index procedures: laparoscopic cholecystectomy, right hemicolectomy, anterior resection, carotid endarterectomy, aortic aneurysm repair, total knee replacement, total hip replacement, coronary artery bypass and aortic valve replacement will be undertaken. Each surgeon will be assessed undertaking the two relevant index procedures on at least two occasions, to help compensate for variation in case complexity. The two operations will preferably be performed on the same day, but otherwise with as little delay as possible, to avoid any significant training effect. To find a significant correlation between two variables that is different from zero can be done with about 28 subjects if that correlation is 0.7, but if it is only 0.3 the sample size goes up to 136. A minimum of 50–60 subjects in each operation group will be required to estimate these curves. This will detect whether the five operations were significantly different in their learning curve characteristics, e.g. the confidence limits would not cross, or do so for only part of the curve, or two linear slopes were different. For multiple regression, the larger the sample size the better.

These major operations are all performed on a daily/weekly basis at all three centres. The lead assessor will be based in Sheffield and visit Nottingham and Leeds 1 day each week. Three days each week will be spent collecting assessments, 1 day spent collating the data and 1 day spent following up the in-hospital outcomes and scheduling the next week's assessments. Support will be provided by the administrative and secretarial assistant, who will also be responsible for maintaining the record of expenditure.

The reason for selecting these nine operations is that the task analyses for these particular operations have already been developed by the principal investigator in Sheffield for the GMC Performance Procedures (Beard). They were each subsequently validated for content by at least 10 specialist consultants and senior trainees from Sheffield, Nottingham and Leeds. PBAs for these operations have since been written by the respective Specialty Advisory Committees for the ISCP. These operations also represent typical index procedures for each specialty.

Methodology

Three different assessment methods will be compared in terms of the parameters outlined in the aims and objectives. These are OSATS, using the global rating scale, PBA which has been adopted as the main workplace-based assessment tool for the ISCP and NOTSS. The assessment forms can be found in *Appendix 1*.

The lead assessor will receive the operating lists for the relevant specialties from the three hospitals each week. Suitable operations will be identified and the theatre sister, surgeon and assistant informed. Surgeons will be asked to provide information on their age, gender, country of qualification, duration of training, the total number of operations previously performed and the number in the last 6 months (plus duration of practice if a consultant) and whether or not they have received any training in assessment, as these have all been shown to have an effect in other WBAs. The patient will also be given an information leaflet explaining the study, and consent for video recording obtained. Prior to the operation the recording equipment will be assembled by the lead assessor, and the assessment forms with written instructions given to the surgeon and assistant. The patient information sheet, consent form and instruction sheet for the surgeon and assistant can be found in *Appendix 2*. During the operation the lead assessor will complete the OSATS, PBA and NOTSS forms as well as recording the ASA status of the patient, the duration of the operation, the difficulty of the operation, blood loss and any intraoperative complications. The surgeon and assistant will complete their assessment forms at the end of the operation and, if the surgeon is a trainee, the supervising consultant will be asked to provide feedback. They will also record how long the forms took to complete and their satisfaction with the new assessment methods, using Kirkpatrick's model for evaluating educational outcomes (Freeth *et al.*, 2003). The lead assessor will discuss any differences between the various scores with the surgeon and assistant and ask for further comments on the assessment and feedback process. The scrub nurse and the anaesthetist will also be asked to complete an NOTSS form. Completion of the forms and the subsequent discussion should not affect service delivery as there is usually plenty of time between cases. After discharge the lead assessor will examine the case records to record any postoperative complications and the length of stay.

Surgeons will be sent a questionnaire by e-mail about 1 month after their assessment, to further evaluate the educational impact of the new assessment methodology, after a period for reflection, again using Kirkpatrick's model.

Some surgeons may be subsequently asked to perform a simulated operation in the skills laboratory (e.g. the carotid endarterectomy model, Limbs & Things, Bristol, UK) to study the correlation between simulation and reality.

Videos of the operating field and operating room will be recorded screen-in-screen, together with sound, onto DVD. Specialty experts will perform the same assessments from the DVDs and will not be informed of the identity, seniority or experience of the surgeon. It is likely that the experts might recognise some of the surgeons and trainees but a previous study showed no evidence of

any halo effect using this method. The recordings can be assessed in fast playback mode, which we have used successfully for the analysis of operative recordings in the past.

Analysis

Satisfaction will be judged according to a simple presentation of the responses from the surgeon and the assistant. This will be presented as a proportion of responses in each response category and a digest of unstructured comments.

Reliability indicates how well an assessor's score of the surgeon's performance (using each assessment method) would reflect any assessor's score when the surgeon undertakes the procedure on any patient. It will be presented as the standard error of measurement of a single score, and as the number of assessors and cases that need to be combined to reach a predetermined level of reliability. Its calculation depends on comparing the effect of assessor-to-assessor variation and case-to-case variation in scores (sources of error) with overall surgeon-to-surgeon variation in scores. The analysis will be conducted using generalisability theory. The *G*-study, or variance component analysis, will be conducted using the VARCOMP procedure in SPSS. The MINQUE method will be used because of its superior handling of unbalanced data. The model will assume that all effects are randomly sampled from an infinite universe, and will estimate the effect on score of surgeon, case (nested within surgeon), assessor (partially crossed with surgeon), assessor designation (lead assessor, assistant, nurse, anaesthetist), and the second-order effects of assessor and surgeon designation. Redundant effects will be excluded by reverse stepwise regression. The variances obtained will be combined using the standard formula for standard error of measurement and using Cronbach's equations to estimate the effect of multiple assessors or cases.

Validity indicates how well the score reflects the intended construct of 'surgical performance'. The study provides many sources of information about validity and these will all be presented in evidence for or against the validity of the two methods. If valid, the following hypotheses will be fulfilled:

1. Scores obtained by each assessment method will correlate with the other assessment method.
2. Scores will increase with duration of training, number of similar procedures performed (experience) and duration of practice if a consultant (seniority).
3. Higher-scoring operations will result in less perioperative blood loss and in fewer perioperative and postoperative complications and shorter length of stay.
4. Mean scores, and scores for each element, will not be significantly different across the nine different procedures.

Each of these hypotheses will be tested. Pearson's method will be used for hypothesis 1 and 2. A cross-tabulation method will be used for hypothesis 3. A one-way ANOVA will be used for hypothesis 4.

Secondary outcomes will include:

1. The relationship between assessed scores and self-scores as a measure of insight of the surgeon. Scores will be compared for correlation using Pearson's method.
2. The validity (fidelity) of video as an indicator of directly observed performance. Scores will be compared for correlation using Pearson's method.
3. The validity (fidelity) of simulators as an indicator of directly observed performance. Scores will be compared for correlation using Pearson's method.

4. The educational impact of the assessment on trainees. The satisfaction of the trainee, immediately after feedback, will be compared with the lead assessor and consultant supervisor's scores using Pearson's method. Links between identified learning objectives, other comments and the scores will also be studied using qualitative methods of analysis. Ideally, we would like to demonstrate that performance on the procedures improved over time in a group being given feedback from the assessments, compared with a control group not being given such feedback. We hope that this will be the subject of a subsequent randomised trial.
5. Educational impact at 1 month in terms of progression to higher levels on Kirkpatrick's model, e.g. have the surgeons used the assessment methods again?

Scheduling

Centre of Research Ethical Campaign and Local Research Ethics Committee approval will be obtained prior to the commencement of the study in April 2007. The Trial Management Committee and Monitoring Committee will also meet or teleconference, initially separately, and then together. Purchase of the audiovisual equipment and training of the lead assessor in its use, and in the assessment methodologies and giving feedback, will also be undertaken before this time. Site visits to the operating theatres, meetings with the specialty departments involved, identification of consultants and trainees plus some preliminary data collection will be undertaken during the first 2 months of the study. This process will be facilitated by the principal investigators at Sheffield, Leeds and Nottingham. Data collection will then continue for 16 months, leaving 6 months for data analysis and writing.

Output of study

Day-to-day management of the trial will be the responsibility of the principal investigator and lead assessor, helped by the other members of the Trial Management Committee when required. The Trial Monitoring Committee will provide overall supervision to ensure that the study is conducted to according to the Department of Health's research governance framework and the Medical Research Council's Guidelines for Good Clinical Practice, including trial progress, adherence to protocol and patient safety. Interim reports on progress will be provided to the sponsor at 6, 12 and 18 months. A final report will be issued within 24 months. The interim and final reports will comply with the requirements for such reports. Presentation to learned societies such as the Association of Surgeons Annual General Meeting and the Annual Ottawa Medical Education Conference, as well as publication in related journals such as the *British Journal of Surgery* and *Medical Education*, are planned, subject to approval.

References

Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. *BMJ* 2005;**330**:1251–3.

Backstein D, Agnidis Z, Regehr G, Reznick RK. The effectiveness of video feedback in the acquisition of orthopedic technical skills. *Am J Surg* 2004;**187**:427–32.

Beard JD, Choksy S, Khan S. Assessment of operative competence during carotid endarterectomy. *Br J Surg* 2007;**94**:726–30.

Beard JD, Jolly BC, Southgate LJ, Newble DI, Thomas WEG, Rochester J. Developing assessments of surgical skills for the GMC Performance Procedures. *Ann R Coll Surg Engl* 2005a;**87**:242–7.

Beard JD, on behalf of the Education and Training Committee of the Vascular Society. Setting standards for the assessment of operative competence. *Eur J Vasc Endovasc Surg* 2005b;**30**:215–18.

Beard JD, Jolly BC, Newble DI, Thomas WEG, Donnelly J, Southgate LJ. Assessing the technical skills of surgical trainees. *Br J Surg* 2005c;**92**:778–82.

Calman KC, Temple JG, Naysmith R, Cairncross RG, Bennett SJ. Reforming higher specialist training in the United Kingdom – a step along the continuum of medical education. *Med Educ* 1999;**33**:28–33.

Department of Health. *European Working Time Directive*. 2003. Available from: www.doh.gov.uk/workingtime/index.htm

Fitzgerald JT, White CB, Gruppen LD. A longitudinal study of self-assessment accuracy. *Med Educ* 2003;**37**:645–9.

Freeth D, Hammick M, Reeves S, Barr H. *Critical review of evaluations of interprofessional education*. London: Higher Education Academy Learning and Teaching Support Network for Health Sciences and Practice; 2003.

General Medical Council. *Good Medical Practice*. London: General Medical Council; 1998.

Galasko C, Mackay C. Unsupervised surgical training: logbooks are essential for assessing progress. *BMJ* 1997;**315**:1306–7.

Hays RB, Jolly BC. Is insight important? Measuring capacity to change performance. *Med Educ* 2002;**36**:967–71.

Joint Committee on Postgrad Training for General Practice. 2004. Available from: www.jcptgp.org.uk/certification/assessment.asp

Katory M, Singh S, Beard JD. Twenty Trent trainees: a comparison of operative competence after BST. *Ann R Coll Surg Engl* 2001;**83**(Suppl.):328–30.

Martin JA, Regehr G, Reznick R, MacRoe H, Murnaghan J, Hutchinson C. Objective structured assessment of technical skill (OSATS) for surgical residents. *Br J Surg* 1997;**84**:273–8.

PMETB. *Guidance from the Examinations in Postgraduate Medicine and the Workplace Based Assessment Subcommittees of the Postgraduate Medical Education and Training Board. Quality Assurance, Quality Control and Assessment Systems.* 2005. Available from: www.pmetb.org.uk/media/pdf/1/a/PMETB_quality_assurance_quality_control_and_assessment_systems_guidance_(1_August_2005).pdf

Prytherch DR, Ridler BMF, Beard JD, Earnshaw JJ, on behalf of the Audit and Research Committee of the Vascular Society of Great Britain & Ireland. A model for national outcome audit in vascular surgery. *Eur J Vasc Endovasc Surg* 2005;**21**:477–83.

Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE format examination. *Acad Med* 1998;**73**:993–7.

Rethans J-J, Norcini JJ, Baron-Maldonado M, Blackmore D, Jolly BC, LaDuca T, *et al.* The relationship between competence and performance: implications for assessing practice performance. *Med Educ* 2002;**36**:901–9.

Reznick RK. Teaching and testing technical skills. *Am J Surg* 1993;**165**:358–61.

Rogers CR. *Freedom to learn*. Columbus, OH: Merrill; 1969.

Scott DJ, Rege RV, Bergen PC, Guo WA, Laycock R, Tesfay ST, *et al.* Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. *J Laparosc Adv Surg Tech* 2000;**10**:183–90.

Szalay D, MacRae H, Regehr G, Reznick R. Using operative outcome to assess technical skill. *Am J Surg* 2000;**180**:234–7.

Thornton M, Donlon M, Beard JD. The operative skills of higher surgical trainees: measuring competence rather than experience undertaken. *Ann R Coll Surg Engl* 2003;**85**:190–3.

Winckel CP, Reznick RK, Cohen R, Taylor B. Reliability and construct validity of a structured technical skills assessment form. *Am J Surg* 1994;**167**:423–7.

Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons' non-technical skills. *Med Educ* 2006;**40**:1098–104.

Health Technology Assessment programme

HTA Commissioning Board *(continued)*

Professor James Raftery,
Chair of NETSCC and Director of the Wessex Institute, University of Southampton

Professor Barney Reeves,
Professorial Research Fellow in Health Services Research, Department of Clinical Science, University of Bristol

Professor Martin Underwood,
Warwick Medical School, University of Warwick

Professor Marion Walker,
Professor in Stroke Rehabilitation, Associate Director UK Stroke Research Network, University of Nottingham

Dr Duncan Young,
Senior Clinical Lecturer and Consultant, Nuffield Department of Anaesthetics, University of Oxford

Observers

Dr Morven Roberts,
Clinical Trials Manager, Health Services and Public Health Services Board, Medical Research Council

HTA Clinical Evaluation and Trials Board

Chair,
Professor Sallie Lamb,
Director,
Warwick Clinical Trials Unit,
Warwick Medical School,
University of Warwick and Professor of Rehabilitation,
Nuffield Department of Orthopaedic, Rheumatology and Musculoskeletal Sciences,
University of Oxford

Deputy Chair,
Professor Jenny Hewison,
Professor of the Psychology of Health Care,
Leeds Institute of Health Sciences,
University of Leeds

Programme Director,
Professor Tom Walley, CBE,
Director, NIHR HTA programme, Professor of Clinical Pharmacology, University of Liverpool

Members

Professor Keith Abrams,
Professor of Medical Statistics, Department of Health Sciences, University of Leicester

Professor Martin Bland,
Professor of Health Statistics, Department of Health Sciences, University of York

Professor Jane Blazeby,
Professor of Surgery and Consultant Upper GI Surgeon, Department of Social Medicine, University of Bristol

Professor Julia M Brown,
Director, Clinical Trials Research Unit, University of Leeds

Professor Alistair Burns,
Professor of Old Age Psychiatry, Psychiatry Research Group, School of Community-Based Medicine, The University of Manchester & National Clinical Director for Dementia, Department of Health

Dr Jennifer Burr,
Director, Centre for Healthcare Randomised trials (CHART), University of Aberdeen

Professor Linda Davies,
Professor of Health Economics, Health Sciences Research Group, University of Manchester

Professor Simon Gilbody,
Prof of Psych Medicine and Health Services Research, Department of Health Sciences, University of York

Professor Steven Goodacre,
Professor and Consultant in Emergency Medicine, School of Health and Related Research, University of Sheffield

Professor Dyfrig Hughes,
Professor of Pharmacoeconomics, Centre for Economics and Policy in Health, Institute of Medical and Social Care Research, Bangor University

Professor Paul Jones,
Professor of Respiratory Medicine, Department of Cardiac and Vascular Science, St George's Hospital Medical School, University of London

Professor Khalid Khan,
Professor of Women's Health and Clinical Epidemiology, Barts and the London School of Medicine, Queen Mary, University of London

Professor Richard J McManus,
Professor of Primary Care Cardiovascular Research, Primary Care Clinical Sciences Building, University of Birmingham

Professor Helen Rodgers,
Professor of Stroke Care, Institute for Ageing and Health, Newcastle University

Professor Ken Stein,
Professor of Public Health, Peninsula Technology Assessment Group, Peninsula College of Medicine and Dentistry, Universities of Exeter and Plymouth

Professor Jonathan Sterne,
Professor of Medical Statistics and Epidemiology, Department of Social Medicine, University of Bristol

Mr Andy Vail,
Senior Lecturer, Health Sciences Research Group, University of Manchester

Professor Clare Wilkinson,
Professor of General Practice and Director of Research North Wales Clinical School, Department of Primary Care and Public Health, Cardiff University

Dr Ian B Wilkinson,
Senior Lecturer and Honorary Consultant, Clinical Pharmacology Unit, Department of Medicine, University of Cambridge

Observers

Ms Kate Law,
Director of Clinical Trials,
Cancer Research UK

Dr Morven Roberts,
Clinical Trials Manager, Health Services and Public Health Services Board, Medical Research Council

Diagnostic Technologies and Screening Panel

Members

Chair,
Professor Lindsay Wilson Turnbull,
Scientific Director of the Centre for Magnetic Resonance Investigations and YCR Professor of Radiology, Hull Royal Infirmary

Professor Judith E Adams,
Consultant Radiologist, Manchester Royal Infirmary, Central Manchester & Manchester Children's University Hospitals NHS Trust, and Professor of Diagnostic Radiology, University of Manchester

Mr Angus S Arunkalaivanan,
Honorary Senior Lecturer, University of Birmingham and Consultant Urogynaecologist and Obstetrician, City Hospital, Birmingham

Dr Stephanie Dancer,
Consultant Microbiologist, Hairmyres Hospital, East Kilbride

Dr Diane Eccles,
Professor of Cancer Genetics, Wessex Clinical Genetics Service, Princess Anne Hospital

Dr Trevor Friedman,
Consultant Liason Psychiatrist, Brandon Unit, Leicester General Hospital

Dr Ron Gray,
Consultant, National Perinatal Epidemiology Unit, Institute of Health Sciences, University of Oxford

Professor Paul D Griffiths,
Professor of Radiology, Academic Unit of Radiology, University of Sheffield

Mr Martin Hooper,
Service User Representative

Professor Anthony Robert Kendrick,
Associate Dean for Clinical Research and Professor of Primary Medical Care, University of Southampton

Dr Anne Mackie,
Director of Programmes, UK National Screening Committee, London

Mr David Mathew,
Service User Representative

Dr Michael Millar,
Consultant Senior Lecturer in Microbiology, Department of Pathology & Microbiology, Barts and The London NHS Trust, Royal London Hospital

Mrs Una Rennard,
Service User Representative

Dr Stuart Smellie,
Consultant in Clinical Pathology, Bishop Auckland General Hospital

Ms Jane Smith,
Consultant Ultrasound Practitioner, Leeds Teaching Hospital NHS Trust, Leeds

Dr Allison Streetly,
Programme Director, NHS Sickle Cell and Thalassaemia Screening Programme, King's College School of Medicine

Dr Alan J Williams,
Consultant Physician, General and Respiratory Medicine, The Royal Bournemouth Hospital

Observers

Dr Tim Elliott,
Team Leader, Cancer Screening, Department of Health

Dr Catherine Moody,
Programme Manager, Medical Research Council

Professor Julietta Patrick,
Director, NHS Cancer Screening Programme, Sheffield

Dr Kay Pattison,
Senior NIHR Programme Manager, Department of Health

Professor Tom Walley, CBE,
Director, NIHR HTA programme, Professor of Clinical Pharmacology, University of Liverpool

Dr Ursula Wells,
Principal Research Officer, Policy Research Programme, Department of Health

Disease Prevention Panel

Members

Chair,
Professor Margaret Thorogood,
Professor of Epidemiology, University of Warwick Medical School, Coventry

Dr Robert Cook,
Clinical Programmes Director, Bazian Ltd, London

Dr Colin Greaves,
Senior Research Fellow, Peninsula Medical School (Primary Care)

Mr Michael Head,
Service User Representative

Professor Cathy Jackson,
Professor of Primary Care Medicine, Bute Medical School, University of St Andrews

Dr Russell Jago,
Senior Lecturer in Exercise, Nutrition and Health, Centre for Sport, Exercise and Health, University of Bristol

Dr Julie Mytton,
Consultant in Child Public Health, NHS Bristol

Professor Irwin Nazareth,
Professor of Primary Care and Director, Department of Primary Care and Population Sciences, University College London

Dr Richard Richards,
Assistant Director of Public Health, Derbyshire Country Primary Care Trust

Professor Ian Roberts,
Professor of Epidemiology and Public Health, London School of Hygiene & Tropical Medicine

Dr Kenneth Robertson,
Consultant Paediatrician, Royal Hospital for Sick Children, Glasgow

Dr Catherine Swann,
Associate Director, Centre for Public Health Excellence, NICE

Professor Carol Tannahill,
Glasgow Centre for Population Health

Mrs Jean Thurston,
Service User Representative

Professor David Weller,
Head, School of Clinical Science and Community Health, University of Edinburgh

Observers

Ms Christine McGuire,
Research & Development, Department of Health

Dr Kay Pattison,
Senior NIHR Programme Manager, Department of Health

Professor Tom Walley, CBE,
Director, NIHR HTA programme, Professor of Clinical Pharmacology, University of Liverpool

External Devices and Physical Therapies Panel

Members

Chair,
Dr John Pounsford,
Consultant Physician North Bristol NHS Trust

Deputy Chair,
Professor E Andrea Nelson,
Reader in Wound Healing and Director of Research, University of Leeds

Professor Bipin Bhakta,
Charterhouse Professor in Rehabilitation Medicine, University of Leeds

Mrs Penny Calder,
Service User Representative

Dr Dawn Carnes,
Senior Research Fellow, Barts and the London School of Medicine and Dentistry

Dr Emma Clark,
Clinician Scientist Fellow & Cons. Rheumatologist, University of Bristol

Mrs Anthea De Barton-Watson,
Service User Representative

Professor Nadine Foster,
Professor of Musculoskeletal Health in Primary Care Arthritis Research, Keele University

Dr Shaheen Hamdy,
Clinical Senior Lecturer and Consultant Physician, University of Manchester

Professor Christine Norton,
Professor of Clinical Nursing Innovation, Bucks New University and Imperial College Healthcare NHS Trust

Dr Lorraine Pinnigton,
Associate Professor in Rehabilitation, University of Nottingham

Dr Kate Radford,
Senior Lecturer (Research), University of Central Lancashire

Mr Jim Reece,
Service User Representative

Professor Maria Stokes,
Professor of Neuromusculoskeletal Rehabilitation, University of Southampton

Dr Pippa Tyrrell,
Senior Lecturer/Consultant, Salford Royal Foundation Hospitals' Trust and University of Manchester

Dr Sarah Tyson,
Senior Research Fellow & Associate Head of School, University of Salford

Dr Nefyn Williams,
Clinical Senior Lecturer, Cardiff University

Observers

Dr Kay Pattison,
Senior NIHR Programme Manager, Department of Health

Professor Tom Walley, CBE,
Director, NIHR HTA programme, Professor of Clinical Pharmacology, University of Liverpool

Dr Ursula Wells,
Principal Research Officer, Policy Research Programme, Department of Health

Interventional Procedures Panel

Members

Chair,
Professor Jonathan Michaels,
Professor of Vascular Surgery, University of Sheffield

Deputy Chair,
Mr Michael Thomas,
Consultant Colorectal Surgeon, Bristol Royal Infirmary

Mrs Isabel Boyer,
Service User Representative

Mr David P Britt,
Service User Representative

Mr Sankaran Chandra Sekharan,
Consultant Surgeon, Breast Surgery, Colchester Hospital University NHS Foundation Trust

Professor Nicholas Clarke,
Consultant Orthopaedic Surgeon, Southampton University Hospitals NHS Trust

Ms Leonie Cooke,
Service User Representative

Mr Seumas Eckford,
Consultant in Obstetrics & Gynaecology, North Devon District Hospital

Professor Sam Eljamel,
Consultant Neurosurgeon, Ninewells Hospital and Medical School, Dundee

Dr Adele Fielding,
Senior Lecturer and Honorary Consultant in Haematology, University College London Medical School

Dr Matthew Hatton,
Consultant in Clinical Oncology, Sheffield Teaching Hospital Foundation Trust

Dr John Holden,
General Practitioner, Garswood Surgery, Wigan

Professor Nicholas James,
Professor of Clinical Oncology, School of Cancer Sciences, University of Birmingham

Dr Fiona Lecky,
Senior Lecturer/Honorary Consultant in Emergency Medicine, University of Manchester/Salford Royal Hospitals NHS Foundation Trust

Dr Nadim Malik,
Consultant Cardiologist/Honorary Lecturer, University of Manchester

Mr Hisham Mehanna,
Consultant & Honorary Associate Professor, University Hospitals Coventry & Warwickshire NHS Trust

Dr Jane Montgomery,
Consultant in Anaesthetics and Critical Care, South Devon Healthcare NHS Foundation Trust

Professor Jon Moss,
Consultant Interventional Radiologist, North Glasgow Hospitals University NHS Trust

Dr Simon Padley,
Consultant Radiologist, Chelsea & Westminster Hospital

Dr Ashish Paul,
Medical Director, Bedfordshire PCT

Dr Sarah Purdy,
Consultant Senior Lecturer, University of Bristol

Professor Yit Chiun Yang,
Consultant Ophthalmologist, Royal Wolverhampton Hospitals NHS Trust

Observers

Dr Kay Pattison,
Senior NIHR Programme Manager, Department of Health

Dr Morven Roberts,
Clinical Trials Manager, Health Services and Public Health Services Board, Medical Research Council

Professor Tom Walley, CBE,
Director, NIHR HTA programme, Professor of Clinical Pharmacology, University of Liverpool

Dr Ursula Wells,
Principal Research Officer, Policy Research Programme, Department of Health

Pharmaceuticals Panel

Members

Chair,
Professor Imti Choonara,
Professor in Child Health, University of Nottingham

Deputy Chair,
Dr Yoon K Loke,
Senior Lecturer in Clinical Pharmacology, University of East Anglia

Dr Martin Ashton-Key,
Medical Advisor, National Commissioning Group, NHS London

Mr John Chapman,
Service User Representative

Dr Peter Elton,
Director of Public Health, Bury Primary Care Trust

Dr Ben Goldacre,
Research Fellow, Division of Psychological Medicine and Psychiatry, King's College London

Dr James Gray,
Consultant Microbiologist, Department of Microbiology, Birmingham Children's Hospital NHS Foundation Trust

Ms Kylie Gyertson,
Oncology and Haematology Clinical Trials Manager, Guy's and St Thomas' NHS Foundation Trust London

Dr Jurjees Hasan,
Consultant in Medical Oncology, The Christie, Manchester

Dr Carl Heneghan,
Deputy Director Centre for Evidence-Based Medicine and Clinical Lecturer, Department of Primary Health Care, University of Oxford

Dr Dyfrig Hughes,
Reader in Pharmacoeconomics and Deputy Director, Centre for Economics and Policy in Health, IMSCaR, Bangor University

Dr Maria Kouimtzi,
Pharmacy and Informatics Director, Global Clinical Solutions, Wiley-Blackwell

Professor Femi Oyebode,
Consultant Psychiatrist and Head of Department, University of Birmingham

Dr Andrew Prentice,
Senior Lecturer and Consultant Obstetrician and Gynaecologist, The Rosie Hospital, University of Cambridge

Ms Amanda Roberts,
Service User Representative

Dr Martin Shelly,
General Practitioner, Silver Lane Surgery, Leeds

Dr Gillian Shepherd,
Director, Health and Clinical Excellence, Merck Serono Ltd

Mrs Katrina Simister,
Assistant Director New Medicines, National Prescribing Centre, Liverpool

Professor Donald Singer,
Professor of Clinical Pharmacology and Therapeutics, Clinical Sciences Research Institute, CSB, University of Warwick Medical School

Mr David Symes,
Service User Representative

Dr Arnold Zermansky,
General Practitioner, Senior Research Fellow, Pharmacy Practice and Medicines Management Group, Leeds University

Observers

Dr Kay Pattison,
Senior NIHR Programme Manager, Department of Health

Mr Simon Reeve,
Head of Clinical and Cost-Effectiveness, Medicines, Pharmacy and Industry Group, Department of Health

Dr Heike Weber,
Programme Manager, Medical Research Council

Professor Tom Walley, CBE,
Director, NIHR HTA programme, Professor of Clinical Pharmacology, University of Liverpool

Dr Ursula Wells,
Principal Research Officer, Policy Research Programme, Department of Health

Psychological and Community Therapies Panel

Members

Chair,
Professor Scott Weich,
Professor of Psychiatry, University of Warwick, Coventry

Deputy Chair,
Dr Howard Ring,
Consultant & University Lecturer in Psychiatry, University of Cambridge

Professor Jane Barlow,
Professor of Public Health in the Early Years, Health Sciences Research Institute, Warwick Medical School

Dr Sabyasachi Bhaumik,
Consultant Psychiatrist, Leicestershire Partnership NHS Trust

Mrs Val Carlill,
Service User Representative

Dr Steve Cunningham,
Consultant Respiratory Paediatrician, Lothian Health Board

Dr Anne Hesketh,
Senior Clinical Lecturer in Speech and Language Therapy, University of Manchester

Dr Peter Langdon,
Senior Clinical Lecturer, School of Medicine, Health Policy and Practice, University of East Anglia

Dr Yann Lefeuvre,
GP Partner, Burrage Road Surgery, London

Dr Jeremy J Murphy,
Consultant Physician and Cardiologist, County Durham and Darlington Foundation Trust

Dr Richard Neal,
Clinical Senior Lecturer in General Practice, Cardiff University

Mr John Needham,
Service User Representative

Ms Mary Nettle,
Mental Health User Consultant

Professor John Potter,
Professor of Ageing and Stroke Medicine, University of East Anglia

Dr Greta Rait,
Senior Clinical Lecturer and General Practitioner, University College London

Dr Paul Ramchandani,
Senior Research Fellow/Cons. Child Psychiatrist, University of Oxford

Dr Karen Roberts,
Nurse/Consultant, Dunston Hill Hospital, Tyne and Wear

Dr Karim Saad,
Consultant in Old Age Psychiatry, Coventry and Warwickshire Partnership Trust

Dr Lesley Stockton,
Lecturer, School of Health Sciences, University of Liverpool

Dr Simon Wright,
GP Partner, Walkden Medical Centre, Manchester

Observers

Dr Kay Pattison,
Senior NIHR Programme Manager, Department of Health

Dr Morven Roberts,
Clinical Trials Manager, Health Services and Public Health Services Board, Medical Research Council

Professor Tom Walley, CBE,
Director, NIHR HTA programme, Professor of Clinical Pharmacology, University of Liverpool

Dr Ursula Wells,
Principal Research Officer, Policy Research Programme, Department of Health

Expert Advisory Network

Members

Professor Douglas Altman,
Professor of Statistics in Medicine, Centre for Statistics in Medicine, University of Oxford

Professor John Bond,
Professor of Social Gerontology & Health Services Research, University of Newcastle upon Tyne

Professor Andrew Bradbury,
Professor of Vascular Surgery, Solihull Hospital, Birmingham

Mr Shaun Brogan,
Chief Executive, Ridgeway Primary Care Group, Aylesbury

Mrs Stella Burnside OBE,
Chief Executive, Regulation and Improvement Authority, Belfast

Ms Tracy Bury,
Project Manager, World Confederation of Physical Therapy, London

Professor Iain T Cameron,
Professor of Obstetrics and Gynaecology and Head of the School of Medicine, University of Southampton

Professor Bruce Campbell,
Consultant Vascular & General Surgeon, Royal Devon & Exeter Hospital, Wonford

Dr Christine Clark,
Medical Writer and Consultant Pharmacist, Rossendale

Professor Collette Clifford,
Professor of Nursing and Head of Research, The Medical School, University of Birmingham

Professor Barry Cookson,
Director, Laboratory of Hospital Infection, Public Health Laboratory Service, London

Dr Carl Counsell,
Clinical Senior Lecturer in Neurology, University of Aberdeen

Professor Howard Cuckle,
Professor of Reproductive Epidemiology, Department of Paediatrics, Obstetrics & Gynaecology, University of Leeds

Professor Carol Dezateux,
Professor of Paediatric Epidemiology, Institute of Child Health, London

Mr John Dunning,
Consultant Cardiothoracic Surgeon, Papworth Hospital NHS Trust, Cambridge

Mr Jonothan Earnshaw,
Consultant Vascular Surgeon, Gloucestershire Royal Hospital, Gloucester

Professor Martin Eccles,
Professor of Clinical Effectiveness, Centre for Health Services Research, University of Newcastle upon Tyne

Professor Pam Enderby,
Dean of Faculty of Medicine, Institute of General Practice and Primary Care, University of Sheffield

Professor Gene Feder,
Professor of Primary Care Research & Development, Centre for Health Sciences, Barts and The London School of Medicine and Dentistry

Mr Leonard R Fenwick,
Chief Executive, Freeman Hospital, Newcastle upon Tyne

Mrs Gillian Fletcher,
Antenatal Teacher and Tutor and President, National Childbirth Trust, Henfield

Professor Jayne Franklyn,
Professor of Medicine, University of Birmingham

Mr Tam Fry,
Honorary Chairman, Child Growth Foundation, London

Professor Fiona Gilbert,
Consultant Radiologist and NCRN Member, University of Aberdeen

Professor Paul Gregg,
Professor of Orthopaedic Surgical Science, South Tees Hospital NHS Trust

Bec Hanley,
Co-director, TwoCan Associates, West Sussex

Dr Maryann L Hardy,
Senior Lecturer, University of Bradford

Mrs Sharon Hart,
Healthcare Management Consultant, Reading

Professor Robert E Hawkins,
CRC Professor and Director of Medical Oncology, Christie CRC Research Centre, Christie Hospital NHS Trust, Manchester

Professor Richard Hobbs,
Head of Department of Primary Care & General Practice, University of Birmingham

Professor Alan Horwich,
Dean and Section Chairman, The Institute of Cancer Research, London

Professor Allen Hutchinson,
Director of Public Health and Deputy Dean of ScHARR, University of Sheffield

Professor Peter Jones,
Professor of Psychiatry, University of Cambridge, Cambridge

Professor Stan Kaye,
Cancer Research UK Professor of Medical Oncology, Royal Marsden Hospital and Institute of Cancer Research, Surrey

Dr Duncan Keeley,
General Practitioner (Dr Burch & Ptnrs), The Health Centre, Thame

Dr Donna Lamping,
Research Degrees Programme Director and Reader in Psychology, Health Services Research Unit, London School of Hygiene and Tropical Medicine, London

Professor James Lindesay,
Professor of Psychiatry for the Elderly, University of Leicester

Professor Julian Little,
Professor of Human Genome Epidemiology, University of Ottawa

Professor Alistaire McGuire,
Professor of Health Economics, London School of Economics

Professor Neill McIntosh,
Edward Clark Professor of Child Life and Health, University of Edinburgh

Professor Rajan Madhok,
Consultant in Public Health, South Manchester Primary Care Trust

Professor Sir Alexander Markham,
Director, Molecular Medicine Unit, St James's University Hospital, Leeds

Dr Peter Moore,
Freelance Science Writer, Ashtead

Dr Andrew Mortimore,
Public Health Director, Southampton City Primary Care Trust

Dr Sue Moss,
Associate Director, Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton

Professor Miranda Mugford,
Professor of Health Economics and Group Co-ordinator, University of East Anglia

Professor Jim Neilson,
Head of School of Reproductive & Developmental Medicine and Professor of Obstetrics and Gynaecology, University of Liverpool

Mrs Julietta Patnick,
Director, NHS Cancer Screening Programmes, Sheffield

Professor Robert Peveler,
Professor of Liaison Psychiatry, Royal South Hants Hospital, Southampton

Professor Chris Price,
Director of Clinical Research, Bayer Diagnostics Europe, Stoke Poges

Professor William Rosenberg,
Professor of Hepatology and Consultant Physician, University of Southampton

Professor Peter Sandercock,
Professor of Medical Neurology, Department of Clinical Neurosciences, University of Edinburgh

Dr Philip Shackley,
Senior Lecturer in Health Economics, Sheffield Vascular Institute, University of Sheffield

Dr Eamonn Sheridan,
Consultant in Clinical Genetics, St James's University Hospital, Leeds

Dr Margaret Somerville,
Director of Public Health Learning, Peninsula Medical School, University of Plymouth

Professor Sarah Stewart-Brown,
Professor of Public Health, Division of Health in the Community, University of Warwick, Coventry

Dr Nick Summerton,
GP Appraiser and Codirector, Research Network, Yorkshire Clinical Consultant, Primary Care and Public Health, University of Oxford

Professor Ala Szczepura,
Professor of Health Service Research, Centre for Health Services Studies, University of Warwick, Coventry

Dr Ross Taylor,
Senior Lecturer, University of Aberdeen

Dr Richard Tiner,
Medical Director, Medical Department, Association of the British Pharmaceutical Industry

Mrs Joan Webster,
Consumer Member, Southern Derbyshire Community Health Council

Professor Martin Whittle,
Clinical Co-director, National Co-ordinating Centre for Women's and Children's Health, Lymington

Current and past membership details of all HTA programme 'committees' are available from the HTA website (www.hta.ac.uk)

Feedback

The HTA programme and the authors would like to know your views about this report.

The Correspondence Page on the HTA website (www.hta.ac.uk) is a convenient way to publish your comments. If you prefer, you can send your comments to the address below, telling us whether you would like us to transfer them to the website.

We look forward to hearing from you.

NETSCC, Health Technology Assessment
Alpha House
University of Southampton Science Park
Southampton SO16 7NS, UK
Email: hta@hta.ac.uk
www.hta.ac.uk

ISSN 1366-5278