Improved Methods for the Assessment of Surgical Trainees.

Peter J. Driscoll

MD

University of Edinburgh
2009

Declaration.

I hereby declare that this thesis has been composed by myself and contains my own work.

I also declare that the work within this thesis has not been submitted for any other degree or professional qualification.

signed

Peter J. Driscoll.

Acknowledgements.

I would like to acknowledge the support of a number of people.

My darling wife, Karen, and my two beautiful children, Angus and Lily, without whom the sky would have fallen in long ago.

My supervisor, Mr. Simon Paterson-Brown, who has supported me throughout this research and who, through his direction and patience, has taught me the kind of surgeon I wish to be.

Drs. Nicola Maran and Ronnie Glavin at the Scottish Clinical Simulation Centre in Stirling, for their expertise and advice.

Dr. Jeremy Walker for his statistical advice and translation services.

All the consultants, specialist registrars, basic surgical trainees and nursing staff who took part in the studies herein.

The Royal College of Surgeons of Edinburgh and NHS Education for Scotland (formerly The Scottish Council for Postgraduate Medical and Dental Education) for their financial support of this project.

ABSTRACT.

BACKGROUND:

The Edinburgh Basic Surgical Trainee Assessment Form (EBSTAF) was developed from the consensus opinion of 111 consultants across all surgical specialties to assess a total of 70 skills and attributes felt to be necessary in a Basic Surgical Trainee (BST) for a successful surgical career. It was subsequently shown to be reliable and valid in its application to longitudinal multi-disciplinary assessment of the in-post everyday performance of BSTs. This thesis addresses the subsequent clinical application of EBSTAF and other methods in the assessment of BSTs in the southeast Scotland region.

METHODS & RESULTS:

Trainees are adult learners. If they do not recognise an assessment as relevant to their future practice, they may dismiss its findings and any feedback based upon it. Trainees were therefore asked their opinions on the fields examined by EBSTAF in order to determine the acceptability of the form and thus it's potential value to formative assessment. Response rate from 33 trainees was 100%. 68 (97%) of the 70 fields were considered to be of equal (44: 63%) or greater (24: 34%) importance than the original consultant consensus opinion. This supports the application of EBSTAF to formative assessment for BSTs.

For the purposes of trainee selection, an assessment must not only be robust (i.e. reliable and valid) but must identify individuals who are likely to do

well or struggle subsequently. The predictive validity of EBSTAF was therefore examined by the comparison of trainees' scores with their subsequent career progression. EBSTAF showed potential in this area by yielding lower scores for BSTs who took longer to progress than their peers or who left the specialty altogether. However, due to the structure of the EBSTAF form, which assessed 'competence', the identification of outstanding trainees was not possible.

EBSTAF assessments of a new BST cohort were used to provide structured and anonymous feedback of trainee performance. The highly-detailed nature of the feedback proved popular with the trainees who reported it to be very helpful in directing their efforts during their training. Despite initial reservations concerning their assessment by the nursing staff, they found their comments to be particularly insightful. Comparison of EBSTAF scores was also made between the original validation cohort and the feedback group. However, no statistically significant differences were demonstrated between the two groups. This study in combination with the BSTs' agreement on the importance of the fields within EBSTAF suggests that they will value future multidisciplinary assessment and the detailed feedback on performance it provides.

High-fidelity human patient simulation offers the opportunity to practice high-stress critical care scenarios in complete safety for both trainees and patients. This modality was applied to trainee assessment using purpose-written scenarios with performances rated using EBSTAF. This study was the first to demonstrate construct validity for closely-related levels of surgical

trainee and highlighted the influence of non-technical skills on clinical practice.

Finally, the validity of assessment of BST tissue-handling skills using edited video was examined. Hernia repairs were judged using a modified Technical Skills domain from EBSTAF (coined EBSTAF-Tech) in conjunction with the well-validated Global Ratings Scale of Operative Performance. This demonstrated highly-sensitive validity by allowing consultant assessors to discriminate between trainees separated by only 6 months training. Trainees were also able to identify good operative skills when provided with such a structured framework on which to base their assessments, suggesting that such assessment methods may improve BSTs self-assessment skills.

CONCLUSIONS:

The Edinburgh Basic Surgical Trainee Assessment Form offers a robust longitudinal assessment of the everyday performance of BSTs that is acceptable to them, has the formative benefit to guide training and the summative potential to identify those who may later struggle in their surgical career. Video-assessment of basic tissue-handling skills is valid and should be considered for future selection processes. High-fidelity human patient simulation of critical care scenarios is valid and highlights the need for formal classification, assessment and training of non-technical skills required for surgical practice.

PUBLICATIONS FROM THIS PROJECT.

Trainees' Opinions of the Skills Required of Basic Surgical Trainees (BSTs).
P.J.Driscoll, A.M.Paisley, S.Paterson-Brown.
American Journal of Surgery 186(1) (2003); 77-80

High Fidelity Patient Simulators – The Surgical Rabbit-Hole?
(Invited Contribution)
P.J.Driscoll, N.J.Maran.
Association of Surgeons in Training Yearbook 2003-2004; 48-49

Challenges in Surgical Training.
(Invited Contribution)
P.J.Driscoll
Surgeons News 3(2) (2004); 56-57

Video Assessment of Basic Surgical Trainess' Operative Skills.
Abstract
P.J.Driscoll, A.M.Paisley, S.Paterson-Brown
British Journal of Surgery 92(S1) (2005); 24

Video Assessment of Basic Surgical Trainees' Operative Skills.
P.J.Driscoll, A.M.Paisley, S.Paterson-Brown
American Journal of Surgery, In Press

PRESENTATIONS FROM THIS PROJECT:

And So Alice Happens Upon the Surgical Rabbit Hole.
(INVITED SPEAKER)
P.J.Driscoll, N.J.Maran
National Association for Medical Simulation
Liverpool, 7th October 2005.

Video-Assessment of Basic Surgical Trainees' Operative Skills.
P.J.Driscoll, A.M.Paisley, S.Paterson-Brown
Association of Surgeons of Great Britain & Ireland
Glasgow, 13th-15th April 2005

The High-Fidelity Patient Simulator and Surgical Critical Care.
(INVITED SPEAKER)
P.J.Driscoll, N.J.Maran, R.J. Glavin, S. Paterson-Brown
Immersion Simulator Users Meeting
Addenbrooke's Hospital, Cambridge, 17th September 2003

The High-Fidelity Patient Simulator and Surgical Critical Care.
P.J.Driscoll, N.J.Maran, R.J. Glavin, S. Paterson-Brown
The Association for the Study of Education in Medicine Annual Conference
Edinburgh, 8th-10th September 2003

A High-Fidelity Patient Simulator Model of Critical care for Basic Surgical Trainees (BSTs).
P.J.Driscoll, S.Paterson-Brown, R.J.Glavin, N.J.Maran.
University of Edinburgh School of Surgery, 8th November 2002

High-Fidelity Human-Patient Simulation in a New Critical Care Course for Basic Surgical Trainees (BSTs).
(INVITED SPEAKER)
P.J.Driscoll, S. Paterson-Brown, R.J. Glavin, N.J.Maran
Scottish Clinical Skills Network Annual Meeting
Ninewell's Hospital, Dundee, 27th September

High-Fidelity Human-Patient Simulation in a New Critical Care Course for Basic Surgical Trainees (BSTs).
P.J.Driscoll, S. Paterson-Brown, N.J.Maran
The Association for the Study of Education in Medicine Annual Conference,
Norwich, 10th-12th September 2002

High-Fidelity Human-Patient Simulation in a New Critical Care Course for Basic Surgical Trainees (BSTs).
(POSTER)
P.J.Driscoll, S. Paterson-Brown, R.J. Glavin, N.J.Maran
'The Ottawa Conference',
Ottawa, Canada, 13th-16th July 2002

Trainees' Opinions of the Skills Required of Basic Surgical Trainees (BSTs).
P.J.Driscoll, A.M.Paisley, S.Paterson-Brown
Association for Surgical Education Spring Meeting,
Baltimore, USA, 4th – 6th April 2002

Trainees' Opinions of the Skills Required of Basic Surgical Trainees (BSTs).
(POSTER PRIZE)
P.J.Driscoll, A.M.Paisley, S.Paterson-Brown
Association of Surgeons in Training Annual General Meeting
Sheffield, 23rd / 24th February 2002

INDEX.

LIST OF TABLES

Section I.

Section II.

Section III.

Section IV.

Section V.

Section VI.

Section VII.

LIST OF FIGURES.

Section I.

Section II.

Section III.

Section IV.

Section V.

Section VI.

Section VII.

LIST OF ABBREVIATIONS.

A&E	Accident & Emergency
A.D.	anno domini
ABSITE	American Board of Surgery In-Training Examination
ACRM	Anaesthesia Crisis Resource Management
ANOVA	Analysis of Variance
ANTS	Anaesthetists Non-Technical Skills
BST	Basic Surgical Trainee
CAA	Civil Aviation Authority
CAR	Consultant with Administrative Responsibility
CARMA	Crisis Avoidance and Resource Management in Anaesthesia
CASE	Comprehensive Anaesthesia Simulation Environment
CCST	Certificate of Completion of Specialist Training
CCT	Certificate of Completion of Training
CRM	Crew Resource Management
CVP	Central Venous Pressure
DBU	Day Bed Unit
EBSTAF	Edinburgh Basic Surgical Trainee Assessment Form
EBSTAF-Tech	Technical Skills portion of EBSTAF applied to Video and Real-Time Assessment

EBSTAF-TS	Technical Skills portion of EBSTAF applied to In-Post Assessment
ECG	Electrocardiograph
ENT	Ear, Nose & Throat Surgery
EWTD	European Working Time Directive
FRCS	Fellowship of the Royal College of Surgeons
FY1	Foundation Year 1
FY2	Foundation Year 2
FYs	Foundation Year Doctors (FY1 & FY2)
GA	Global Assessment
GAS	Gainsville Anaesthesia Simulator
GMC	General Medical Council
GP	General Practice
GT	Generalisability Theory
HDU	High Dependency Unit
HPS	Human Patient Simulation
HPS-GA	Global Assessment applied to High-fidelity Patient Simulation
HST	Higher Surgical Training
ICC	Intra-Class Correlation
ICSAD	Imperial College Surgical Assessment Device
ICU	Intensive Care Unit
IQR	Inter-Quartile Range

ISCP	Intercollegiate Surgical Curriculum Project
JCHST	Joint Committee on Higher Surgical Training
JCPTGP	Joint Committee on Postgraduate Training for General Practice
KW	Kruskal-Wallis
LEP	Longitudinal Evaluation of Performance
MCQs	Multiple Choice Questions
METI	Medical Education Technologies Inc., Sarasota, Florida, United States
mini-CEX	Mini Clinical Evaluation Exercise
MIST-VR	Minimally Invasive Surgical Trainer - Virtual Reality
MMC	Modernising Medical Careers
MRCGP	Membership of the Royal College of General Practitioners
MRCS	Membership of the Royal College of Surgeons
MSF	Multi-Source Feedback
MWU	Mann-Whitney U-test
NASA	North American Space Administration
NHS	National Health Service
NIBP	Non-invasive Blood Pressure
NJM	Dr. Nicola J. Maran, Scottish Clinical Simulation Centre
NOTECHS	Non-Technical Skills Taxonomy as developed in aviation.
NOTSS	Non-Technical Skills in Surgery

OPD	Outpatients' Department
OSATS	Objective Structured Assessment of Technical Skill
OSCE	Objective Structured Clinical Examination
PA	Pulmonary Artyery
PAME	Patient Assessment and Management Examination
PhD	Doctorate of Philosophy
PJD	Peter Driscoll, Principal Researcher
PLAB	Professional Linguistics and Assessment Board
PMETB	Postgraduate Medical Education and Training Board
PRHO	Pre-Registration House Officer
PseudoSim	In-post EBSTAF-derived simulator assessment
RCSEd	Royal College of Surgeons of Edinburgh
RCSEng	Royal College of Surgeons of England
RITA	Registrar In-Training Assessment
RTA	Real Time Assessment
RTA-VAS	Real-Time Assessment Visual Analogue Scale
SAQs	Short Answer Questions
SCSC	Scottish Clinical Simulation Centre
SES	Southest Scotland
SHO	Senior House Officer
SPB	Mr. Simon Paterson Brown
SpR	Specialist Registrar

SPRAT	Sheffield Peer Review Assessment Tool
SPs	Standardised Patients
ST	Specialist Trainee
STA	Specialist Training Authority
Toronto	Global Rating Scale of Operative Performance as developed at the University of Toronto.
UK	United Kingdom
US	United States (of America)
VA	Video Assessment
VAS	Visual Analogue Score
VASComm	Visual Analogue Scale addressing Communication
VASCS&M	Visual Analogue Scale addressing Clinical Skills & Management
Video-VAS	Visual Analogue Scale for Video Assessment
VR	Virtual Reality
Wilcoxon	Wilcoxon signed-rank matched-pairs test
WoS	West of Scotland

Section I.

INTRODUCTION.

I.1. BACKGROUND.

High quality patient care in surgery, like any other medical specialty, depends upon the effective selection of trainees combined with repeated assessment as they pass through training to ensure they attain and maintain the required levels of expertise for consultant practice. The selection, training and assessment of surgeons have received considerable attention over recent years but the issues are not new. Celsus suggested in around 30 A.D that surgeons should be young with a strong and steady hand, ambidextrous, have clear vision and be unmoved by the cries of the patient (Jackson 1998)! Thankfully things have moved on, yet high-profile cases of poorly performing doctors continue to raise public concern and highlight a lack of rigorous self-regulation within the profession as a whole (Davies *et al*, 1999; Bauchner *et al*, 2001; Baker, 2004).

The practice of surgery in the UK continues to evolve due to an ever-increasing knowledge base and public expectations of accountability from surgeons who practice with a comparatively high degree of autonomy when compared to other healthcare systems such as the US (Brearley, 1994). The modern surgeon requires appropriate judgement based upon broad surgical knowledge and experience in combination with an ability to act thoughtfully yet decisively (Wanzel *et al*, 2002b). The surgeon must be compassionate, perceptive, dedicated, a good communicator and display high levels of technical skill (Wanzel *et al*, 2002b), morality and personal reflection (Schon ,

1987; Treasure, 1998). Requirements to document an individual's skills demand equally robust assessment methods. The assumption of skill is no longer acceptable in the absence of proof (Darzi *et al*, 1999).

Concerns regarding surgical training in the UK and the competence of future surgeons (Reed, 1993; Jones, 1993; Collins, 1995; Livingstone *et al*, 1996) highlight the need for good assessment to ensure trainees attain accepted standards. The training of doctors has faced intense scrutiny over recent years. Failures to prepare medical students for their future practice have been highlighted and blamed on competing service demands within a system that is reliant upon unrewarded teaching from full-time NHS staff (Catto, 2000; Jolly, 2001). Subsequent deficiencies are seen to propagate throughout all levels of postgraduate training (Calman *et al*, 1991; Richards, 1992a; Richards, 1992b; Toogood *et al*, 1996; Gillard *et al*, 2000; Lambert *et al*, 2000) with particular concern being raised with regard to the SHO grade (Bulstrode *et al*, 1993) who until recently constituted 47% of doctors in training (Donaldson 2002).

I.2. THE HISTORY OF ASSESSMENT IN SURGERY

I.2.a. Past.

Modern-day surgery has its origins in the Barber Surgeons of the 16th century, a time when surgery was a trade far removed from the profession of medicine. It was taught in a purely clinical setting by apprenticeship, the trainee bound by legal agreement to the trainer (his employer) for the period of time it took to become proficient. The trainee's role was to copy the master (de Cossart & Fish 2005c) who in turn determined when the apprentice was proficient.

Over time, surgery became allied with medicine. It was no longer simply a trade and it was recognised that some form of certification was required to maintain the standards of a profession. The Royal Colleges of Surgeons of Edinburgh followed by that of London introduced the examination of the "Fellowship" (FRCS) in 1854 and 1884 respectively, signifying the attainment of surgical expertise in what was at that time a very general profession.

The emergence of subspecialties within surgery (such as orthopaedics or neurosurgery) during the first half of the 20th century led trainees to sit FRCS increasingly early in their training, being passable with little actual practical experience (Stotter *et al*, 1986), before embarking upon subspecialty

practice. The FRCS quickly became an entry requirement for higher surgical training, the trainee progressing to consultant some years after this formal examination. As a result, the maintenance of standards was once again dependent upon the subjective assessments of the trainer. In response, the specialty-specific Intercollegiate Part III or "exit" examination was introduced by the Royal Colleges of Surgeons of England, Edinburgh, Glasgow and Ireland to signify successful completion of specialist training (MacLaren, 1988). However this still failed to address operative ability.

During this time surgical training was of uncertain direction and duration; fifteen years or more to attain a consultant post was commonplace. 30,000 clinical hours (Donaldson 2002) almost inevitably produced highly-experienced surgeons (de Cossart & Fish 2005a) but Calman highlighted career progression as frequently hindered by personality clashes and independent of actual surgical competence (The Department of Health , 1993). Furthermore, experience was gained under variable levels of supervision with trainees learning by their own mistakes at the expense of patients (Reynolds, 1999).

I.2.b. Present.

At the time of this study, doctors wishing to follow a career in surgery took up SHO posts in regional basic surgical training programmes that gave them exposure to a number of surgical specialties. Standards of posts were set

and monitored by the Surgical Royal Colleges and Postgraduate Deans who were able to withdraw educational approval (and part or all of the funding) from those posts that fell below an acceptable standard. During this period of time trainees sat examination for Membership of their chosen Royal College of Surgeons (MRCS), roughly equivalent to the old FRCS. This two-part examination, which remains in place, examines theoretical knowledge, its application and rudimentary clinical examination.

Having chosen a specialty, trainees entered Higher Surgical Training (HST) as Specialist Registrar (SpR) for a minimum of 5 years, often with an additional 1-3 year period of research. The SpR grade remains in place but will be progressively replaced by Specialist Trainee (ST) posts as a result of Modernising Medical Careers (MMC) (Department of Health 2003).Towards the end of the SpR grade, trainees sit the final Part III Intercollegiate Fellowship Examination (FRCS) in their chosen specialty. Once confirmed that the trainee has satisfactorily completed specialist training by their respective Royal College, a Certificate of Completion of Specialist Training (CCST) is issued by the General Medical Council (GMC) and the trainee becomes eligible to join the Specialist Register and apply for consultant posts within the NHS (Department of Health , 1998).

Current surgical training employs two principles; increasing exposure as principal surgeon leading to technical competence (Cuschieri *et al*, 1997) with graded supervision protecting patients from harm (Isbister, 2002).

However, supervision remains variable and training opportunities are frequently not realised (Hurley *et al*, 1999). Surgery in particular now faces changes as a result of MMC (Department of Health 2003) that challenge the continued training of competent surgeons.

It has been estimated that the reorganisation of postgraduate medical training as a result of The New Deal for Junior Doctors Hours (NHS Management Executive , 1991), The Calman Report (The Department of Health , 1993), and the European Working Time Directive (European Council, 1993) resulted in surgical training being reduced from 30,000 to as little as 6,000 clinical hours (MacIntyre, 1996; Beecham, 1996; Donaldson 2002). Trainees are now exposed to 60% less surgical cases (Bulstrode *et al*, 1996), leading trainers (Dawson, 1998; Reynolds, 1999) and trainees (Bourne *et al*, 1999) to question whether they will be sufficiently prepared for independent Consultant practice. The almost universal change towards shift-working in order to meet the limited hours legislation (NHS Management Executive , 1991; European Council, 1993), along with cross-cover between specialties, has further fragmented patient care, particularly at the junior level (Reynolds, 1999; Hilton *et al*, 2002) Training has inevitably also suffered as it is intimately related to both exposure and continuity of care. As a result, training must now become qualitative and not simply quantitative (Hurley *et al*, 1999). While courses such as "Training the Trainers" (Bulstrode *et al*, 1996) address a trainer's teaching skills and are to be commended, they may fail to alter practice in the face of increasing clinical, managerial and political pressures.

The value of training to service must be acknowledged and rewarded if training is to be optimised.

SHOs made up 47% of doctors in training in 2002 and became the focus of particular concern. Donaldson recognised extensively documented criticisms regarding the educational experience, assessment and subsequent competence of SHOs whose role remained unclear (Donaldson 2002) and whose future was uncertain (Bulstrode *et al*, 1993), frequently embarking upon postgraduate research simply to compete for entry into the SpR grade. While changes to PRHO (General Medical Council 1997) and SpR (The Department of Health , 1993) grades were successfully implemented, the SHO grade remained in need of radical reform (Donaldson 2002). Interestingly, it is worthy of note that the same criticisms were made of HST by Calman (The Department of Health , 1993) prior to its reform.

In response, MMC (Department of Health 2003) aimed to streamline the training of doctors and remains in evolution. Initial postgraduate training is broad-based over a *foundation* period of two years to give trainees a broad base of clinical experience. Specialist Training (ST) programmes will then provide fixed term Specialist Training planned to be time-capped single-grade "run-through" training to Consultant employing competence-based in-post assessment. Details of the training pathway continue to evolve and were the subject of a recent enquiry (Tooke *et al* 2007).

The regulation of the medical profession has also undergone reform. The Postgraduate Medical Education and Training Board (PMETB) is the amalgamation of the Specialist Training Authority (STA) and the Joint Committee on Postgraduate Training for General Practice (JCPTGP), formerly responsible for training in hospital and primary care settings respectively. It was established by Government to address the inconsistency of postgraduate medical education, training and assessment throughout the UK. The remit of PMETB differs little from its predecessors: the approval of curricula and the setting and maintenance of standards of basic and higher specialist training; and the regulation of entry into the Specialist and General Practitioner Registers (Griffiths, 2006). It does not cover undergraduate or foundation education, both of which remain the responsibility of the General Medical Council (GMC). PMETB differs from the STA and JCPTGP in that it reports to and remains under the overall control of office of The Secretary of State for Health whilst acting independent of Government (Postgraduate Medical Education and Training Board. 2004a). PMETB "went live" in September 2005 but its impact in an already highly dynamic system remains unclear.

Revalidation procedures introduced in parallel by the GMC aim to ensure continued "Fitness to Practice" (General Medical Council , 2003) with practitioners required to demonstrate their continued adherence to the principles set out within *Good Medical Practice* (General Medical Council , 2001). Annual appraisals for consultants and the RITA process for SpRs fulfil

these requirements but again there remained no standardised formal assessment or appraisal process for SHOs. With the introduction of Foundation, assessment processes were put in place but their use remains variable. Revalidation of other grades further fuels demands for objective measures of the competence of SHOs / FYs in clinical, technical and non-technical skills. Despite these recent changes in career structure, the assessment and selection of trainee doctors (surgeons) remains one of the biggest challenges faced by modern medicine in the UK.

I.3. ASSESSMENT IN GENERAL.

Jolly defines assessment as "the measurement of an individual's (or group's) performance against external criteria" (Jolly *et al* , 1997). Although assessment is often regarded as specifically related to examinations, it encompasses any educational process that records the development or progress of the learner and/or provides evidence of progression within an educational programme or career (de Cossart & Fish 2005a). In surgery, assessment most commonly involves the measurement of a trainee's knowledge or skills in relation to the surgical curriculum as prescribed by the Royal Colleges. The products of assessment may then be used to guide a trainee's progress within the appraisal process wherein trainer and trainee agree future goals by informal, non-threatening and confidential discussion, aimed at encouragement and support. Alternatively, it may form the basis for 'high-stakes' assessment, taking on a disciplinary or gate-keeping role with potentially major consequences for the individual or institution.

I.3.a. Reasons for assessment.

Robust and timely assessment is pleuripotent: it monitors and documents skills acquisition, identifying strengths and weaknesses and so facilitating feedback; it allows the ranking of trainees for selection processes, motivating both trainee and trainer; and permits comparisons between trainers and

training programmes, thereby helping to maintain standards (Jolly *et al* , 1997).

It has been repeatedly recognised that assessment drives learning (Newble *et al*, 1983; Malik *et al*, 1988; Wakeford *et al*, 1992; Fowell *et al*, 1999; Hamdorf *et al*, 2001; Wass *et al*, 2001; Norcini, 2002). If success is defined by examination, trainees will understandably concentrate on the content of the examination. However, assessment must be mapped to actual practice or else it may misdirect the learner (Hamdorf *et al*, 2001). Thus learning also drives assessment (Handfield-Jones *et al*, 2002). It is therefore surprising that the assessment of surgeons has addressed little more than cognitive knowledge until recently.

Failures of professional self-regulation, such as the events at Bristol Royal Infirmary (Kennedy 2001), have reduced public trust and raised concerns of more systematic failings (Davies *et al*, 1999). As a result, the medical profession faces public and political demands for more formal objective assessment of both acquisition and maintenance of competence throughout a doctor's (surgeon's) career (Wass *et al*, 2001).

I.3.b. Assessment categories.

Assessment may be described as "formative" or "summative", terms that describe how the product of assessment is applied rather than the assessment itself.

I.3.b.i. Formative assessment.

Formative assessment focuses on the individual and is designed to give constructive feedback of both strengths and weaknesses (Wanzel *et al*, 2002b), enhancing learning and promoting reflective practice. It bears no relation to the performance of the peer group and is therefore not dissimilar to appraisal, providing the discussion remains low stakes i.e. without penalty. It therefore guides learning and has more recently been coined as "assessment FOR learning".

I.3.b.ii. Summative assessment.

Summative assessment aims to accumulate all relevant information to make a decision as to whether or not a pre-determined standard has been achieved (Jolly *et al* , 1997; Wanzel *et al*, 2002b). It is commonly used at the end of a defined period of training for high-stakes purposes, such as trainee selection or rejection, career progression or licensing. These 'barrier' assessments address minimum acceptable standards (Hamdorf *et al*, 2001;

Bulstrode *et al* , 2001) in relation to peers or criteria (Ritchie, 2001; Wanzel *et al*, 2002b) but are frequently delivered too late to benefit the development of the individual trainee (Pietroni, 1993b). Summative assessment is therefore a high-stakes assessment OF learning. At present, MRCS, FRCS and the RITA process form the main summative assessments in surgery.

Formative and summative assessments should ideally utilise the same instruments to allow the trainee to become familiar with, and therefore unaffected by, the instrument itself. The only difference between the two should be intent (Rolfe *et al*, 1995). However, the design of tools suitable for both summative and formative assessment has proved difficult (Wass *et al*, 2001).

I.3.c. Timing of assessments.

Assessments are often intrusive to everyday clinical practice, time-consuming and expensive, both financially and in terms of trainer hours. They should therefore be performed often enough to assess and guide skill acquisition in real time while at the same time not impairing the acquisition process itself.

I.3.d. "Good" assessment.

The consequences of failing high-stakes summative assessment may be far-reaching for trainee, trainer and the training scheme. It is therefore imperative to demonstrate that an assessment satisfies accepted criteria of "good" assessment before widespread use (Jolly *et al* , 1997).

The focus and purpose of an assessment must drive its design (Crossley *et al*, 2002b). Previous judgements of competence as evidenced by written examination exemplifies the assessment of what is *easily* assessed rather than what needs to be assessed (i.e. *fit for purpose*). Three frameworks are frequently applied to categorise professional activity and thus aid in defining the focus of an assessment. Miller is probably the most widely quoted and describes 2 cognitive and 2 behavioural levels of competence (Miller, 1990). The trainee first acquires the knowledge ("knows") and then learns how to apply it ("knows how"). Demonstration, at the first behavioural level ("shows how"), is then followed by, but does not predict, day-to-day performance ("does") (Rethans *et al*, 1991). Bloom's taxonomy of "knowledge", "skills" and "attitudes" (Bloom *et al* , 1956) is then frequently used to subdivide Miller's cognitive domain and highlights the fact that a test of cognitive knowledge does not examine skills or attitudes. Finally, Donabedian defines the level of assessment in terms of "structure" (the programme or curriculum), "process" (formative) or "outcome" (summative) (Donabedian , 1980).

Once the assessment focus has been determined, the critical elements of professional activity at the relevant stage of development (the curriculum) must then be systematically defined to produce a "blueprint" (Newble *et al*, 1994). These are most frequently gleaned from expert opinion, as used for the development of the GMC's Good Medical Practice (General Medical Council , 1998a). The curriculum should then be published (Pietroni, 1993a) and it should be noted that a defined surgical curriculum has not been published until recently (Intercollegiate Surgical Curriculum Project ,2005; Intercollegiate Surgical Curriculum Project ,2006).

So in summary, a "good" assessment needs to be designed specifically for the purpose for which it is intended based upon clearly set objectives. However, the application of an assessment in the real world inevitably involves compromise between rigour and practicality. Examination of everything that a doctor ever does would give a full picture of practice, and therefore the level of competence, but is not practical. The assessment must therefore be *feasible* (i.e. deliverable within reasonable financial and time constraints), frequently resulting in processes that differ significantly from the ideal (Bligh, 2001; Manogue *et al*, 2001).

Since assessment drives learning, an assessment must be *acceptable* to all those involved. Using an example from surgery, a junior trainee performing an assessment procedure without guidance may offer an accurate picture of that trainee's skill but the inferior end result would clearly be unacceptable to

the patient. Similarly, surgical training should be thought of as a form of contract, a concept that may on occasion be reinforced by the collective signing of just such a document, in which trainers agree to impart the knowledge and skills to allow the trainee to progress while trainees agree to follow the trainer's guidance and value to training they are given. Without this acceptance of training on the part of the trainee it is likely to be ignored and come to nothing, wasting considerable effort on both sides. The acceptability of assessments to trainees is similarly essential, yet it has been overlooked by the medical literature, despite its acceptance in industry (McEvoy *et al*, 1987; Fedor *et al*, 1989; Yuki *et al*, 1995; Wimer *et al*, 1998). Only in anaesthetics in Denmark have trainees been approached to comment upon newly imposed assessment processes, although their responses were limited to general impressions (Ringsted *et al*, 2003).

A fundamental principle of any scientific method, including assessment, is that experiments must be reproducible to allow meaningful interpretation (Downing *et al*, 2004). This *reliability* is mandatory but not in itself sufficient for good assessment [see Section I.8.a.].

Good assessment data must also be shown to reflect the construct it is designed to measure. The degree to which this is achieved is termed the *validity*. It is not an absolute term (i.e. assessments are not valid or invalid); rather, assessments have a degree of validity evidence to support the proposed interpretation (Downing *et al*, 2003). Assessment data must, at the

very least, display good *face* validity, by appearing to measure what is intended at first impression (Thomas *et al*, 1992), and *content* validity by including relevant performance criteria (Crossley *et al*, 2002b). As a result of the numerous methods that have been described to address the validity of assessment data, further sub-categories are now recognised [see Section 1.8.b.].

Just like any other scientific instrument, the validity and reliability of an assessment of competence or skill must be empirically demonstrated within pilot studies prior to its application to high-stakes decision-making (Jolly *et al* , 1997)

I.3.e. Application of assessment.

Once an assessment has been rigorously developed and proven, it is still not fail-safe in its application. A number of further issues must still be considered:

Reliable assessment requires broad sampling (Case *et al*, 1988; van der Vleuten *et al*, 1991; Swanson *et al*, 1995) since performance in one area of practice does not predict that in another (Kassirer *et al*, 1978; Swanson *et al*, 1995) – it is *content specific*. This is recognised as the most critical aspect of the assessment of clinical competence (Wass *et al*, 2001)

Assessment is also *time specific*; performance at one point in time does not predict performance in the future unless it is actively maintained by regular reinforcement (Guest *et al*, 2001). A surgeon competent in procedures in which he/she has been trained may not be competent in newer techniques and the unregulated uptake of laparoscopic cholecystectomy in association with the subsequent rise in bile duct injuries remains a case in point (Rogers *et al*, 2001).

The choice of assessor is also vital to reliable assessment - it is *assessor specific*. Different assessors may perceive the candidate differently and thus be more lenient (the so-called 'dove') or harsh (the so-called 'hawk') and this inter-rater reliability is a major source of error in assessment, reduced by the use of multiple assessors (Downing *et al*, 2004).

Other seemingly unrelated factors also interfere with objective assessment. Trainee social skills (Kassebaum *et al*, 1999), a trainer's knowledge of their performance previously (Rolfe *et al*, 1995) or the trainee's level of seniority (Winckel *et al*, 1994) may affect assessment ratings. This *cognitive bias* is termed the "Halo Effect" and may act favourably or unfavourably for the subject of the assessment (Thorndike, 1920). Responses may also reflect what the assessor feels is expected of them rather than their honest opinion of the trainee since they may perceive their assessment to reflect as much upon themselves as on the trainee. This *response bias* can be reduced by anonymity but this is unworkable in the assessment of individuals when their

identity is a prerequisite (Tweed *et al*, 2001a). Further, the very fact that a trainee is being assessed may result in modified performance that differs from everyday practice; this is termed the "Hawthorne Effect" (Roethlisberger *et al* , 1939).

I.3.f. Legal Considerations.

Assessment now spans the whole of medical practice from undergraduate education, through postgraduate training to revalidation of the consultant grade. UK law allows an Examination Board's assessment processes to be legally challenged (Tweed *et al*, 2001b) and this has been described both in nursing and medicine. While good assessment serves to protect both trainer(s) and trainee, a successful challenge of a less than robust assessment incurs costs and loss of standing to the Examining Board. A deficient trainee may be identified and redirected as appropriate without fear of legal challenge only if there has been robust assessment documenting their deficiencies. Conversely, a trainee treated unfairly may use former assessments to prove their case and gain re-instatement or recompense. In order to resist such legal challenges, an assessment process must be fully evaluated before its application to high-stakes assessment (Fowell *et al*, 1999). Widespread use of an assessment tool is not in itself a legal defence (Tweed *et al*, 2001b).

I.3.g. Standards.

Acceptable and deficient levels of performance within an assessment must be determined. These pass/fail standards should be clearly defined and published, then monitored and enforced (Pietroni, 1993a). In the UK, surgical standards are currently set by The Surgical Royal Colleges and the GMC. The standard of an assessment must be influenced by the purpose for which it is intended.

Summative assessment standards may take two general forms. To ensure a minimum level of competence, such as that required for registration, a simple pass/fail decision based upon fixed criteria may be all that is required. This *criterion referenced* standard is independent of the performance of the trainee cohort and forms the basis for the assessment of competence, with the potential for all of a particularly good cohort to pass (Bulstrode *et al* , 2001). In contrast, competitive selection requires a degree of ranking above the minimum standard. Each individual's result sets the standard for others (Crossley *et al*, 2002b) and success or failure is determined by position within the cohort. This *norm referenced* assessment is typical of selection procedures which allow a set number of individuals pass through while the empirical pass mark may vary (Bulstrode *et al* , 2001). It thus identifies excellence within the cohort rather than the minimum standard.

Formative assessment may be used to generate information on an individual's performance for subsequent feedback aimed at encouraging and guiding progress. It differs from summative assessment in that trainees benefit most from a broad profile of their strengths and weaknesses across each area of practice rather than a single overall score.

I.3.h Appraisal.

Appraisal encompasses all areas of a doctor's (or surgeon's) activity with the express purpose of supporting their learning and personal development thereby maintaining their professional performance (General Medical Council , 1998b). In contrast to assessment, it does not itself generate new evidence or information on performance but instead captures information that already exists to facilitate the identification of areas in need of attention (NHS Management Executive ,2001).

I.4. COMPETENCE IN SURGERY

I.4.a. Competent – a definition.

The cardinal ethic of medicine is *competence*. *Competent* is defined by The Oxford Dictionary as "properly qualified for a task" (Oxford University Press , 2006) but a better definition may be "the possession and application of the requisite knowledge, technical skill and humanism" (Jonsen , 1990). It is an absolute term; an individual is either competent or not competent and cannot be more or less competent than another. The determination of competence is therefore criterion referenced and holistic. It recognises that professionals engage in intelligent and wise conduct rather than protocol-driven behaviour, applying professional judgement and creative thinking (de Cossart & Fish 2005a).

I.4.b. Competencies.

Surgical competence requires cognitive knowledge, skills, attitudes (Bloom *et al* , 1956) and behaviours (Miller, 1990). These *competencies* may be described relatively, for example with reference to the stage of training. Thus an individual may be a competent SpR but not (yet) a competent Consultant. This competency-based approach to training (and assessment) deconstructs competence into a collection of skills to infer that an individual who has all the competencies is competent (de Cossart & Fish 2005a). However, it has been

criticised since behaviour is also driven by personal values. Espoused values (those that we put in to words) rarely equate to values-in-use (as demonstrated by our actions) (de Cossart & Fish 2005b) and as a result an individual may have a particular skill but use it badly, if at all, in their everyday practice (Wood *et al*, 2000; Bligh, 2001). Performance in individual competencies does not predict competence (Rethans *et al*, 1990) equating to the difference between Miller's behavioural levels of demonstration and everyday practice.

I.5. CURRENT ASSESSMENT STRATEGIES.

Conventional examinations allow a one-time and usually summative assessment of a candidate's abilities. No single assessment addresses all that is required to be competent (Wilkinson *et al*, 2002) and numerous strategies have been described, each with their own strengths and weaknesses when examined in terms of assessment criteria. Assessment of individuals at the higher levels of Miller's competence pyramid ("shows how" and "does" (Miller, 1990)) is more applicable to actual practice but is more difficult to achieve and requires increasing clinical authenticity (Swanson *et al*, 1995). This performance-based assessment becomes more important with increasing experience (Postgraduate Medical Education and Training Board. 2004b)

I.5.a. Assessment of cognitive skills.

The assessment of cognitive skills is commonly carried out by written examination. The choice of the most appropriate type, however, may prove difficult (Schuwirth *et al*, 2003).

MCQs examine a broad knowledge base ("knows") in the absence of inter-examiner variation (Bulstrode *et al* , 2001) but it has been suggested that the use of negative marking may discriminate on the basis of risk-taking behaviour and should be abandoned (Fowell *et al*, 2000).

SAQs and Essays cover a far narrower knowledge base but may be constructed to examine a candidate's understanding or reasoning rather than just simple knowledge ("knows how"). However scoring is highly subjective and unreliable even when combined with highly-structured marking schemes, although they remain useful in norm-referenced competitions where the candidate has the chance to excel (Bulstrode *et al* , 2001)

Oral examinations (the *viva voce*) allow assessment of a candidate's reasoning or deep understanding on a subject. They are highly valid but of low reliability and may be disproportionately affected by good or bad communication skills for whatever reason (Jolly *et al* , 1997).

Assessment of academic achievement (presentations, publications or postgraduate research by dissertation) shows a narrow-based validity (Jolly *et al* , 1997) that is unrelated to clinical skills, professionalism or competence. It is open to fraud or plagiarism and may actually reflect the input of supervisors rather than the abilities of the trainee. This may be countered to an extent by submission in advance, allowing examiners to formulate appropriate questions to confirm the material to be the candidate's own work.

I.5.b. Assessment of behavioural skills.

An individual that is aware of being assessed may, consciously or sub-consciously, modify their performance - the Hawthorne Effect (Roethlisberger *et al* , 1939)). As a result, assessment with the knowledge of the trainee will likely only address level III of the Miller pyramid (i.e. shows how) (Miller, 1990). The challenge in the assessment of competence is to assess (or at least predict) day-to-day performance.

I.5.b.i. Assessment of clinical skills.

I.5.b.i. (a) Assessment of clinical skills in the absence of observation.

(i) Assessment by Examination.

The traditional real-patient "long case" assesses the presentation of an unobserved structured interview and related knowledge and interpretation (Hardy *et al*, 1998). It has high face validity since it comes very close to a candidate's actual practice (Thomas, 1992; Wass *et al*, 2004), but is highly unreliable due to case (Norcini, 2002) and examiner specificity (Newble *et al*, 1980). The use of standardised patients gives some gains in reliability (Wass *et al*, 2004) but these are costly and still do not address the issue of case-specificity, overcome only by observing multiple cases (Norcini, 2001;

Norman, 2002; Hamdy *et al*, 2003). Long cases remain useful however for formative feedback.

(ii) Assessment in the Workplace.

a. Assessment by Case Note Review.

The objective review of case notes has the potential to examine a trainee's history-taking, examination, investigation and management skills and would clearly relate to everyday practice. However, it is a skilled and time-consuming task to extract the relevant information from notes containing entries from multiple individuals and it may be further confounded by difficulties in distinguishing poor patient management skills from simply poor record keeping.

b. Continuous Assessment.

The documentation of mistakes and inappropriate behaviours in addition to the assessment of desirable attributes provides continuous assessment and promotes a trainee's reflective practice. Data collected, either by the trainees or other members of the multidisciplinary team, may prove valuable within feedback processes but trainees are unlikely to willingly reveal their weaknesses within more formal assessment (Newble, 1983).

c. Portfolios.

A portfolio is simply a collection of material brought together for a defined purpose (Wilkinson *et al*, 2002). Within undergraduate medicine they are well established, offering a useful focus for formative assessment, documenting an individual's learning and giving a more complete picture of a person's ability (Wilkinson *et al*, 2002). They are a component of the GMC's revalidation process (General Medical Council , 1998b; General Medical Council , 2000a; NHS Management Executive ,2001) where they aim to illustrate a doctor's continuing professional development based on the logic that a doctor who fails to adapt their practice in response to ongoing developments will, at some point, become incompetent (Wilkinson *et al*, 2002). Portfolios allow assessors to review material that might not normally be available, such as patients' letters or reflections upon critical incidents, but reliability is directly affected by the candidates themselves who select what is to be included. Portfolios therefore run the risk of being overly positive by the omission of unfavourable material (Bulstrode *et al* , 2001). In addition, they do not involve direct observation of the trainee and are therefore of questionable validity in the determination of the quality of practice. However their benefits to learning and quality improvement for the profession should not be sacrificed in the search for reliability as they do link assessment and performance (Handfield-Jones *et al*, 2002) and, when applied effectively, can aid the restoration of competence by alerting all concerned to areas of weakness (Wilkinson *et al*, 2002).

d. Assessment on the basis of Clinical Outcomes.

Assessment of individual surgeons on the basis of clinical outcomes is a highly controversial issue, principally because no surgeon works in isolation. Instead, they form part of a multidisciplinary team and are therefore unable to claim sole credit for a good result. Neither should an individual take all the blame for a poor outcome which is likely to involve a chain of events involving multiple clinicians (Gawande *et al*, 2003), not to mention the patients themselves. However, directly resulting from issues surrounding complex paediatric cardiac surgery at Bristol Royal Infirmary and the subsequent public inquiry (Kennedy 2001), cardiac surgeons were the first to publish 30-day mortality rates in 2004, with many other specialties doing so subsequently. Although broadly welcomed, the profession was keen to highlight the hazards in comparing like with like when differences in case-mix and operative risk threaten to mislead the public and unfairly criticise pioneering surgical teams. Furthermore, the publication of results comes as a consequence of surgeons' own practices of auditing their results in an effort to further the surgical craft, while non-surgical specialties remain contentiously safe from scrutiny.

In contrast, trainees work under the auspices of their trainer(s) with only limited control over their working practices. Assessment of trainees on the basis of clinical outcome is therefore even less valid.

e. Record of In-Training Assessment (RITA).

UK SpRs are currently assessed using a written record of their progress through the grade, the RITA. A specialty-based RITA committee reviews each trainee annually, in combination with the trainee's logbook and curriculum vitae. Assessment forms, completed by the trainer(s), grade the trainee's everyday performance in the work place as *unsatisfactory requiring repeat training*, *unsatisfactory requiring targeted training* or *satisfactory* and trainers have the option to make additional comments. The form examines 27 competencies within the broad categories of clinical skills (7), knowledge (2), postgraduate activities (7) and attitudes (11).

The RITA process is not in itself a means of assessment. It is designed to document and support a trainee's progress through the SpR grade (Department of Health 1998b) while linking the Postgraduate Dean and the Surgical Royal Colleges to the training programmes, assessing out-of-programme activities and, once FRCS has been obtained, providing final certification of successful completion of higher surgical training (Department of Health 1998a). It has clear guidelines for the support of trainees progressing more slowly than might be expected combined with clearly defined appeals procedures. However, the lack of validity of the in-service assessment forms has been demonstrated, suggesting that the process could be improved (Paisley *et al*, 2001a).

Although local strategies may have emulated the RITA process, no such nationwide assessment process was put in place for SHOs. More widespread assessment processes for FYs remain variable in their application.

I.5.b.i. (b) Assessment of clinical skills by direct observation.

This approach relies on the direct observation of a trainee's behaviour in the area of practice being assessed. A skill that is not observed cannot be reliably assessed.

(i) Assessment by Examination.

a. Objective Structured Clinical Examination (OSCE).

The OSCE (Harden *et al*, 1979) allows reliable and valid examination of a broad range of clinical skills by the use of multiple (8 to 20) stations addressing history-taking, examination, communication, deductive reasoning and even cardio-pulmonary resuscitation (Bulstrode *et al* , 2001). However, OSCEs require expert design to ensure curriculum objectives are addressed and, because candidates are assessed by a different examiner at each station, they are labour-intensive and expensive to run (Cusimano *et al*, 1994). Further, the limited time at each case has raised concerns as to validity in the assessment of deeper understanding and excellence,

questioning their application to higher levels of competence (Hodges *et al*, 1999; Ben David, 2003; Hodges, 2003).

b. Patient Assessment and Management Examination (PAME).

This is a surgical adaptation of an OSCE which mimics general surgical referrals by examining history, examination, investigation, discussion with the patient and operative knowledge (MacRae *et al*, 1997). It has the potential to probe the candidate's deeper understanding of the case as well as observing the clinical interaction but remains labour-intensive and case-specific.

c. Standardised Patients (SPs).

SPs offer the valid assessment of fundamental clinical skills: taking a relevant history, examining the patient appropriately, communicating with the patient, documenting findings, differential diagnoses and planned investigations. They are reliant upon careful patient recruitment and training, detailed checklists or ratings scales, and require multiple cases to counter content specificity (van der Vleuten *et al*, 1990). Widely used in undergraduate education, they have also been used as a component of The Medical Council of Canada licensing examination (Reznick *et al*, 1993; Reznick *et al*, 1996) and the GMC's Professional Linguistics and Assessment Board (PLAB) examination (Tombleson *et al*, 2000).

(ii) Assessment in the Workplace.

a. Mini-Clinical Evaluation Exercise (Mini-CEX) / Longitudinal Evaluation of Performance (LEP).

The mini-CEX (Norcini *et al*, 1995) and the LEP (Prescott *et al*, 2002) involve limited observation of history and examination of real patients in the medical or dental work-place respectively. Encounters are brief, allowing the resident to be evaluated on several occasions by different assessors and they have been demonstrated to achieve acceptable reliability (Durning *et al*, 2002) and validity (Grossman *et al*, 1992; Norcini *et al*, 1995).

b. Multi-Source Feedback (MSF).

MSF is also referred to as '360 degree Assessment ', a term which emphasises the use of assessments not just from colleagues but also from juniors, seniors, nurses and, in some cases, patients.

First applied in industry, MSF assesses professional behaviours in the workplace, having been shown to be practical, reliable and reasonably valid. MSF will, however, never be totally safe from challenge since each assessment is, by its very nature, subjective. Like any other assessment tool, MSF needs to be applied correctly for the resulting assessment to mean anything. It is therefore important to learn from other industries that have successfully applied these techniques. In a recent review of the literature, Wood quotes McCarthy & Garavan who detail six applications of MSF from

industry: identification of strengths and weaknesses of both individual and organisation; enhancement of culture change; summative assessment of performance; evaluation of an individual's potential for selective purposes; enhancement of teamwork by allowing members to comment; and identification of training needs for the benefit of the system (McCarthy *et al*, 2001). The healthcare literature almost exclusively aims to identify sub-standard individuals, but such assessments must be handled with care since those identified as below standard may suffer disadvantage. MSF assessment must therefore be just as robust as any other strategy if it is to be used in high stakes. However, if trainees identified as sub-standard are to receive skilled feedback and/or directed training then a less 'safe' assessment may be acceptable with the potential to both safeguard patients and rescue poorer performers (Wood *et al*, 2006).

Just as assessment strategies previously described serve to drive learning, MSF makes accepted behaviours explicit and drives professional behaviour. The descriptors therefore need to cover the full range of desired behaviours since they will, effectively, act as the curriculum. The development of MSF systems is therefore time-consuming and highly skilled, calling upon expertise in psychology and behavioural sciences rather than a knowledge of surgery itself. Many MSF systems have been described prior to and since the GMC's Good Medical Practice document (General Medical Council , 2001). However, as far back as 1975, Linn raised concerns as to what MSF was actually assessing. By factor analysis of their own 16-item four-point scale

they showed that 40% of the total variance was due to what they termed an 'interpersonal or relationship factor'. They also noted that a further third of their variance resulted from a second 'knowledge or skill factor' that they were keen to point out may be better assessed by other methods (Linn *et al*, 1975). Many have found similar findings, demonstrating the overwhelming 'halo effect' of knowledge and interpersonal ability on the final outcome of MSF assessments (Davidge *et al*, 1980; Dielman *et al*, 1980; Maxim *et al*, 1987; Risucci *et al*, 1989; Ramsey *et al*, 1993). The implication is that if an assessee is knowledgeable and a good 'people person' their failings may not be revealed by MSF assessment.

However, MSF may offer unique potential. Unlike other strategies, MSF may prove able to assess the fourth level of Miller's pyramid, that of everyday performance i.e. does (Miller, 1990). A trainee who is knowingly being assessed may exhibit the Hawthorne Effect (Roethlisberger *et al* , 1939), adapting their practice for the purposes of the assessment. In contrast, one who is assessed unknowingly cannot modify their behaviours, thus allowing the observation and assessment of everyday practice. However, assessing an individual without their knowledge raises issues of both ethics and employment law, challenging the assessment of what requires assessing the most. The field of anthropology may offer a solution in the form of *ethnography*, the study of a population or culture. Most frequently described in aboriginal tribes, the observer becomes integrated into the group whilst repeatedly making observations until eventually he/she is ignored by those

being assessed. Everyday behaviour may now be observed without either demonstration or the influence of the observer (Atkinson *et al*, 2005). MSF assessments must therefore be repeated and ongoing so that assessees learn to ignore them, cease demonstration and revert to everyday practice that may now be observed and assessed. It is clearly impractical to shadow every surgeon in everything they do (the direct equivalent of ethnography). However, by obtaining assessments from every area of clinical practice, in the full knowledge of those being assessed, it is possible to obtain a fuller picture of their day-to-day practice than would otherwise be possible (Anwar *et al*, 1981). The technique has been applied to the assessment of trainees by The Royal Australasian College of Physicians (Paget *et al*, 1996), The Royal College of Obstetricians and Gynaecologists in the UK and the Canadian Medical Licensing Authority (Hall *et al*, 1999) but the psychometric properties of these individual assessments remain undescribed. The Sheffield peer review assessment tool (SPRAT) was developed as a voluntary appraisal tool for paediatric consultants, consisting of 24 field-tested fields mapped to Good Medical Practice (General Medical Council , 2001) and found to be reliable (Archer and Davies 2003). It was subsequently applied to paediatric middle grades (92 SHOs and 20 SpRs) and found to offer valid assessment of progress (Archer *et al*, 2005). Further development away from the specialty of paediatrics resulted in an abridged and more generic version of SPRAT, termed the mini Peer Assessment Tool or mini-PAT (Archer *et al*, 2008) which is further examined in Section X.

The Edinburgh Basic Surgical Trainee Assessment Form (EBSTAF) addresses the performance of SHOs in the workplace using multiple multi-disciplinary assessors (Baldwin *et al*, 1999). It was developed using a modified Delphi technique consulting 111 consultant surgeons in the Southeast of Scotland with each main surgical subspecialty represented. Stage I requested the anonymous identification of skills and qualities required for a successful training in surgery and the likely technical skills of a BST at the end of 6 months in their own unit. 68 (61%) responded and identified qualities falling into five domains: *communication* with patients and relatives; application of *knowledge*; *team-working* skills; *clinical skills* and *technical skills*. Stage II listed the identified attributes and returned them to all 111 consultants who were now required to rank each on the basis of *essential* (4), *important* (3), *useful* (2) or *irrelevant* (1). Responses were received from 78 consultants (70%) across all specialties and the resultant form addressed a total of 70 skills. All but the 19 qualities addressed by the technical skills domain were generic attributes likely to be of value across multiple medical settings. Surgeons were therefore shown to value well-rounded doctors as well as accomplished technicians. Similar studies have yielded similar results (Albo *et al*, 1976; Clark *et al*, 1993; Martella, 1995), supported by the fact that the most common reason cited for disciplinary action against a doctor is that of unprofessional behaviour (Anwar *et al*, 1981; Bergen *et al*, 2000), rather than a lack of clinical (or technical) skill.

Assessment of SHOs' in-post performance by multi-disciplinary assessors (including medical, nursing and secretarial staff) was then evaluated amongst

all BSTs in the Southeast of Scotland basic surgical training programme over an 18-month period. EBSTAF was shown to be feasible, reliable, internally consistent and construct valid following detailed analysis (Paisley *et al*, 2001a). EBSTAF differed from previously validated MSF, such as SPRAT, in that it examined overall impressions of individual trainees' performance over a prescribed period (subsequently termed 'long-loop' feedback) rather than within individual encounters or procedures (referred to as 'short-loop' assessment).

The use of non-medical assessors was first described by Butterfield & Pearson who gave nurses the task of assessing the 'humanistic behaviours' of doctors. In doing so they illustrated distinct differences in opinion as to what qualities were desirable, leading them to question whether assessment of doctors by nurses was appropriate (Butterfield *et al*, 1990). Similar issues have since been raised in the assessment of doctors by patients, whose assessments may be unduly influenced by issues such as the timeliness with which they see the doctor or whether they leave the consultation satisfied with the outcome. Such concerns may be addressed by increasing the number of assessors and drawing from multiple disciplines, thus reducing the subjectivity and bias integral to single source assessment. However, this is logistically challenging and increases costs so that it is equally important not to use more assessors than are required. The evidence would suggest that the optimum number depends on the assessors themselves; whereas between 5 and 10 peers / colleagues may be required to get a representative result, this may need to be increased to 10 to 20 nurses and over 50 patients

(Butterfield *et al*, 1991; Ramsey *et al*, 1993; Wenrich *et al*, 1993; Woolliscroft *et al*, 1994; Ramsey *et al*, 1996). However, the techniques used vary widely in both the training of the assessors and the focus of the assessment instruments, making them difficult to objectively compare.

EBSTAF also illustrated the importance of rater identity in multidisciplinary assessment. Nursing staff tended to be more lenient than medical staff, awarding significantly higher grades in all but the HDU setting. SpRs appeared consistently more demanding than consultants (although this may reflect increased exposure to areas of weakness rather than higher standards) and SHOs awarded themselves the lowest ratings in the fields of knowledge and teamwork (Paisley 2002). Previous studies have failed to demonstrate such differences between nursing and medical raters (Crosbie *et al*, 1961; Risucci *et al*, 1989; Ramsey *et al*, 1993). Trainees, however, have repeatedly been shown to demonstrate poor self-assessment skills, although self-assessments generally tend to be higher than those of MSF (Morton *et al*, 1977; Mabe *et al*, 1982; Arnold *et al*, 1985; Risucci *et al*, 1989; Gordon, 1991; Woolliscroft *et al*, 1993; Das *et al*, 1998; Johnson *et al*, 1998; Fletcher, 1999; Ward *et al*, 2002; Ward *et al*, 2003; Van der Heijden *et al*, 2004). It has been suggested that this illustrates a lack of insight that could be corrected by MSF as it identifies so-called 'blind spots'. However, in management at least, MSF from subordinates may be ignored despite potentially identifying important failings (Bernadin *et al*, 1993), countering the beneficial effects and even leading to ill-feeling. It is therefore vital that the

criteria examined by MSF are acceptable to those being assessed; if they feel an assessment to be irrelevant to their day-to-day practice, they are likely to ignore the outcome. This may be countered by the use of a penalty if future assessments fail to show improvement (akin to the proverbial stick). However, if the assessee recognises the assessment to be relevant, he/she is likely to actively strive to improve (the proverbial carrot) with far better results. This has been repeatedly shown in industry (McEvoy *et al*, 1987; Fedor *et al*, 1989; Yuki *et al*, 1995; Wimer *et al*, 1998) but medicine has been slow to recognise the importance of assessee opinion on assessment processes. This acceptance is vital to MSF and the shared understanding and development of a positive and supportive culture within which professional behaviours are recognised as being as important as technical prowess, and where deficient individuals welcome the insight that MSF may afford in an effort to improve. The acceptability of the fields examined by EBSTAF to the trainees themselves is the focus of Section IV of this study.

To date, no study has demonstrated a lasting relationship between excellence within an assessment and similarly outstanding performance in practice. However the application of MSF in industry, where descriptors of performance are better developed, suggests evidence for the superior predictive validity of MSF over other strategies (Church, 2000). The potential of EBSTAF to predict subsequent career progression, and thus potentially aid in the selection of HSTs, is therefore addressed within Section III of this study.

I.5.b.i. (c) Assessment of clinical skills by indirect observation (video).

The use of video has proven highly effective in sport for both acquisition of technical skills (Lounsbery *et al*, 1996) and self-assessment (Winfrey *et al*, 1996) but its application to medicine has so far been limited, except in the field of GP where it has been extensively applied. Currently in the UK, consultation skills of GP registrars are regularly assessed using a portfolio of videotaped consultations on real patients selected by the trainees themselves. They are used in the determination of minimal competence (formerly administered by the JCPTGP, now PMETB) and as part of membership examinations. (Royal College of General Practitioners. 2006). The MRCGP applies a marking scheme based upon expert-determined criteria deemed to be important to patients and the technique has been widely accepted, despite a lack of demonstrated reliability or validity in the literature.

Simultaneous introduction of assessment of GP consultations by patients themselves has produced a number of reliable (but not as yet validated) instruments (Baker, 1990; Campbell *et al*, 1996; Howie *et al*, 1998; Howie *et al*, 1999). Ratings from these appear to correlate well with each other but a lack of agreement between video and patient-based assessment has also been demonstrated (McKinstry *et al*, 2004). This may result from the inadvertent assessment of different competencies selected by experts and

patients respectively, unrepresentative video selection by the trainee (again highlighting differences between demonstration and practice) or inherently invalid methods. Further work is required.

I.5.b.i. (d) Assessment of clinical skills by simulation.

Simulation has been successfully applied to complex high-reliability industries such as aviation, nuclear power and the military where the recreation of critical high-risk situations allows the development and rehearsal of strategies to deal with them should they occur for real. Similarly, medical simulation replicates real life situations for the purposes of education or assessment while maintaining both patient and participant safety.

Technology is not a pre-requisite for simulation (as in role play or case discussion) although many simulators are highly technological. Simulation may also involve a variable degree of learner participation and may occur in a variety of locations, ranging from multimedia at home to a purpose-built facility.

A *medical simulator* is any device that presents the participant(s) with a simulated patient (or part thereof) with sufficient realism as to facilitate learning or the display of skills or behaviours that may otherwise be difficult to capture safely (Krummel, 1998). The *fidelity* of the simulation must reflect the complexity of the real-life task (Swanson *et al*, 1995) if the participant is

to believe in the scenario (*immersion*) and demonstrate everyday behaviours (Barnes, 1987).

(i) The history of medical simulation.

The first medical simulator was "Resusci-Annie" which allowed training in mouth-to-mouth resuscitation of a dying patient whose airway could obstruct requiring neck-extension before successful lung insufflation. Although first created in the 1960's (Winchell *et al*, 1966) it continues to be the basis of present-day cardiopulmonary resuscitation training.

"Sim One" was the first computer-controlled high-fidelity simulator (Denson *et al*, 1969) with a chest that moved with breathing, blinking eyes and reactive pupils, causing Abrahamson to first realise the potential for training (in this case for endotracheal intubation) away from the patient (Abrahamson *et al*, 1969). Unfortunately the cost of the technology involved prevented its further development.

"Harvey", developed around the same time, was a full-size mannequin capable of accurately simulating 27 cardiac conditions. Harvey is of proven educational efficacy (Gordon *et al*, 1980; Ewy *et al*, 1987) and has been applied to assessment (Jones *et al*, 1997), now offering a comprehensive general curriculum in cardiology (Issenberg *et al*, 2001).

The development of mathematical models of physiology and pharmacology allowed increased fidelity. Gaba developed the Comprehensive Anaesthesia Simulation Environment (CASE) within a real operating theatre with real monitoring which allowed the manipulation of physiological parameters by the instructor to simulate critical events with a high degree of realism (Gaba *et al*, 1988). In parallel, Good designed the Gainesville Anaesthesia Simulator (GAS) to diagnose faults with anaesthetic machines incorporating software capable of detecting and reacting to drugs and interventions in real-time (Good *et al*, 1989; Good, 1990). Further development of GAS resulted in the highest fidelity to date – the Medical Education Technologies Inc. (METI – Sarasota, Florida) Human Patient Simulator (HPS).

(ii) METI-HPS - The state of the art in simulation.

Today's METI HPS is a life-size, life-like manikin that breathes, has heart sounds, breath sounds and palpable pulses. Its eyes open and close and its pupils are reactive to light, drugs and intracranial events. It passes urine and has interchangeable parts allowing it to be male or female. It has even been known to be pregnant. It has pleural spaces allowing acute needle thoracocentesis or chest drain insertion, and pericardiocentesis or diagnostic peritoneal lavage can also be carried out. It can speak, vomit and can even be made to move its arms and legs. The manikin is driven by a computer that applies dynamic physiological models allowing it to react in real time to interventions, including the administration of drugs and fluids, and produce

changes in physiological parameters that can be picked up and displayed on routine unmodified monitoring equipment. In order to increase the reality of the simulation, the manikin may be housed in a realistic clinical environment. Scenarios are commonly programmed and 'driven' from an adjoining control room where a composite of events and physiological signals can be recorded and replayed for the purposes of debriefing or assessment.

The strength of HPS is in the emulation of pathophysiology in acutely unwell patients, allowing trainees to manage the critically ill in a totally patient-safe and trainee-safe environment. Simulations remain limited by the fact that the manikin does not alter its appearance according to its physiological condition (for example becoming pale, sweaty and cool peripherally when simulating shock) but participants repeatedly report a sufficiently high degree of realism for effective learning (Howard *et al*, 1992; Chopra *et al*, 1994; Holzman *et al*, 1995).

Studies have demonstrated HPS to be reliable and valid in anaesthesia (Good *et al*, 1989; Devitt *et al*, 1998; Devitt *et al*, 2001) but evidence of learning transfer to everyday clinical practice is scarce, although recent applications to undergraduate teaching (Murray *et al*, 2002; Boulet *et al*, 2003), critical care (Lighthall *et al*, 2003), accident & emergency (Small *et al*, 1999; Reznek *et al*, 2003) and trauma (Lee *et al*, 2003) have been favourable. Despite increasing research interest and recommendations for its integration into mainstream training (Vozenilek *et al*, 2004), there remains a

lack of hard evidence that HPS reduces actual patient risk or improves clinical outcomes. However, it has been suggested that this expectation may be unrealistic in such a multifactorial system and should not limit its application and potential for learning (Gaba, 1992; Blum *et al*, 2004). Furthermore, increasing recognition of the importance of non-technical aspects of both individual and team performance (termed *human factors*) initially in aviation and latterly medicine has resulted in an area of research in which HPS may have a pivotal role [see Section I.5.b.iii].

I.5.b.ii. Assessment of technical skills.

Technical competence is one of the fundamental prerequisites for successful surgery and yet it's formal training and assessment has been neglected until recently (Hamdorf *et al*, 2000) with training programmes generally following a Halstedian approach (see, do, teach). The unregulated introduction of minimally invasive surgery illustrated the relationship between experience and outcome (termed "the learning curve") by the dramatic increase in bile duct injuries at laparoscopic cholecystectomy (Rogers *et al*, 2001). Sub-optimal results as a result of the learning curve were quickly recognised as unacceptable (Hasan *et al*, 2000) and demanded reliable methods of operative training and assessment to protect the public, assist training programmes and maintain the standards of the profession.

The first British surgical skills workshops were held in the 1970s using animal tissue to train basic surgical techniques. The potential of such workshops to train new techniques safely away from the operating room was quickly recognised (Apley, 1980; Stotter *et al*, 1986; Bevan, 1986; Greenhalgh *et al*, 1987) and they now address every specialty at every training level with the use of increasingly realistic methods and materials.

Robust assessment of technical skills has proved difficult, encountering similar problems to those of clinical skills assessment: inevitable case variation independent of operator skill (patient anatomy, operative findings and complications); the highly subjective nature of unstructured scoring methods (observer bias – "doves" versus "hawks"); the halo effect; and the setting of suitable standards. With the general acceptance of surgical skills workshops, assessment was also able to move away from the operating theatre to the laboratory, aided by the development of increasingly realistic simulations.

I.5.b.ii. (a) Assessment of technical skills in the absence of observation.

Historically, a trainee's exposure to surgical cases has been recorded within a surgical logbook. This details a trainee's experience whilst encouraging them to audit their work and reflect upon it. Logbooks remain an integral part of the assessment of trainees in surgery, obstetrics & gynaecology and anaesthetics in the UK where their regular inspection is felt to ensure

satisfactory exposure and progress. They also allow some small assessment of the post itself (Galasko *et al*, 1997). Logbooks are of limited summative value however since exposure (quantity) does not equate to competence (quality) (Jolly *et al* , 1997). The structure and use of the logbook is currently under review by the Royal College of Surgeons of England as part of the Intercollegiate Surgical Curriculum Project (ISCP).

I.5.b.i. (b) Assessment of technical skills by direct observation.

The assessment of technical skills within the operating theatre equates to workplace-based assessment while the use of other environments (such as simulations, wet-lab or virtual reality) is the technical equivalent of examination. The latter has allowed the development of valid assessment tools that have subsequently been applied to real-life operating and although there are differences between the two broad methods, they should be considered a continuum for the purposes of this section.

One of the first attempts to objectively assess surgical ability directly observed and graded bowel anastamoses performance using specific criteria (Stotter *et al*, 1986) and was shown to demonstrate improvement of skill over time (Steele *et al*, 1992). Numerous methods and instruments have subsequently been developed to objectively score technical skill. The most widely validated is the Objective Structured Assessment of Technical Skill

(OSATS) (Reznick *et al*, 1997; Martin *et al*, 1997) developed at the University of Toronto. This performance-based examination, modelled closely upon the OSCE, uses global assessments of 7 items using a 5-point Likert scale anchored at extremes and midpoints by specific behavioural descriptors [see Table I.1] combined with dichotomous task-specific checklists. Trainees are assessed as they perform standardised surgical tasks under the direct observation of expert examiners. OSATSs, in common with OSCEs, are labour intensive and expensive (Cusimano *et al*, 1994) but have repeatedly demonstrated validity and reliability sufficient for summative purposes (Winckel *et al*, 1994; Jansen *et al*, 1995; Reznick *et al*, 1997; Martin *et al*, 1997; Regehr *et al*, 1998; Anastakis *et al*, 1999). Criticisms of OSATS not assessing operative skills in the operating theatre prompted their application to the assessment of real-life operating where good correlation with ratings from the laboratory further validated the bench-station assessment of technical ability (Datta *et al*, 2004). OSATS has also been shown to be transferable between establishments (such as London to Hong Kong) without loss of reliability or validity (Bann *et al*, 2003b), but limitations of fidelity may make OSATS most useful for simpler tasks with more junior trainees (Darzi *et al*, 2001).

Motion analysis offers further objective analysis of surgical skill. As an operator becomes more competent, so his/her dexterity and efficiency of hand-movements improves (Rosenbaum 1992). With this in mind the Imperial College Surgical Assessment Device (ICSAD) was developed

(Taffinder *et al*, 1999a). This maps hand movements within a generated electro-magnetic field. The raw data (total path length, total number of movements and time taken) obtained from real, simulated and virtual reality procedures has been shown to be valid and reliable in the assessment of surgical dexterity (Taffinder *et al*, 1998b; Taffinder *et al*, 1999a; Smith *et al*, 1999b; Taffinder *et al*, 1999b; Datta *et al*, 2001; Datta *et al*, 2002a). ICSAD data has also been shown to correspond directly to hand kinematics captured from video and real-time procedures (Dosis *et al*, 2005).

The combination of OSATS with PAME also demonstrates reliability and validity in an assessment suggested to address the overall competence (both clinical and technical) of senior surgical trainees (MacRae *et al*, 2000).

Further workplace-based assessment resulted in the method of Direct Observation of Procedural Skills (DOPS), developed by The Royal College of Physicians, and its theatre equivalent of surgical Procedural Based Assessment (PBA). These involve the direct observation and assessment of generic clinical and specialty-specific operative procedures using the combination of checklist and global assessment methodology. They are inherently both procedure-specific and assessor-specific, equating to a technical version of the short and long case respectively. As such, each offers only a one-time snap-shot assessment of trainee ability across a heterogeneous range of procedures that are difficult to compare with one

another. However, their use in combination has been shown to be valid and reliable (Wragg *et al*, 2003).

Directly observed assessments (OSCEs, OSATSs, DOPSs and PBAs) are limited by their requirement for expert examiners to be physically present. This is often difficult to organise around clinical commitments resulting in difficulties in assessor recruitment. An alternative is that of indirect observation.

I.5.b.i. (c) Assessment of technical skills by indirect observation.

(i) Assessment of the final product.

Szalay and colleagues examined the value of assessing solely the final surgical product of an OSATs-type examination without observing performance at the stations themselves. They applied a 4-item 5-point global rating scale to assess completeness, aesthetics, function and overall quality and demonstrated good inter-rater and inter-station reliability, construct validity and ranking agreement. However procedures were of an advanced nature (choledochojejunostomy, rectal anastomosis and femoral artery anastamosis) and the authors admit that preceding pilot studies had proven difficult on more junior trainees who were unable to complete the specified tasks. Furthermore, no attempt was made to relate product scores to actual performance in the OSATS examination (Szalay *et al*, 2000). The ability to

detect purposely placed errors within an end-product may also discriminate between varying levels of surgical experience (Bann *et al*, 2003a) and has recently been shown to predict subsequent surgical skill on bench tasks (Bann *et al*, 2005) However, neither of these methods are directly applicable to everyday practice.

(ii) Assessment using video.

The use of video in sport is commonplace, with demonstrable benefits on skills acquisition (Guadagnoli *et al*, 2002). Cameras are often incorporated into modern operating theatres to offer bi-directional real-time teaching at a distance (Rafiq *et al*, 2004) or the recording of procedures for later review. The feasibility and reliability of assessing videotaped surgical procedures has been examined but remains unclear. One study found insufficient correlation between assessments obtained from edited videotape versus real-life laparoscopic cholecystectomy, with poor inter-rater reliability in the videotaped arm (Scott *et al*, 2000). In contrast, Dath demonstrated that blinded and independently rated videotapes of senior surgical residents performing laparoscopic low anterior resection and Nissen fundoplication maintained high inter-rater reliability despite the videos remaining unedited and assessors being permitted to use the "fast-forward" facility (Dath *et al*, 2004).

Indirect observation using video has a number of potential advantages. Examiners can assess when convenient to themselves and their schedules with greater efficiency and this may limit "examiner burnout" from fatigue and loss of concentration. If 'everyday' procedures are reviewed rather than 'assessment' procedures, indirect observation may also have higher face validity than direct observation by capturing what a candidate actually does and eliminating the Hawthorne Effect (Roethlisberger *et al* , 1939).The potential for anonymity allows blinded and more objective assessment by eliminating examiner bias while reliability may also be improved if multiple assessors are used (Dath *et al*, 2004). However, the reliability of video-methods is likely to be dependent on optimal views and standardised editing. The value of such methods requires further clarification.

(iii). Assessment by simulation.

The application of simulation to surgery has great potential for technical skills training and assessment. Motor skills may be acquired in a patient-safe, learner-safe and non-stressful environment where the rate of learning is determined by the learner rather than theatre time or case-mix (MacIntyre *et al*, 1990; Gardiner *et al*, 1996; Gaba, 2004). Simulators can address a wide range of applications ranging from the acquisition of a trainee's basic skills to a "dry run" of a complicated procedure on a particular patient (Meier *et al*, 2001), as well as rare situations that retain the potential for catastrophe (Dunnington *et al*, 1994). They allow repetitive practice away from patients

and may offer objective metric feedback (Coleman *et al*, 1994; Taffinder *et al*, 1998a; Datta *et al*, 2002a). But a competent surgeon is more than a collection of technical skills (Baldwin *et al*, 1999; Thomas, 2000). For simulators to fulfil their potential, they must also address important issues around the procedure itself, such as that of surgical access (MacIntyre *et al*, 1990), and the clinical environment. Simulation of surgical procedures in isolation removes the trainee from the theatre or ward where skill in communication and teamwork, knowledge of indications, contraindications and complications and ultimately decision-making have far greater influence on surgical outcome than pure technical skills (Spence *et al*, 1987). A perfect operation, when carried out for the wrong reasons, can only give a poor outcome (McDonald, 1998).

Surgical simulators remain almost exclusively procedural or part-procedural at the present time. Since the description of the first endoscopy trainer in the late 1980's, simulators have been developed across almost all specialties. They are diverse in nature, but fall within 3 basic categories.

The optimal simulation of surgery would be offered by the use of fresh human cadavers but this is precluded by moral and legal issues (Stotter *et al*, 1986), potential hazards of infectious disease and cost. The alternative use of live-animal models is prohibited in the UK by the Cruelty to Animals Act of 1876 but is common outside the UK within animal laboratories or "wet labs". Though they may mimic real tissues, comparative anatomy may differ

substantially from that in the human. Described examples include a porcine model for laparoscopic cholecystectomy (Bailey *et al*, 1991) and canine vascular bypass (Saifi *et al*, 1990).

Freshly procured animal tissue simulates the handling characteristics of human tissue and has been used to train isolated skills (Razaboni *et al*, 1980; Woods *et al*, 1980; Stotter *et al*, 1986; Rogers *et al*, 1986; Gardiner *et al*, 1996). However, they are costly, may require time-consuming preparation and cannot be kept safely for long. They must be used in designated sites and disposed of in strict accordance with health and safety guidelines. Religious and moral issues may also limit their use.

Synthetic materials, usually latex or silicone, avoid the above problems and can often be used and reused (McMahon *et al*, 1995; Thomas *et al*, 1996). They have been successful in teaching skills across surgical specialties (Allen, 1990; Hamdorf *et al*, 2000; Kneebone *et al*, 2001). Initial devices were very simple and taught knotting or suturing techniques (Barnes *et al*, 1989; Munro *et al*, 1994b; Thomas *et al*, 1996). Development allowed simulation of surgical tasks and procedures and they have repeatedly demonstrated validity in the assessment of procedural skills within processes such as OSATSs.

Laparoscopic surgery requires different skills from open surgery with even an accomplished surgeon experiencing a steep 'learning curve' when first adopting laparoscopic techniques (Dent, 1991; Cuschieri, 1992; Cundiff,

1997; Edwards *et al*, 2000). Simulators have been developed to allow surgeons to learn and practice laparoscopic skills with some providing objective scores of performance: the simple Borinquen Ring (Medina, 1993; Medina, 2002); the laparoscopic box trainer (Munro *et al*, 1994a; Rosser *et al*, 1997; Derossis *et al*, 1998); the Berci-Sackier Laparoscopic Trainer (Karl Storz-Endoscopy America Inc.) (Bailey *et al*, 1991; Sackier *et al*, 1991); and the Advanced Dundee Endoscopic Psychomotor Tester (ADEPT) (Hanna *et al*, 1997; Macmillan *et al*, 1999; Francis *et al*, 2001b).

Virtual reality (VR) facilitates the interaction of humans and computers within artificial computer-generated environments that simulate the physical world. These have been widely applied to the military and aviation industry and advances in computing technology now allow their application to surgical training but the complexity of what can be emulated remains limited. Human anatomy, normal or pathological, is highly variable while the deformation of tissues with realistic haptic feedback to the operator challenges even the most advanced systems (Coleman *et al*, 1994; Raibert *et al*, 1998). That said, virtual reality has been successfully applied to endoscopy (Sedlack *et al*, 2002; Shah *et al*, 2002; Datta *et al*, 2002b; Moorthy *et al*, 2003; Moorthy *et al*, 2004), laparoscopic surgery (Wilson *et al*, 1997), intravenous cannulation (Prystowsky *et al*, 1999), arthroscopy (Ziegler *et al*, 1995; Muller *et al*, 1995; Smith *et al*, 1999a), endonasal surgery (Ecke *et al*, 1998) and vascular anastomoses (Playter *et al*, 1997).

MIST-VR (Mentice Medical Simulation, Gothenburg, Sweden) has been most widely examined. Taffinder described construct validity based upon a comparison of naïve and experienced (>100 cases) laparoscopic surgeons but little difference in score was demonstrated between the more closely comparable groups of trainees and non-surgeons (Taffinder *et al*, 1998a). Around the same time, Chaudhry described similar results but these were based upon a critically small sample size (Chaudhry *et al*, 1999). These failings prompted others to question the validity of such assessment methods having failed to demonstrate validity (Paisley *et al*, 2001b). Subsequent studies lend further support to the validity of MIST-VR and skills transfer to the operating theatre has subsequently been demonstrated (Seymour *et al*, 2002) but Ro recently found that on first exposure to a very similar system (Lap-Sim, Immersion Medical, MD, USA), experienced laparoscopic surgeons actually scored less highly than the surgically naïve suggesting the need for careful application for the purposes of assessment (Ro *et al*, 2005). However, issues of face validity are likely to be questioned by the trainees themselves until such time that they are successfully combined with high-fidelity simulation of the whole operating experience.

Surgical simulation is now widespread, with most trainees having access to skills laboratories. Their use in the assessment of trainees' operative and clinical skills remains controversial (Vanchieri, 1999) although they are gaining acceptance as validity (Seki, 1987; Reznick *et al*, 1997; Taffinder *et al*, 1998a; Chaudhry *et al*, 1999) and skills transfer (Faulkner *et al*, 1996;

Macmillan *et al*, 1999; Seymour *et al*, 2002) are repeatedly demonstrated. However, like all assessments, caution is needed during their incorporation into the summative assessment of surgical skill (Prystowsky *et al*, 1999; Paisley *et al*, 2001b).

I.5.b.iii. Assessment of non-technical skills.

I.5.b.iii. (a) Lessons from aviation.

In 1979, the North American Space Administration (NASA) sponsored a review of commercial aviation accidents. It concluded that 70% of air-crashes resulted not from mechanical failures, as had previously been assumed, but from failures of interpersonal communication, teamwork, decision-making and leadership (Cooper *et al* , 1980); these were collectively termed *human factors*. In response, *crew resource management* (CRM) courses were developed to reduce "pilot error" by training better use of resources on the flight deck. Based on Reason's premise that human error is inevitable and ubiquitous (Reason , 1990), they aimed to reduce accidents by providing pilots with countermeasures to avoid, trap and mitigate the consequences of error before disaster (Helmreich *et al*, 1999). CRM training was made mandatory in 1993 by the UK Civil Aviation Authority (CAA) for all UK pilots and has subsequently been adopted in other high-reliability environments including air-traffic control, the military, off-shore oil, fire services, and nuclear power (Flin *et al*, 2002). The benefits of human factors training are now

widely accepted despite a lack of direct evidence in such multifactorial systems.

I.5.b.iii. (b) Errors in medicine.

In 1993, a review of 2000 medical critical incidents found 70-80% were caused by human factors related to communication (Williamson *et al*, 1993). This was inevitably compared with findings in other high-reliability industries and triggered a similar approach to that adopted by aviation. This was led by anaesthetics, the specialty most frequently likened to the flight deck owing to periods of high-intensity (induction and reversal of anaesthesia being comparable to take-off and landing) in association with the longer and less intense operative phase requiring constant vigilance (analogous to the flight itself). In the USA, Anaesthesia Crisis Resource management (ACRM) courses were developed that applied HPS to simulate crises and train technical and teamwork skills using a combination of experience and feedback (Howard *et al*, 1992; Gaba *et al* , 1994). Similar Crisis Avoidance and Resource Management in Anaesthesia (CARMA) courses developed in the UK (Maran *et al*, 2001).

I.5.b.iii. (c) Behavioural markers of non-technical skills.

In common with aviation, ACRM training was initially purely formative due to the absence of assessment instruments to evaluate an individual's non-

technical skills. However, it was recognised that some individuals displayed behaviours contributing to superior or sub-standard non-technical performance, suggesting that individuals might be assessed by observing what were termed *behavioural markers* within the work environment. The first behavioural marker system was again developed in the airline industry and called NOTECHS (Avermaete *et al* 1998). It identified 4 categories of behaviour: cooperation; leadership & managerial skills; situation awareness; and decision making [Table I.2] along with their constituent elements. Following a validation period, the assessment of non-technical skills using NOTECHS has now become mandatory for all UK pilots (CAA 2004).

Behavioural marker systems developed in aviation or abroad have previously been applied to medicine (Gaba *et al*, 1998; Small *et al*, 1999) but this practice must be questioned since behavioural marker systems are context-specific and do not transfer across domains (i.e. aviation to medicine) (Klampfer *et al* 2001). On this basis, the Anaesthetists' Non-Technical Skills (ANTS) project used cognitive task analysis to develop an anaesthesia-specific behavioural marker system that differed considerably from NOTECHS. The ANTS system identified 4 behavioural categories of task management, team working, situation awareness and decision making [Table I.3], again with constituent elements. ANTS was subsequently evaluated and found to have satisfactory validity, reliability and usability, provided assessors received adequate training (Fletcher *et al*, 2003) and full integration into the anaesthetic curriculum is currently under discussion.

In surgery, the literature repeatedly identifies communication, leadership, teamworking and decision making as desirable non-technical aspects of surgical practice (Greenburg *et al*, 1982; Baldwin *et al*, 1999; de Leval *et al*, 2000; Giddings *et al* , 2000; General Medical Council , 2000b; Satish *et al*, 2001; Healey *et al*, 2004). Good communication and interpersonal skills improve patient satisfaction (Suchman *et al*, 1993), clinical outcome (Stewart, 1995) and decrease the risk of malpractice litigation (Beckman *et al*, 1994; Levinson *et al*, 1997) while failures in these same areas may be causal in up to half of surgical errors (de Leval *et al*, 2000; Giddings 2001; Gawande *et al*, 2003). The GMC's process of revalidation, based on Good Medical Practice (General Medical Council , 2001), includes communication, teamwork and leadership among the necessary attributes of the competent doctor (or surgeon). Determination of a surgeon's competence in these areas demands robust assessment but the necessary specialty-specific behavioural marker systems are yet to be developed.

I.6. SELECTION OF SURGICAL TRAINEES.

Effective selection of employees is critical to any profession and surgery is no different. Traditionally, surgery has relied upon performance at unstructured interview, academic achievement to date and personal references. This practice has been heavily criticised (Gough *et al*, 1988; Gough, 1988; Gough, 1993). The unstructured interview is unreliable and invalid with poor correlation between interview ratings and subsequent performance (Wood *et al*, 2000), although modifications of the interview process have shown promise (Gilbart *et al*, 2001; Eva *et al*, 2004). Academic achievement prior to medical school has repeatedly been shown to be unrelated to subsequent performance (Schueneman *et al*, 1984; Schueneman *et al*, 1985; Vickers *et al*, 1990) but a recent and powerful 20 year study by McManus following 511 medical students who passed through Westminster Medical School would suggest otherwise, having demonstrated the long-term predictive validity of A-level results for both undergraduate and postgraduate careers (McManus *et al*, 2003). Once in medical school or residency, the two may (Grossman *et al*, 1992; Martin *et al*, 2002; Boyse *et al*, 2002)or may not (Kron *et al*, 1985; Papp *et al*, 1997) be related.

The use of psychometric testing (standardised psychological measures of cognitive ability and aptitude) in combination with the assessment of personality is widespread in the selection of personnel outside the field of medicine. The Royal Air Force was one of the first to apply these techniques

in the selection of its pilots and their introduction was followed by a significant reduction in pilots who failed training (Bell, 1998).

In surgery, the examination of factors predictive of success has been hampered by a lack of performance criteria that define the surgical role (Wanzel *et al*, 2002b). Thus the retrospective examination of existing trainees dominates the literature.

I.6.a. Aptitudes.

An aptitude is an innate inborn physical or mental ability to do a certain kind of work or task (Oxford University Press , 2006). Many aptitudes are identified and testable but research in this field has concentrated on cognitive testing, personality, manual dexterity and visual-spatial ability,

I.6.a.i. Cognitive tests.

Cognitive testing appears unhelpful with correlation to subsequent clinical performance being found to be significant (Kron *et al*, 1985), weak (Schwartz *et al*, 1973) or absent (Ansell *et al*, 1979; Lazar *et al*, 1980; Papp *et al*, 1997). At best, cognitive aptitudes may only help to predict future cognitive performance (Erlandson *et al*, 1982), suggesting that the assessment of aptitudes may also be content specific.

I.6.a.ii Personality.

The influence of personality has also been examined. The ability to tolerate stressful situations appears to confer some advantage (Schueneman *et al*, 1984; Linn *et al*, 1984) and successful candidates have been found to be conscientious (Deary *et al*, 1992), competitive and practical (Schwartz *et al*, 1994a), decisive, honest, and motivated with the ability to be flexible and work well in a team (Greenburg *et al*, 1982). However, none of these characteristics are specific to surgery and until the "surgical personality" is more clearly defined, personality testing is unlikely to be helpful.

1.6.a.iii. Manual Dexterity.

Some studies have supported the relationship between good manual dexterity and performance in surgery (Dashfield *et al*, 2001; Francis *et al*, 2001a). Other studies refute this (Squire *et al*, 1989; Steele *et al*, 1992; Shah *et al*, 2003) . Electromagnetic tracking of hand-movements by ICSAD objectively relates hand-motion efficiency to surgical skill (Taffinder *et al*, 1999a; Datta *et al*, 2001) but its predictive value has not been assessed.

1.6.a.iv. Visual-spatial ability.

The psychomotor skill of visual-spatial ability appears to be the best predictor of operative ability (Schueneman *et al*, 1984; Gibbons *et al*, 1986; Steele *et*

al, 1992; Murdoch *et al*, 1994; Risucci, 2002; Wanzel *et al*, 2002a; Wanzel *et al*, 2003) but this relationship also appears inconsistent (Deary *et al*, 1992; Van Rij *et al*, 1995; Francis *et al*, 2001a) with poor performance potentially overcome by training and experience (Francis *et al*, 2001a; Wanzel *et al*, 2003).

Overall, there appears to be little difference between surgical trainees and those of other specialties (Harris *et al*, 1994; Gilligan *et al*, 1999) other than their aspirations.

The application of reliable and valid competency-based in-post assessment may identify those trainees who excel or falter but if it is to be used in selection it must also demonstrate the ability to predict future performance. In the absence of proven *predictive validity* it should not be applied in isolation. To date, few studies have prospectively examined subsequent career progression and further work is required in this regard.

I.7. ASSESSMENT FEEDBACK IN SURGICAL TRAINING

No one is more concerned about the current changes to surgical training than the trainees themselves. With the *quantity* of surgical training being reduced to one fifth of previous levels (Donaldson 2002), the *quality* of surgical training must be significantly improved (Hargreaves, 1996; Hurley *et al*, 1999). "Over-experienced and under-trained" (Bottomley 1992) is no longer acceptable, if it ever was, in the training of competent surgeons.

Safe graduated practice (as applied in surgical training) is described in terms of its principal components; commitment on both sides to optimise training opportunities and close supervision by a trainer who is available to support, advise and intervene as appropriate, allowing the trainee to acquire technical expertise and confidence by graduated practice and regular feedback (Hargreaves, 1996). While many of these features are integrated into training programmes, there remains a need for detailed feedback using defined objectives to accurately direct trainees' efforts.

I.7.a. Good feedback.

Kolb and Fry's model of how adults learn (Kolb & Fry 1975) is analogous to that of surgical audit and the two are depicted in combination in Figure I.1 to illustrate the features of good feedback. For trainees to optimise their training they must have clear objectives, the tools to compare actual and desired

performance, and the means to lessen the gap between the two (Sadler, 1989; Gipps , 1994). However, learners are poor self-assessors. Their ratings show low correlation with those of experts (Morton *et al*, 1977; Risucci *et al*, 1989; Gordon, 1991; Das *et al*, 1998; Johnson *et al*, 1998; Ward *et al*, 2002; Ward *et al*, 2003) unless benchmarked to the performance of others (Martin *et al*, 1998) and as a result they may be unable to recognise their own failings. They therefore require repeated structured feedback involving a 3-stage process which has been shown to give substantial learning gains in the field of educational research (Black *et al*, 1998).

First, the attributes or qualities that make a good surgeon must be identified to provide documented objectives for the trainee. EBSTAF offers just such a framework of desirable qualities and aptitudes as determined by the consensus view of consultant surgeons (Baldwin *et al*, 1999).

Second, assessment should be applicable to in-post performance and demonstrate subsequent improvement (construct validity) while at the same time allowing comparison with the agreed standard. EBSTAF has been demonstrated to be reliable and construct valid for the assessment of in-post performance of BSTs (Paisley *et al*, 2001a).

Finally, the means to address the trainee's strengths and weaknesses comes from directed training that is itself guided by accurate assessment.

A truly robust assessment process will:

- document the acquisition of skills thereby maintaining standards;
- facilitate the formative process of feedback;
- motivate both trainee and trainer to take an active role in the trainee's professional development and;
- may subsequently be used within selection processes (Jolly *et al* , 1997).

Initial examination of in-post assessment by EBTAF appears highly encouraging (Paisley *et al*, 2001a) but its acceptability, predictive validity and use in structured feedback are yet to be examined.

1.8. DEFINITIONS.

I.8.a. Reliability.

Reliability indirectly indicates the random error of assessment data, estimating its consistency (Downing *et al*, 2004). Inconsistent assessment data with a large random error threatens to misdirect learning or selection, lowering standards in the long run.

Reliability is estimated by a number of methods and is expressed as a reliability coefficient ranging in value from 0.0 (totally unreliable) to 1.0 (totally reliable).

Reliability is improved by the use of multiple examiners, cases or assessment episodes, as demonstrated by the OSCE [see section I.5.b.i.(b).(i).a]. For the purposes of high-stakes assessment, it is generally agreed that reliability should be greater than 0.9 (total agreement being expressed as 1.0). However, 0.8 to 0.89 may be accepted for more moderate-stakes assessments while formative applications may accept 0.7 or above. The acceptable value is thus determined by the consequences of a falsely positive or negative (i.e. incorrect) outcome (Downing *et al*, 2004).

I.8.a.i. Estimations of Reliability.

I.8.a.i. (a) Test-Retest Reliability.

This is based on the repetition of the same assessment using the same judges performed on the same candidates on two different occasions. If the assessment is reliable it should yield the same result, assuming no additional learning has taken place in the interim. Commonly used in the medical education literature, it is rarely performed in practice since repeating many assessments (particularly written examinations) is likely to see trainees attaining higher scores as a result of learning in the interim, thereby reducing the apparent reliability.

I.8.a.i. (b) Internal Consistency.

This is based upon the logic that an assessment that examines a single construct may be randomly split in half to give two half-tests that bear a reasonable approximation to two separate tests administered to the same group. Subsequent correlation between the two halves is thus a purely statistical representation of test-retest analysis. However, this correlation has the potential to be affected by the position of the statistical split. This is then countered by taking the logic further and determining correlations for all possible ways that the test may be split into two equal halves. These

correlations are then averaged and the resulting single mean correlation is Cronbach's *alpha* coefficient (Cronbach, 1951).

I.8.a.i. (c) Reliability of the Raters.

The largest threat to the reproducibility of data from assessments dependent on human raters is that between the raters themselves (Downing *et al*, 2004). This is termed *inter-rater reliability*. A simple percentage agreement does not take account of the occurrence of such agreement purely by chance. However, it may be useful to indicate the direction of disagreement if several options have been supplied to raters.

The kappa statistic (Cohen, 1960) is a correlation coefficient that takes account of the effects of chance and is therefore frequently applied as an inter-rater reliability estimate.

Generalisability theory (GT) analysis is an analysis of variance (ANOVA) that considers all aspects of an assessment's design to produce a generalisability coefficient (Crossley *et al*, 2002a). However, individual variances can be difficult to estimate making GT difficult to apply. An acceptable alternative is the estimation of *intraclass correlation coefficient* (Ebel, 1951), another analysis of variance of factors in the reliability of an assessment's design. It is commonly available in statistical software and permits the estimation of both inter-rater and individual rater reliability while coping with missing ratings.

I.8.b. Validity.

Validity addresses the evidence presented to support or refute an interpretation of assessment data without which the assessment has no intrinsic meaning (Downing *et al*, 2003). Validity takes many forms drawing evidence from multiple sources.

I.8.b.i. Face Validity.

Face validity is a test that looks good for a particular purpose on first inspection. It is not validity in any technical sense but instead refers, not to what the test actually measures, but what it appears to measure. Its importance is that without face validity, an assessment is unlikely to be taken seriously.

I.8.b.ii. Content Validity.

Content validity is a subjective judgement of how well an assessment reflects the aims and weightings of a syllabus. It cannot be assigned a numerical value but is an important concept since assessment drives learning (Hamdorf *et al*, 2001; Wass *et al*, 2001) and misrepresentation may have a detrimental effect upon what trainees feel to be important. Content validity therefore reflects the quality of the test or assessment.

I.8.b.iii. Construct Validity.

The most commonly applied type of validity is that of *construct validity*, equating to whether an assessment assesses what it is designed to assess (i.e. the hypothesis or construct) (Jolly *et al* , 1997). In general, assessments address the hypothesis that individuals performing better will score higher ratings in an appropriately designed assessment. Thus, the determination of construct validity within surgical assessment usually addresses the hypothesis that an individual acquires greater knowledge and skills as they pass through their training, achieving higher ratings in association with greater experience.

I.8.b.iv. Criterion Validity.

Criterion validity is the degree of correlation between an assessment method and another criterion. The ability of a test to distinguish between individuals that it should theoretically be able to distinguish between is the *concurrent* validity, and often equates to a comparison with a current gold standard, even if that gold standard may be relatively subjective. If no gold standard is available then the examination of the ability of an assessment to predict future performance may be addressed (the *predictive* validity).

I.9. HYPOTHESES.

- In-post assessment of BSTs using EBSTAF is predictive of subsequent career progression in Surgery.

- BSTs find the skills examined within EBSTAF to be acceptable.

- Detailed and structured feedback of in-post performance as assessed by EBSTAF improves in-post performance.

- The assessment of critical care skills using human patient simulation reflects in-post clinical performance as determined by EBSTAF.

- The use of video for the remote assessment of BST's basic tissue-handling skills is valid, reliable and sensitive.

I.10. AIMS.

- To determine the predictive validity of EBSTAF by examining subsequent career progression in a previously studied BST cohort.

- To confirm the acceptability to BSTs of the qualities addressed by EBSTAF.

- To apply EBSTAF as a formative assessment tool and to examine the impact of detailed and structured feedback of BSTs' in-post performance on subsequent assessments

- To investigate the construct and concurrent validity of critical care skills assessment of BSTs' using human patient simulation.

- To examine issues of concurrent validity of video assessment of basic tissue-handling skills by correlation with current gold standards in technical skills assessment and EBSTAF in-post assessments.

Table I.1.

GLOBAL RATING SCALE OF OPERATIVE PERFORMANCE.

Please circle the number corresponding to the candidate's performance.

Respect for Tissue.				
1	**2**	**3**	**4**	**5**
Frequently used unnecessary force on tissue or caused damage by inappropriate instrument use.		Careful handling of tissue but occasionally caused inadvertent damage.		Consistently handles tissue appropriately with minimal damage to tissue
Time and Motion.				
1	**2**	**3**	**4**	**5**
Many unnecessary moves.		Efficient use of time / motion but some unnecessary moves.		Clear economy of movement and maximum efficiency
Instrument Handling.				
1	**2**	**3**	**4**	**5**
Repeatedly made tentative or awkward moves with instruments through inappropriate use.		Competent use of instruments but occasionally appeared stiff or awkward.		Fluid movements with instruments and no stiffness or awkwardness
Knowledge of Instruments.				
1	**2**	**3**	**4**	**5**
Frequently asked for wrong instrument or used inappropriate instrument.		Knew names of most instruments and used appropriate instrument		Obviously familiar with instruments and their names
Flow of Operation.				
1	**2**	**3**	**4**	**5**
Frequently stopped operating and seemed unsure of next move.		Demonstrated some forward planning and reasonable progression of procedure.		Obviously planned course of operation with effortless flow from one move to the next
Use of Assistant.				
1	**2**	**3**	**4**	**5**
Consistently placed assistant poorly or failed to use them.		Appropriate use of assistant most of the time.		Strategically used assistant to the best advantage
Knowledge of Procedure.				
1	**2**	**3**	**4**	**5**
Deficient knowledge: required specific instructions at most steps.		Knew all important steps of operation.		Demonstrated familiarity with all steps of operation.
OVERALL PERFORMANCE.				
1	**2**	**3**	**4**	**5**
Very Poor.		Competent.		Clearly Superior
QUALITY OF FINAL PRODUCT.				
1	**2**	**3**	**4**	**5**
Very Poor		Competent.		Clearly Superior

(Martin *et al*, 1997)

Table I.2.
Behavioural Markers of Non-Technical Skills as Developed in Aviation (NOTECHS).

Category	**Element**
Cooperation	Team building and maintaining
	Considering others
	Supporting others
	Conflict solving
Leadership & Managerial Skills	Using authority and assertiveness
	Maintaining standards
	Planning and coordinating
	Workload management
Situation Awareness	System awareness
	Environmental awareness
	Anticipation
Decision Making	Problem definition / diagnosis
	Option generation
	Risk assessment / option choice
	Outcome review

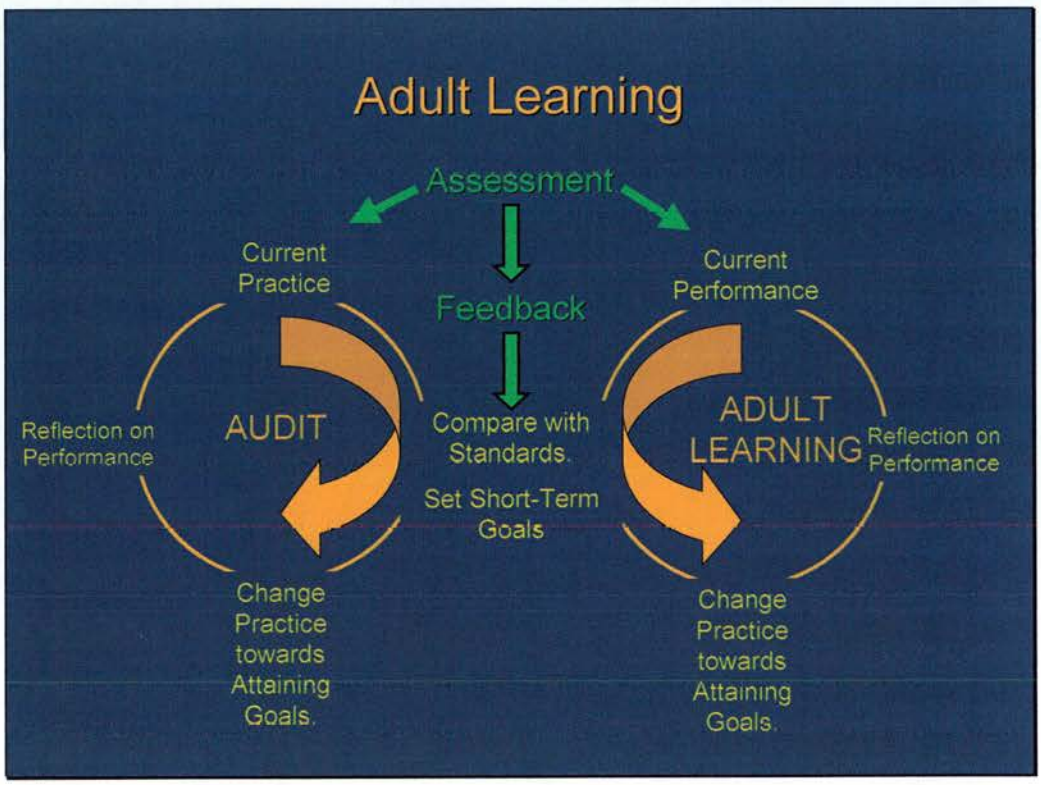
(Avermaete *et al* 1998)

Table I.3.
Anaesthetists Non-Technical Skills (ANTS).

Category	**Element**
Task Management	Planning and preparing
	Prioritising
	Providing and maintaining standards
	Identifying and utilising resources
Team Working	Co-ordinating activities with team members
	Exchanging information
	Using authority and assertiveness
	Assessing capabilities
	Supporting others
Situation Awareness	Gathering information
	Recognising and understanding
	Anticipating
Decision Making	Identifying options
	Balancing risks and selecting options
	Re-evaluating

(Fletcher *et al* , 2001; Fletcher *et al*, 2003)

Figure I.1.
Adult Learning and the Audit Cycle – One and the Same.



modified from (Kolb & Fry 1975)

Section II.

MATERIALS AND METHODS.

II.1. PRECEDING WORK.

The Edinburgh Basic Surgical Trainee Assessment Form (EBSTAF) was developed using a modified two-stage Delphi technique by the consultation of a total of 111 consultant surgeons in the Southeast of Scotland (Baldwin *et al*, 1999). Each main surgical subspecialty was represented.

Stage I required them to anonymously identify technical skills and cognitive or personal attributes that they would expect of a trainee who had been working in their own unit for a period of six months. They were also asked to identify specific procedures that such a trainee should be capable of completing unsupervised at the end of the same time period. Responses were received from 68 consultants (61% response rate). The identified qualities were separated into five domains: communication with patients and relatives; application of knowledge; team-working skills; clinical skills; and technical skills.

Stage II employed a second anonymous questionnaire, returned to all the original 111 consultants, listing the identified attributes and asking them to rank each as 'essential' (4), 'important' (3), 'useful' (2) or 'irrelevant' (1). Responses were received from 78 consultants (70% response rate) across all specialties and all skills deemed to be 'irrelevant' were removed from the final list.

The EBSTAF form consisted of nine sections, as seen in Appendix Section 1. Sections 1-5 [I. Communication; II. Application of Knowledge; III. Teamwork, IV; Clinical Skills; V. Technical Skills] addressed the 70 skills and attributes common to all surgical specialties and considered reasonable to expect of the surgical trainee. Section 6 addressed specialty-specific operative exposure whilst sections 7 and 8 consisted of visual analogue scales allowing the assessor to indicate their overall impression of the trainee and their working relationship respectively. A final section allowed additional comments to be made.

Assessment of trainee in-post performance by multi-disciplinary assessors (comprising medical, nursing and secretarial staff) using EBSTAF was then evaluated amongst all BSTs in the Southeast of Scotland basic surgical training programme over the subsequent 18-month period. EBSTAF was shown to be feasible, reliable, internally consistent and construct valid following detailed analysis (Paisley *et al*, 2001a; Paisley 2002).

II.2. PREDICTION OF CAREER PROGRESS.

II.2.a. Data Collection.

The original cohort of 36 BSTs were located and contacted by letter and/or telephone in an effort to determine their career pathway since leaving the Southeast Scotland BST programme.

For the purposes of this study, career progression was defined as gaining an SpR number or postgraduate research likely to lead to the same (an appropriate MD or PhD). This was based upon McManus's premise that medical (and therefore surgical) careers are hierarchical with the speed of progression and attainment of postgraduate qualifications being indicative of success with those who are slow to progress being likely to realise their potential less (McManus *et al*, 2003). Career progression was analysed at both 1 year and 2 ½ years and trainees were stratified into 'Fast Track' (successful attainment of SpR number or research) and 'Slow Track' (not yet attained SpR number but still in surgery and those leaving surgery) groups at the two time-points. Trainees who were no longer following a career in the surgical (or allied) specialties were also separately identifiable within the slow track group.

The original trainee codes were then broken to identify trainees and allow corresponding scores from assessments by EBSTAF to be reviewed. Median

EBSTAF scores for each domain and EBSTAF overall were calculated from medical assessments (i.e. consultant & SpR), nursing staff, multidisciplinary assessment by all assessors (except the trainee themselves) and trainee self-assessment. Median visual-analogue scores for 'overall impression of trainee' were also determined as above.

II.2.b. Data Analysis.

Analysis was completed using two strategies: first, that slow track trainees would score significantly lower than those in the fast track group (i.e. lower EBSTAF scores predict slow career progression); second, that trainees who subsequently left surgical training were previously identifiable using EBSTAF.

Differences between slow and fast track groups were examined by domain, overall score and VAS across assessor groups using Mann Whitney U test with significance assigned to $p< 0.05$.

Identification of trainees leaving surgery was by their ranking within their entry cohort expressed by quartile, the lowest quartile arbitrarily denoted by 1 and the highest denoted by 4.

II.3. ACCEPTABILITY OF EBSTAF TO BSTs.

II.3.a. Data Collection.

The second stage of the previous Delphi technique was repeated with ranking questionnaires sent to all BSTs in post on the Southeast Scotland Basic Surgical Training scheme. This was carried out between July and August of 2001 so as to include trainees just about to leave the program and those who had just joined. Experience therefore ranged from completely naïve surgical trainees to those who had completed 2½ years training and who were about to take up research or specialist registrar posts.

Questionnaires listed the fields from the EBSTAF by domain, but in jumbled order to remove any inferred field ranking from the order of the form itself, as seen in Appendix Section 2. In common with the initial consultant questionnaire, trainees were asked to indicate for each skill whether they considered it to be essential (4), important (3), useful (2) or irrelevant (1). They were also requested to list any attributes that they felt had been omitted from the EBSTAF, allowing them to express their own opinion, independent of consultants, on what skills they felt might favour a successful surgical career. Return of completed questionnaires was requested within two weeks of receipt and trainees were reminded as required by mail, by phone-call and finally personal visit at four, six and eight weeks respectively.

II.3.b. Data Analysis.

II.3.b.i. Estimation of domain importance.

Responses for each task were summed to create a single score for each domain, regarded as an indication of an individual's overall assessment of the importance of that domain. Although having no arithmetic meaning (i.e. a score of 36 is not twice as good as a score of 18), this did allow a summary score to be developed for each domain. This was then expressed as a percentage of the maximum possible score for that domain (i.e. number of fields therein multiplied by 4 (essential weighting)). The median scores from consultant and trainee groups (designated *Median %*) were then compared by Mann-Whitney statistics.

II.3.b.ii. Internal consistency.

Internal consistency for each domain and for EBSTAF overall for both trainee and consultant groups was determined by the estimation of Cronbach's alpha (α).

II.3.b.iii. Agreement.

For each field, a median ranking from BST and Consultant groups (the latter taken from the previous study data (Baldwin *et al*, 1999)) was determined and compared in turn. The number of fields where median ranking by

consultants and trainees agreed precisely was expressed as a percentage of the total, designated *Exact % Agreement*. Similarly when opinion differed, the number of fields assigned greater or lesser value by trainees was designated *Exact % Greater* and *Exact % Lesser* respectively. This equates to examination of a simple frequency table but may be criticised since it takes no account of the effect of chance. Thus further analysis was made by the estimation of weighted kappa (κ) with the limitations of this technique subsequently discussed.

II.4. ASSESSMENT OF PERFORMANCE IN PRACTICE.

II.4.a. Trainees.

EBSTAF was applied to all SHOs in-post or joining the Southeast Scotland BST programme over a 2-year period from August 2000 to August 2002. Stand-alone SHOs not on the BST rotation were not included in the study.

II.4.b. Southeast Scotland BST programme.

Over the above time period, a number of changes occurred in the structure of the Southeast Scotland BST rotation with the number of SHO posts being increased from an initial 30 to 36. All major surgical specialties were represented: general surgery, orthopaedics, cardiothoracic, neurosurgery, paediatric surgery, plastics surgery, urology, and vascular surgery, with the later addition of ENT surgery. The rotation encompassed a total of nine different hospitals: (Old) Royal Infirmary, Edinburgh; New Royal Infirmary, Edinburgh; Western General Hospital, Edinburgh; Princess Margaret Rose Orthopaedic Hospital, Edinburgh; Royal Hospital for Sick Children, Edinburgh; City Hospital, Edinburgh; St. John's Hospital, Livingston; Queen Margaret Hospital, Dunfermline; and Victoria Hospital, Kirkcaldy.

The Southeast Scotland BST programme lasted 2½ years and was divided into five separate six-month specialty posts. All trainees entering the

programme would spend their first year in general surgery, including a period of 2 or 3 months in the intensive care unit (ICU). The first year was then followed by six months in orthopaedics. The final year comprised two specialties determined by trainee career intentions and availability. Individual posts were subdivided according to unit structure with trainees spending a minimum of two months in each subunit.

II.4.c. Assessment.

Each six-month post denoted an individual assessment period for each SHO. The assessment protocol was identical throughout the rotation. Each time a trainee changed unit, an assessment episode involving evaluation by a number of multi-disciplinary assessors was completed using EBSTAF. If during a six-month post the trainee worked in more than one sub-unit within that specialty, a full assessment episode was completed at the end of his/her time in each sub-unit. Trainees therefore underwent 1, 2 or 3 assessment episodes in every six-month specialty post.

Assessment forms were individualised according to trainee and assessor discipline (i.e. trainee (self-assessment), surgical (consultant or SpR) or nursing staff from each clinical area. Each form was completed by a single assessor drawing upon their own observations of the trainee and specifically not the opinion of others. If the assessor had not observed the trainee performing a particular assessment task it was emphasised that they should

select the appropriate 'cannot be assessed by me' response. This was explained during the initial personal visit and reiterated both in the letter accompanying each form and on the form itself.

All assessment forms were anonymous to trainee identity once completed, assessors being required to remove the front sheet upon which the trainee was identified. The trainee was thereafter identifiable only by trainee code number. Each form also included assessor information in the form of name, grade and clinical area along with date of assessment completion.

In contrast to previous work, assessors were encouraged to make detailed additional comments throughout the assessment form. They were advised that these would be fed back to the trainee in a wholly anonymous fashion in order to help them identify their shortcomings and address them accordingly. In order to maintain this anonymity, no discussion of the form was permitted between assessor and trainee (in direct contrast to the intercollegiate assessment form). Assessors were also assured that no documentation of EBSTAF assessments would enter the trainee's official training record.

II.4.d. Assessors.

Potential assessors were identified from the previous study and through discussion with the 'Consultant with Administrative Responsibility' (CAR) and

senior nursing staff from each clinical area. Assessor identity remained unknown to trainees.

Each assessment episode involved a consultant and SpR who had had close contact with the trainee and felt able to complete a fair and accurate assessment. Similarly, one nursing staff assessor (F-grade or above wherever possible) was drawn from each area of trainee clinical practice to include home ward, home theatre, emergency ward, emergency theatre, intensive care unit (ICU), high-dependency unit (HDU), outpatients' departments (OPD) and daybed unit (DBU).

II.4.e. Study protocol.

At the beginning of each six-month period, the names of Southeast Scotland BSTs and their attachments was obtained from the chairman of the BST rotation (SPB). Each new trainee was then contacted either in person or by telephone in order to (i) provide information on the study and the study protocol, (ii) clarify how the next six-months would run for each individual and (iii) obtain their consent for inclusion in the study. Potential assessors were also approached early on in the six-month period to confirm that they were willing to take part in the study and had knowledge of their respective trainee(s). The completion of the assessment form was explained, if unfamiliar, and a reference copy provided.

Assessment forms then were distributed to all assessors three weeks prior to the trainee leaving the unit with the request that they be returned by the end of the trainee's attachment. Reminders were made to non-responders by telephone, letter and personal visit, with those forms being returned more than four weeks late being excluded from subsequent analysis.

II.4.f. Score generation.

Each field received a transformed score based on the median importance weighting from the original consultant survey ('essential' = 4, 'important' = 3, 'useful' = 2). Ratings' scores ('competent' = 3, 'more practice needed' = -2, 'unable to complete task' = -3, 'not observed' = 0) also corresponded to the original study to allow comparison. Fields were summed to give domain and overall scores, expressed as a percentage of maximum possible score for directly observed tasks. The maximum score therefore depended upon how many tasks were directly observed by the assessor with only scores generated from >75% observation considered to be a valid assessment to be included in subsequent analysis.

II.5. STRUCTURED FEEDBACK.

II.5.a. Pre-existing appraisal process.

At the time of the study, SHOs joined the Southeast Scotland BST programme on a twice-yearly basis in February and August. SHOs already within the programme rotated post / specialty on the same day. Four to six weeks later, the chairman of the rotation appraised all SHOs. This addressed how the trainee was settling in, past performance (evidenced by operative logbook and intercollegiate trainee assessment form) and future specialty preferences, whilst targeting any problems that may have become apparent.

II.5.b. Structured feedback document generation.

Completed EBSTAF forms returned within four weeks of the trainee leaving the unit were combined into a single feedback document for each trainee comprising three parts.

II.5.b.i. General domains.

All individual ratings for each field were included in the feedback document. Assessor ratings were denoted by a black cross (✕) whilst the trainee's own self-assessments were denoted by a red cross (✕) to clearly illustrate any difference between the two.

II.5.b.ii. Visual analogue scales.

Sections 7 and 8 of EBSTAF addressed the overall impression of the trainee and their working relationships using visual analogue scales. Ratings of overall impression were summarised as mean percentage scores from all assessors (exclusive of self), consultant assessor(s) only and self-assessment. Ratings of working relationships of the trainee with consultant(s) and all assessors were also provided.

II.5.b.iii. Comments.

Assessor comments were collected together within the final part of the feedback document along with those of the trainee. Assessors' identities were safeguarded by the removal of elements that may have identified the clinical area concerned whilst still retaining the underlying point of the comment. Comments addressing a specific event were avoided. Comments were not duplicated on the feedback document, an example of which may be seen in Appendix Section 3.

II.5.c. Appraisal and structured feedback.

SHO appraisals took place as before the study period. Only once the standard appraisal, including decisions with regard to future posts, was

complete was additional time then taken to examine the EBSTAF Feedback Document and identify areas of strength and weakness in the trainee's performance. Goals were then set for the next six-month period. No information from the EBSTAF Feedback Document was entered into the trainee's official record.

II.5.d. Examination of the effect of the feedback process.

It was considered both impossible and inappropriate to randomise the current trainees to feedback or no feedback by EBSTAF. However, the cohort of trainees that took part in the initial development of EBSTAF had received no feedback from their assessments in an effort to accurately examine the psychometric properties of the form. The original cohort was therefore used as the no-feedback group, while the present cohort made up the feedback group. The two groups were compared by examining the distributions of assessment scores after 12 months surgical training using the Mann-Whitney U test. A significant result (denoted by $p< 0.05$) would suggest that any such difference would be due, at least in part, to the structured feedback process.

II.5.e. Trainee assessment of the structured feedback process.

All trainees attending appraisal at the end of the study period were asked to complete an evaluation form addressing the feedback process itself (Appendix Section 3). Visual analogue scales addressed the usefulness,

fairness and level of the feedback and whether trainees found it threatening to be assessed by colleagues or nursing staff. They were also asked to indicate which part of the feedback document (general domains, visual analogue scales or comments) they found most informative along with any suggestions as to how the feedback might be improved.

II.6. HUMAN PATIENT SIMULATION (HPS)

II.6.a. The Scottish Clinical Simulation Centre

The Scottish Clinical Simulation Centre (SCSC) is a purpose-built national training facility located within Stirling Royal Infirmary. It houses Scotland's only high-fidelity human patient simulator, a METI-HPS manikin, which is capable of simulating programmed clinical scenarios with a high degree of realism. The manikin produces physiological signals that are detectable by routine and unmodified monitoring equipment; thus electrocardiograph (ECG), peripheral oxygen saturations (SpO_2), non-invasive blood pressure (NIBP) and arterial blood pressure, central venous pressure (CVP) and even pulmonary artery (PA) pressures may be measured. The simulator is housed in an environment capable of mimicking appropriate and realistic surroundings such as Accident and Emergency (A&E), the ward space, the High-Dependency Unit (HDU) or the operating theatre. A computer controls the manikin with in-built physiological and pharmacological programs allowing it to react to drugs, intravenous fluids and other interventions in a dynamic real-time manner.

Multiple cameras and radio-microphones are used to record all events during a clinical scenario to videotape as a composite with simultaneous physiological monitoring. This is then used as part of the post-scenario debriefing.

Originally set up as an anaesthetic training centre, SCSC has diversified to offer scenario-based training to doctors, dentists, nurses and other allied health professional involved in the delivery of acute medical, dental or surgical care. The course described was the first to be introduced for surgical trainees. In keeping with the philosophy of the simulation centre, it was primarily designed as an educational exercise to allow surgical trainees the opportunity to manage acute perioperative problems in a safe environment with structured debriefing using video to allow reflection on action thereafter. However, the fact that multiple trainees attended the course and took part in the same scenarios allowed the assessment of the reliability and construct validity of HPS itself in the assessment of BST critical care skills. The impact of the debriefing process could also be examined. Further, comparison with parallel multidisciplinary in-post assessment of the same trainees would allow a determination of concurrent validity.

II.6.b. Trainees.

The course was originally designed for 35 southeast Scotland (SES) BSTs to allow examination of concurrent validity by parallel in-post assessment of everyday practice using EBSTAF. However, news of the course quickly spread by word-of-mouth and a number of trainees from West of Scotland (WoS) and Tayside regions independently approached the primary researcher (PJD) and requested the opportunity to attend. With the full

support of the Chairs of their training programmes, extra courses were made available.

It was emphasised to participants that the courses were part of on-going research and that their performances would remain confidential from programme directors. No documentation of attendance, non-attendance or performance would enter trainees' official records. As scenarios were to be repeated for each course, trainees were asked not to divulge scenario details to future participants.

A minimum of 2 and maximum of 4 trainees were to attend each course. Places were offered on a first-come first-served basis with priority given to trainees from the southeast Scotland to maximise numbers for comparison with parallel in-post assessment.

II.6.c. Faculty.

Faculty was made up of consultants drawn on a non-regional basis from anaesthetic and surgical disciplines. A minimum of 1 consultant surgeon and 1 consultant anaesthetist was required for each day along with a research fellow (PJD), SCSC personnel and a number of observers who were involved in scenarios when appropriate.

II.6.d. Course structure – Figure II.1.

II.6.d.i. Orientation.

Trainees were first introduced to the faculty and fully orientated to the simulation centre, the HPS and its environment. The confidentiality of both the contents of the course and trainees' performances was again emphasised.

II.6.d.ii. Clinical Scenario.

Each trainee was required to manage their own scenario taking on their usual role of middle-grade surgeon-on-call in a district general hospital with houseman and nursing assistance. For the purpose of the course this was termed the "hot seat". Their consultant, played by the faculty consultant surgeon within the control room, was contactable by phone within the simulator room. Help and advice from other secondary specialties was available on request while tertiary specialties, such as neurosurgery, were 30 miles away and also contactable by phone.

A second trainee was available to help but remained unaware of events until contacted, termed the "jump seat". The remaining trainees observed proceedings via a live composite display of events and patient vital signs in an adjoining room.

Faculty were positioned behind one-way glass in the control room from where they were involved in both the running of the scenario and subsequent assessment of the trainee's performance.

II.6.d.iii. Pre-debriefing assessment.

Upon completion of the scenario and before any discussion, trainees (other than the jump seat trainee) and faculty completed a pre-debriefing assessment form.

II.6.d.iv. Debriefing.

Debriefing of a trainee's performance followed a semi-structured format based upon Pendleton's rules of feedback (Pendleton *et al* 1984). Trainees (hot seat and observers) were first asked what they felt had gone well. They were then asked to detail areas that they felt had gone less well and how they could be improved. Time was then taken to examine the hot-seat trainee's decision-making and situation awareness during the scenario, reviewing the underlying pathophysiology as necessary. Final reflection allowed trainees to comment on how they might better approach a similar clinical situation in the future.

II.6.d.v. Post-debriefing assessment.

After the debriefing, a post-debriefing assessment form was completed by everyone to allow comparison with pre-debriefing views.

II.6.d.vi. Course evaluation.

At the conclusion of the day's course, participants were asked to complete a course evaluation form and state what they would likely change in their everyday clinical practice as a result of having attended the course. Trainee demographics were also obtained including experience level and courses attended before coming to SCSC.

II.6.e. Assessment methods.

Assessment within the simulator was to be based on EBSTAF. A simulator assessment score sheet was constructed from 29 EBSTAF attributes that were felt on prior discussion to be potentially observable on the simulator. Three additional fields (namely 'takes command of the situation', 'identifies the problems appropriately' and 'manages the problems appropriately') were also included. The resulting global assessment (GA) was graded in identical fashion to in-post assessment by EBSTAF as 'competent', 'needs more practice', 'unable to perform'. For a skill to be assessed it had to be observed and so a fourth option of 'not applicable' was offered. Visual-analogue scales

addressed Communication (VASComm) and Clinical Skills and Management (VASCS&M). The simulator assessment form can be seen in Appendix Section 4.

II.7.e. Clinical scenarios.

Four scenarios were designed to illustrate specific learning objectives arising from challenging everyday clinical cases involving critical and perioperative care. Each scenario was planned to last 20 minutes and was developed by a consultant surgeon (SPB), consultant anaesthetist (NJM) and research fellow (PJD). Scenarios were recoverable but had the potential for the death of the patient as a result of major error or inactivity. Two scenarios were based within the resuscitation room of Accident & Emergency (A&E); another two took place in the High Dependency Unit (HDU) (see Appendix Section 4).

II.6.f. Data analysis.

GA assessments were transformed as previously described (section II.4.f.) to produce a percentage based upon observed skills only. Visual analogue scores (VASComm and VASCS&M) were expressed as percentages of total line length. EBSTAF in-post assessments were drawn from the current post for each trainee at the time of their attendance on the course and were transformed in identical fashion.

A breakdown of trainee data analysis is shown in Figure II.2. Data from all trainees were used in the analysis of reliability and the effect of debriefing. Trainee experience differed significantly between training schemes due to different pathways within the three programmes such that construct validity analysis was undertaken on SES trainees only. Only SES trainees underwent in-post assessment by EBSTAF for the examination of concurrent validity.

II.6.f.i. Feasibility.

The feasibility of the new course was determined by the surrogate measures of trainee attendance (and demand) and post-course evaluation.

II.6.f.ii. Reliability.

The reliability of HPS assessment scores was examined by estimation of internal consistency (Cronbach's α) and Inter-rater agreement (Spearman rank).

II.6.f.iii. Construct Validity.

For the construct validity of HPS to be demonstrated, experienced trainees should score more highly than those of less experience. The relationship between trainee scores and experience was therefore examined by Kruskal-Wallis (KW) analysis of variance and Mann-Whitney U (MWU) test.

II.6.f.iv. Concurrent validity.

The relationship between assessment of trainees by HPS (GA, VASComm and VASCS&M) and by in-post assessment by EBSTAF (by domain and overall) was examined by Spearman rank correlation and Kendall's concordance (*tau-b*).

There was some concern that relationships apparent between EBSTAF and GA could result from the fact that GA was itself derived from EBSTAF. In an effort to quantify such an effect, a composite assessment based upon the fields included in GA but taken from in-post assessment by EBSTAF was produced (PseudoSim). This was also included in HPS-clinical comparisons by Spearman rank correlation and Kendall's concordance (*tau-b*).

II.6.f.v. Effect of debriefing.

Trainee and trainer scores before and after debriefing were compared by Wilcoxon signed-rank matched-pairs test.

II.7. ASSESSMENT OF BST OPERATIVE SKILLS

II.7.a. Real-Time Assessment (RTA).

Eleven SHOs from the Southeast Scotland BST programme were video-recorded as they performed elective open mesh-repair of inguinal hernia 6 months and 12 months following commencement of general surgical training. Procedures were supervised and assisted by a Consultant or post-CCST Specialist Registrar who was asked to give minimal assistance unless absolutely necessary, in order to facilitate assessment of the trainee's own performance rather than their ability to follow direction. If the trainee was unable to continue at any point, the supervisor was permitted to take over the procedure until such time that the trainee could resume operating. Immediately following the procedure, trainee and supervisor completed an assessment form (EBSTAF-Tech) with an additional visual-analogue scale (VAS) to denote overall performance.

II.7.b. Video Assessment (VA).

Video recordings of the above procedures were made using a digital video-camera with remote zoom facility (Sony DCR-PC120E, Sony Corporation, Japan) attached to the satellite operating lamp by means of a simple G-clamp. Patients were approached before surgery and asked to give their

specific consent for the video-recording of their hernia repair for the purposes of this study.

Each video recording was subsequently edited using Adobe Premiere 6.0 (Adobe Systems Incorporated, San Jose, California, USA) and a Dell Latitude C800 Laptop Computer (Dell Computer Corporation). The resulting assessment video ran from the time of initial incision to the division of the aponeurosis of *external oblique* and then from closure of the same to skin closure. In this way it was aimed to remove the procedure-specific aspect of the video and facilitate the assessment of generic basic surgical skills such as dissection, tissue handling, haemostasis, suturing and knot tying. Sections allowing the identification of the surgeon were edited out, as was the soundtrack, to maintain anonymity.

In addition to the trainee videos, two consultants were also recorded as they performed the same procedure and these were edited in identical fashion for the determination of construct validity. Videos of trainee (at 6 and 12 months) and consultant operators were chosen at random and duplicated within the video set to allow evaluation of reliability by test-retest correlation.

VHS videocassettes and corresponding spiral-bound scorebooks were supplied to nine consultants and seven trainees. Videos were assessed using EBSTAF-Tech and VAS (as in RTA) with the addition of a modified Toronto Global Rating Scale of Operative Performance (Reznick, 1993;

Reznick *et al*, 1997) (hereafter referred to as Toronto). The use of the fast-forward facility was permitted once the assessor felt they had seen enough to make an accurate assessment of all observable skills.

Assessors were also asked to estimate the training level of each surgeon and state the basis upon which this was made along with the time taken to complete each assessment.

II.7.c. Score Generation.

II.7.c.i. EBSTAF and EBSTAF-Tech.

Scores for in-post assessment by EBSTAF and EBSTAF-Tech in both RTA and VA were generated as previously described (Paisley *et al*, 2001a), being expressed as a percentage of maximum observed score.

II.7.c.ii. Visual-Analogue Scale.

The position of the assessor's mark along the line was measured and expressed as a percentage of the total length.

II.7.c.iii. Toronto.

Seven assessment areas were taken from the Toronto Global Rating Scale. These examined 'Respect for Tissue', 'Time and Motion', 'Instrument Handling', 'Flow of Operation' 'Use of Assistants', 'Overall Performance' and 'Quality of Final Product'. Two areas (namely 'Knowledge of Instruments' and 'Knowledge of Procedure') would have required the use of sound footage and were therefore omitted. Each was graded 1 to 5 in order of increasing competence with mid-points and extreme anchored by specific descriptors to aid criterion referenced assessment. The total score was expressed as a percentage of the maximum possible score from the observed fields (max = 35) (Reznick, 1993).

II.7.d. Psychometric properties of video assessment of BST tissue-handling skills

II.7.d.i. Feasibility.

Feasibility of RTA and VA was determined by the surrogate measures of response rate and assessment time.

II.7.d.ii. Reliability.

Reliability was examined using three well-established methods. Internal consistency was determined using Cronbach's alpha. Inter-rater agreement was examined using intra-class correlation coefficient. Test-retest analysis was performed using Spearman's rank correlation.

II.7.d.iii. Validity.

II.7.d.iii. (a) Construct validity.

It has been hypothesised that performance should improve with training. Operative performance scores at 6 and 12 months surgical training were compared for both RTA and VA. Kruskal-Wallis and Mann-Whitney U tests were used to test for differences across the training levels (6 months, 12 months and consultant) whilst Wilcoxon matched pairs signed rank test was used to test for improvement in operator performance scores between 6 months and 12 months training.

II.7.d.iii. (b) Concurrent validity.

(i) Comparison to a Gold Standard.

The Toronto Global Rating Scale of Operative Performance has been extensively validated (Reznick *et al*, 1997) and may be considered to be a

Gold Standard. It was hypothesised therefore that correlation between this and EBSTAF-Tech would support the validity of EBSTAF-Tech used in video assessment of tissue-handling skills. Relationships were examined using Spearman rank correlation.

(ii) Comparison with in-post assessment.

Trainees involved in the video-assessment study were also evaluated in parallel by multidisciplinary EBSTAF assessment of performance in practice as previously described (section II.4). Consultant ratings of technical skill and overall performance for the six-month period preceding the assessment procedure were compared with those of RTA and VA using Spearman's rank correlation.

II.7.d.iii. (c) Trainer-Trainee agreement.

The relationships between trainer and trainee assessment scores of RTA and VA was examined using both Spearman's rank correlation and Kendall's .concordance (*tau-b*).

II.8. STATISTICAL ANALYSIS.

All statistical analysis was performed using the computer programme Statistical Package for Social Sciences (SPSS) for Windows version 10.0 and 11.0 (SPSS Inc; Chicago, Illinois,USA). Non-parametric methods were applied throughout with a p value <0.05 regarded as significant.

II.8.a. Chi square test.

The chi square test is a frequency table analysis of the relationship between two nominal variables. If the null hypothesis is true then an observation falling into one group of one variable does so independently of its grouping within the other variable. The test calculates the expected proportions for each grouping and compares them with the observed values. A p-value of less than 0.05 denotes a significant difference between the two variables.

II.8.b. Mann-Whitney U test (MWU).

The Mann-Whitney U test compares data from two independent groups. Observations are sequentially ranked as if taken from a single sample and if the null hypothesis is true then the distributions of both groups will be the same. Comparison of rankings rather than absolute values makes this test more resistant to the effects of extreme values. However, MWU is only valid when comparing two groups.

II.8.c. Kruskal-Wallis test (KW).

The Kruskal-Wallis test is a non-parametric analysis of variance that compares two or more independent groups by sequential ranking of all data and comparison of group distributions. It tests the null hypothesis that two or more groups have the same distribution against the alternative that at least one group has a different distribution. By the use of ranks rather than absolute values it is similarly resistant to outliers.

II.8.d. Wilcoxon signed-rank matched-pairs test (Wilcoxon).

This test compares paired continuous data with a non-normal distribution. Observations are paired if collected on a single sample over time or under different conditions. It is the equivalent of MWU for unpaired data and again uses ranks, conveying resistance to outliers.

II.8.e. Spearman's rank order correlation coefficient (Spearman's).

Spearman's allows comparison of paired continuous variables and again uses the ranks of the data rather than the numerical values. It is appropriate if at least one of the variables displays a non-normal distribution. It generates a correlation coefficient *rho* with values ranging from –1 (perfect negative /

inverse linear relationship) through 0 (no linear relationship) to +1 (perfect positive linear relationship).

II.8.f. Kendall's Concordance (*tau-b*).

Kendal *tau-b* is similar to Spearman's and is equivalent in statistical power. However, it expresses the difference between the probability that the observed data are in the same order for the two variables versus the probability that the observed data are in different orders for the two variables. Again, values range from -1 to +1.

II.8.g. Internal Consistency.

This is an indirect measure of reliability, examining the extent to which an individual scores similarly throughout an individual assessment. It is traditionally measured using Cronbach's alpha, which determines the overall correlation between individual items within a scale. A value of greater than 0.8 is considered acceptable for high-stakes assessment. However, α does have its limitations; a high value may reflect assessment of the same skill many times while a low value would result from different skills being assessed by different parts of the assessment tool, something that may actually be desirable when attempting to obtain an overall view of the assessee.

II.8.h. Intra-class correlation coefficient (ICC).

Intra-class correlation as applied to reliability analysis is orientated towards the estimation of inter-rater reliability and thus examines how consistently individuals are scored across multiple raters. A two-way average random effect consistency model with a 95% confidence interval was employed as advised by Nichols (Nichols ,1998).

II.8.i. Kappa statistic (κ).

The kappa statistic provides an accurate measure of absolute agreement whilst taking into account the proportion of agreement that might be expected purely by chance. It ranges from –1 (absolute disagreement) to +1 (perfect agreement) with 0 denoting random agreement no better than that expected by chance. A failing of the simple kappa statistic, however, is that it fails to examine the degree of disagreement such that *weighted kappa* may be used as an alternative. There is, however, no substitute for inspecting the table of frequencies as many different tables may yield a similar value of kappa. Kappa scores are graded as in Table II.1.

TABLE II.1

Grading of Agreement as determined by Kappa

Value of Kappa (κ)	Strength of Agreement
<0.20	Poor
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Good
0.80 – 1.00	Very good

(Altman , 1991)

FIGURE II.1

Surgical Trainee Critical Care Human Patient Simulator Course Structure.

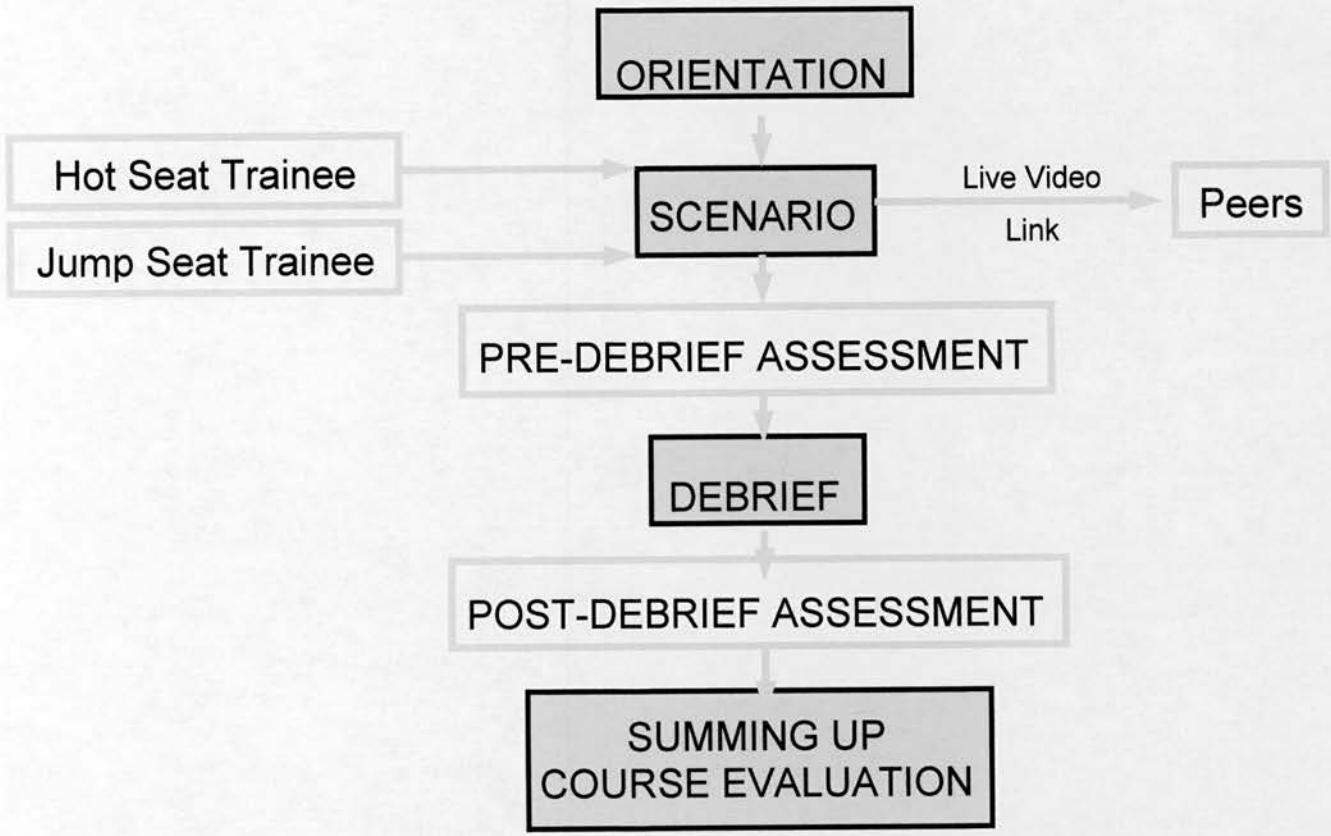
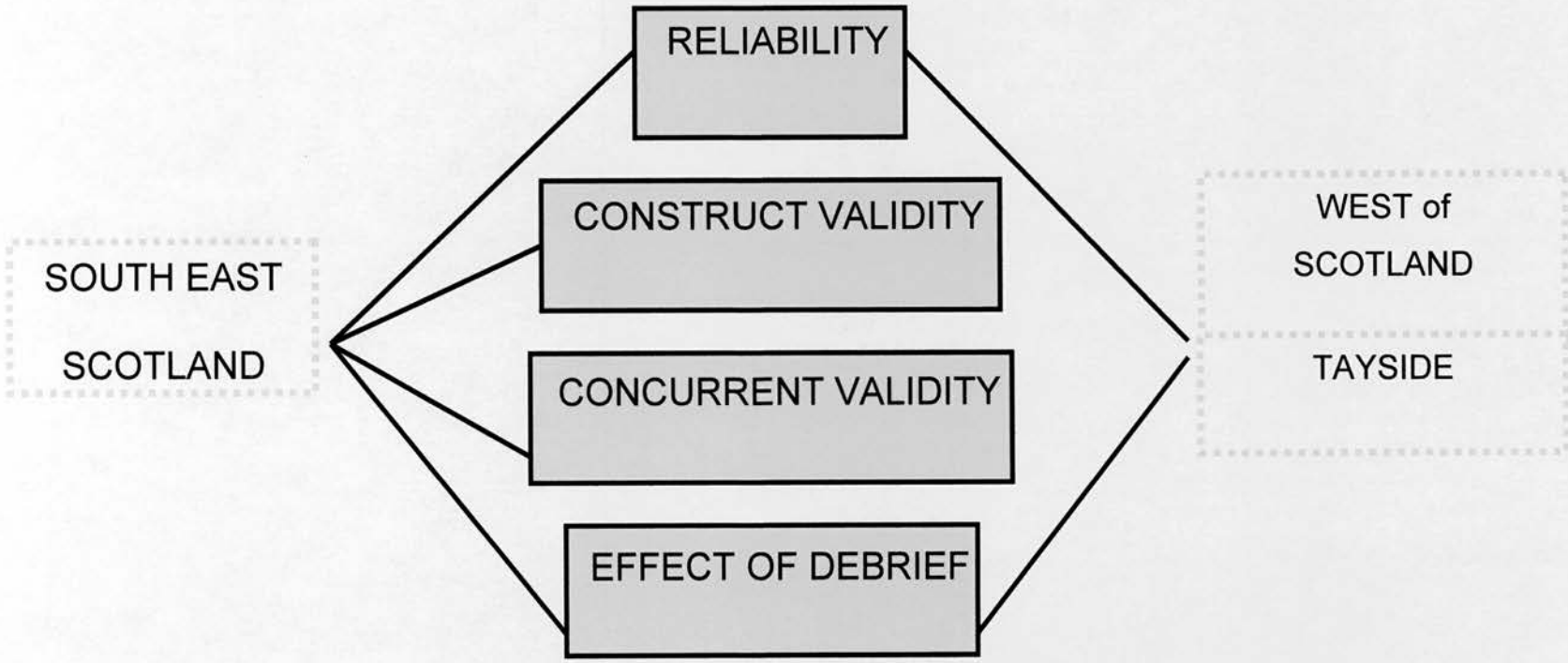


FIGURE II.2

Schematic Breakdown of Trainee Data Analysis.



Section III.

PREDICTIVE VALUE OF MULTI-DISCIPLINARY ASSESSMENT OF SURGICAL TRAINEES.

III.1. INTRODUCTION.

Although surgery demands high levels of cognitive, manual and inter-personal skills, the selection of surgeons remains an inexact science, hampered by a lack of clarity as to what makes a 'good' or 'bad' surgeon. Present methods of selection of trainees, based upon academic prowess, unstructured interview and personal reference appear to bear no relationship to career performance (Wingard *et al*, 1973; Keck *et al*, 1979; Lazar *et al*, 1980; Papp *et al*, 1997) and current cognitive tests may at best only predict future cognitive performance (Erlandson *et al*, 1982). However, EBSTAF offers a consensus view of the skills required of a surgical trainee (Baldwin *et al*, 1999) and may therefore relate better to future success.

In the past, studies were hampered by the lack of robust performance-based evaluations of trainees. Although now available in the form of Objective Structured Clinical Examinations, (OSCEs) and Objective Structured Assessment of Technical Skills (OSATs), these evaluations fail to address trainee performance outside a formal assessment environment. They therefore only examine the third level in Millers competency pyramid (Miller, 1990), that of demonstration. High-profile cases (Treasure, 1998) have highlighted the need for robust assessment of everyday practice but a significant part of a trainee's clinical practice still occurs in the trainer's absence; assessment by trainer alone is therefore inadequate. Multidisciplinary assessment of the everyday practice of surgical trainees by

EBSTAF may offer the best overall evaluation of a trainee's competence, having previously demonstrated the qualities of robust assessment (Paisley *et al*, 2001a), with the potential to uncover behaviours quite different from those displayed during formal assessment.

III.2. AIMS.

- To determine the career pathways of a previously studied cohort of BSTs.

- To examine the relationship between multidisciplinary in-post assessment of performance using EBSTAF and subsequent career progression.

- To examine whether BSTs who subsequently left surgery were identifiable by multidisciplinary in-post assessment using EBSTAF.

III.3. RESULTS.

III.3.a. Demographics of the Assessment Process.

III.3.a.i. Distribution.

Nine hundred and eighty four assessment forms were previously distributed to assessors in 117 distinct assessment episodes. Nine hundred and thirty eight forms (95%) were returned within the 4-week deadline. The SHO was not known to the assessor in 49 cases leaving 889 forms (90%) for further analysis.

Assessments were completed in all eight hospitals comprising the southeast Scotland surgical training programme at that time (Table III.1). No statistically significant difference in response rate was seen between hospitals (p=0.31, Chi square).

Assessments covered 8 different surgical specialties (Table III.2). No statistically significant difference in response rate was seen between specialties (p=0.91, Chi square).

III.3.a.ii. Trainees.

Thirty-six trainees were evaluated. Assessments evaluated 10 trainees (28%) over a single post, 13 trainees (36%) over 2 posts (i.e. 1 year) and 13 trainees (36%) over 3 posts (i.e. 18 months). The median number of assessment forms completed on each trainee was 26 (IQR 16-38), having completed between 2 and 36 months of surgical training.

III.3.a.iii. Assessors.

A total of 332 assessors were involved in the study. 229 assessment forms were distributed to medical staff, 540 forms to nursing staff, 98 to secretarial staff and 117 to the SHOs themselves. A further breakdown of assessors is seen in Table III.3.

III.3.a.iv. Assessment Episodes.

At each assessment episode, trainees were evaluated by between 4 and 12 assessors (median 9, IQR 7-10), determined by the structure of the unit. Twenty-six episodes (151 assessments) related to the preceding 2 months, 38 (287 assessments) to the preceding 3 months, 17 (185 assessments) to the preceding 4 months and 36 (361 assessments) to the preceding 6 months. A total of 75 six-month posts were examined.

III.3.a.v. Response Rates and Validity of Assessments.

There was no significant difference in response rate by the 4-week deadline between the assessor groups (Table III.4). However the validity of assessments varied widely between assessor groups and EBSTAF domains (Table III.5).

III.3.b. Career Progression.

All thirty-six SHOs were contacted by personal visit, telephone, email or post to determine current position and specialty. As a result, follow-up was 100%.

III.3.b.i. At One Year.

Twelve months after leaving the southeast Scotland basic surgical training programme, 25 trainees (70%) had attained either an SpR training post or postgraduate research with a view to completing a PhD or MD. Eight trainees (22%) remained within surgery but had yet to achieve SpR or postgraduate research posts. Three trainees (8%) had left surgery altogether. There were therefore 11 trainees that were considered non-achievers at this time point.

III.3.b.ii. At Two and a Half Years.

Two and a half years after leaving the southeast Scotland basic surgical training programme, 27 trainees (75%) had attained an SpR training post or remained in postgraduate research. Six trainees (17%) remained within surgery but had still not achieved an SpR number and the same 3 trainees (8%) had left surgery altogether. The 'non-achiever' group therefore contained 9 trainees at this time point.

III.3.c. Relationship Between Career Progression and In-Post Assessment Scores.

III.3.c.i. Career Progression At One Year – Figures III.1 a to g.

III.3.c.i. (a) Assessment by Medical Staff – Table III.6a

Trainees not achieving SpR or postgraduate research posts by one year post BST scored significantly lower in 4 EBSTAF domains (namely Knowledge, Teamwork, Clinical Skills and Technical Skills) as well as in EBSTAF Overall Score and Visual-Analogue Score. The EBSTAF domain of Communication failed to reach significance ($p=0.055$).

III.3.c.i. (b) Assessment by Nursing Staff – Table III.6b.

No difference was seen between the two groups of trainees in their assessments by nursing staff.

III.3.c.i. (c) Multi-Disciplinary Assessment – Table III.6c.

Multi-disciplinary assessment demonstrated a significant difference in scores between the two groups in the EBSTAF domain of Teamwork and for EBSTAF Overall Score and Visual-Analogue Score. Other EBSTAF domains failed to reach significance.

III.3.c.i. (d) Self-Assessment – Table III.6d.

No difference was seen between the two groups of trainees in their assessments of themselves.

III.3.c.ii. Career Progression At 2 ½ Years – Figures III.2a-g.

III.3.c.ii. (a) Assessment by Medical Staff – Table III.7a.

Trainees not achieving SpR or research posts by 2 ½ years scored significantly lower in 3 EBSTAF domains (namely Communication, Knowledge and Technical Skills) as well as in EBSTAF Overall Score and

Visual-Analogue Score. The EBSTAF domain of Teamwork failed to reach significance (p=0.074), as did Clinical Skills (p=0.076).

III.3.c.ii. (b) Assessment by Nursing Staff – Table III.7b.

No difference was seen between the two groups of trainees in their assessments by nursing staff.

III.3.c.ii. (c) Multi-Disciplinary Assessment – Table III.7c.

Multi-disciplinary assessment demonstrated a significant difference in scores between the two groups in the EBSTAF domains of Communication, Teamwork and Technical Skills. EBSTAF Overall Score and Visual-Analogue Score also demonstrated significant differences between the two groups.

III.3.c.ii. (d) Self-Assessment – Table III.7d.

A significant difference between the two groups in self-assessment of Technical Skills but this was not seen in other EBSTAF domains or EBSTAF Overall Score.

III.3.c.iii. Identification of Trainees Subsequently Leaving Surgery.

Trainees were ranked by quartile within each six-month BST post. The position of those trainees subsequently leaving surgery was then examined.

III.3.c.iii. (a) Assessment by Medical Staff – Table III.8a.

All three trainees who subsequently left surgery were placed in the bottom 25% (quartile 1) by EBSTAF overall and by VAS when first evaluated by medical staff. They were also consistently placed within the lower 50% (quartiles 1 and 2) across individual EBSTAF domains; only Trainee C was placed in the third quartile for the domain of Communication.

III.3.c.iii. (b) Assessment by Nursing Staff – Table III.8b.

Nursing staff assessments were not discriminatory with the majority of returned forms falling short of the 75% completion requirement for inclusion in further analysis. A particularly low rate was found for the assessment of Knowledge by nursing staff (see Table III.5).

III.3.c.iii. (c) Multi-Disciplinary Assessment – Table III.8c.

Multi-disciplinary assessment again placed all three trainees in the bottom quartile for EBSTAF overall and VAS. Across EBSTAF domains

multidisciplinary assessment strongly identified Trainees B and C while Trainee A was placed in the first quartile for Communication, Knowledge and Clinical Skills but in the third quartile for Teamwork and Technical Skills.

III.3.c.iii. (d) Self-Assessment – Table III.8d.

Self-assessment showed Trainees B and C to be self-critical; Trainee C placed himself within the lowest quartile for all but the Teamwork domain while Trainee B consistently scored himself as below average. In stark contrast, Trainee A consistently placed himself in the top 50%.

III.4. DISCUSSION.

Medical (consultant and SpR) and multidisciplinary assessments of BST performance using EBSTAF were able to identify trainees slow to progress in their future surgical career. Assessments made by nursing staff alone showed no relationship to career progression, suggesting their assessments to be unnecessary for such a purpose. It also suggests that the demonstrated relationship between career progression and multidisciplinary assessment may result purely from the incorporation of medical assessments. BSTs' self-assessments demonstrated no relation to career progression.

Looking at the medical assessments more closely, progression at 1 year appears to be related to all domains of EBSTAF *except* that of Communication, which just failed to reach significance ($p=0.055$). Overall and VAS scores also appear to be predictive at 1 year. This differs considerably from previous MSF studies where the biggest source of variance was interpersonal skills and communication (Linn *et al*, 1975; Davidge *et al*, 1980; Dielman *et al*, 1980; Maxim *et al*, 1987; Risucci *et al*, 1989; Ramsey *et al*, 1993). By 2 ½ years, only Communication, Knowledge and Technical skills appear to show any predictive value, suggesting that failings of Knowledge, Teamworking and Clinical Skills have either been addressed by further training or are no longer determinants of career progression. At this time point, Communication and Knowledge appear related to career progress, along with Technical Skill, in line with previous studies.

These results are encouraging since not only may EBSTAF demonstrate predictive validity but it may do so using far fewer assessors than was originally envisaged, considerably reducing the administrative burden. However, three issues need to be considered in accepting this conclusion:

Firstly, two of the trainees leaving surgery were assessed just before their departure while the third remained for only one further placement. The decision to leave was therefore made before the assessment and scores may reflect a lack of effort and motivation resulting from that decision rather than an inability to perform. Furthermore, it is difficult to know with certainty whether the decision to leave surgery was made on the basis of poor performance or simply a change of mind. It will no doubt prove more challenging to identify those individuals who lack insight into their failings (such as Trainee A who consistently awarded high self-assessment scores) and yet are destined to fail despite their persistent efforts. The demonstration of a predictive relationship between career progression and medical assessment suggests EBSTAF may be useful in identifying trainees who might benefit from additional guidance and feedback. This may include guidance away from a surgical career if no improvement is forthcoming before wasting further time, effort and NHS resources. However, the relationship is insufficiently strong to warrant selection (or expulsion) on the basis of EBSTAF assessments alone since other trainees who also fell into

the lowest quartile went on to progress satisfactorily in their subsequent surgical career.

Secondly, all trainees identified as 'slow to progress' but who remained in surgery eventually attained SpR. One has to therefore question whether slow progress as defined by this study has been to the trainees' detriment. Although the attainment of an SpR post is a reasonable marker of success for a basic surgical training scheme, it is no guarantee of becoming a consultant, or being successful once established in consultant practice. It may therefore be more valuable to consider follow-up at 10 years to determine whether BST performance assessed by EBSTAF truly relates to career success by determining whether trainees have indeed been slow to achieve consultancy and therefore realise their full potential, as defined by McManus (McManus *et al*, 2003). However, the implementation of Modernising Medical Careers (Department of Health 2003) will result in early selection to specialty-specific training programmes. This will certainly mean that slow progression early on will have a direct effect on a trainee's career progression if not their specialty. The early identification of trainees with the potential to succeed but who may benefit from additional support will clearly be even more valuable, both to them as individuals and to the NHS, if we are to avoid incorrect selection and the loss of individuals who may, in time, make a significant contribution to the profession.

The lack of predictive value to nursing assessments may simply reflect their lack of involvement in determining surgical trainee selection. Indeed, one might expect medical assessments to predict progression as one determines the other outside of the EBSTAF study. However, although It has been suggested that nurses perceptions of desirable characteristics differ considerably from those of doctors (Butterfield *et al*, 1990; Butterfield *et al*, 1991), this does not equate to them being irrelevant in a trainee's future practice, or to the well-being of his / her patients. Indeed, in Section V we demonstrate that trainees value the comments of nurses, feeling that they provide insight into their everyday practice. Initial concerns that nursing staff might be reluctant to assess doctors were not upheld in this study and this area of assessment certainly warrants further study.

Finally, the lack of predictive validity of trainee self-assessments comes as no surprise since they are well-known to be unreliable. It is likely, however, that the value of EBSTAF self-assessments lies in providing trainees with insight into their own performance and this is an area further examined in Section V.

Previous studies of the predictive validity of assessments of performance in practice have been disappointing. The majority have centred upon performance in A-levels or as medical students rather than surgical trainees and examine relationships between faculty evaluations and final examinations rather than subsequent career performance.

Only Schueneman addressed clinical performance ratings of surgical residents in relation to completion of surgical training. By examining over 4000 assessments of 199 surgical residents over 15 years, they were able to identify 78% of those trainees who were eventually ejected from surgical training whilst predicting final scores and rankings (Schueneman *et al*, 1994). However, this study can be criticised for the fact that the study assessments were completed by the same faculty as provided the final outcome measures. Furthermore, such a high drop-out rate would clearly be unacceptable to the structure of the NHS.

Ward evaluations of 32 American general surgical residents by supervisors and peers were shown by Risucci to moderately correlate with subsequent American Board of Surgery In-Training Examination (ABSITE) scores but interestingly factor analysis suggested the main determining factor in faculty ratings to be the interpersonal skills of the resident above ability (Risucci *et al*, 1989). In examining the psychometric properties of an OSCE for junior surgical residents, Cohen demonstrated only a low correlation with ward assessments (Cohen *et al*, 1990) while Schwartz found no such relationship between faculty ward evaluations and performance in either ABSITE, OSCE or structured oral examinations (Schwartz *et al*, 1994b).

The general lack of agreement between ward assessments and tests of knowledge highlights the difference between cognitive and behavioural

performance. Although knowledge is the foundation of competence, it does not predict behaviour. Ward assessments address the behavioural levels of clinical competence and it is at this level that trainees should be selected. A lack of knowledge can be easily identified and directly addressed. However, undesirable behavioural traits may be lifelong and detrimental to both patients and the profession, as so graphically illustrated by cases such as Shipman (Baker, 2004).

EBSTAF measures broad-based clinical performance in practice. It has been shown to demonstrate acceptable reliability and construct validity (Paisley *et al*, 2001a) and assesses aspects of surgical practice that are valued by both trainers (Baldwin *et al*, 1999) and trainees (see Section IV). This study suggests EBSTAF's ability to predict failure in the short term and suggests a degree of predictive validity that further supports the application of EBSTAF to the assessment of basic surgical trainees. Full evaluation of the predictive validity of EBSTAF requires follow-up in the longer term, relating EBSTAF assessments in BST to performance at consultant level. However, robust and generally accepted measures of consultant performance remain illusive and until available, such a study will be easy to criticise.

III.5. SUMMARY.

- EBSTAF assessment of in-post performance of BSTs by medical staff (consultants and SpRs) appears to identify trainees who subsequently struggle to progress in their surgical career. Nursing and trainee self-assessments were not similarly predictive.

- EBSTAF assessments by medical staff identified the 3 trainees who subsequently left surgery. Nursing and trainee self-assessments were not discriminative.

- Multidisciplinary assessment may not be necessary for the purposes of prediction of career success.

Table III.1

Distribution of Assessments by Hospital.

HOSPITAL	EPISODES	ASSESSMENTS
Royal Infirmary, Edinburgh	45	402
Western General Hospital, Edinburgh	30	214
St. John's Hospital, Livingston	17	175
Princess Margaret Rose Orthopaedic Hospital, Edinburgh	9	61
Royal Hospital for Sick Children, Edinburgh	5	44
Victoria Hospital, Kirkcaldy	5	40
Queen Margaret Hospital, Dunfermline	4	32
Eastern General Hospital, Edinburgh	2	16
TOTAL	**117**	**984**

Episode: Assessment of a single trainee at the end of a single post by multiple assessors.
Assessment: Evaluation of a single trainee at the end of a single post by a single assessor.

Table III.2

Distribution of Assessments by Specialty.

SPECIALTY	EPISODES	ASSESSMENTS
General Surgery	60	508
Orthopaedic Surgery	25	194
Neurosurgery	10	70
Plastic Surgery	6	63
Cardiothoracic Surgery	6	53
Vascular Surgery	5	50
Paediatric Surgery	4	36
Urology	1	10
TOTAL	**117**	**984**

Episode: Assessment of a single trainee at the end of a single post by multiple assessors.
Assessment: Evaluation of a single trainee at the end of a single post by a single assessor.

Table III.3

Distribution of Assessments by Assessor.

ASSESSOR	n =	ASSESSMENTS
Medical Staff	**113**	**229**
Consultants	51	117
Registrars	62	112
Nursing Staff	**143**	**540**
Home Ward	42	117
Emergency Ward	2	16
HDU	14	66
Theatre	34	120
Day Bed Unit	6	46
Outpatients'	30	99
A&E	15	76
Secretarial Staff	**40**	**98**
SHOs	**36**	**117**
TOTAL	**332**	**984**

Episode: Assessment of a single trainee at the end of a single post by multiple assessors.
Assessment: Evaluation of a single trainee at the end of a single post by a single assessor.

Table III.4

Response Rate for Assessor Groups.

ASSESSOR	Distributed	Returned by Deadline	Response Rate (%)
Medical Staff	229	213	93.0
Nursing Staff	540	522	96.7
Multidisciplinary	867	827	95.4
SHOs	117	111	94.9

Table III.5

Validity of Assessments by EBSTAF Domain for Different Assessor Groups.

EBSTAF DOMAIN	Medical	Nursing	Multidisciplinary	SHO
	n = 213	n = 474	n = 778	n = 111
Communication	160	360	532	107
	75.1%	75.9%	68.4%	96.4%
Knowledge	126	28	157	105
	59.2%	5.9%	20.2%	94.6%
Teamwork	146	200	367	105
	68.5%	42.2%	47.2%	94.6%
Clinical Skills	161	158	321	111
	75.6%	33.3%	41.3%	100%
Technical Skills	153	103	356	95
	71.8%	21.7%	32.9%	85.6%

The validity of each domain within an individual assessment was determined by greater than 75% of the fields therein being directly observed.

Table III.6a

Assessment by Medical Staff Grouped According to Career Progression at 1 Year Post BST

		Median	I Q R	M-W p =
COMMUNICATION	Group I	76.2	19.5 – 100	0.055
	Group II	35.2	-4.2 – 79.2	
KNOWLEDGE	Group I	50.6	7.2 – 87.1	0.040
	Group II	20.2	-48.4 – 54.1	
TEAMWORK	Group I	94.7	63.0 – 100	0.014
	Group II	68.0	23.0 – 91.7	
CLINICAL SKILLS	Group I	84.4	51.0 – 98.3	0.011
	Group II	41.7	18.2 – 74.8	
TECHNICAL SKILLS	Group I	48.2	14.6 – 87.4	0.035
	Group II	31.6	-28.1 – 59.1	
EBSTAF Overall	Group I	73.6	39.7 – 91.4	0.015
	Group II	42.5	-6.0 – 70.1	
Visual Analogue Score Overall Impression of Trainee	Group I	83.0	74.3 – 89.8	0.006
	Group II	77.5	63.0 – 81.8	

I Q R: Inter-Quartile Range. **M-W:** Mann Whitney U.

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

Table III.6b

Assessment by Nursing Staff Grouped According to Career Progression at 1 Year Post BST

		Median	I Q R	M-W p =
COMMUNICATION	Group I	100	81.5 – 100	ns
	Group II	100	72.2 – 100	
KNOWLEDGE	Group I	100	100 – 100	ns
	Group II	100	72.8 – 100	
TEAMWORK	Group I	100	86.1 – 100	ns
	Group II	94.7	54.9 – 100	
CLINICAL SKILLS	Group I	96.2	81.1 – 100	ns
	Group II	93.9	57.0 – 98.4	
TECHNICAL SKILLS	Group I	76.4	22.0 – 91.9	ns
	Group II	95.93	46.3 – 100	
EBSTAF Overall	Group I	87.7	75.0 – 96.0	ns
	Group II	93.9	31.5 – 100	
Visual Analogue Score Overall Impression of Trainee	Group I	79.0	75.0 – 86.0	ns
	Group II	77.0	74.3 – 84.8	

I Q R: Inter-Quartile Range. **M-W**: Mann Whitney U.

Group I: Trainees achieving SpR / Research. **Group II**: Trainees not achieving SpR / Research

Table III.6c

Multi-Disciplinary Assessment Grouped According to Career Progression at 1 Year Post BST

		Median	I Q R	M-W p =
COMMUNICATION	Group I	100	81.5 – 100	ns
	Group II	100	28.2 – 100	
KNOWLEDGE	Group I	58.3	28.6 – 88.1	ns
	Group II	30.6	-47.2 – 86.1	
TEAMWORK	Group I	96.7	78.3 – 100	0.028
	Group II	83.6	63.3 – 98.1	
CLINICAL SKILLS	Group I	91.2	80.2 – 100	0.056
	Group II	77.9	57.0 – 98.4	
TECHNICAL SKILLS	Group I	72.3	34.6 – 91.9	ns
	Group II	44.4	-14.0 – 87.9	
EBSTAF Overall	Group I	78.3	51.5 – 91.4	0.013
	Group II	42.5	-7.0 – 70.1	
Visual Analogue Score Overall Impression of Trainee	Group I	82.0	75.0 – 88.0	0.036
	Group II	78.5	68.5 – 83 0	

I Q R: Inter-Quartile Range. **M-W**: Mann Whitney U.

Group I: Trainees achieving SpR / Research. **Group II**: Trainees not achieving SpR / Research

Table III.6d

SHO Self-Assessment Grouped According to Career Progression at 1 Year Post BST

		Median	I Q R	M-W p =
COMMUNICATION	Group I	81.5	44.4 – 100	ns
	Group II	58.3	-15.2 – 100	
KNOWLEDGE	Group I	26.2	0.6 – 70.2	ns
	Group II	10.7	-37.5 – 48.6	
TEAMWORK	Group I	80.0	66.7 – 90.0	ns
	Group II	80.0	4.1 – 96.7	
CLINICAL SKILLS	Group I	80.2	57.6 – 94.8	ns
	Group II	30.6	-7.2 – 95.0	
TECHNICAL SKILLS	Group I	43.1	-9.25 – 70.0	ns
	Group II	7.36	-42.5 – 79.5	
EBSTAF Overall	Group I	64.4	36.0 – 80.9	ns
	Group II	22.7	-6.6 – 87.3	
Visual Analogue Score Overall Impression of Trainee	Group I	-	-	-
	Group II	-	-	

I Q R: Inter-Quartile Range. **M-W**: Mann Whitney U.

Group I: Trainees achieving SpR / Research. **Group II**: Trainees not achieving SpR / Research

Table III.7a

Assessment by Medical Staff Grouped According to Career Progression at 2½ Years Post BST

		Median	I Q R	M-W p =
COMMUNICATION	Group I	76.2	16.7 – 100	0.026
	Group II	35.2	-45.83 – 79.2	
KNOWLEDGE	Group I	50.0	15.3 – 86.1	0.040
	Group II	20.0	-51.4 – 53.2	
TEAMWORK	Group I	92.8	55.0 – 100	0.074
	Group II	74.9	42.7 – 94.9	
CLINICAL SKILLS	Group I	77.7	48.1 – 100	0.076
	Group II	57.6	23.8 – 86.6	
TECHNICAL SKILLS	Group I	51.2	14.5 – 87.5	0.006
	Group II	33.2	-44.3 – 45.6	
EBSTAF Overall	Group I	74.8	38.2 – 91.8	0.008
	Group II	42.5	-20.5 – 72.9	
Visual Analogue Score Overall Impression of Trainee	Group I	83.0	75.0 – 91.0	0.001
	Group II	77.0	60.0 – 81.5	

I Q R: Inter-Quartile Range. **M-W**: Mann Whitney U.

Group I: Trainees achieving SpR / Research. **Group II**: Trainees not achieving SpR / Research

Table III.7b

Assessment by Nursing Staff Grouped According to Career Progression at 2½ Years Post BST

		Median	I Q R	M-W p =
COMMUNICATION	Group I	100	81.5 – 100	ns
	Group II	100	47.9 – 100	
KNOWLEDGE	Group I	100	100 – 100	ns
	Group II	100	72.8 – 100	
TEAMWORK	Group I	100	89.1 – 100	0.098
	Group II	94.2	24.6 – 100	
CLINICAL SKILLS	Group I	94.3	81.1 – 100	ns
	Group II	100	66.8 – 100	
TECHNICAL SKILLS	Group I	83.7	18.0 – 100	ns
	Group II	78.7	44.6 – 97.0	
EBSTAF Overall	Group I	91.5	82.7 – 100	ns
	Group II	51.1	19.8 – 93.9	
Visual Analogue Score Overall Impression of Trainee	Group I	79.5	75.0 – 85.3	ns
	Group II	75.0	72.5 – 83.0	

I Q R: Inter-Quartile Range. **M-W**: Mann Whitney U.

Group I: Trainees achieving SpR / Research. **Group II**: Trainees not achieving SpR / Research

Table III.7c

Multidisciplinary Assessment Grouped According to Career Progression at 2½ Years Post BST

		Median	I Q R	M-W p =
COMMUNICATION	Group I	100	79.8 – 100	0.007
	Group II	59.5	27.1 – 100	
KNOWLEDGE	Group I	61.3	28.6 – 88.1	ns
	Group II	35.5	-49.3 – 86.4	
TEAMWORK	Group I	96.8	76.1 – 100	0.048
	Group II	86.7	63.4 – 97.5	
CLINICAL SKILLS	Group I	90.5	75.4 – 100	ns
	Group II	82.7	65.5 – 94.0	
TECHNICAL SKILLS	Group I	75.6	34.3 – 91.9	0.028
	Group II	44.4	-12.6 – 77.6	
EBSTAF Overall	Group I	81.6	46.2 – 92.3	0.002
	Group II	51.1	-20.4 – 70.6	
Visual Analogue Score Overall Impression of Trainee	Group I	82.0	76.5 – 88.0	0.013
	Group II	77.0	67.5 – 83.0	

I Q R: Inter-Quartile Range. **M-W:** Mann Whitney U.

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

Table III.7d

SHO Self-Assessment Grouped According to Career Progression at 2½ Years Post BST

		Median	I Q R	M-W p =
COMMUNICATION	Group I	81.5	44.4 – 100	ns
	Group II	46.4	-15.2 – 100	
KNOWLEDGE	Group I	28.6	-18.0 – 76.2	ns
	Group II	16.7	-10.9 – 37.5	
TEAMWORK	Group I	83.3	66.7 – 93.0	ns
	Group II	66.7	53.3 – 93.1	
CLINICAL SKILLS	Group I	80.4	57.6 – 95.8	ns
	Group II	59.3	1.3 – 94.0	
TECHNICAL SKILLS	Group I	47.2	5.0 – 74.6	0.028
	Group II	-17.9	-46.3 – 67.0	
EBSTAF Overall	Group I	66.0	35.4 – 84.8	0.064
	Group II	34.7	6.3 – 82.5	
Visual Analogue Score Overall Impression of Trainee	Group I	-	-	-
	Group II	-	-	

I Q R: Inter-Quartile Range. **M-W**: Mann Whitney U.

Group I: Trainees achieving SpR / Research. **Group II**: Trainees not achieving SpR / Research

Table III.8a

Medical Staff Assessment of Trainees Who Were To Subsequently Leave Surgery.

	Trainee A	Trainee B	Trainee C
COMMUNICATION	1	1	3
KNOWLEDGE	1	1	1
TEAMWORK	2	2	1
CLINICAL SKILLS	2	2	1
TECHNICAL SKILLS	1	1	2
EBSTAF Overall	1	1	1
VAS	2	1	1

Numbers denote ranking by Quartile within the entry cohort

(1 = lowest / worst, 4 = highest / best)

Table III.8b

Nursing Staff Assessment of Trainees Who Were To Subsequently Leave Surgery.

	Trainee A	Trainee B	Trainee C
COMMUNICATION	3	1	1
KNOWLEDGE	-	-	-
TEAMWORK	3	-	1
CLINICAL SKILLS	3	2	1
TECHNICAL SKILLS	3	3	-
EBSTAF Overall	-	-	-
VAS	3	3	1

Numbers denote ranking by Quartile within the entry cohort

(1 = lowest / worst, 4 = highest / best)

Table III.8c

Multidisciplinary Assessment of Trainees Who Were To Subsequently Leave Surgery.

	Trainee A	Trainee B	Trainee C
COMMUNICATION	1	1	1
KNOWLEDGE	1	1	1
TEAMWORK	3	1	1
CLINICAL SKILLS	1	2	1
TECHNICAL SKILLS	3	1	1
EBSTAF Overall	1	1	1
VAS	3	1	1

Numbers denote ranking by Quartile within the entry cohort

(1 = lowest / worst, 4 = highest / best)

Table III.8d

SHO Self-Assessment by Trainees Who Were To Subsequently Leave Surgery.

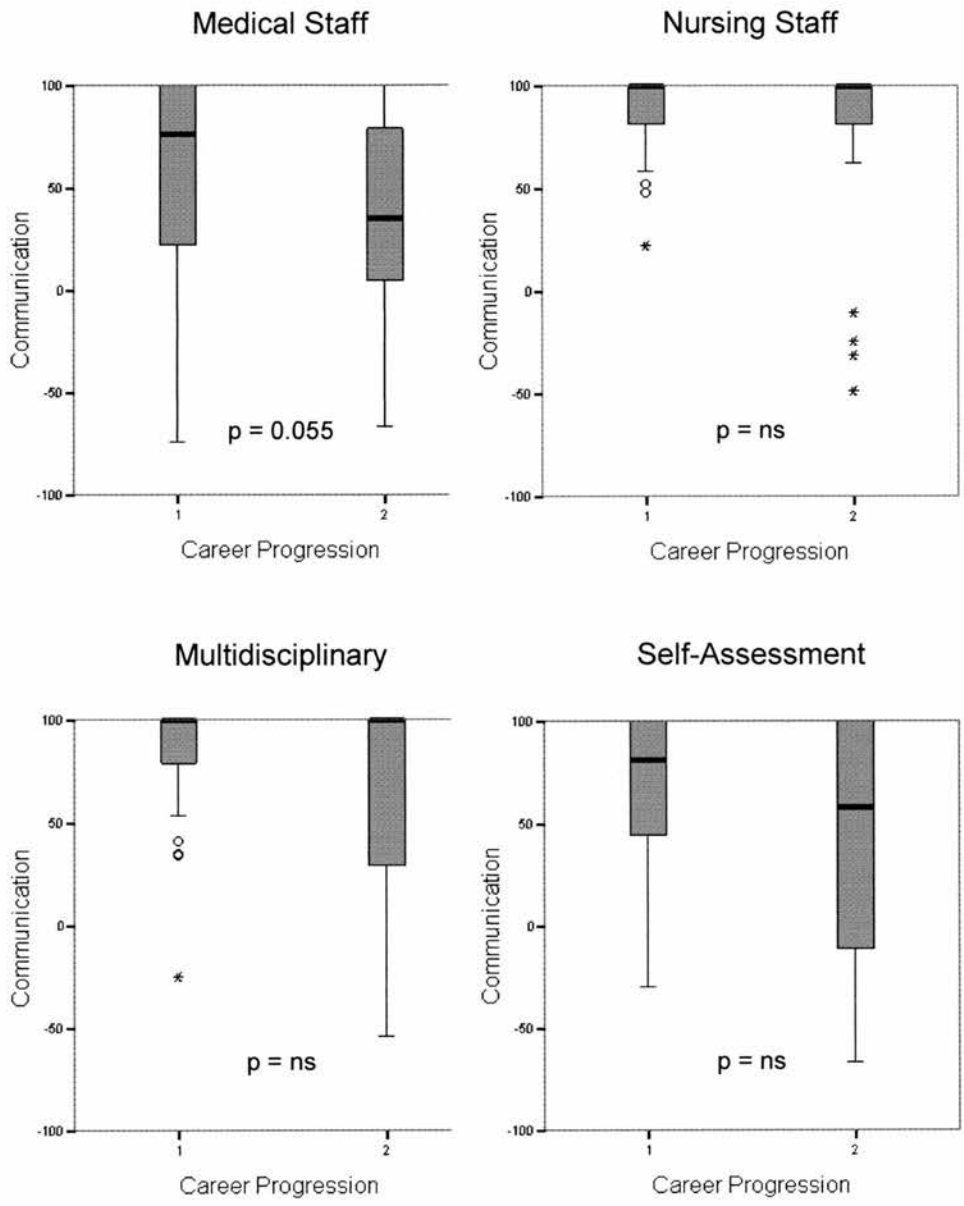
	Trainee A	Trainee B	Trainee C
COMMUNICATION	4	1	1
KNOWLEDGE	3	2	1
TEAMWORK	3	2	3
CLINICAL SKILLS	3	2	1
TECHNICAL SKILLS	3	2	1
EBSTAF Overall	3	2	2
VAS	n/a	n/a	n/a

Numbers denote ranking by Quartile within the entry cohort

(1 = lowest / worst, 4 = highest / best)

Figure III.1a

Career Progression at 1 Year related to EBSTAF Assessment of Communication



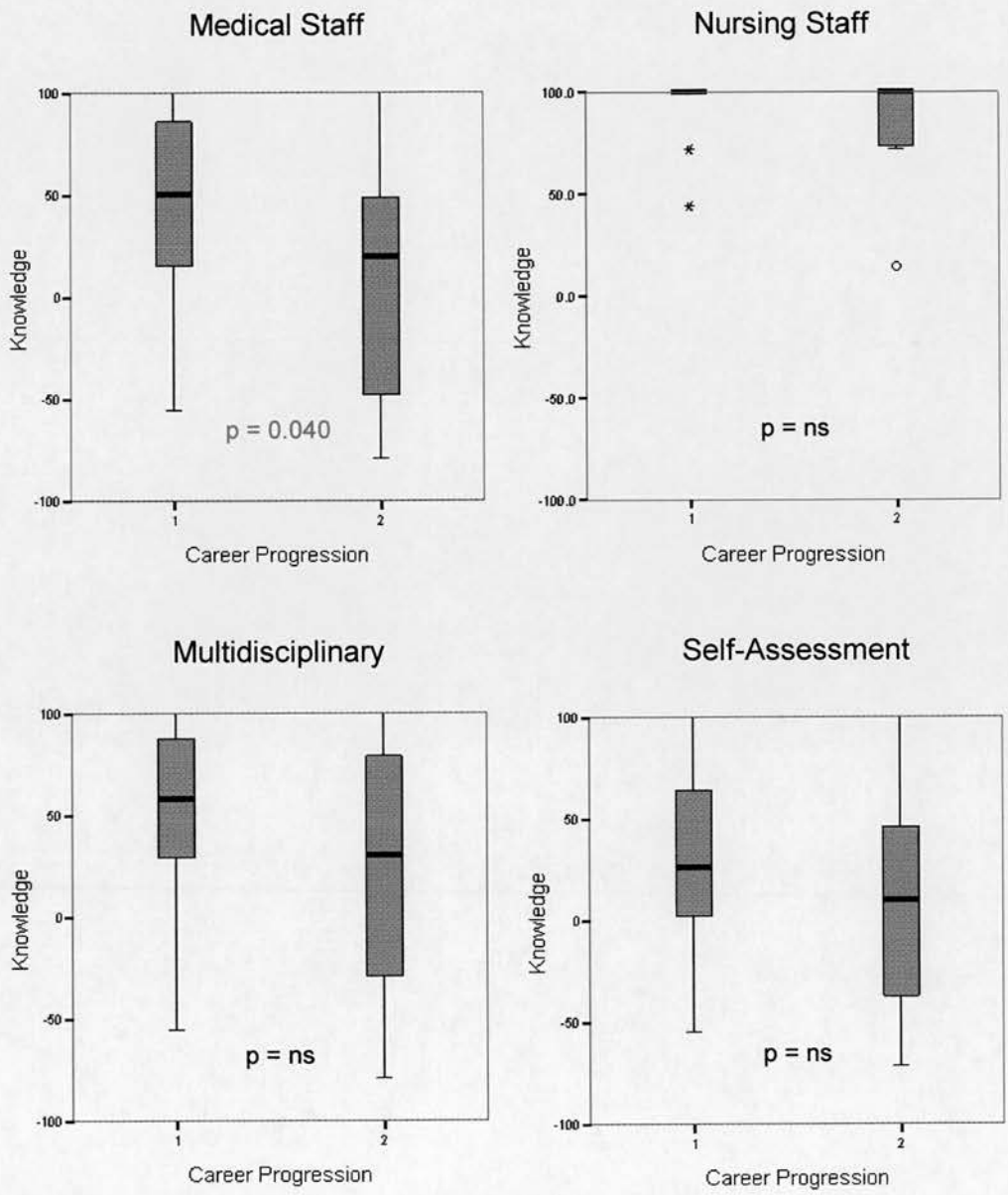
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (•)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.1b

Career Progression at 1 Year related to EBSTAF Assessment of Knowledge



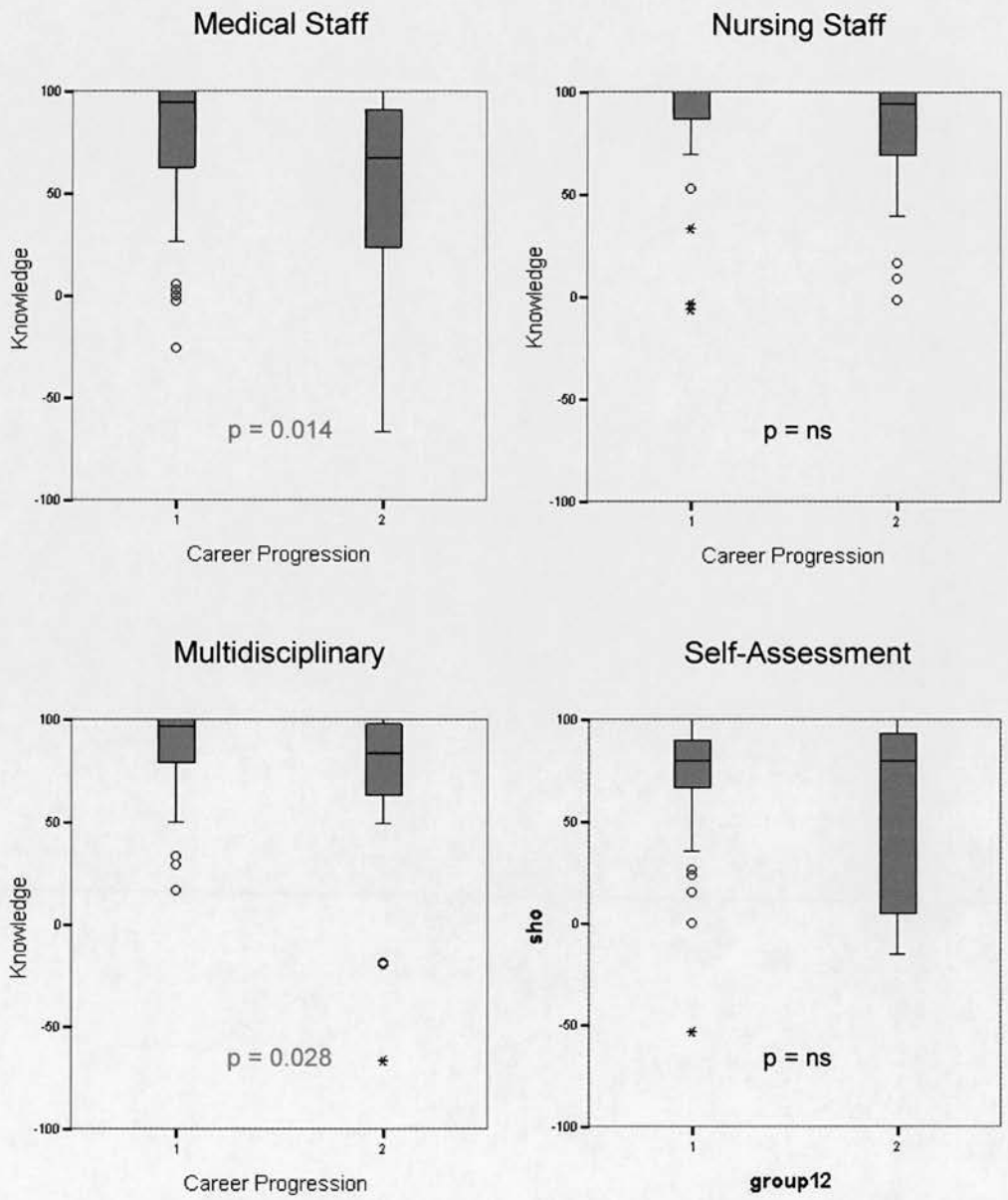
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (·)

Group I: Trainees achieving SpR / Research. **Group II**: Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.1c

Career Progression at 1 Year related to EBSTAF Assessment of Teamwork



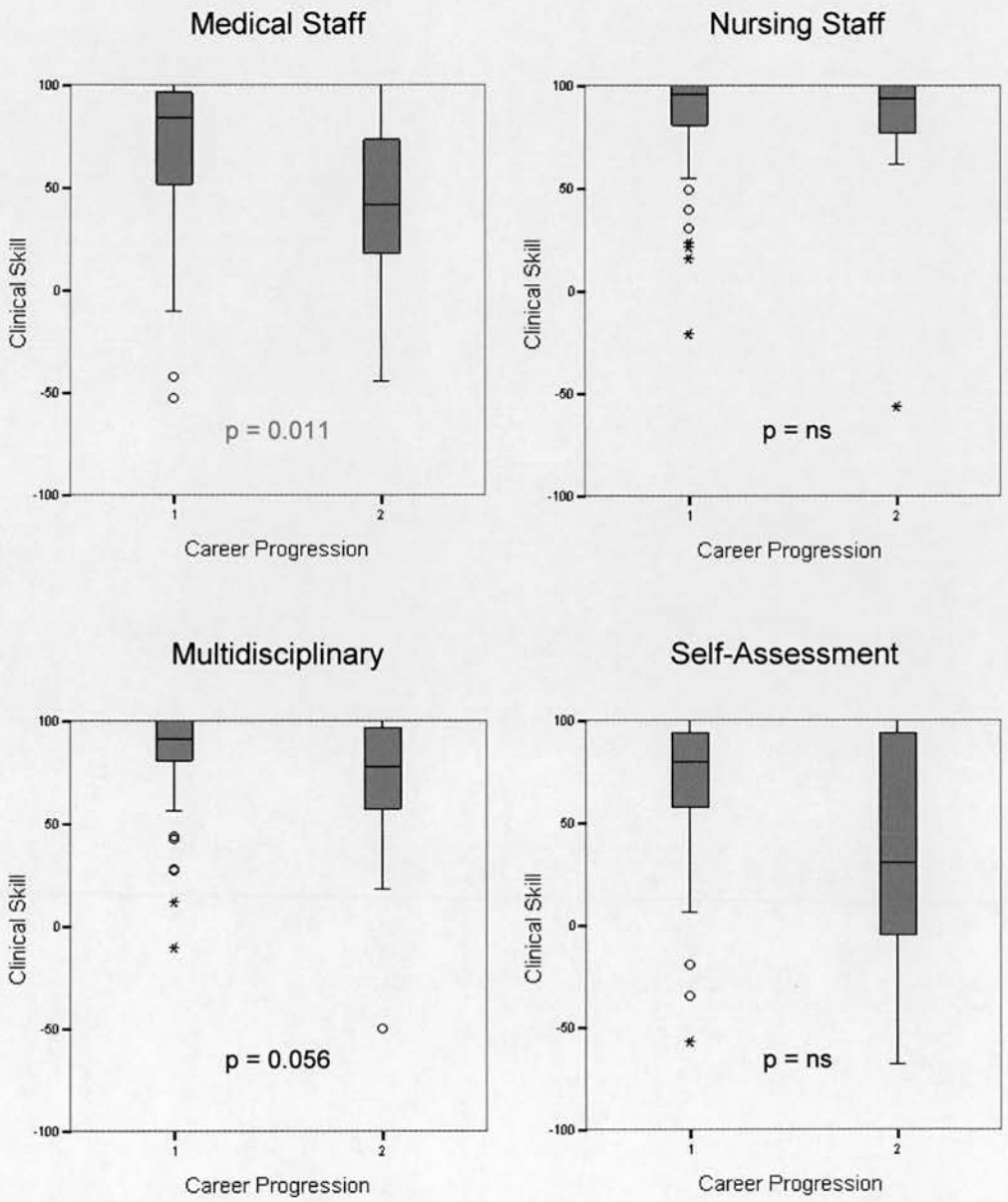
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (*)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.1d

Career Progression at 1 Year related to EBSTAF Assessment of Clinical Skill



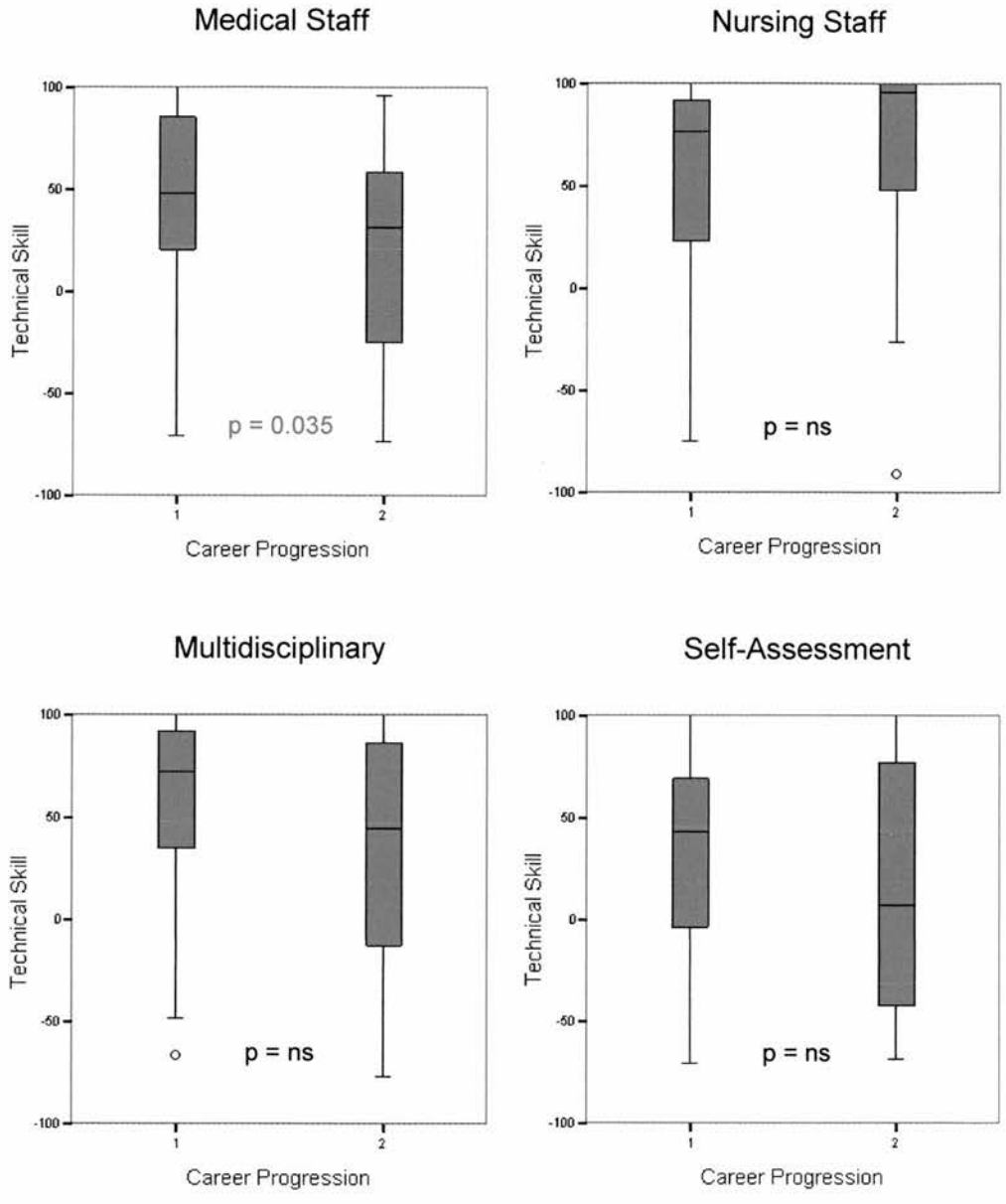
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (*)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.1e

Career Progression at 1 Year related to EBSTAF Assessment of Technical Skill



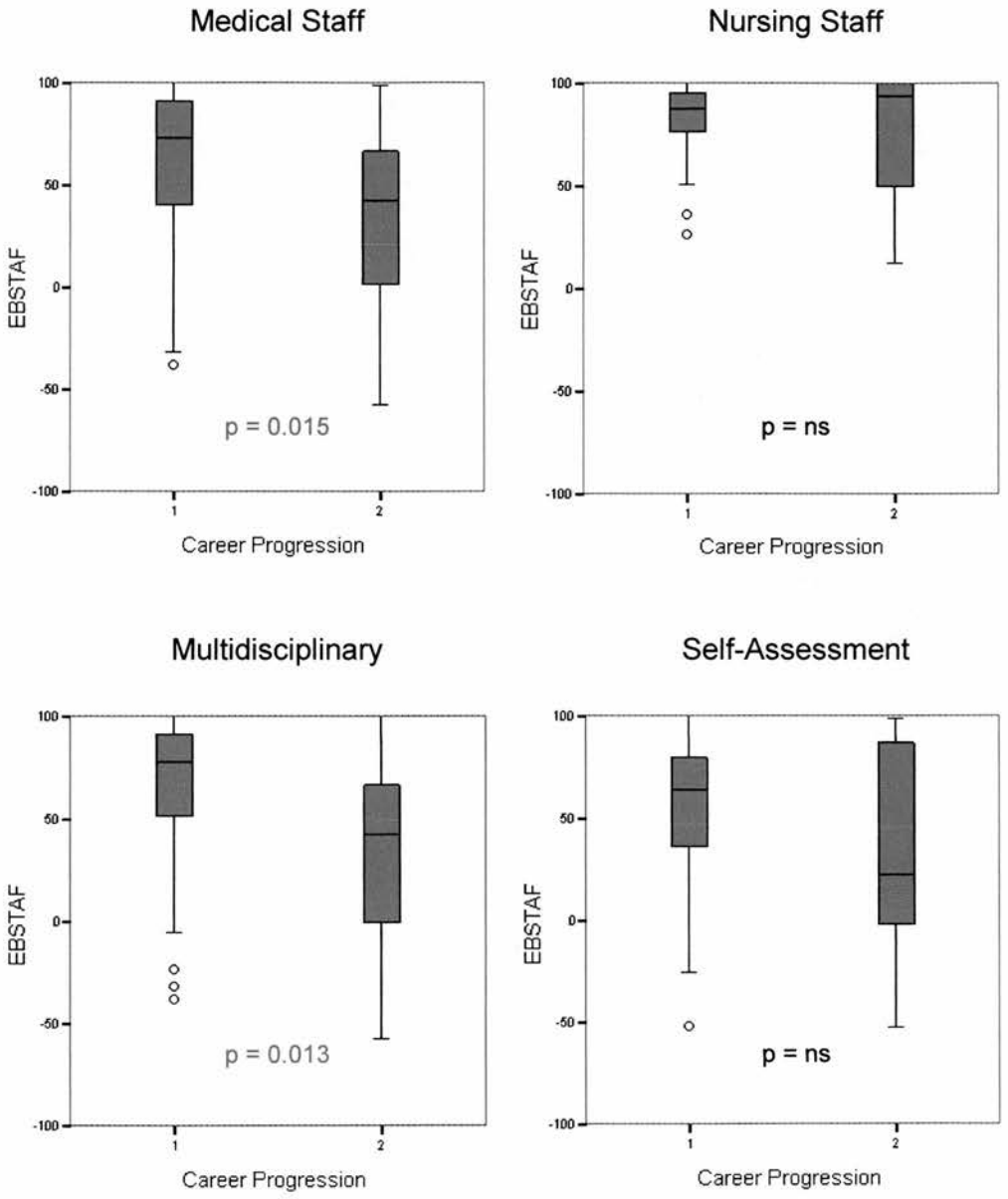
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (•)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.1f

Career Progression at 1 Year related to EBSTAF Overall Assessment



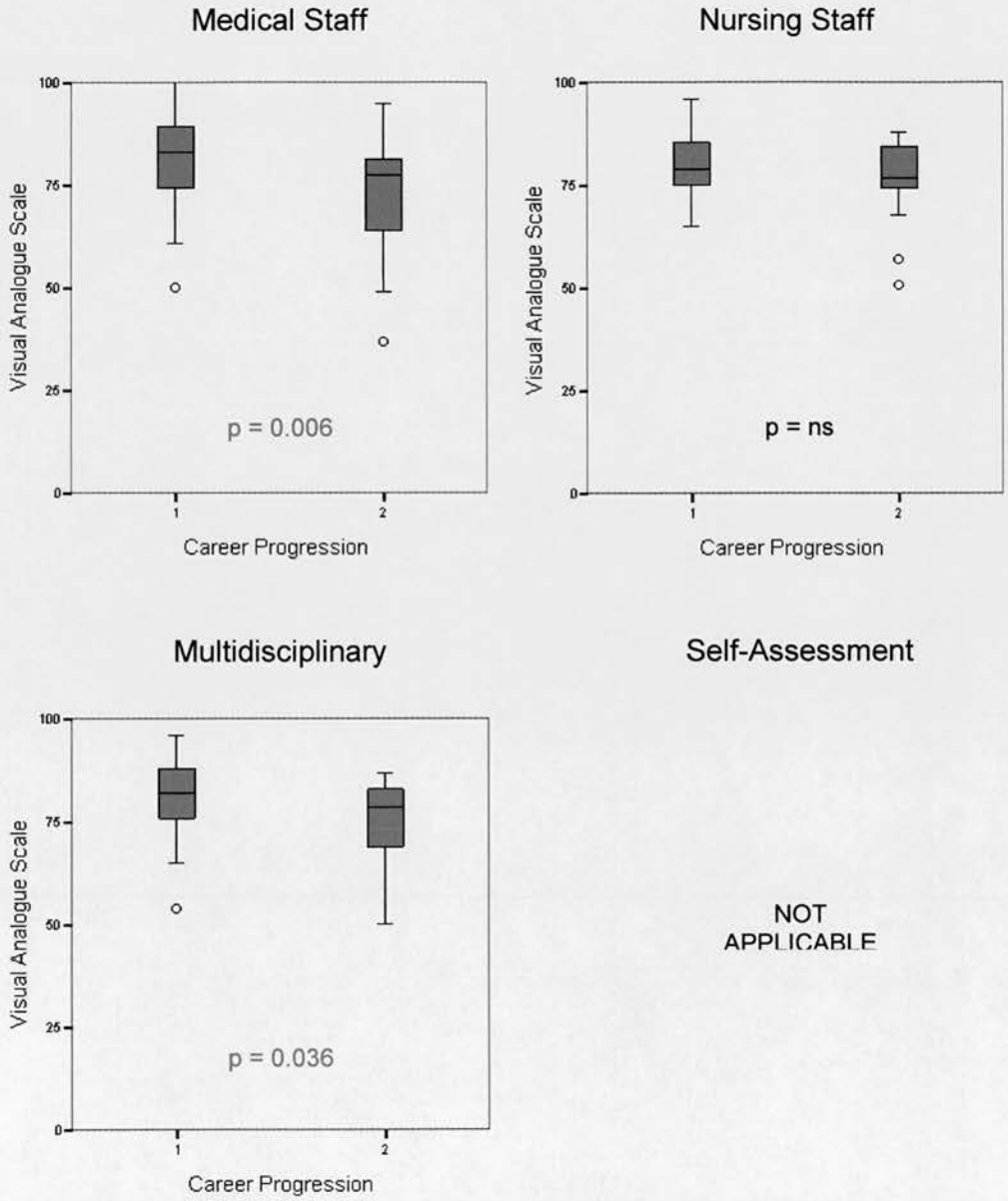
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (•)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.1g

Career Progression at 1 Year related to Visual Analogue Score



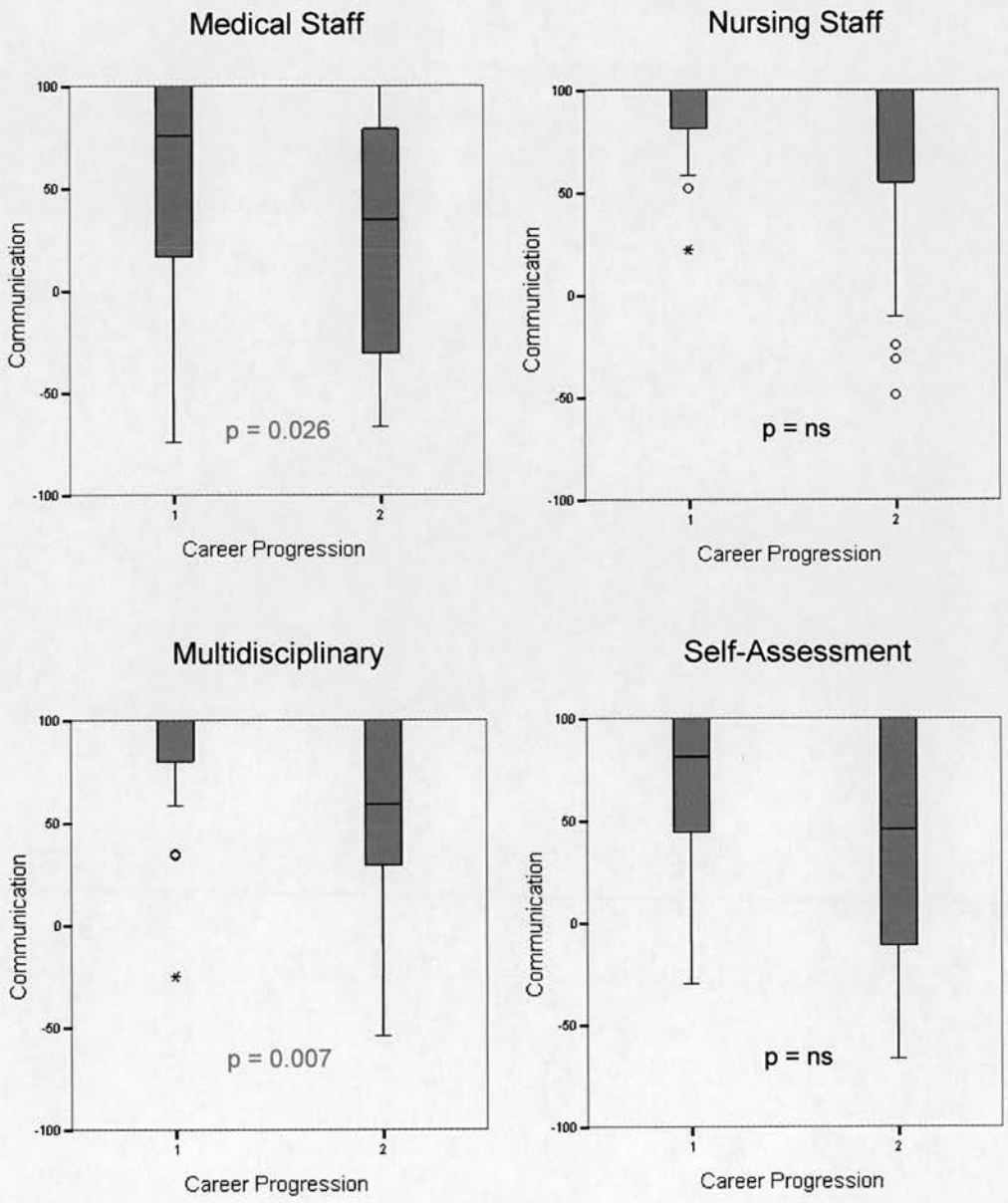
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (•)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.2a

Career Progression at 2½ Years related to EBSTAF Assessment of Communication



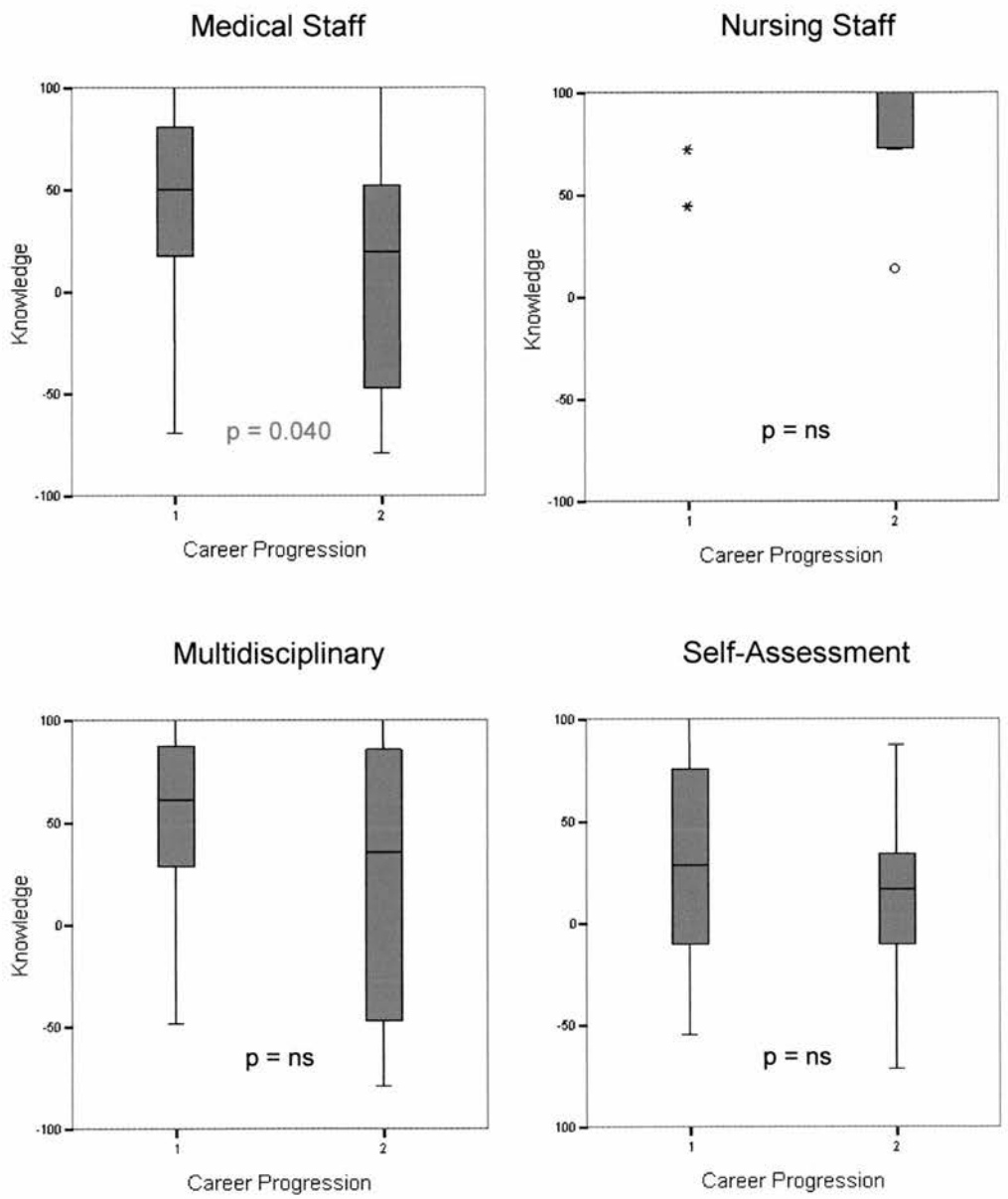
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (•)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.2b

Career Progression at 2½ Years related to EBSTAF Assessment of Knowledge



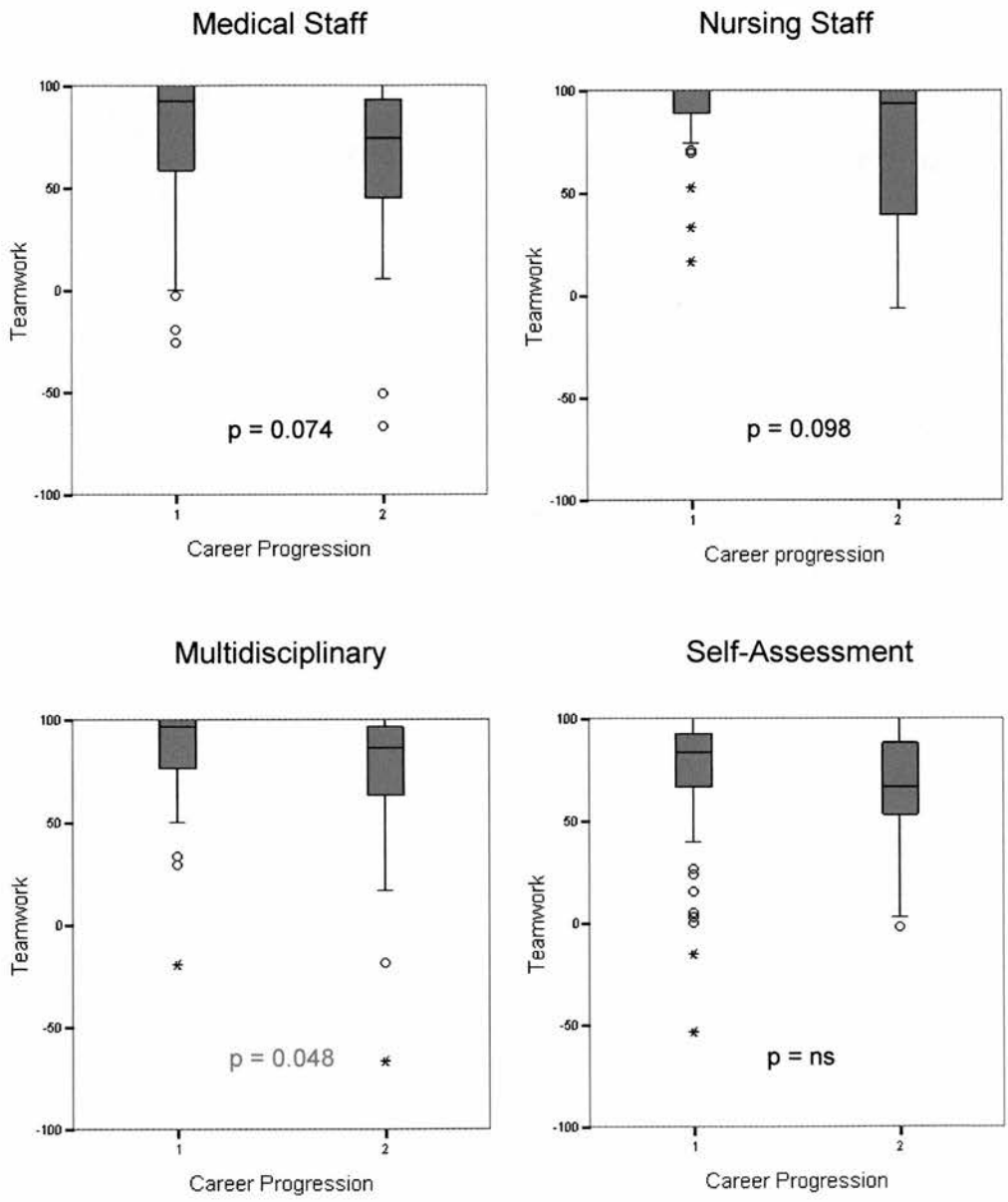
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (•)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.2c

Career Progression at 2½ Years related to EBSTAF Assessment of Teamwork



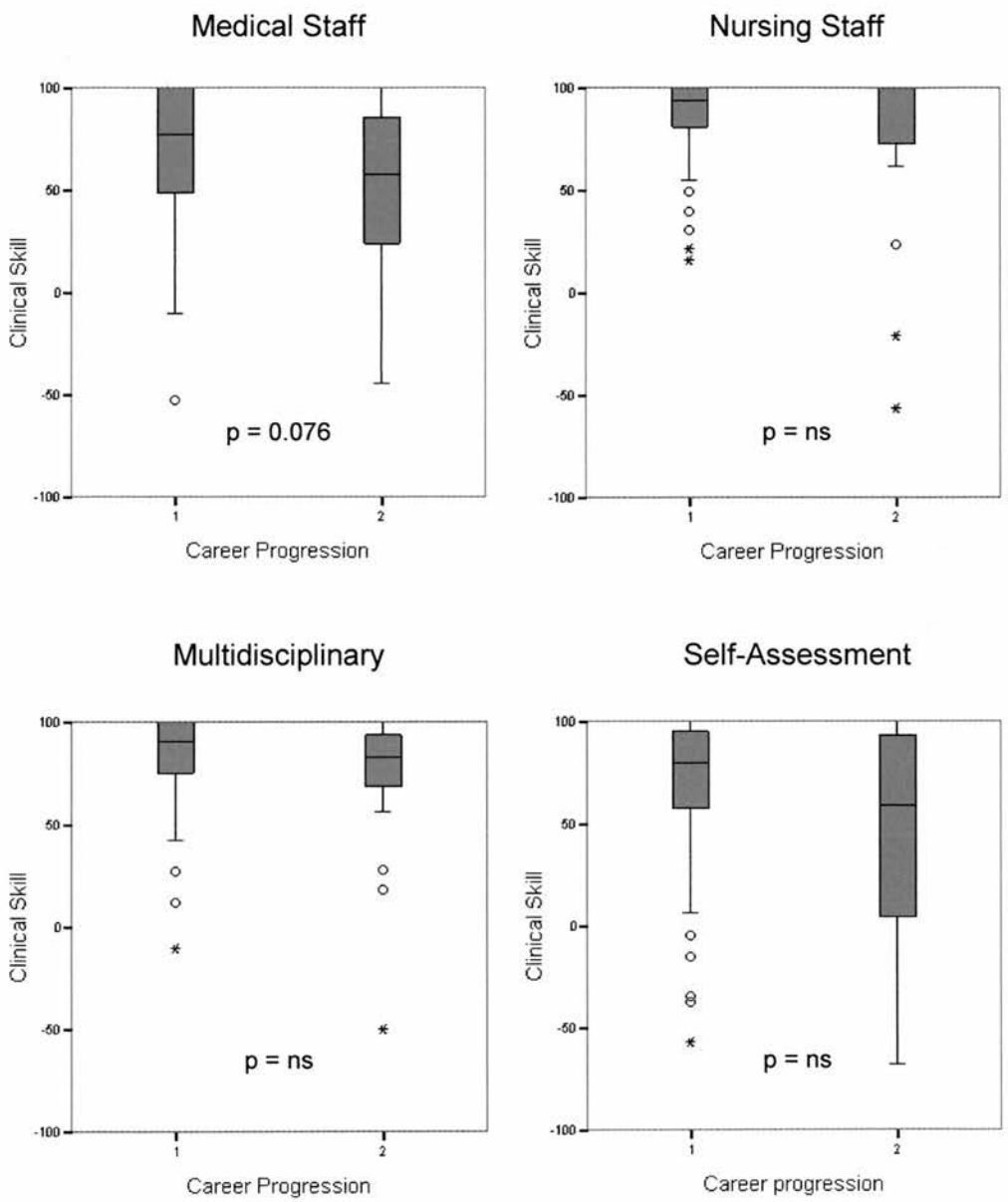
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (·)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.2d

Career Progression at 2½ Years related to EBSTAF Assessment of Clinical Skill



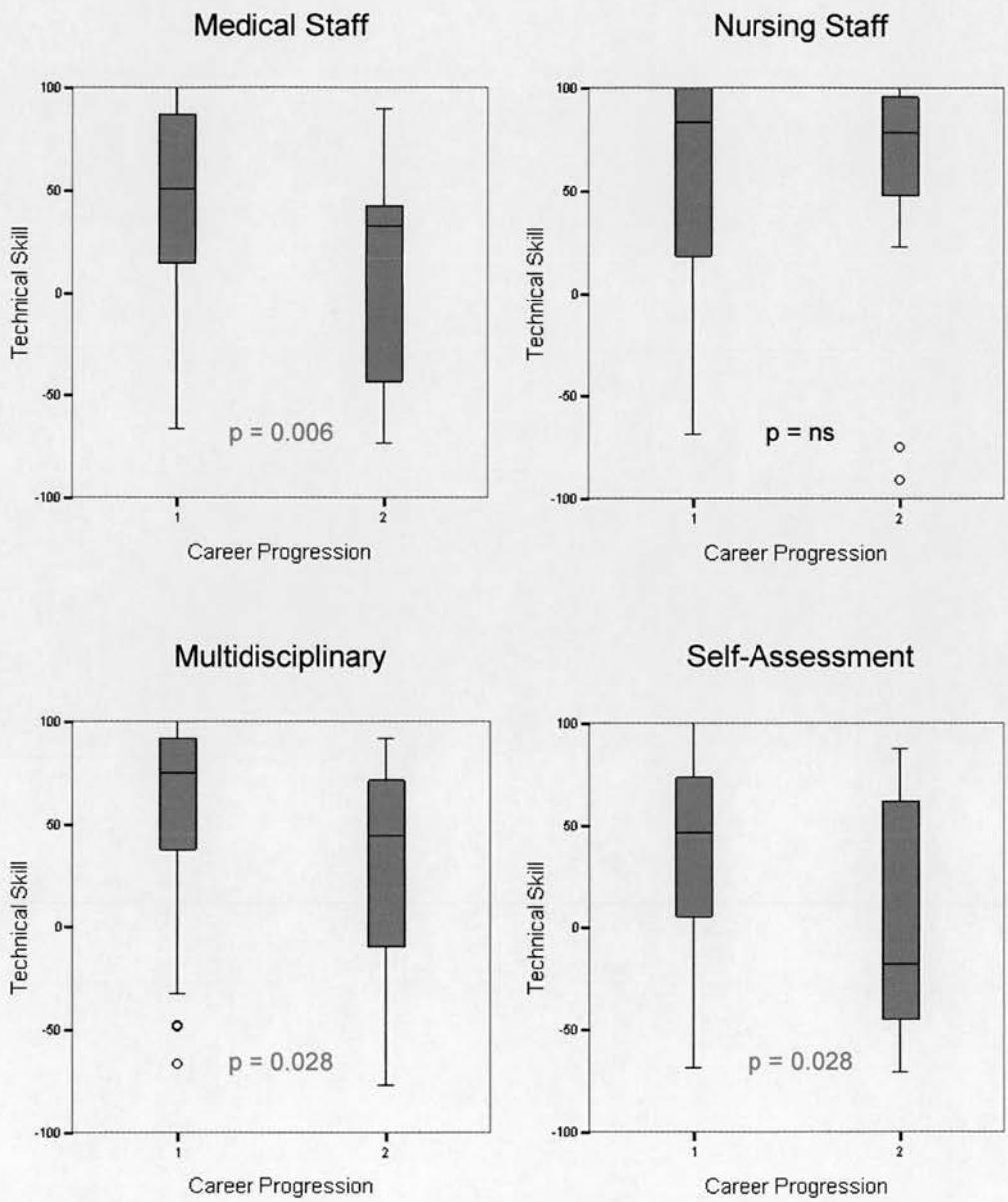
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (*)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.2e

Career Progression at 2½ Years related to EBSTAF Assessment of Technical Skill



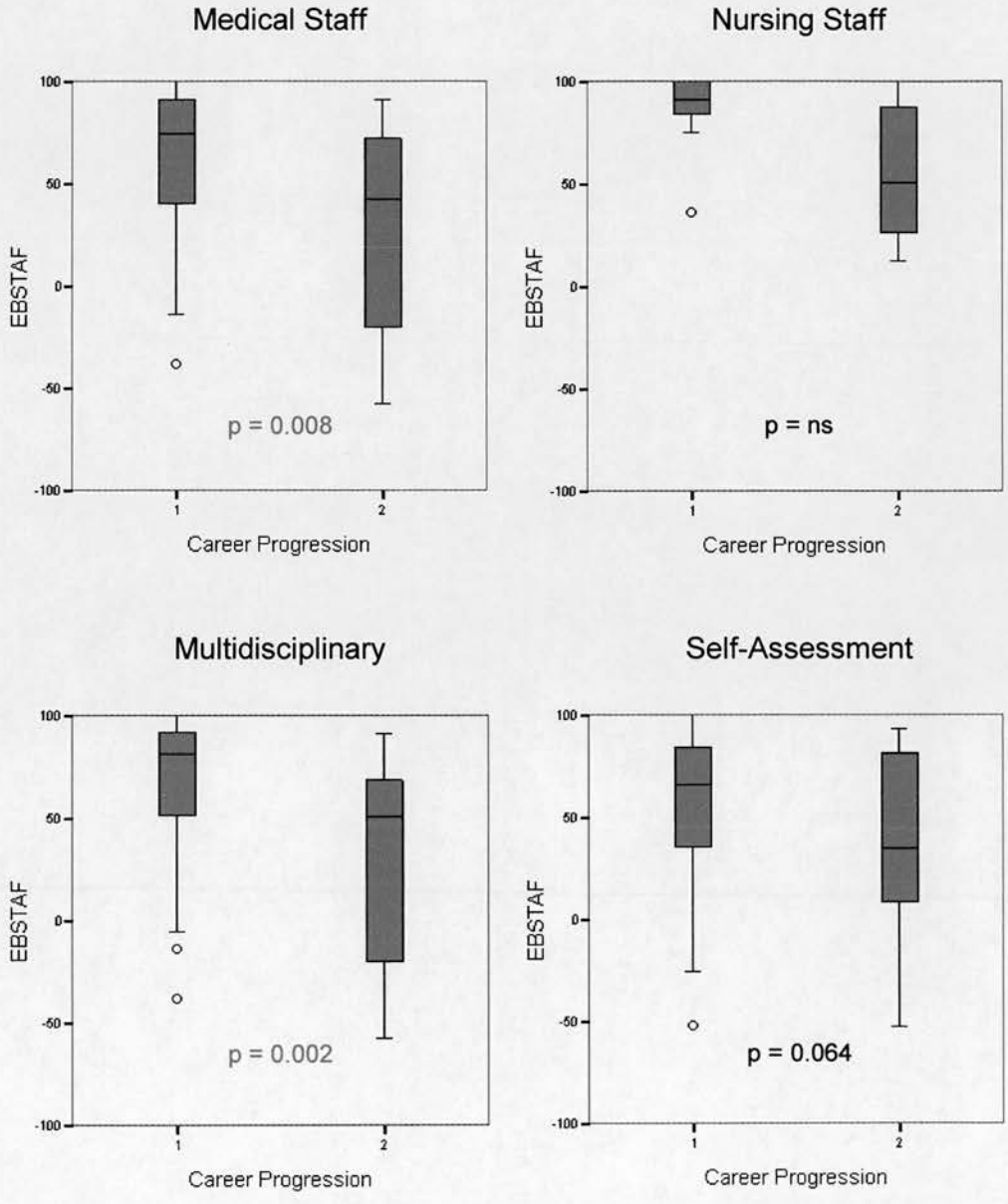
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (•)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.2f

Career Progression at 2½ Years related to EBSTAF Overall Assessment



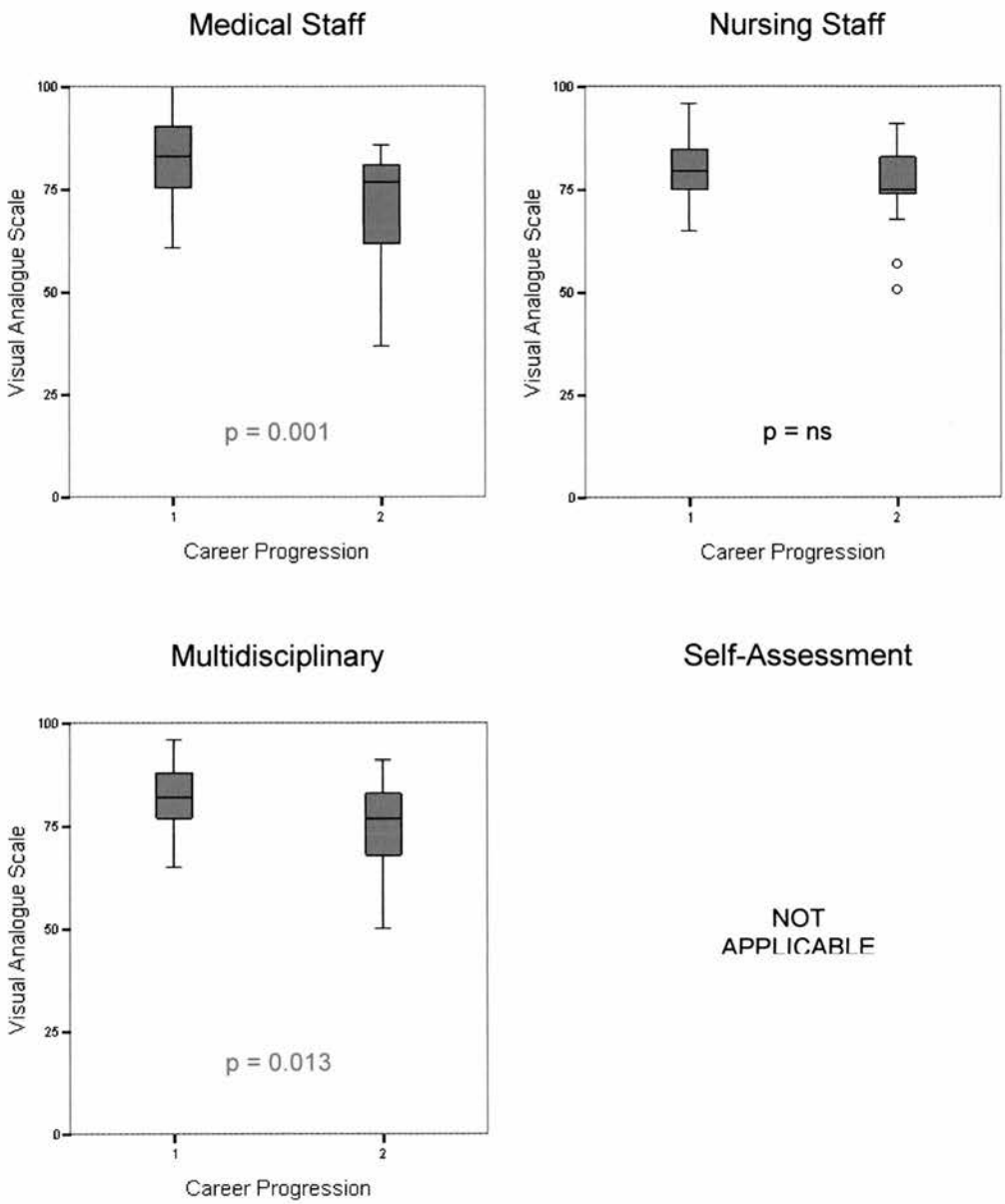
Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (•)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Figure III.2g

Career Progression at 2½ Years related to Visual Analogue Scale



Boxplots illustrate Median, Inter-Quartile Range, Range, Outliers (o) and Extremes (•)

Group I: Trainees achieving SpR / Research. **Group II:** Trainees not achieving SpR / Research

p: comparison by Mann-Whitney U.

Section IV.

TRAINEES' OPINIONS OF THE SKILLS REQUIRED OF BASIC SURGICAL TRAINEES.

IV.1 INTRODUCTION.

The training of a surgical trainee should be regarded as a form of contract between trainer and trainee. The trainer agrees to teach the trainee what is considered to be important for his/her career progression toward a consultant post while the trainee agrees to value the training they are given. Without such a commitment on the part of the trainee, training is likely to be ignored and come to nothing, wasting the efforts of both parties. Thus, the acceptability to the trainee of the criteria examined by a formative assessment tool is vital to its application. The same is true of a truly transparent summative assessment tool if the trainee is to accept the final outcome, particularly if it should prove unfavourable. Trainees' own opinions of how they are assessed have been largely ignored in the literature. In response, this study was designed to determine the acceptability to trainees of the fields examined by EBSTAF within a process of assessment and structured feedback of their clinical performance.

IV.2. AIMS.

- To determine the importance assigned by trainees to the individual fields within the EBSTAF assessment tool.

- To compare trainee opinion with that of trainers involved in the construction of EBSTAF in order to determine the acceptability of the contents of EBSTAF to the trainees.

IV.3. RESULTS.

IV.3.a. Response Rate.

All 33 trainees on the southeast Scotland training programme were enrolled into the study and returned their questionnaires within the study period, a 100% response rate. Forms were returned promptly by 23, following reminder by mail or e-mail by 7 and following telephone call by 3. No additional attributes were suggested by any trainee.

IV.3.b. Internal consistency.

Estimations of Cronbach's α are shown in Table IV.1. Consultant and Trainee internal consistency for each domain and overall was determined as "good" to "excellent" in all but consultant opinion of team-working skills (domain III) where it was found to be "acceptable".

The high response rate and internal consistency together were felt to support the validity of subsequent analysis.

IV.3.c. Domain ratings.

Comparisons of median percentage scores by domain and overall are shown in Table IV.2. Trainees consistently assigned significantly greater importance than consultants to the domains examined by EBSTAF.

IV.3.d. Field ratings.

Group ratings of individual fields are shown in Table IV.3. There was total agreement between BST and consultant in 44 of 70 fields (63%) with a further 24 of 26 fields (34%) being assigned greater importance by BSTs than the consultants. Thus, trainees ranked 68 of 70 fields (97%) to be of equal or greater importance when compared to consultant responses. Only 2 of 70 fields (3%), namely "maintains accurate notes" and "obtains additional information from relatives", were considered to be of lower import by the trainee group.

IV.3.e. Statistical determination of agreement (κ).

Estimations of weighted κ are shown in Table IV.4. Determination of κ was not valid for domains I (Communication) and II (Application of Knowledge) as the test requires variation within both comparison groups and the median scores for all fields here were assigned by the consultant group as "important (3)". Overall agreement between consultant and trainee groups was

determined as "fair" ($\kappa = 0.39$) while agreement for individual domains ranged from "fair" to "moderate".

IV.4. DISCUSSION.

This study demonstrates the acceptability to trainees of the fields examined by the EBSTAF; trainees value the same qualities previously determined to be important by consultants and attach greater or equal value to all but 2 of the 70 fields. The 100% response rate, aided considerably by the involvement of SPB (also chairman of the training program), would support the validity of our results.

The determination of agreement between consultants and trainees is not as straightforward as it might at first appear. Clearly, when both groups assign the same importance to a particular quality this may be regarded as agreement. However, the κ statistic will class a trainee ranking of 'essential' compared with a consultant one of 'important' as a *dis*-agreement, despite the fact that both groups examined consider value in the quality concerned. Thus, examination of simple agreement percentages may be more valid if one considers a trainee ranking of *at least* that of consultants to be an agreement. If this is carried out, as shown in Figure IV.1, overall agreement as to the qualities assessed in EBSTAF would be 97%, with only 2 fields (3%) being felt to be of less value by trainees. However, even this may suffer from response bias since trainees may be expressing what they perceive their consultants want them to think, rather than what they themselves value and employ in everyday clinical practice. However, the degree to which this may be occurring is impossible to quantify.

The fact that the trainees suggested no additional skills would support the validity of the conclusions of this study but this may not be wholly relied on since abstention does not equate to agreement.

Bearing the above reservations in mind, the finding that EBSTAF is so acceptable to trainees is encouraging. They recognise that a good trainee must employ high levels of skill across the fields examined by EBSTAF, the majority being generic to good medical practice rather than specific to surgery (Baldwin *et al*, 1999). Trainees recognise that although technical ability is fundamental to successful surgery, they also require a number of non-technical skills.

EBSTAF is not alone in highlighting the importance of generic non-technical skills in surgery. The literature repeatedly identifies communication, leadership, the ability to work well within a team and decisiveness as among the desirable characteristics of the successful surgeon (Greenburg *et al*, 1982; Galasko, 2000; de Leval *et al*, 2000; Giddings *et al* , 2000; General Medical Council , 2000a; Healey *et al*, 2004). The GMC's revalidation process based on "Good Medical Practice includes communication, teamworking and leadership as requirements of the competent doctor (or surgeon) (General Medical Council , 2001) while the JCHST currently examines SpR judgement, leadership, teamworking and communication, albeit using flawed methods (Paisley *et al*, 2001b). However, it is noticeable

in the literature that the opinions of trainees themselves on the content of their assessments have received little attention, despite recognition as an important factor in the acceptance of feedback from MSF in other environments (McEvoy *et al*, 1987; Fedor *et al*, 1989; Yuki *et al*, 1995; Wimer *et al*, 1998; Wood *et al*, 2006).

This study suggests that trainees will value structured feedback of the detailed assessments of their performance offered by EBSTAF and this may help to optimise surgical learning and promote reflective practice at the BST level.

IV.5. SUMMARY.

- Trainees attach equal or greater value than their consultants to the qualities assessed in EBSTAF, supporting the acceptability of the form to BSTs.

- This suggests that trainees will value detailed structured feedback of performance as determined by EBSTAF, lending support to its application to formative and summative assessment processes.

Table IV.1

Estimation of Internal Consistency within consultant and trainee groups by the application of Cronbach's alpha (α).

Domain	α	
	Consultant	**Trainee**
Communication	0.91	0.93
Knowledge	0.81	0.89
Teamwork	0.74	0.87
Clinical Skills	0.90	0.93
Technical Skills	0.89	0.90
All	0.85	0.89

$\alpha > 0.7$ = acceptable, $\alpha > 0.80$ = good and $\alpha > 0.90$ = excellent (Altman , 1991).

Table IV.2

Comparison of Median Percentage Scores between consultant and trainee groups.

Domain	Mean % Consultant	Mean % Trainee	Mann-Whitney p =
Communication	78	86	0.006
Knowledge	67	75	0.001
Teamwork	77	84	0.025
Clinical Skills	83	86	0.05
Technical Skills	79	84	0.004
All	78	83	<0.001

Mann-Whitney $p \leq 0.05$ is taken to be significant.

Table IV.3a

Group median weightings of individual EBSTAF fields:

For Table 4.3 a-e

blue shading : trainee-consultant agreement

no shading : trainees assigning greater importance to a field than consultant group

black shading : trainees assigning less importance to a field than consultant group

COMMUNICATION.

Quality being assessed	Trainee	Consultant
Establishes a rapport with patients	Essential	Important
Sensitive and empathic towards patients	Important	Important
Explains any potential risks in treatment	Essential	Important
Able to explain management in layman's terms	Essential	Important
Able to explain diagnosis in layman's terms	Essential	Important
Able to allay anxiety	Important	Important
Able to diffuse anger and hostility	Important	Important
Relates management to individual patient's needs	Important	Important
Aware of patient's social history	Important	Important

Table IV.3b

Group median weightings of individual EBSTAF fields:

APPLICATION OF KNOWLEDGE.

Quality being assessed	Trainee	Consultant
Knows the natural history of disease	Important	Important
Actively seeks out further information	Important	Important
Knows the relative merits of different management plans	Important	Important
Can co-ordinate available information on a case	Important	Important
Can present material clearly	Important	Important
Critically evaluates published work	Important	Useful
Can teach or explain with enthusiasm	Important	Useful
Can complete research	Important	Useful
Can initiate research	Important	Useful

Table IV.3c

Group median weightings of individual EBSTAF fields:

TEAMWORK.

Quality being assessed	Trainee	Consultant
Seeks advice when beyond limits of competence	Essential	Essential
Can be trusted to carry out instructions	Essential	Essential
Able to communicate clearly with other staff members	Essential	Important
Accepts feedback on own performance	Essential	Important
Can keep to time	Important	Important
Keeps GP informed	Important	Important
Understands other staff members' points of view	Important	Important
Delegates when appropriate	Important	Important
Aware of the role of other specialties	Important	Important
Able to offer constructive criticism to others	Important	Important
Can cope with unreasonable colleagues	Important	Useful

Table IV.3d

Group median weightings of individual EBSTAF fields:

CLINICAL SKILLS.

Quality being assessed	Trainee	Consultant
Can identify the acutely ill	Essential	Essential
Carries out thorough clinical examination	Essential	Essential
Takes full history	Essential	Essential
Extracts relevant information from history & examination	Essential	Essential
Conscientious in postoperative care	Essential	Essential
Keeps accurate notes	Important	Essential
Pays attention to changes in clinical picture	Essential	Essential
Listens to additional information from relatives	Important	Essential
Reviews diagnosis and management regularly	Important	Important
Uses information in referral letter	Important	Important
Adapts quickly if problems in management arise	Essential	Important
Knows when NOT to intervene	Essential	Important
Remains calm in an emergency	Essential	Important
Can formulate a working diagnosis & give rationale	Essential	Important
Interprets results with reference to other information	Important	Important
Generates & ranks appropriate differential diagnosis	Important	Important
Initiates investigations promptly	Important	Important
Decides quickly in an emergency	Essential	Important
Knows when follow up is appropriate	Important	Important
Knows when discharge is appropriate	Important	Important
Can improvise where necessary	Important	Important
Aware of cost & clinical value of investigations	Important	Useful

Table IV.3e

Group median weightings of individual EBSTAF fields:

TECHNICAL SKILLS.

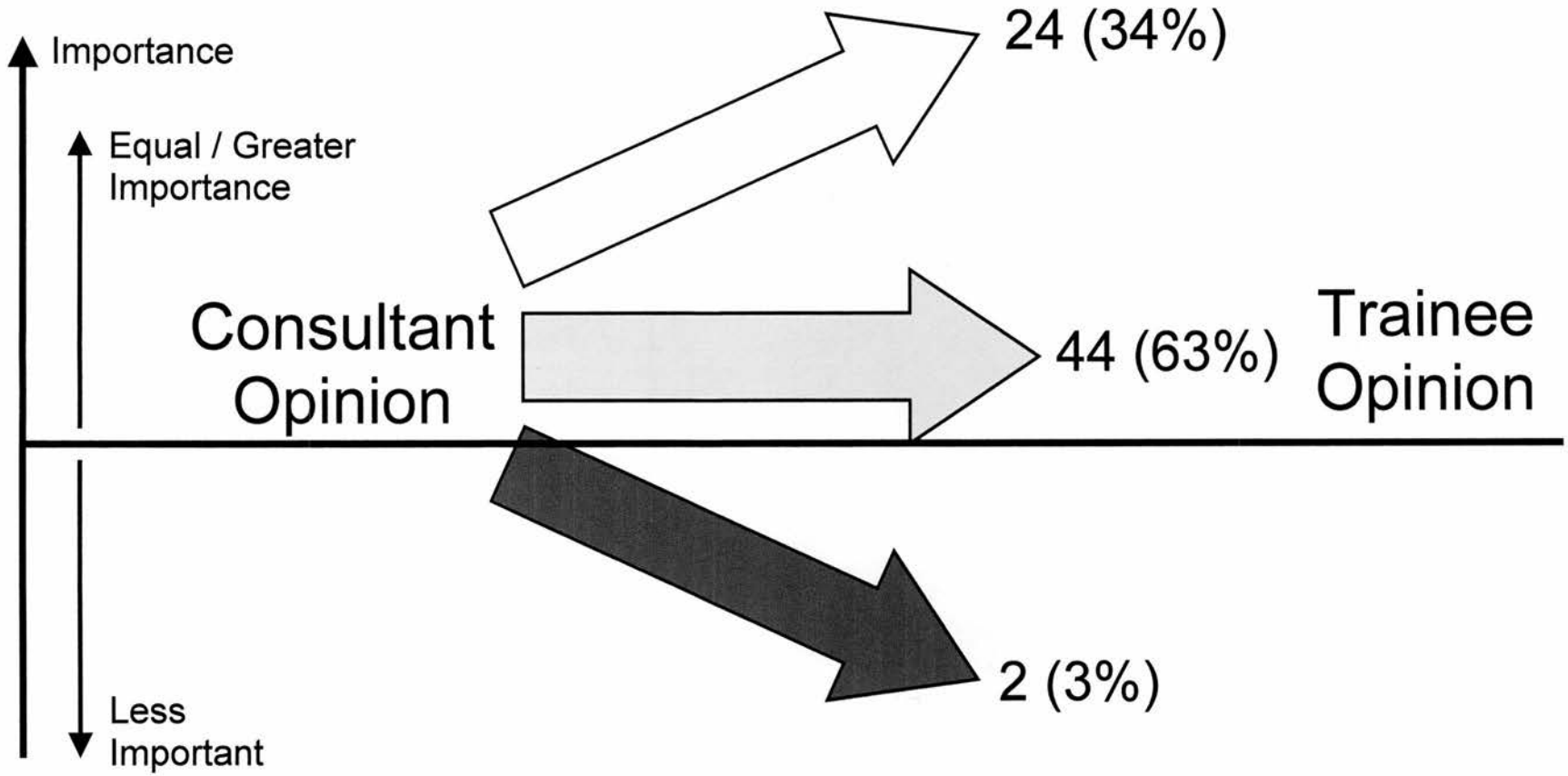
Quality being assessed	Trainee	Consultant
Handles tissue gently	Essential	Essential
Handles dangerous instruments safely	Essential	Essential
Demonstrates sound knowledge of anatomy	Essential	Important
Competent in tying all knots	Essential	Important
Can distinguish normal from abnormal	Essential	Important
Makes incisions appropriately	Essential	Important
Can identify and expose tissue planes	Important	Important
Demonstrates manual dexterity	Important	Important
Able to position patient on operating table	Essential	Important
Able to control bleeding by swab, sucker & clips	Essential	Important
Able to close skin neatly	Important	Important
Can use diathermy techniques	Important	Important
Has good hand eye co-ordination	Important	Important
Able to control bleeding by suturing	Important	Important
Has 3-dimensional spatial awareness	Important	Important
Selects correct instruments	Important	Important
Considers the aesthetic appearance of wound	Important	Important
Anticipates movements during assistance	Important	Important
Is economical in movements	Important	Useful

Table IV.4

Statistical determination of agreement between consultant and trainee groups by the application of weighted Kappa (κ).

Domain.	Kappa (κ).	Degree of Agreement.
Communication	-	-
Knowledge	-	-
Teamwork	0.46	Moderate
Clinical Skills	0.39	Fair-Moderate
Technical Skills	0.25	Fair
All	0.34	Fair

Figure IV.1
Diagrammatic Representation of Consultant : Trainee Agreement



Section V.

THE INFLUENCE OF STRUCTURED FEEDBACK ON TRAINEE PERFORMANCE.

V.1. INTRODUCTION

Junior trainees have repeatedly been shown to possess poor self-assessment skills (Morton *et al*, 1977; Risucci *et al*, 1989; Gordon, 1991; Das *et al*, 1998; Johnson *et al*, 1998; Ward *et al*, 2002; Ward *et al*, 2003) which have recently been shown to improve with experience (Moorthy *et al*, 2006). They therefore find it difficult to estimate their own abilities and complain of a lack of feedback from trainers who either overestimate the amount of feedback they give or underestimate the amount the trainees would like to receive (Fonseka, 1996). In addition, consultant feedback is inevitably incomplete since they are unable to observe every aspect of a trainee's performance.

Multidisciplinary assessment however can observe and evaluate most aspects of a trainee's clinical practice. It therefore has the potential to provide the trainee with detailed and direct feedback of performance, helping them identify and target areas of weakness.

EBSTAF has been shown to provide robust assessment of trainees (Paisley *et al*, 2001a) across 70 fields that trainees accept as being important in their development into a rounded surgeon (see Section IV). By collating 6-monthly EBSTAF assessments for each trainee into an anonymous feedback document, this study examines the effect of providing detailed structured feedback on performance to a cohort of surgical trainees and compares their progress to a previous cohort provided with no such feedback.

V.2. AIMS

- To provide structured feedback of in-post performance as assessed by EBSTAF to a cohort of BSTs.

- To examine the effect on in-post performance of the provision of structured feedback to a cohort of trainees by comparison with a previous cohort receiving no such feedback.

- To demonstrate the effect of structured feedback on poorly performing individuals within the feedback cohort.

- To determine the acceptability to trainees of the structured feedback process.

V.3. RESULTS

V.3.a. Demographics of the Assessment Process

V.3.a.i. Distribution

One thousand and thirty nine assessment forms were distributed to 284 assessors in 155 distinct assessment episodes, 847 (82%) of which were returned within the 4-week deadline. The SHO was not known to the assessor in 61 cases leaving 786 assessments (76%) for further analysis.

Assessments took place at all 7 hospitals included in the rotation at the time of the study (Table V.1).

Assessments covered 9 surgical specialties (Table V.2), differing slightly from the previous cohort with Breast surgery and Ear, Nose and Throat surgery replacing Urology.

V.3.a.ii. Trainees

A total of 43 trainees were evaluated. Eleven trainees (26%) were evaluated over a single post, 9 trainees (21%) over 2 posts (i.e. 1 year), 13 trainees

(30%) over 3 posts (i.e. 18 months) and 10 trainees (23%) over four posts (i.e. 2 years). The median number of assessment forms completed on each trainee was 23 (IQR 14-31), having completed between 2 and 36 months of surgical training (1 trainee had completed 6 months of paediatric surgery prior to starting on the training programme).

V.3.a.iii. Assessors

A total of 284 assessors were involved in the current study. 286 assessment forms were distributed to medical staff, 615 forms to nursing staff and 138 to the SHOs themselves. A further breakdown of assessors is shown in Table V.3.

V.3.a.iv. Assessment Episodes

At each assessment episode, trainees were evaluated by between 4 and 14 assessors (median 7, IQR 6-9), determined by the structure of the unit concerned. 24 episodes (96 assessments) related to the preceding 2 months, 60 (259 assessments) to the preceding 3 months, 24 (225 assessments) to the preceding 4 months and 61 (459 assessments) to the preceding 6 months. A total of 108 six-month posts were examined.

V.3.a.v. Response Rates and Validity of Assessments

There was no significant difference in response rate by the 4-week deadline between the assessor groups (Table V.4) but response rates were

significantly reduced compared to the original study ($p < 0.0001$ df3, Chi square). Once more, the validity of assessments varied widely between assessor groups and EBSTAF domains (Table V.5). This variation in validity differed significantly from the original study ($p = 0.019$ df12, Chi square).

V.3.b. Structured Feedback of Performance

Trainees attended up to four appraisals with the director of the surgical training programme (SPB) during the study period. Each appraisal took place between 4 and 6 weeks after the end of the six-month post being examined by EBSTAF.

Initial appraisal remained unchanged from before the study period. This involved a discussion of how the trainee felt they were doing, their achievements to date and the setting of goals for the next six months with decisions being made as to their preferred post(s) in 5 months time. The EBSTAF Structured Feedback Form was not provided during this process so as to not influence the appraisal or post allocation.

Fifteen minutes was then allocated for the examination and discussion of structured feedback provided by EBSTAF. Each trainee was given their personal EBSTAF Structured Feedback Form which illustrated every assessor's mark for each field as an anonymised black cross (**X**). The trainee's own self-assessment marks were also provided, indicated by a red

cross (X), to allow them to compare their own assessment with those from their assessors. Mean scores on visual analogue scales for overall impression and working relationship were also provided, with the latter split for consultant alone, overall and the trainee's own score. Comments from trainee and all assessors were anonymised and included within a final section. No immediate problems were reported with the summary forms. An example of a completed Structured Feedback Form is included in Appendix Section 3.

Trainees who had completed their final post and had left the training programme were provided with the summary form but were not brought back for appraisal. Sixteen trainees were appraised at the first time point (March 2001), 23 at the second (Sept 2001), 22 at the third (March 2002) and 23 at the final time point (Sept 2002).

Eleven trainees were appraised only once, 9 were appraised twice, 14 were appraised three times and 4 trainees were appraised 4 times.

Twenty six trainees were assessed in their first BST post (general surgery), 23 in their second (also general surgery), 20 in their third (orthopaedics), 13 in their fourth (specialty 1) and 4 in their fifth and final post (specialty 2).

V.3.c. Examination of the Effect of Structured Feedback on Performance by Comparison of Feedback and No Feedback Trainee Cohorts.

This part of the study compared EBSTAF scores achieved by the current cohort of trainees receiving detailed structured feedback with those BSTs examined in the original study who received no feedback of performance (Paisley *et al*, 2001a; Paisley 2002).

V.3.c.i. Assessments by Medical Staff – Tables V.6a-g

No differences were demonstrated between the two groups for the domains of Communication, Knowledge or Clinical Skills.

In Teamwork the feedback group scored significantly better in the second post only, with scores equalising again in the third and subsequent posts.

The feedback cohort scored significantly better for Technical Skills in posts 3 and 5 with no differences demonstrated at other time points.

EBSTAF Overall scores were significantly greater in all but post 4 for the feedback cohort while VAS Overall demonstrated the opposite relationship in all but post 5.

V.3.c.ii. Assessments by Nursing Staff – Tables V.7a-g

No differences were seen between the two groups in the assessment of Technical skills.

The feedback group scored significantly better in Communication (posts 1 and 2), Knowledge (post 1), Clinical Skills (post 2) and EBSTAF Overall (posts 1 and 2). Although the feedback group scored better in Teamwork post 2, this was then reversed in post 3.

VAS Overall scores were consistently lower for the feedback group.

V.3.c.iii. Multidisciplinary Assessments – Tables V.8a-g

No differences were demonstrated between the two groups for the domains of Knowledge or Technical Skill.

The feedback group scored significantly better in Communication (posts 1 and 2), Teamwork (post 2), Clinical Skills (posts 1 and 2) and EBSTAF Overall (posts 1, 2 and 5).

VAS Overall scores were consistently lower for the feedback group.

V.3.c.iv. SHO Self Assessments – Tables V.9a-g

No differences were demonstrated between the two groups in the assessment of Clinical and Technical Skills.

The feedback group scored significantly better in Communication (post 1), Knowledge (post 5) and Teamwork (post 2).

EBSTAF Overall scores were significantly higher in the feedback group across all 5 posts.

V.3.d. Examination of the Effect of Structured Feedback on Performance by the Effect of Feedback on Individual Trainees

It was not possible to quantify the effect of structured feedback on individual trainees. However, poor performance was identified by EBSTAF and structured feedback highlighted areas of weakness, allowing trainees to target the areas of their practice that most needed attention. Discussions during the appraisals showed that trainees were commonly unaware of their weaknesses and subsequent appraisals showed them to have been addressed and to have improved. However, due to the small numbers of trainees and initially large spread of ratings, it was not possible to statistically identify poor performers as outliers. It was therefore not possible to statistically demonstrate improvement when examining individual performances.

V.3.e. Trainee Evaluation of Structured Feedback of Performance

V.3.e.i. Form Distribution and Return.

Twenty-three trainees attended appraisal at the end of the study period. Each was supplied with a Structured Feedback Evaluation Form (Appendix Section 4) with a request for its return within one week. Twenty-one of 23 forms were returned completed within one week of the appraisal, equating to a 91% response rate.

V.3.e.ii. Trainee Evaluations.

All visual analogue scales were completed by trainees returning evaluation forms (i.e. 100% completion). Responses are summarised in Figure V.1.

Trainees found the structured feedback to be fair, at just about the right level of detail and useful to their training. Trainees did not find assessment by either assessor group to be threatening and although assessment by nursing staff scored slightly higher in this respect than by medical colleagues, the difference did not reach statistical significance.

Trainees were asked to indicate which component of the structured feedback process they found particularly useful. Fourteen trainees (66%) identified the detailed field summary while 15 trainees (71%) indicated the visual analogue

scales. Twelve trainees (57%) identified both as being useful. All 21 trainees valued the comments made anonymously by assessors.

15 trainees offered comments as to how the structured feedback might be improved (Table V.10).

V.4. DISCUSSION

Trainees are generally very positive about the feedback they receive during hands-on operating, being told immediately and very clearly when they do something wrong or inappropriate (Fonseka, 1996). However, there is more to being a surgeon than simply operating, and trainees are usually required to infer how well they are progressing overall from indirect evidence. Having been developed by expert consensus (Baldwin *et al*, 1999) and subsequently validated (Paisley *et al*, 2001a), EBSTAF is able to provide robust and detailed evidence of an individual's performance over a six-month period of everyday practice. The collation of EBSTAF assessments into a structured feedback form and its provision to trainees proved feasible and extremely popular with trainees who, despite initial reservations about being assessed by the nursing staff, found it very helpful in directing their efforts during their basic surgical training. In fact, it was often the detailed comments from the nursing staff that provided them with a real insight into their strengths and weaknesses. This reinforces the value of assessment of surgical trainees by nursing staff, despite its apparent irrelevance to career progression (see section III.4), and highlights the fact that they may observe everyday trainee behaviours not demonstrated in the presence of consultants or SpRs.

Criticisms levelled at the feedback process by the trainees were relatively few but there was a general feeling that the provision of an overall ranking within each entry group would have been informative. It was also suggested that

the identification of assessor type (i.e. medical or nursing) associated with each comment would be desirable although it was agreed that this may have affected the frank nature of the comments themselves from assessors who were previously assured of anonymity.

Individual trainees whose EBSTAF assessments highlighted areas of poor performance were able to target these areas and subsequently improved. However, due to the small numbers involved in this study it was not possible to demonstrate this statistically.

Comparison of trainee cohorts with and without feedback showed significant differences in median scores across the assessor groups favouring feedback in the early stages of surgical training (posts 1 and 2) that were not seen later on in training (posts 3 onwards). However, ratings in the first post were prior to any structured feedback. Differences at this time point cannot therefore be attributed to the feedback process. There may be a number of explanations: they may result from a generally superior cohort of trainees or reflect more lenient assessments from different assessor groups. However, corresponding visual-analogue scores do not support either of these explanations since medical, nursing and multidisciplinary VAS ratings were consistently lower for the feedback group, implying that assessors were actually rating more harshly. Alternatively, the involvement of assessors within the study may have increased the amount of informal feedback given to the trainees, confounding the intended control group at post 1. It is therefore difficult to

draw any overall conclusions from these results. However, these issues may have been countered by randomly splitting the current trainee cohort into feedback and non-feedback groups to be assessed in parallel by the same assessor group blinded to the provision of structured feedback (analogous to a randomised controlled trial). However this would have critically reduced numbers in each study arm and would have been unacceptable to the trainees themselves who, on questioning, perceived a significant benefit inherent in the receiving of structured feedback.

In the assessment of technical skills, a *ceiling effect* has been described (Munz *et al*, 2004). This is the point at which an assessment method no longer discriminates between levels of trainee. EBSTAF ratings tended to show a ceiling effect as they approached 100% in posts 3 or 4, although the feedback group appeared to reach this level one post before the non-feedback group. However it is difficult to attribute this difference to the process of structured feedback. It may reflect the higher initial score for the feedback group (effectively a "head start", as discussed above) or else result from the structure of EBSTAF itself which is effectively criterion-referenced by setting "competent" as the end point. EBSTAF does not offer assessors the opportunity to grade a trainee as excellent and although the demonstration of competence is the goal of current assessment processes, it reduces the discriminatory potential and therefore the value in trainee selection by awarding the excellent trainee the same as the trainee who is simply competent. Selection processes are best-achieved using norm-

referenced ratings where each individual's performance sets the standard for others (Bulstrode *et al* , 2001; Crossley *et al*, 2002b). It may therefore be worth considering the addition of a fourth grading level above that of competent, allowing the potential to score greater than 100% while not altering the structure (and therefore the reliability and validity) of EBSTAF.

One should not totally discount the apparent early attainment of the score ceiling. On a few occasions no difference was seen between the two groups in post 1 and yet the feedback group was seen to score significantly higher than the non-feedback group in subsequent posts before the ceiling effect intervened. This was seen in medical assessments of Technical Skills, nursing assessments of Knowledge and was seen consistently across medical, nursing and multidisciplinary assessments of Teamwork. A similar finding was seen in BST self-assessment of Teamwork. In these instances the feedback group attained the score ceiling (and therefore competence) earlier than the non-feedback group. Although limited, these findings are encouraging as they suggest that feedback may have a positive effect on trainee performance that has simply not been demonstrated in other domains for the reasons previously discussed. Furthermore, the feedback group was consistently seen to reduce the interquartile range of scores faster than the non-feedback group, implying that as a whole the feedback group was attaining competence at an increased rate.

It is interesting that an effect of feedback is demonstrated in the teamwork domain. Teamworking has been repeatedly identified in the literature as a critical non-technical skill for safety in high-risk industries (Cooper *et al* , 1980). The impact of good (or poor) teamworking skills can only increase as surgical practice moves away from traditional firm structures to working within *ad hoc* teams as dictated by rotas and working patterns. However, although individuals commonly think of themselves as having good teamworking skills, they often lack a clear understanding of what is involved and are thus particularly likely to be unaware of failings in this area. The teamwork domain of EBSTAF clearly defines what is expected of a surgical trainee in this area of practice and this may explain the maximal effect observed in this domain.

Looking at BST self-assessment scores, no differences were observed between feedback and no-feedback groups except for EBSTAF overall scores where the feedback group scored consistently higher. Although there remained a discrepancy in post 1, with all its implications, it is noteworthy that the ceiling for the feedback group was also raised and maintained above that of the non-feedback group. Trainees appeared to be more confident in their abilities overall as a result of structured feedback of their performance in practice.

Ward states that "the competent physician" (or surgeon) "pursues lifelong learning through the recognition of deficiencies and the formulation of appropriate learning goals" (Ward *et al*, 2002). While this may be true of

experts, learners have been repeatedly shown to be unable to accurately assess their own performance due to poor self-assessment skills (Morton *et al*, 1977; Risucci *et al*, 1989; Gordon, 1991; Das *et al*, 1998; Johnson *et al*, 1998; Ward *et al*, 2002). In order to optimise their training, learners must be provided with clear objectives and comparisons of actual and desired performance (Kolb & Fry 1975). EBSTAF provides just such a framework of desirable qualities and aptitudes (Baldwin *et al*, 1999) and reliably assesses improvement in performance in the workplace (Paisley *et al*, 2001a) across 70 fields that are acceptable to the trainees themselves (Section IV). This study demonstrates that structured feedback of both numerical ratings and assessor comments from EBTAF assessments of in-post performance is highly valued by trainees and suggests a demonstrably positive effect on the training of a fully competent surgeon. It is clearly an area that demands further study.

V.5. SUMMARY

- The collation and provision of detailed and structured feedback of multidisciplinary assessment of in-post performance of surgical trainees using EBSTAF was feasible.

- Detailed structured feedback demonstrated a positive effect on the everyday performance of surgical trainees by identifying areas of poor performance, allowing them to be addressed and improved.

- No overall statistical difference between the current feedback group and the historical trainee cohort was demonstrable. However, differences between the two groups in isolated domains suggest a positive effect that may have been confounded by study design.

- Trainees rated the provision of feedback of performance as highly valuable to their basic surgical training.

Table V.1

Distribution of Assessments by Hospital.

HOSPITAL	EPISODES	ASSESSMENTS
Royal Infirmary, Edinburgh	72	519
Western General Hospital, Edinburgh	31	166
St. John's Hospital, Livingston	24	168
Princess Margaret Rose Orthopaedic Hospital, Edinburgh	12	78
Royal Hospital for Sick Children, Edinburgh	8	57
Queen Margaret Hospital, Dunfermline	5	33
City Hospital, Edinburgh	3	18
TOTAL	**155**	**1039**

Episode: Assessment of a single trainee at the end of a single post by multiple assessors.

Assessment: Evaluation of a single trainee at the end of a single post by a single assessor.

Table V.2

Distribution of Assessments by Specialty.

SPECIALTY	EPISODES	ASSESSMENTS
General Surgery	73	533
Orthopaedic Surgery	45	244
Plastic Surgery	8	67
Paediatric Surgery	8	57
Cardiothoracic Surgery	8	48
Vascular Surgery	4	32
Neurosurgery	4	28
Ear, Nose & Throat	3	18
Breast Surgery	2	12
TOTAL	**155**	**1039**

Episode: Assessment of a single trainee at the end of a single post by multiple assessors.
Assessment: Evaluation of a single trainee at the end of a single post by a single assessor.

Table V.3

Distribution of Assessments by Assessor.

ASSESSOR	n =	ASSESSMENTS
Medical Staff	**128**	**286**
Consultants	60	143
Registrars	68	143
Nursing Staff	**112**	**615**
Home Ward	34	166
Emergency Ward	4	48
ICU	11	30
HDU	9	60
Theatre	30	153
Day Bed Unit	3	41
Outpatients'	21	117
SHOs	**44**	**138**
TOTAL	**284**	**1039**

Episode: Assessment of a single trainee at the end of a single post by multiple assessors.
Assessment: Evaluation of a single trainee at the end of a single post by a single assessor.

Table V.4

Response Rate for Assessor Groups.

ASSESSOR	Distributed	Returned by Deadline	Response Rate (%)
Medical Staff	286	202	70.6
Nursing Staff	615	462	75.1
Multidisciplinary	901	664	73.7
SHOs	138	119	82.2

Table V.5

Validity of Assessments by EBSTAF Domain for Different Assessor Groups.

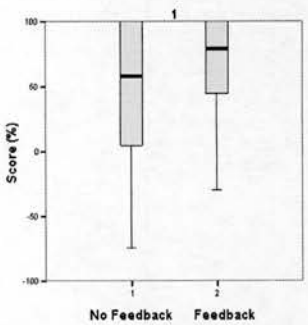
EBSTAF DOMAIN	Medical	Nursing	Multidisciplinary	SHO
	n = 202	n = 462	n = 664	n = 119
Communication	162	337	377	119
	80.2%	72.9%	57%	100%
Knowledge	108	58	166	111
	53.5%	12.6%	25%	93.3%
Teamwork	133	219	352	116
	65.8%	47.4%	53%	97.5%
Clinical Skills	158	212	370	118
	78.2%	45.9%	55.7%	99.2%
Technical Skills	140	102	242	109
	69.3%	22.1%	36.5%	91.6%

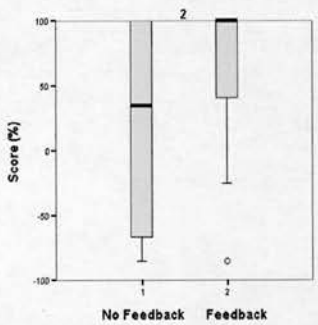
The validity of each domain within an individual assessment was determined by greater than 75% of the fields therein being directly observed.

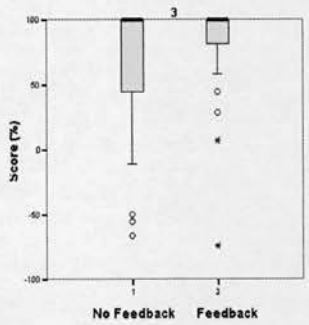
Table V.6a

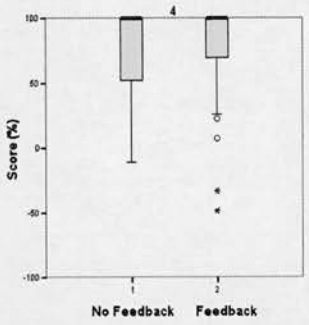
Comparison of Non-Feedback and Feedback. COMMUNICATION - Medical.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**58.3**	0.3 to 100	**78.8**	44.4 to 100	ns
2	**35.2**	-66.7 to 100	**100**	37.5 to 100	0.079
3	**100**	40.5 to 100	**100**	81.5 to 100	ns
4	**100**	50.4 to 100	**100**	63.0 to 100	ns
5	**48.4**	-25.7 to 100	**100**	81.5 to 100	0.067









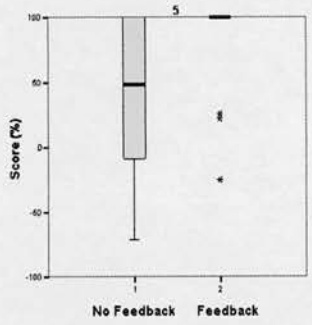


Table V.6b

Comparison of Non-Feedback and Feedback. KNOWLEDGE - Medical.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**29.6**	-44.1 to 76.2	**35.9**	7.3 to 100	ns
2	**16.7**	-38.9 to 82.1	**46.4**	13.7 to 100	ns
3	**34.5**	-26.2 to 83.2	**64.3**	-11.1 to 100	ns
4	**74.4**	40.5 to 100	**61.9**	16.7 to 93.6	ns
5	**44.1**	-17.3 to 97.0	**88.1**	28.6 to 100	ns

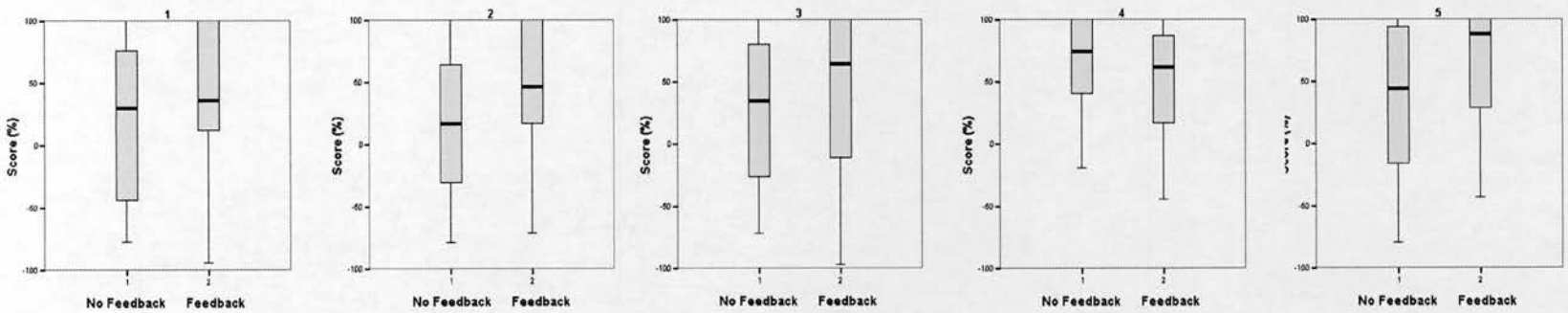
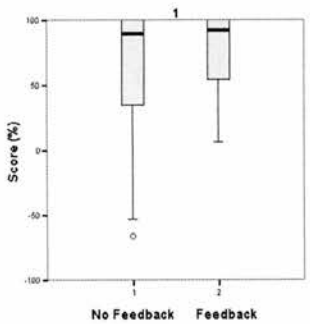
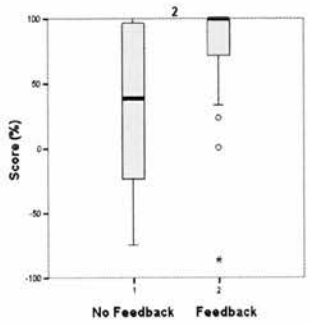


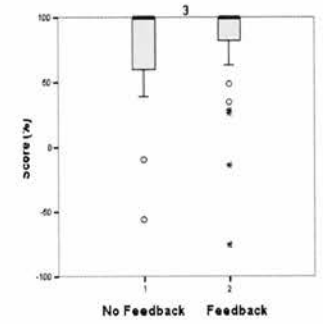
Table V.6c

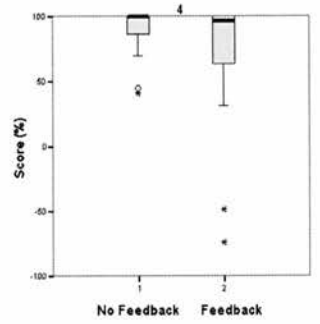
Comparison of Non-Feedback and Feedback. TEAMWORK - Medical.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**89.4**	34.4 to 100	**92.4**	54.5 to 100	ns
2	**38.7**	-32.8 to 98.3	**100**	69.3 to 100	0.015
3	**100**	53.3 to 100	**100**	78.3 to 100	ns
4	**100**	86.1 to 100	**96.7**	62.8 to 100	0.047
5	**100**	60.9 to 100	**100**	86.7 to 100	ns









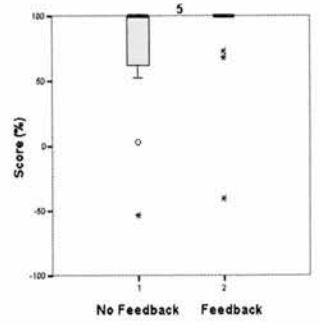


Table V.6d

Comparison of Non-Feedback and Feedback. CLINICAL SKILLS - Medical.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**62.8**	6.6 to 89.9	**75.8**	26.4 to 100	ns
2	**74.6**	11.5 to 98.6	**79.9**	17.3 to 100	ns
3	**84.3**	40.7 to 100	**93.3**	62.7 to 100	ns
4	**91.4**	56.5 to 100	**91.4**	70.6 to 100	ns
5	**86.7**	56.6 to 100	**100**	82.2 to 100	ns

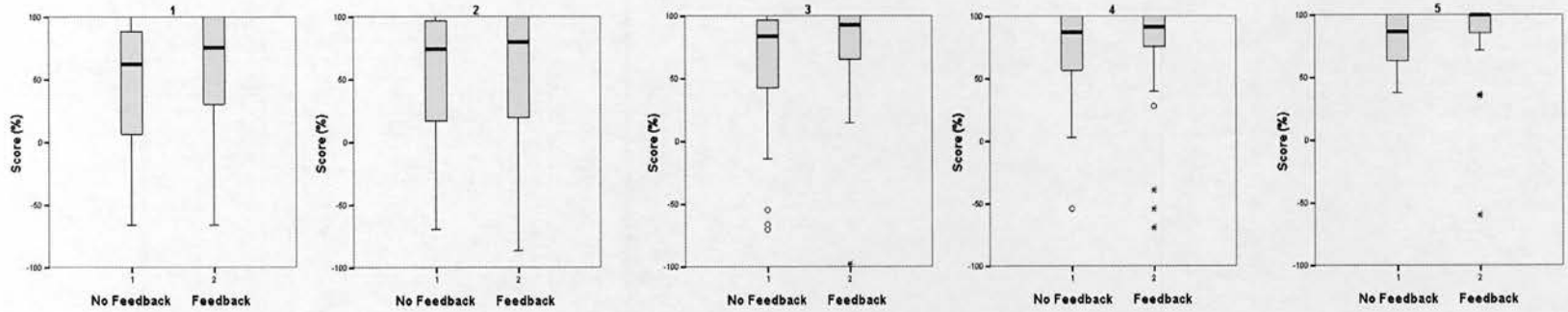


Table V.6e

Comparison of Non-Feedback and Feedback.
TECHNICAL SKILLS - Medical.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**14.6**	-47.6 to 44.4	**21.6**	-9.7 to 51.0	ns
2	**1.7**	-39.9 to 77.0	**36.9**	-62.6 to 85.8	ns
3	**58.3**	18.5 to 83.7	**82.9**	60.4 to 100	0.029
4	**79.7**	26.0 to 100	**91.9**	7.0 to 100	ns
5	**55.3**	-15.4 to 90.5	**100**	83.3 to 100	0.003

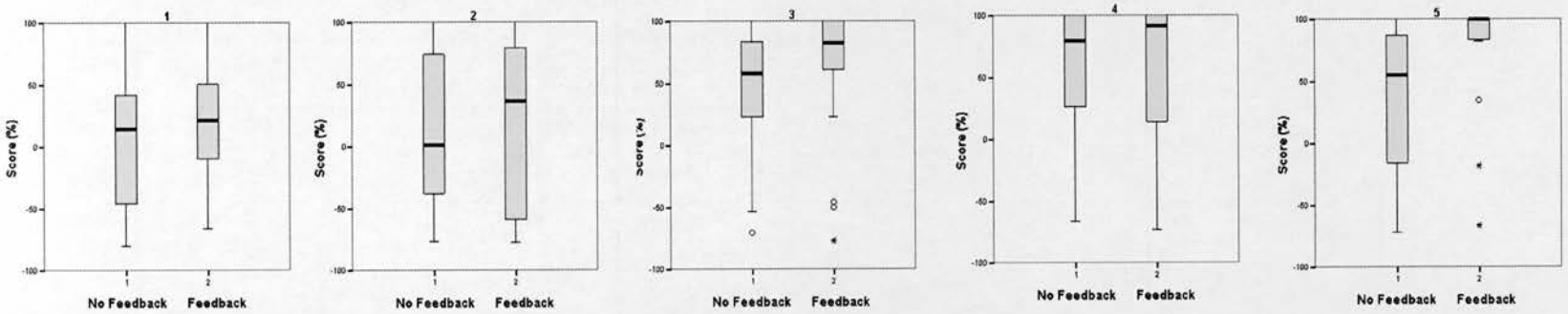


Table V.6f

Comparison of Non-Feedback and Feedback. EBSTAF OVERALL - Medical.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**46.7**	-1.8 to 82.9	**84.2**	64.9 to 98.0	<0.001
2	**34.2**	-13.5 to 87.9	**83.3**	55.9 to 96.4	0.007
3	**74.9**	30.5 to 91.2	**93.7**	79.7 to 99.1	<0.001
4	**79.7**	54.5 to 97.7	**95.0**	78.2 to 98.5	ns
5	**52.3**	5.3 to 87.1	**72.6**	27.5 to 93.6	<0.001

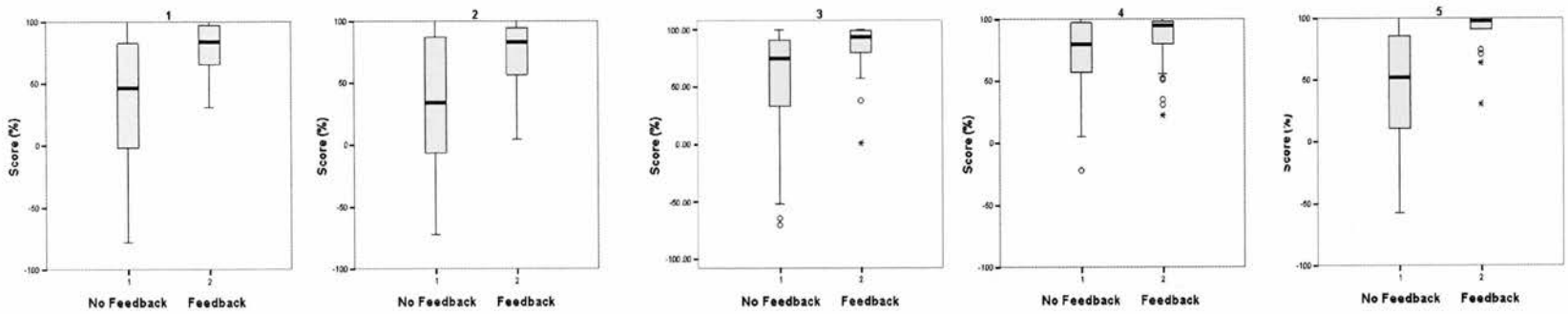
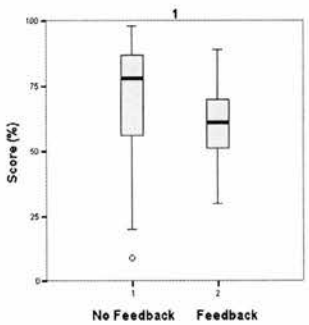
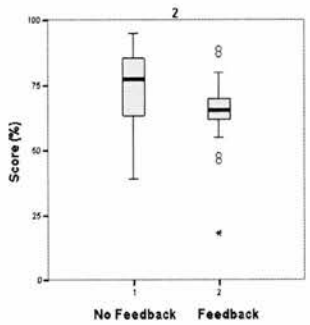


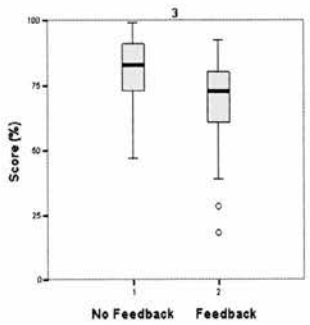
Table V.6g

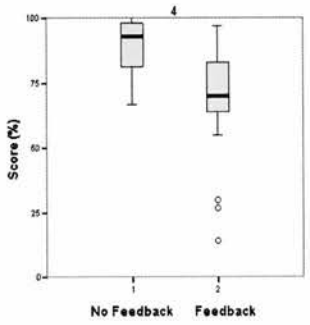
Comparison of Non-Feedback and Feedback.
VAS OVERALL - Medical.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**78.0**	55.5 to 87.0	**61.0**	51.1 to 70.0	<0.001
2	**77.5**	63.3 to 85.8	**65.7**	61.5 to 70.0	0.039
3	**83.0**	73.0 to 91.0	**73.0**	60.8 to 81.0	<0.001
4	**93.0**	81.3 to 98.0	**70.0**	64.0 to 84.3	<0.001
5	**80.5**	65.3 to 92.0	**79.0**	56.3 to 94.8	ns









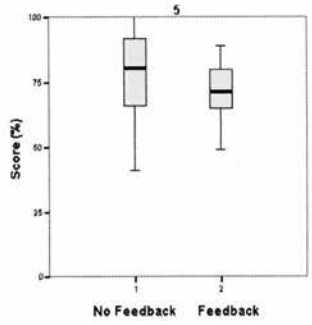


Table V.7a

Comparison of Non-Feedback and Feedback. Communication - Nursing.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**100**	16.7 to 100	**100**	81.5 to 100	0.022
2	**100**	47.9 to 100	**100**	100 to 100	0.004
3	**100**	100 to 100	**100**	95.4 to 100	ns
4	**100**	95.4 to 100	**100**	100 to 100	ns
5	**100**	71.1 to 100	**100**	56.9 to 100	ns

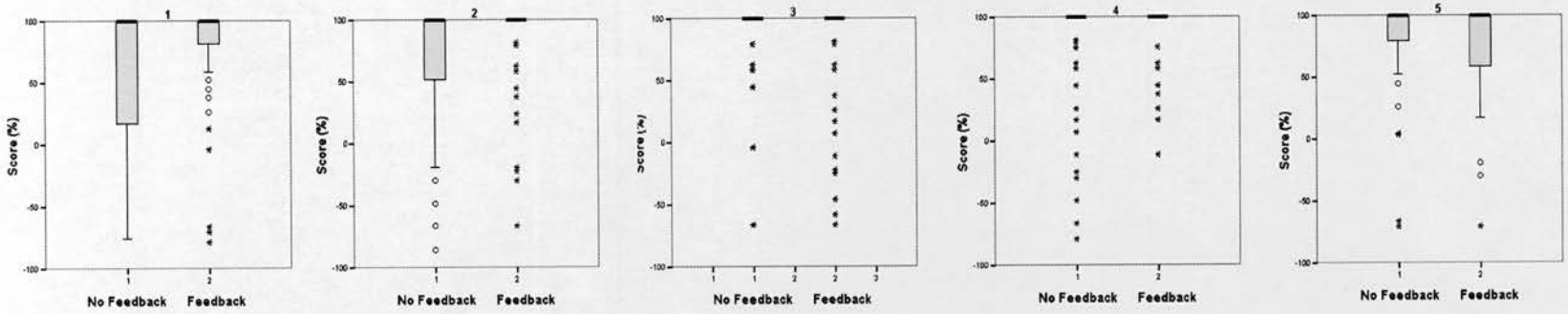


Table V.7b

Comparison of Non-Feedback and Feedback. Knowledge - Nursing.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**100**	72.2 to 100	**100**	-4.76 to 100	ns
2	**100**	79.2 to 100	**100**	100 to 100	0.035
3	**100**	100 to 100	**100**	79.2 to 100	ns
4	**100**	44.3 to 100	**100**	100 to 100	ns
5	**100**	-38.9 to 100	**100**	85.7 to 100	ns

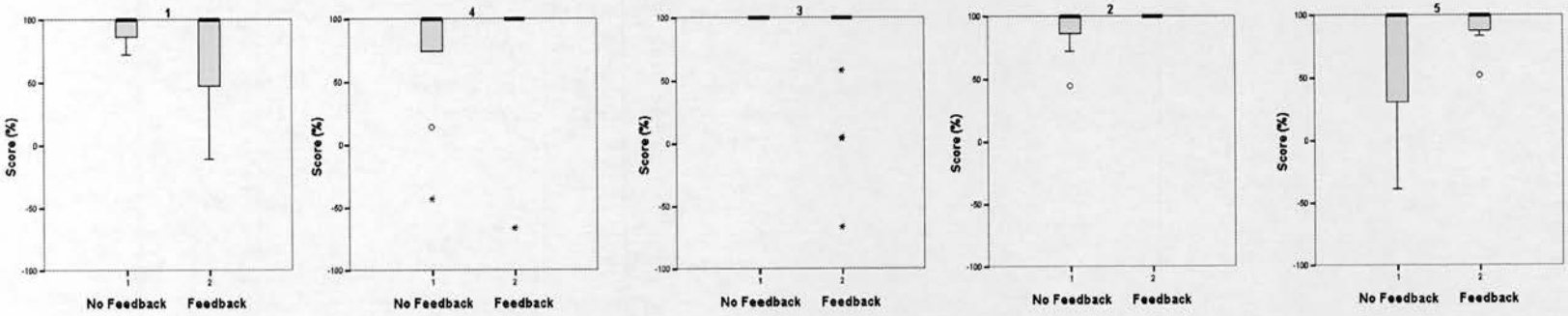


Table V.7c

Comparison of Non-Feedback and Feedback. Teamwork - Nursing.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**100**	76.7 to 100	**100**	84.1 to 100	ns
2	**100**	60.0 to 100	**100**	100 to 100	0.041
3	**100**	92.8 to 100	**100**	34.8 to 100	0.005
4	**100**	100 to 100	**100**	100 to 100	ns
5	**100**	100 to 100	**100**	48.1 to 100	ns

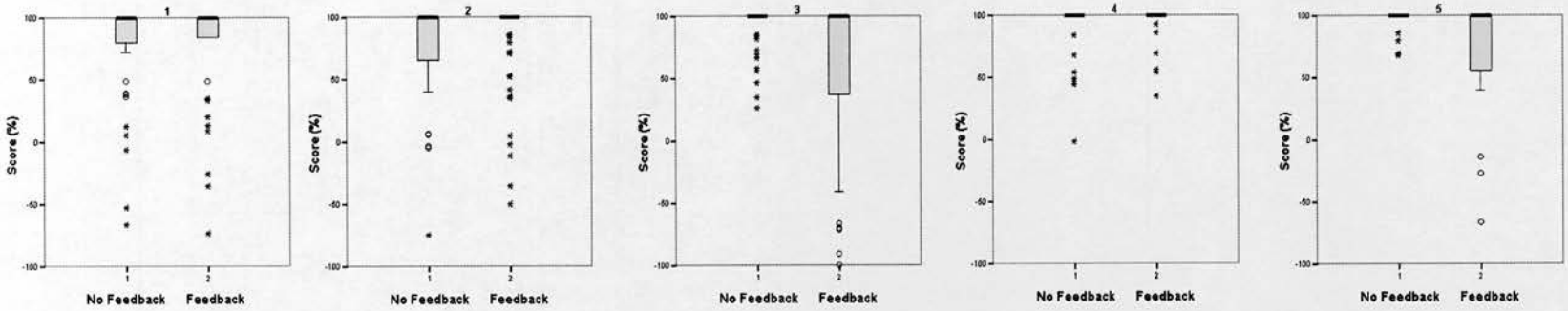


Table V.7d

Comparison of Non-Feedback and Feedback. Clinical Skills - Nursing.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**87.2**	31.0 to 100	**100**	74.3 to 100	0.058
2	**100**	86.9 to 100	**100**	100 to 100	0.028
3	**100**	93.0 to 100	**100**	82.1 to 100	ns
4	**100**	87.2 to 100	**100**	86.5 to 100	ns
5	**91.7**	42.1 to 100	**100**	59.9 to 100	ns

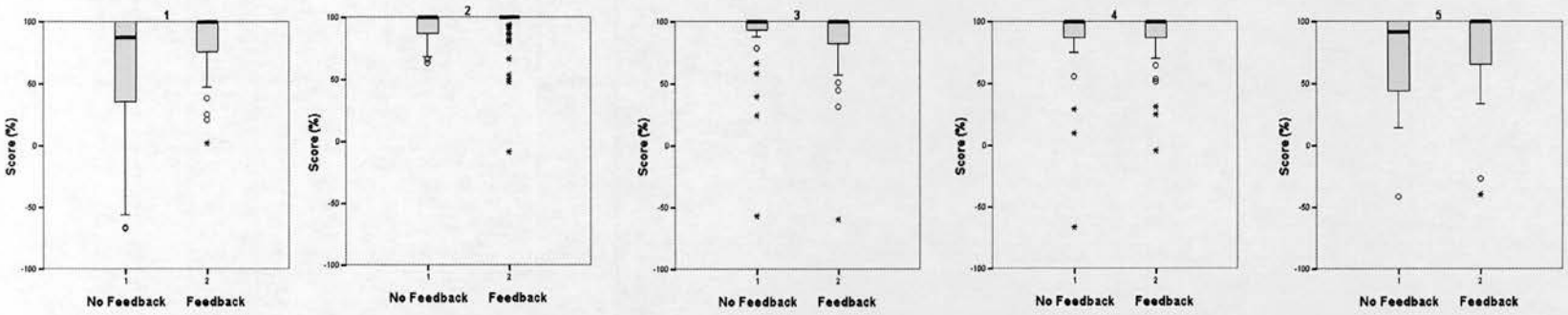


Table V.7e

Comparison of Non-Feedback and Feedback. Technical Skills - Nursing.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**53.0**	-2.9 to 100	**59.4**	7.8 to 100	ns
2	**44.4**	-28.6 to 82.9	**59.6**	12.2 to 100	ns
3	**77.6**	41.2 to 91.9	**53.3**	-38.4 to 98.0	ns
4	**91.9**	91.9 to 100	**91.9**	57.4 to 100	ns
5	**100**	79.7 to 100	**82.1**	35.0 to 97.5	ns

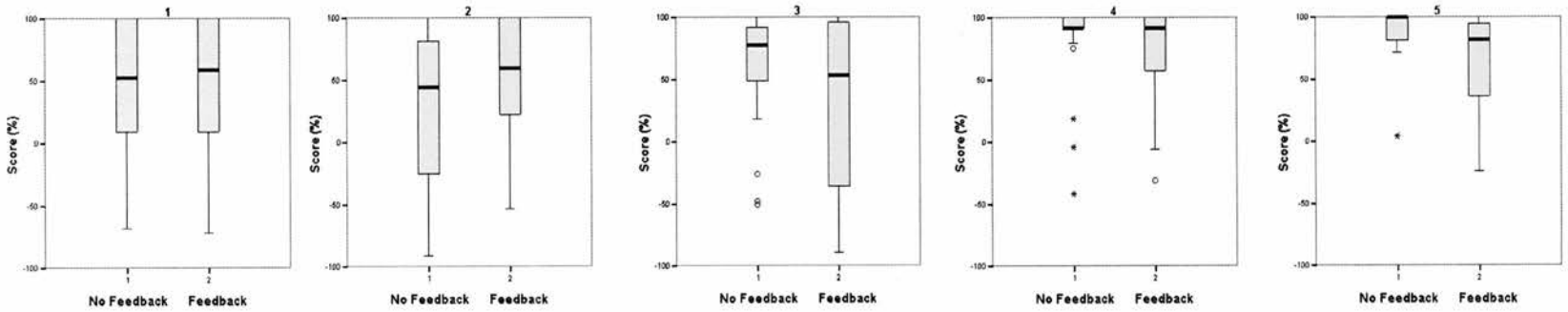


Table V.7f

Comparison of Non-Feedback and Feedback. EBSTAF Overall - Nursing.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**88.1**	38.0 to 100	**96.0**	72.7 to 100	0.036
2	**93.8**	58.6 to 100	**100**	87.4 to 100	0.016
3	**100**	85.8 to 100	**97.8**	81.1 to 100	ns
4	**100**	86.6 to 100	**100**	86.5 to 100	ns
5	**95.6**	70.3 to 100	**97.0**	78.9 to 100	ns

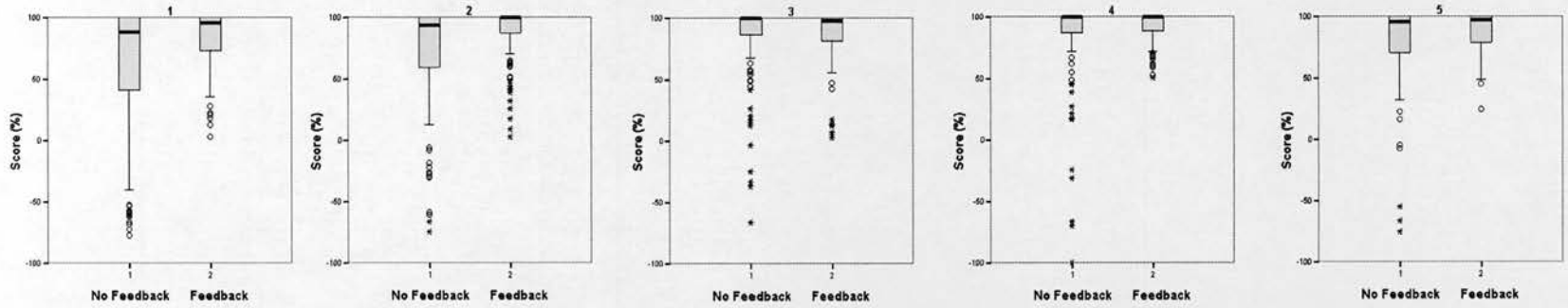
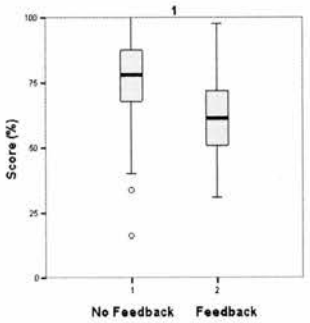
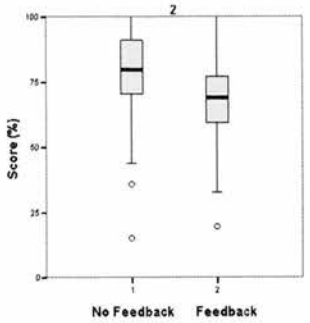


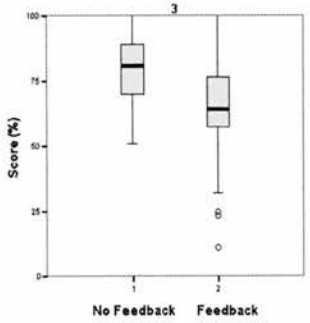
Table V.7g

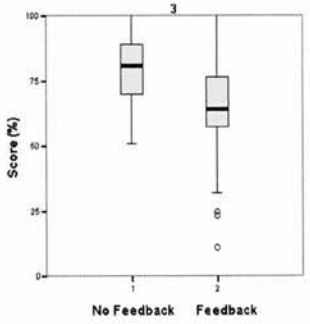
Comparison of Non-Feedback and Feedback. VAS Overall - Nursing.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**78.0**	68.0 to 88.0	**61.5**	51.0 to 72.3	<0.001
2	**80.0**	70.0 to 91.0	**69.0**	59.3 to 77.4	<0.001
3	**81.0**	70.0 to 89.0	**64.2**	57.0 to 78.0	<0.001
4	**85.0**	72.0 to 92.0	**69.3**	58.4 to 78.0	<0.001
5	**75.0**	64.0 to 91.0	**61.5**	52.3 to 74.3	<0.001









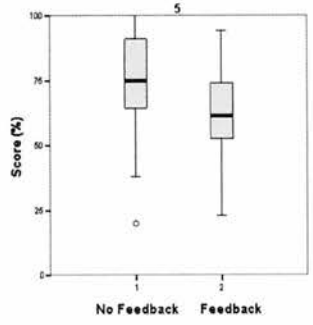


Table V.8a

Comparison of Non-Feedback and Feedback. COMMUNICATION - Multidisciplinary.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**100**	12.0 to 100	**100**	63.0 to 100	0.010
2	**100**	44.4 to 100	**100**	79.2 to 100	0.002
3	**100**	81.5 to 100	**100**	81.5 to 100	ns
4	**100**	66.0 to 100	**100**	76.2 to 100	ns
5	**100**	27.3 to 100	**100**	58.3 to 100	ns

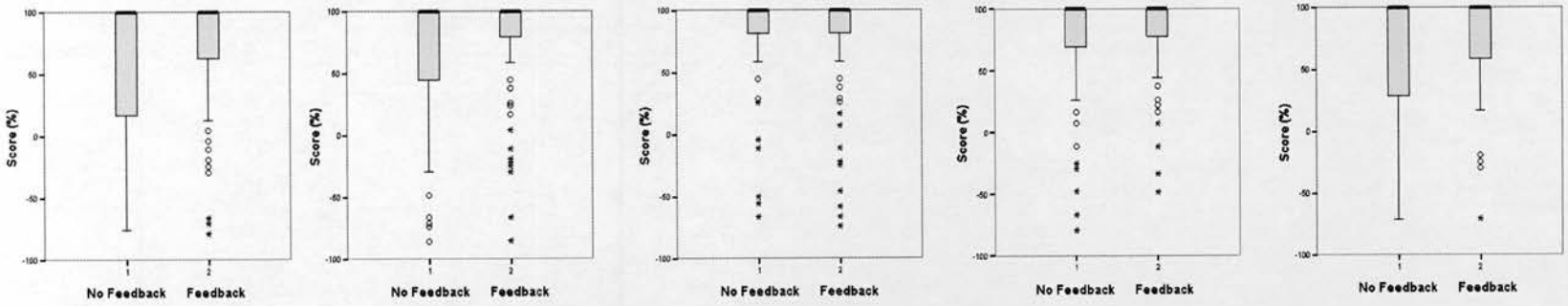


Table V.8b

Comparison of Non-Feedback and Feedback. KNOWLEDGE - Multidisciplinary.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**50.0**	-42.9 to 100	**46.0**	5.1 to 100	ns
2	**44.4**	-9.53 to 100	**100**	30.6 to 100	0.059
3	**61.5**	-14.3 to 100	**100**	24.2 to 100	ns
4	**81.2**	40.5 to 100	**67.9**	17.0 to 100	ns
5	**52.4**	-19.1 to 100	**100**	52.4 to 100	0.069

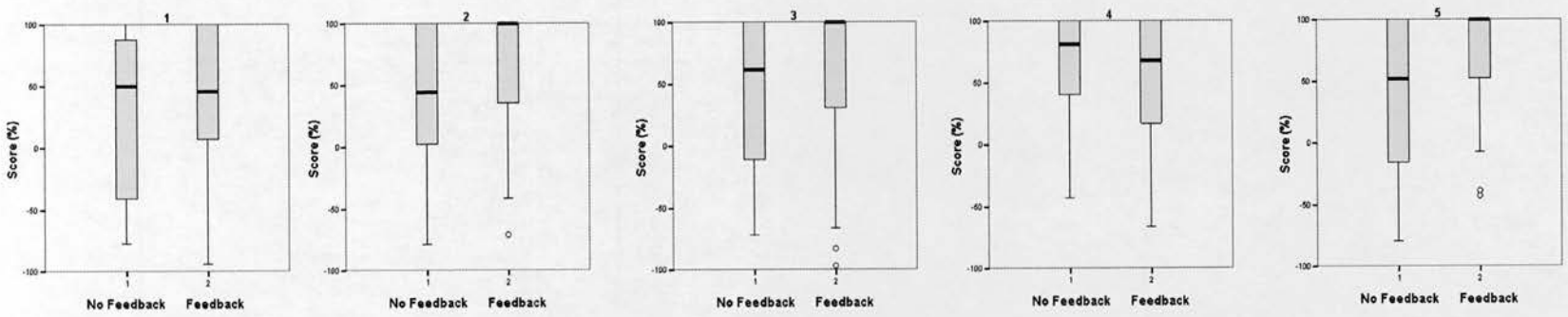


Table V.8c

Comparison of Non-Feedback and Feedback. TEAMWORK - Multidisciplinary.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**100**	49.3 to 100	**100**	60.6 to 100	ns
2	**100**	50.7 to 100	**100**	86.7 to 100	0.001
3	**100**	80.2 to 100	**100**	80.2 to 100	0.081
4	**100**	100 to 100	**100**	75.0 to 100	0.067
5	**100**	81.7 to 100	**100**	73.3 to 100	ns

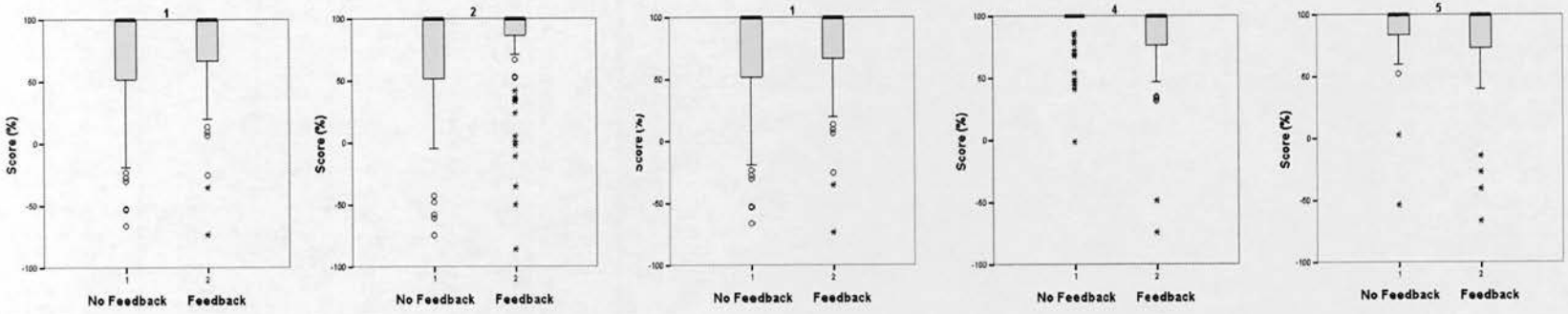
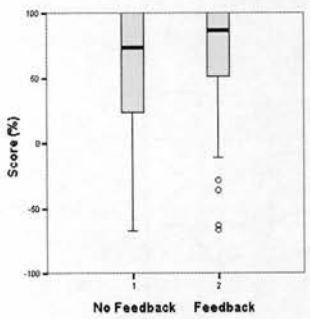
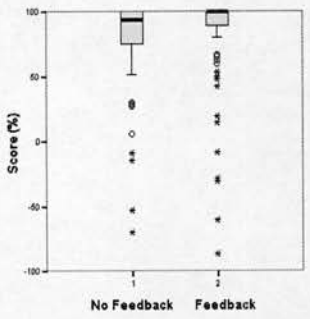


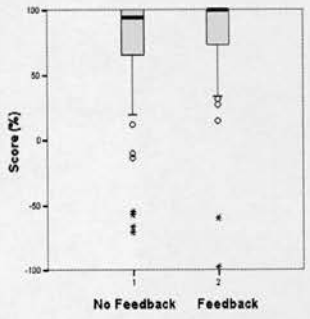
Table V.8d

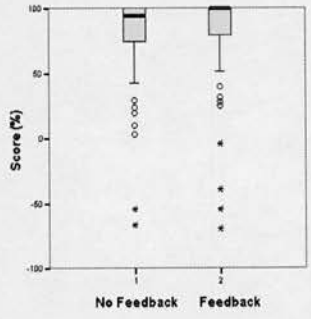
Comparison of Non-Feedback and Feedback. CLINICAL SKILLS - Multidisciplinary.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**73.3**	23.3 to 100	**86.7**	50.8 to 100	0.023
2	**93.6**	74.7 to 100	**100**	88.3 to 100	0.013
3	**94.4**	61.9 to 100	**100**	73.4 to 100	ns
4	**94.4**	74.6 to 100	**100**	79.0 to 100	ns
5	**89.4**	54.0 to 100	**100**	71.4 to 100	ns









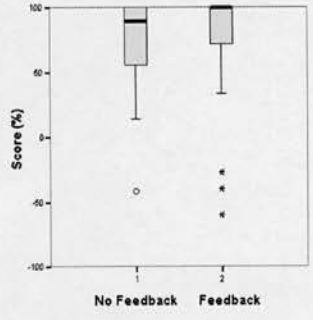


Table V.8e

Comparison of Non-Feedback and Feedback. TECHNICAL SKILLS - Multidisciplinary.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**30.1**	-31.4 to 71.6	**40.2**	-3.8 to 72.0	ns
2	**32.1**	-29.1 to 80.3	**47.3**	-19.7 to 100	ns
3	**60.4**	23.2 to 88.4	**71.5**	39.0 to 100	ns
4	**91.9**	51.2 to 100	**91.9**	32.5 to 100	ns
5	**81.3**	27.0 to 100	**83.3**	54.6 to 100	ns

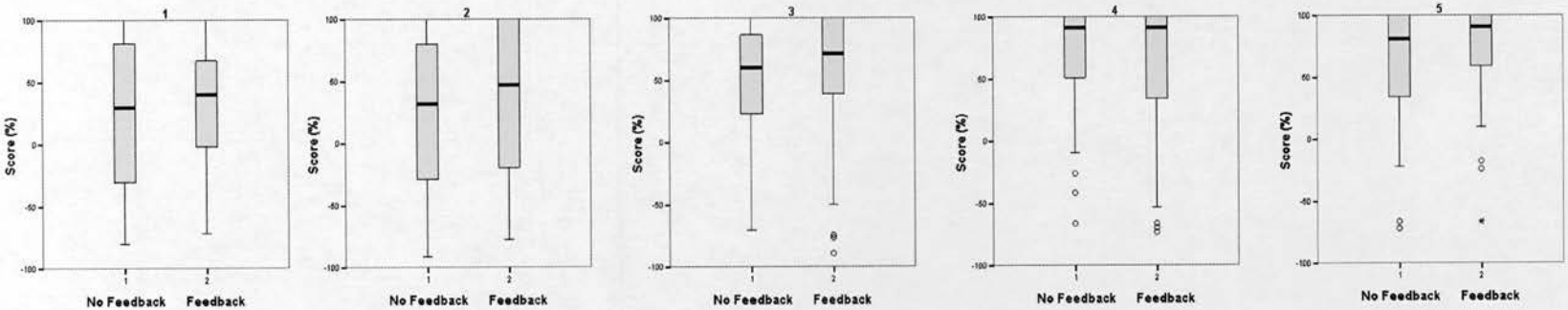


Table V.8f

Comparison of Non-Feedback and Feedback. EBSTAF OVERALL - Multidisciplinary.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**74.6**	19.2 to 100	**91.8**	70.0 to 100	<0.001
2	**88.7**	51.0 to 100	**98.1**	81.4 to 100	0.002
3	**94.7**	63.5 to 100	**98.0**	80.7 to 100	ns
4	**98.1**	72.4 to 100	**97.6**	85.9 to 100	ns
5	**87.4**	50.2 to 100	**97.2**	79.5 to 100	0.015

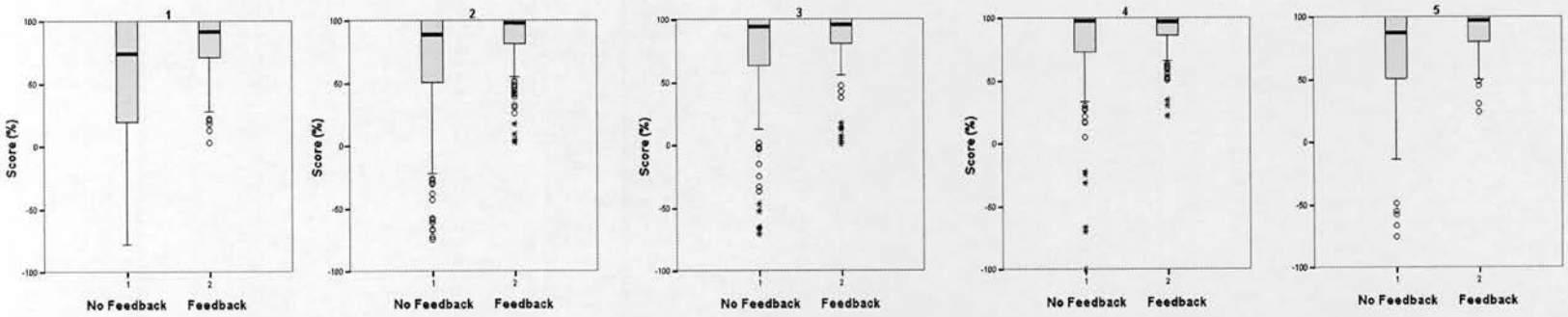
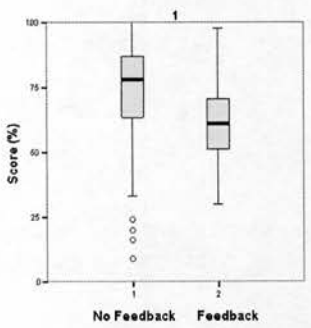
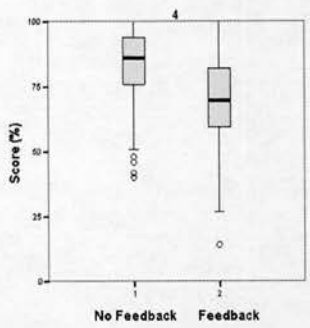


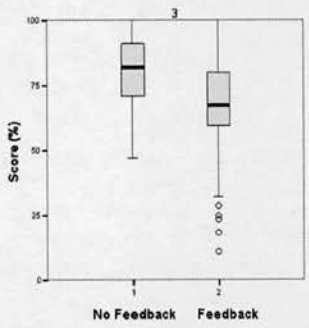
Table V.8g

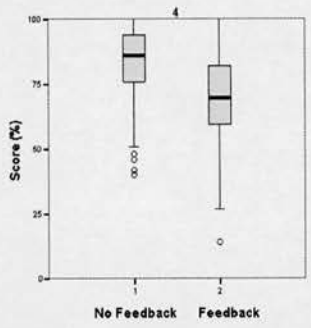
Comparison of Non-Feedback and Feedback. VAS OVERALL - Multidisciplinary.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**78.0**	63.3 to 87.0	**61.0**	51.1 to 70.9	<0.001
2	**80.0**	68.0 to 91.0	**68.0**	60.0 to 76.3	<0.001
3	**82.0**	71.0 to 91.0	**67.5**	59.2 to 80.0	<0.001
4	**86.0**	76.0 to 94.0	**70.0**	59.3 to 82.3	<0.001
5	**76.0**	64.5 to 91.5	**67.0**	54.5 to 75.0	<0.001









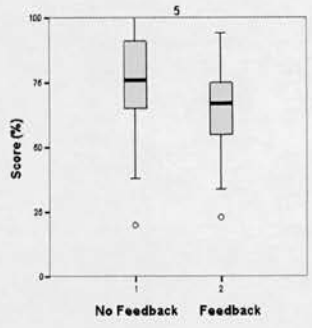


Table V.9a

Comparison of Non-Feedback and Feedback. Communication – SHO Self Assessment.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**44.4**	7.4 to 81.5	**81.5**	39.8 to 100	ns
2	**81.5**	49.1 to 100	**100**	72.2 to 100	ns
3	**100**	100 to 100	**100**	81.5 to 100	ns
4	**100**	44.4 to 100	**100**	63.0 to 100	ns
5	**63.0**	-27.3 to 90.7	**100**	44.4 to 100	ns

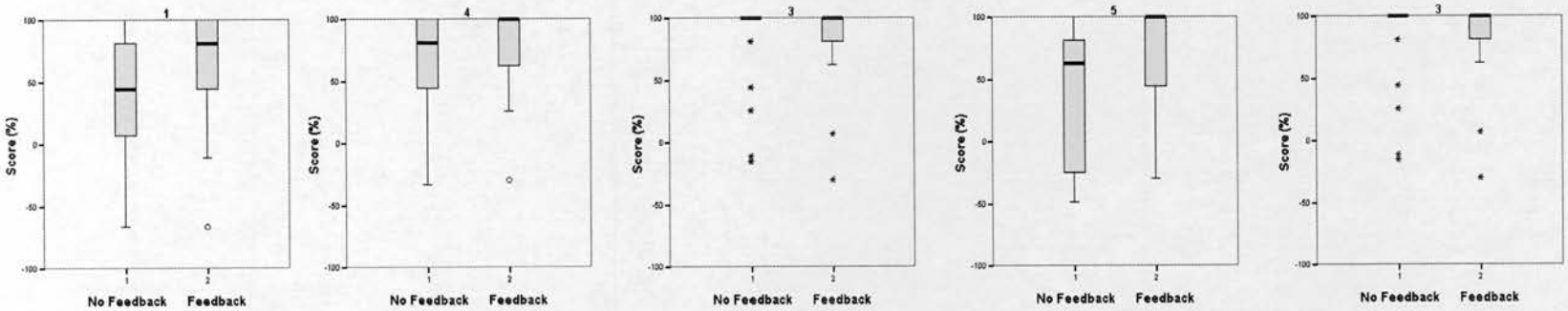


Table V.9b

Comparison of Non-Feedback and Feedback. Knowledge - SHO Self Assessment.

BST Post	**Non-Feedback Cohort**		**Feedback Cohort**		**p =**
	Median	IQR	Median	IQR	Mann-Whitney U
1	**2.4**	-19.1 to 28.6	**40.5**	-10.1 to 86.6	ns
2	**16.7**	4.8 to 76.2	**64.3**	37.5 to 88.1	0.087
3	**40.5**	-7.1 to 72.2	**40.5**	1.8 to 67.3	ns
4	**34.5**	-25.2 to 61.3	**40.5**	-7.1 to 86.1	ns
5	**22.6**	4.8 to 67.3	**82.1**	52.4 to 100	0.04

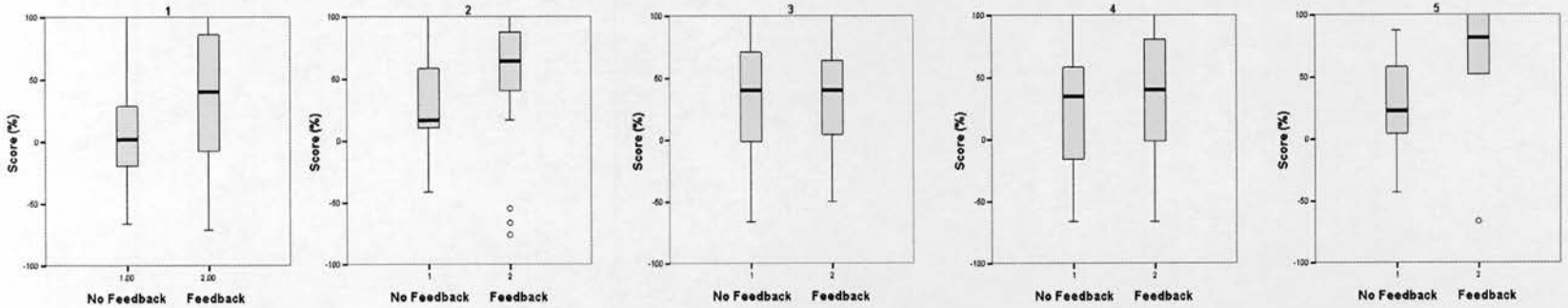


Table V.9c

Comparison of Non-Feedback and Feedback. Teamwork - SHO Self Assessment.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**79.1**	533.3 to 86.7	**80.0**	48.9 to 100	ns
2	**66.7**	53.3 to 96.7	**93.3**	73.3 to 100	0.04
3	**86.4**	66.7 to 100	**80.0**	36.7 to 100	ns
4	**86.7**	35.0 to 100	**82.8**	65.9 to 100	ns
5	**76.7**	30.3 to 86.7	**90.0**	52.9 to 100	ns

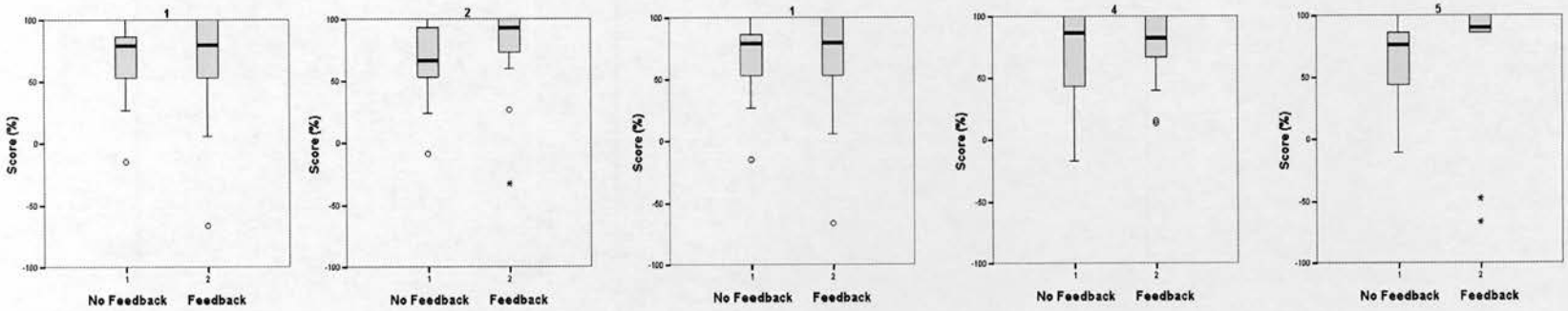
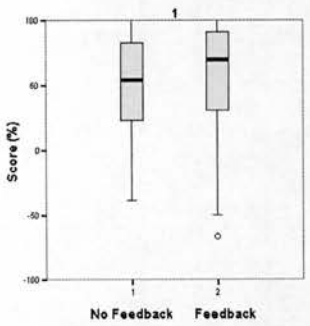
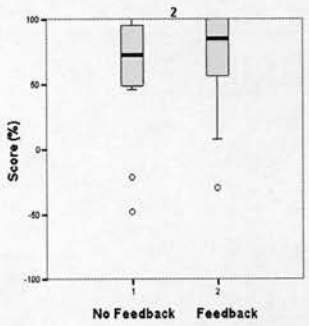


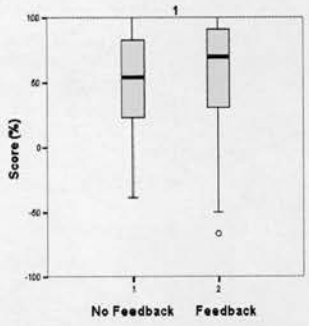
Table V.9d

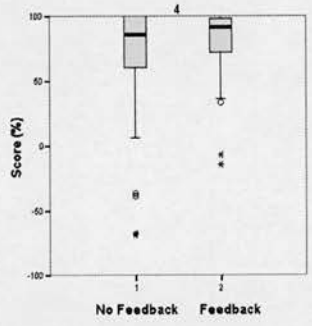
Comparison of Non-Feedback and Feedback. Clinical Skills - SHO Self Assessment.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**53.8**	21.9 to 83.5	**69.8**	31.0 to 92.1	ns
2	**72.7**	47.7 to 96.5	**85.6**	56.2 to 100	ns
3	**88.7**	72.5 to 100	**85.6**	66.6 to 95.6	ns
4	**85.9**	33.6 to 100	**91.4**	71.3 to 100	ns
5	**68.9**	-1.7 to 91.5	**98.6**	69.1 to 100	ns









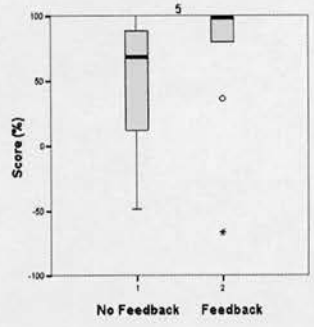


Table V.9e

Comparison of Non-Feedback and Feedback. Technical Skills - SHO Self Assessment.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**53.0**	-2.9 to 100	**59.4**	7.8 to 100	ns
2	**44.4**	-28.6 to 82.9	**59.6**	12.2 to 100	ns
3	**77.6**	41.2 to 91.9	**53.3**	-38.4 to 98.0	ns
4	**91.9**	91.9 to 100	**91.9**	57.4 to 100	ns
5	**100**	79.7 to 100	**82.1**	35.1 to 97.5	ns

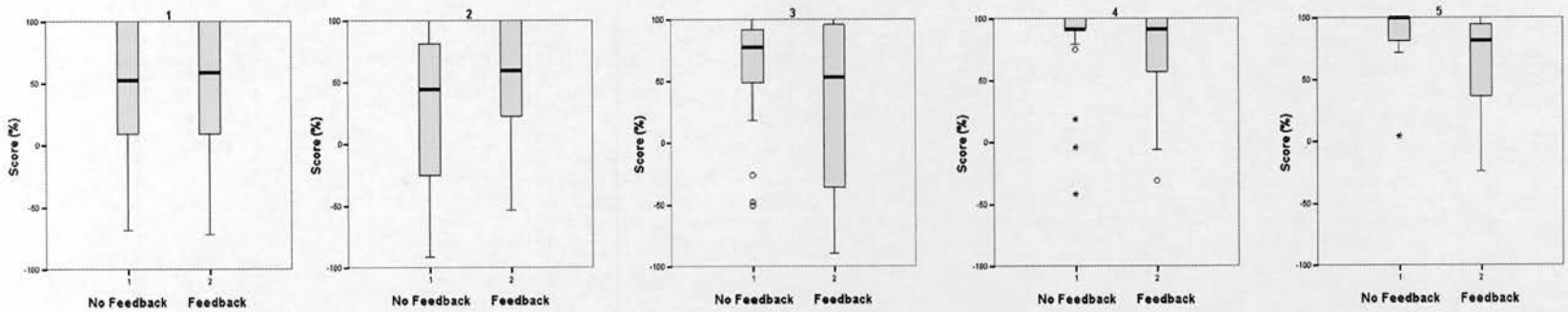
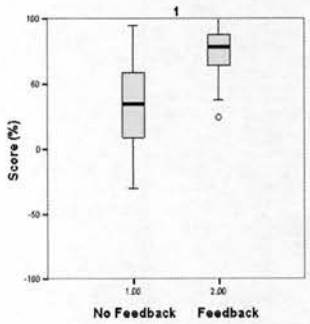
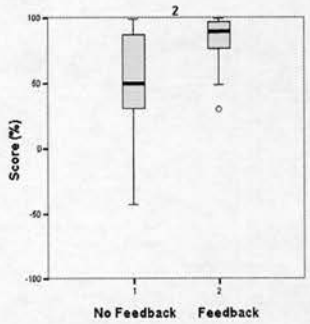


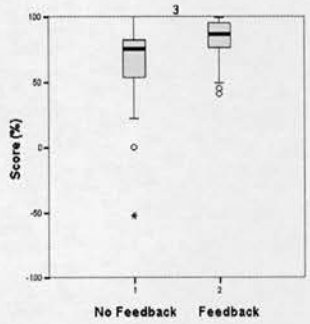
Table V.9f

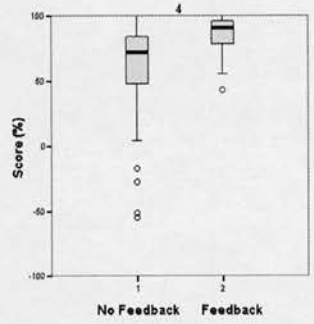
Comparison of Non-Feedback and Feedback. EBSTAF Overall - SHO Self Assessment.

BST Post	Non-Feedback Cohort		Feedback Cohort		p =
	Median	IQR	Median	IQR	Mann-Whitney U
1	**34.8**	8.7 to 58.9	**78.4**	62.8 to 90.3	<0.001
2	**49.7**	28.3 to 87.25	**89.3**	76.2 to 97.0	0.003
3	**75.4**	53.3 to 83.2	**87.0**	76.1 to 95.4	0.006
4	**72.2**	26.1 to 84.1	**91.0**	73.7 to 96.2	0.012
5	**65.0**	-15.2 to 85.1	**97.2**	80.8 to 98.8	0.005









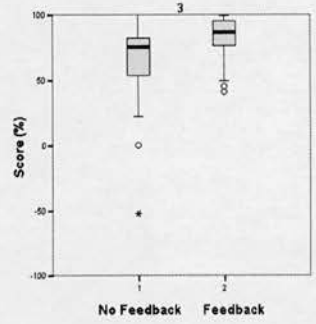


Table V.10

Trainees' Suggestions / Comments.

Comments were made freely on the evaluation sheet within a box entitled :

"How might the assessment / feedback be improved ?"

If it became recognised by training bodies / colleges / other hospital it would be very useful to carry forward.

Very thorough and fair assessment process.

Splitting the comment summaries for each section might be useful.

I really feel it was very worthwhile. Very helpful. Occasionally forms may have been sent to people not ideally placed to assess me.

Not sure it can be (improved). I would like to know who made the comments (but I know I can't !).

Some indication of who is making comments – not wanting names but knowing whether nursing or medical would help to put comments into context.

Gives a good feel of colleagues' impressions of you. Most useful.

Assessments from 3 or more consultants.

Feedback was very detailed and useful. I know there will always be non-responders, but for consultants to not feed back information to trainees is very poor. I find the form very useful and cannot really suggest any improvements.

More time to discuss would be desirable.

Perhaps visual analogue scales are more useful than the detailed field summaries – they are certainly more "eye-friendly".

Assessors must be people who have spent enough time observing the trainee to give a fair assessment. Otherwise good.

Occasionally nursing staff used their assessor status to reinforce their power over you. I don't know if the questionnaire could be targeted in such a way as to discourage this.

Tick box (Field Summary) too defined – if visual analogue scale used this would give more freedom. Comments interesting and surprising.

If profoundly negative comments are made the assessor should be identifiable either by name or position so that problems with particular individuals could be addressed.

Figure V.1.

Trainees' Evaluation of Structured Feedback By Visual Analogue Scale.

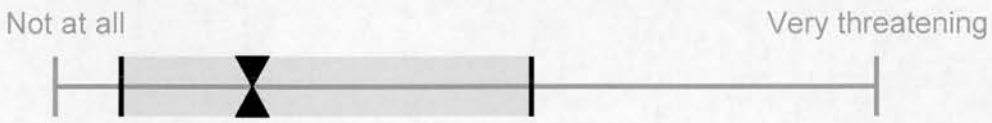
1. Did you find the formal feedback useful in your training ?

2. Did you find the feedback given to be fair ?

3. How did you find the level of feedback given ?

4. Did you find being assessed by your colleagues to be threatening ?

5. Did you find being assessed by nursing staff to be threatening ?



Visual-analogue scales represent Median and Inter-Quartile Range

Section VI.

ASSESSMENT OF BASIC SURGICAL TRAINEES' CRITICAL CARE SKILLS USING HUMAN PATIENT SIMULATION (HPS)

VI.1. INTRODUCTION.

Care of the critically ill patient is the most challenging area in clinical practice, irrespective of a doctor's level of training. Concerns have been raised as to the competence of junior doctors to deal with such patients (Toogood *et al*, 1996) and yet it is in the critical care domain of clinical practice that training and assessment is most difficult. Critical events cannot be left untreated while awaiting an appropriate response from the trainee being assessed (Devitt *et al*, 2001) and the necessary early involvement of more senior trainees or consultants confounds attempts to assess a junior trainee's true abilities.

Human patient simulation (HPS) offers absolute safety for both patients and trainees while facilitating learning and assessment of critical care skills by the creation of relevant standardised scenarios (Devitt *et al*, 2001; Beyea, 2004).

VI.2. AIMS.

To design a purely formative one-day course for BSTs to address specific learning objectives within simulated perioperative critical care scenarios on the human patient simulator and to examine the feasibility, reliability and validity of assessing BSTs within such a course using a modification of EBSTAF.

- To examine the effect of feedback on trainees' self-assessment skills when provided within a structured debriefing.

- To determine the opinions of trainees attending the course as to the value of such a course.

VI.3. RESULTS.

VI.3.a. Study Participants.

VI.3.a.i. Surgical Trainees.

The course was originally designed for the 35 southeast Scotland BSTs to allow examination of concurrent validity by parallel in-post assessment of everyday practice using EBSTAF. However, news of the course quickly spread by word-of-mouth and a total of 21 trainees from West of Scotland (n=17) and Tayside (n=4) regions independently requested the opportunity to attend. With the full support of their training programme chairmen, extra courses were made available. This unexpected demand suggests a perceived need amongst trainees for such a course addressing critical care.

A total of 56 BSTs attended 17 courses over a 13-month time period (Oct 2001 – November 2002). Surgical experience ranged from 1 to 40 months and is summarised in Table VI.1.

The additional data from WoS and Tayside trainees was included in the analysis of reliability and the effect of the debriefing process. However, differences in programme structure precluded their use for analysis of construct validity since their experience of critical care differed considerably

between programmes. Since only SES BSTs were assessed in-post by EBSTAF, WoS and Tayside BSTs were also excluded from the examination of concurrent validity. The breakdown of data analysis is shown in Figure VI.1.

VI.3.a.ii. Faculty.

Faculty totalled eight consultant surgeons and three consultant anaesthetists. A minimum of one consultant surgeon and one consultant anaesthetist made up the faculty for each one-day course, although there were frequently more.

VI.3.b. Assessments.

VI.3.b.i. HPS.

One course resulted in incomplete assessments on four BSTs due to the unavoidable absence of the principal researcher (PJD). These were excluded, leaving a total of 412 assessments on 52 trainees for further analysis.

VI.3.b.ii. Clinical Assessment by EBSTAF.

Thirty-five BSTs on the SES surgical training programme underwent parallel independent multi-disciplinary assessment of their everyday practice using

EBSTAF as previously described. A total of 368 assessments were completed with each trainee receiving between 6 and 14 ratings within the assessment episode corresponding to their attendance on the course.

VI.3.c. Estimation of Reliability.

Completed assessments on all 52 trainees were included in the analysis of reliability.

VI.3.c.i. HPS.

HPS – Global Assessment (HPS-GA) demonstrated acceptable internal consistency (α = 0.95 ; table VI.2a).

Significant inter-rater correlation (Spearman rank) was demonstrated between surgeon and anaesthetist ratings for HPS-GA (*rho* r_s = 0.69, p<0.001) and VASCSM (*rho* r_s = 0.447, p=0.002). Such correlation was not seen for VASComm. An examination of concordance (Kendall's tau-b) gave similar results: HPS-GA tau-b = 0.514, p < 0.001; VASCSM tau-b = 0.295, p = 0.004; VASComm tau-b = 0.155, p = ns. (Figures VI.2.a to c)

VI.3.c.ii. Clinical Assessments by EBSTAF (Table VI.2b).

In-post EBSTAF assessments demonstrated high internal consistencies (α = 0.80 to 0.96) for individual domains and also overall. The virtual simulator score ('PseudoSim') also scored highly (α = 0.92).

VI.3.d. Analysis of Construct Validity (Tables VI.3 a-c, Figs VI.3 a-c)

Significant differences in the structures of West of Scotland and Tayside programmes resulted in some trainees gaining General Surgical and Critical Care experience comparatively late in their training. Such diverse career pathways were not felt to be fairly comparable within the scope of this study. As a result, only data from the 35 SES trainees was analysed for construct validity since each followed a common pathway for the first 18 months of their surgical training.

HPS-GA and VASCS&M scores significantly improved with increasing surgical experience (Kruskal Wallis). This was not seen for VASComm.

HPS-GA demonstrated a significant improvement between the first and second time points equivalent to 0 to 6 and 6 to 12 months of surgical training (Mann-Whitney U). No other intervals showed significant improvement in their own right.

VI.3.e. **Concurrent Validity** (Tables VI.4 a and b)

Only SES trainees underwent parallel in-post assessment by EBSTAF resulting in data on 35 trainees for the analysis of concurrent validity.

VI.3.e.i. HPS

There was both excellent rank correlation (Spearman *rho*) and concordance (Kendall's *tau-b*) between all three modalities employed within the simulator assessment process.

VI.3.e.ii. Clinical Assessment by EBSTAF.

Rank correlation (Spearman *rho*) and concordance (Kendall's *tau-b*) was demonstrated between the five EBSTAF domains and visual analogue overall score. Only ratings within the technical skills domain failed to relate to all five other domain ratings, showing no relationship to communication and team working skills.

VI.3.e.iii. Clinical-Simulation Correlation

Significant rank correlation was demonstrated between HPS-GA and in-post ratings of clinical skills within the EBSTAF assessment process (EBSTAF-

CS). There was similar correlation between EBSTAF-CS and VASComm but this was not seen for VASCS&M.

VI.3.f. Effect of the Debriefing Process.

Completed assessments for all 52 trainees were used in the analysis of the effect of debriefing process upon self-assessment.

VI.3.f.i. Self-Assessment by Trainees (Table VI.4a)

Self-assessment scores were consistently improved following the debriefing process.

VI.3.f.ii. Trainer vs. Trainee vs. Peer Assessment (Table VI.4b)

Trainees consistently scored themselves significantly below faculty ratings in both visual-analogue scales (VASComm and VASCS&M). This finding was abolished by the debriefing, suggesting a recalibration of trainees' self-assessment skills.

Trainer and trainee global assessment ratings showed no significant difference prior to debriefing. Debriefing resulted in a similar increase in HPS-GA self-assessment scores to raise them significantly above those of trainers.

Peer assessments were consistently above those of both trainee and faculty with a significant positive effect of debriefing seen only in VASComm.

VI.3.g. Course Evaluation by Trainees

VI.3.g.i. Overall Impression of the Course (Table VI.6a)

Response rate to this section of the evaluation form was 51 out of 56 (91%). Trainees graded the course at a median of 9 as for all four questions addressing the course being well organised, interesting, enjoyable and meeting their perceived educational needs.

VI.3.g.ii. Impression of Trainee's Own Scenario (Table VI.6b)

Response rate was variable across the four questions.
When asked whether they felt better prepared to deal with a similar clinical situation as a result of their scenario, 33 trainees (59%) gave a median response of 8.
Thirty-six trainees (64%) responded to whether they had found the debriefing helpful with a median evaluation of 9.
When asked whether they had learnt something useful from their scenario, 34 trainees (61%) responded with a median score of 8.

When asked how threatening they had found their own scenario, 48 trainees (86%) gave a median response of 6, 10 being most threatening. This would suggest they found the scenario challenging but not overly threatening.

VI.3.g.iii. Impression of Observed Scenarios (Table VI.6c)

When asked whether they felt better prepared to deal with a similar clinical situation as a result of their scenario, 33 trainees (59%) gave a median response of 7.

Thirty-six trainees (64%) responded to whether they had found the debriefing helpful with a median evaluation of 8.

When asked whether they had learnt something useful from the observed scenarios, 34 trainees (61%) responded with a median score of 8.

Regarding how threatening they had found observed scenarios, 48 trainees (86%) gave a median response of 5.

VI.3.g.iv. Take Home Learning Points (Table VI.6d)

Trainees were asked if their attendance on the HPS critical care course wound likely alter their everyday clinical practice. 44 trainees (79%) responded to this question with 43 trainees (98%) stating that it would whilst citing specific areas that they felt would change. These were varied with a number of non-technical aspects of clinical practice independently highlighted

across the areas of communication, situation awareness and decision-making.

VI.3.g.v. Comments & Suggestions for Future Courses (Table VI.6e)

General comments were all extremely positive and trainees made several suggestions as to how to improve future courses.

VI.4. DISCUSSION.

HPS-GA demonstrated excellent internal consistency and acceptable inter-rater correlation for HPS-GA and VASCS&M but not for VASComm. This is perhaps surprising since only 2 assessors rated each participant over one of four possible scenarios, while poor inter-rater correlation in VASComm may illustrate differing expectations regarding communication between anaesthetic and surgical assessors. Gaba *et al* previously suggested that due to the cognitive, psychomotor, inferential and deductive skills being assessed, reliable assessment of anaesthetist performance in the simulator demanded multiple expert raters (Gaba *et al*, 1998). However, several other studies have achieved acceptable reliability using only two or three judges (Devitt *et al*, 1997; Schwid *et al*, 2002; Weller *et al*, 2003).

Boulet *et al* subsequently examined this issue using 24 medical students and 13 junior residents each completing 6 of a possible 10 scenarios rated by four independent raters using well-defined scoring rubrics. They then analysed their data using Generalisability Theory and determined that multiple raters per scenario gave only minimal reliability gains. They concluded that the content of scenarios was the major determinant of score variation (Boulet *et al*, 2003). This is reassuring since it agrees with both real-life practice and other clinical assessments (such as OSCEs). They also found that performance in a single simulated scenario does not predict that in another (Boulet *et al*, 2003). Assessment within HPS is therefore content

specific and reliable high stakes assessment of an individual will require broad sampling. Weller *et al* suggested that 10 to 15 cases, each with a single rater, would be optimal (Weller *et al*, 2005). The reliability demonstrated in this study may therefore be misleading, resulting from the use of a single scenario per trainee eliminating inter-case variation that would be seen across multiple scenarios. The assessment of individual trainees across all four scenarios was considered during the development of this study but was not felt to be possible due to the intended structure of the course and the limited time available at SCSC.

Even having considered the above limitations, the reliability demonstrated in this study is still higher than expected. This may reflect the highly structured nature of the assessment, HPS-GA, derived from EBSTAF. However, the above concerns regarding reliability must be borne in mind when interpreting subsequent data from this study.

The completion of a post in Accident & Emergency was a pre-requisite to joining the southeast Scotland BST. For the majority therefore, the first post on the programme equated to the second post-registration post, with trainees not being totally unaccustomed to the care of the critically ill. Statistically significant construct validity was demonstrated between first and second BST posts. Thereafter, no significant improvement was seen. These findings may be explained by a number of factors:

Firstly, decreased ratings accuracy, and therefore reliability, increases the range of ratings. HPS may therefore have greater validity than was demonstrable in this study due to the overlap of ratings ranges. This is supported by the demonstration of construct validity in similar studies involving anaesthetists (Byrne *et al*, 1997; Devitt *et al*, 1998; Devitt *et al*, 2001; Schwid *et al*, 2002).

Secondly, more experienced clinicians increasingly employ subjective cues (for example, the appearance of the patient "from the end of the bed") in addition to the purely numerical observations provided by HPS. The manikin was of limited fidelity in this respect, being unable to change its outward appearance. There may therefore have been a relative reduction of immersion that was unduly detrimental to the performance of more senior trainees.

Thirdly, it may simply illustrate differences in trainees' aptitudes. Not all trainees learn at the same rate, nor employ their skills equally effectively. The assumed relationship between the duration of training and competence is now recognised as flawed and it is this realisation that is the driving force behind the development of competence-based assessment. It may therefore be the construct itself that is at fault rather than the assessments on the simulator. More senior trainees may not, as a group, be better at dealing with critical care scenarios, particularly if their career path takes them away from

the critically ill and it is this assumption that underlies the estimation of construct validity.

Finally, it is interesting that construct validity is demonstrated in HPS-GA and VASCS&M but not VASComm. This may be explained in two ways: trainees may improve their communication skills more slowly than they do their clinical skills or, more likely, this illustrates an understandable limitation of HPS. The assessment of communication skills should therefore use other validated methods, such as OSCE or simulated patients.

Further evaluation of the validity of HPS required comparison between HPS assessments and a suitable 'gold standard' to determine concurrent validity. EBSTAF has been previously validated for the multidisciplinary assessment of in-post performance (Paisley *et al*, 2001a), albeit based upon the same construct of improved BST performance over I year of surgical training. It was therefore reasonable to regarded EBSTAF as such a gold standard.

The demonstration of significant relationships between ratings of trainee performance in HPS and the Clinical Skills domain of EBSTAF is highly favourable when considering HPS and its application to the assessment of surgical trainees. It implies that desirable skills and behaviours displayed in everyday clinical practice , assessable by EBSTAF, are also observable within the simulator. However, the strengths of the relationships are limited

and may result from the course structure as detailed above. Clinical assessment by EBSTAF addresses in-post performance over a six-month period to include multiple individual cases while the simulator course exposed each trainee to only a single scenario. By exposing trainees to multiple HPS scenarios, assessment reliability is improved (Boulet *et al*, 2003; Weller *et al*, 2005) and this would likely strengthen the relationship between the two assessment modalities. However, this cannot be assumed since it is possible that the relationship may be *as a result of* HPS rating error and thus be abolished by increasing its reliability. When considering this, it is reassuring that there were no clearly spurious relationships demonstrated between unrelated assessment domains such as EBSTAF Technical Skills (which contains fields that relate only to operative skill) and HPS-GA. This would therefore support the observed relationship but further work is needed.

There were also concerns that the relationship between EBSTAF Clinical Skills and HPS-GA might result from the fact that HPS-GA was itself derived from EBSTAF. In an effort to examine this causality, *Pseudosim* was created to include the same fields as HPS-GA but drawn from in-post clinical assessments using EBSTAF. Comparison showed no relationship, suggesting the concurrent validity to be unrelated to HPS-GA's origins.

There is no guarantee that performance in the simulator will translate to real-life practice (Boulet *et al*, 2003). The relationship of simulator performance to other assessment methods has been examined in anaesthetic trainees.

Schwid *et al* demonstrated moderate correlations between simulator scores and written American Board of Anaesthesiologists in-training examinations, departmental faculty evaluation and mock oral scores (Schwid *et al*, 2002). However, EBSTAF offers the only fully validated in-post multidisciplinary assessment of clinical performance in surgical trainees. Correlation between EBSTAF and HPS-GA therefore lends considerable support to the use of HPS to the training of clinical and non-technical skills. However, the use of assessments obtained in HPS in high stakes assessment should not be considered at present; further work is required to optimise the validity of HPS.

If HPS is to be used summatively, assessment of trainee performance will likely be limited to the third level of Miller's competency pyramid, that of demonstration (Miller, 1990). However, if sufficient immersion is achieved, the simulator may be unique in offering the opportunity to observe and modify actual practice within a totally patient-safe environment. Trainees may feel able to go further in their patient management (and therefore their learning) in the absence of assessment than if such actions are likely to jeopardise their high-stakes score. This huge potential for excellence in patient care should not be abandoned in the search for robust documentation of simple competence.

Comparisons of trainer and trainee ratings proved interesting, suggesting that the use of a highly structured assessment form (HPS-GA) may in itself improve trainees' self-assessment skills. Gordon states that "the ability to

recognise one's strengths and weaknesses is critical to the enterprise of lifelong learning" (Gordon, 1991). Continuing professional development demands accurate self-assessment and recognition of remediable weaknesses (Evans *et al*, 2002) but trainees are poor self-assessors (Morton *et al*, 1977; Anthoney, 1986; Risucci *et al*, 1989; Gordon, 1991; Das *et al*, 1998; Johnson *et al*, 1998; Ward *et al*, 2002). Accuracy is related to proficiency (Gordon, 1991; Ward *et al*, 2003; MacDonald *et al*, 2003) and the least able trainees are therefore likely to be the most inaccurate self-assessors of technical skills. Interestingly, recent work by Moorthy *et al* might suggest junior trainees to be better self-assessors of non-technical skills than their senior colleagues (Moorthy *et al*, 2006).

In this study, trainees scored themselves significantly less favourably than faculty when using the two visual analogue scoring systems, VASComm and VASCS&M. This agrees with previous work in general (Arnold *et al*, 1985; Woolliscroft *et al*, 1993) and on EBSTAF itself (Paisley 2002) although other studies suggest a tendency toward overestimation of ability by trainees (Evans *et al*, 2002; Weller *et al*, 2005). However, self-assessment is improved by the provision of specific criteria (Gordon, 1991) and when trainees assessed themselves using HPS-GA no such discrepancy was seen. While a visual analogue score estimates the overall impression of a performance, it may be overly influenced by a single area of poor performance within an otherwise proficiently completed scenario. By

dissecting a performance into its constituent skills, HPS-GA improves the accuracy of trainees' ratings and brings them closer to those of faculty.

Debriefing consistently increased trainees' self-evaluations, abolishing differences between trainer and trainee ratings for VASComm and VASCS&M. This may have resulted from putting areas of poor performance back into perspective. However, debriefing also resulted in trainees over-estimating performance using HPS-GA, suggesting the potential of nurturing over-confidence should the balance of the critique be overly positive.

Participant evaluations were highly favourable with trainees rating the course as a whole, their own scenario and observed scenarios extremely highly. Observed scenarios were scored slightly lower, highlighting the perceived value of participation rather than simple observation of scenarios on video, achievable away from the simulator environment.

Trainees' freely made comments proved to be most interesting, raising a number of non-technical issues in addition to those more specific to individual scenarios. Comments repeatedly addressed issues of information gathering, situation awareness, teamworking and communication. These issues were only briefly touched upon during debriefing yet comments illustrate how important the trainees realised them to be as a result of the course which, by the use of composite video, allowed perfect review of their performance. The role of non-technical skills in safety in health care is increasingly recognised

in the literature and a number of courses have been developed to specifically address this area and its relationship to clinical practice. However, for non-technical skills to be assessed accurately, be it formatively or summatively, robust assessment methods must first be developed. Non-technical skills assessment tools developed in other environments, such as commercial aviation (Helmreich *et al*, 1999), have been successfully applied to surgeons (Moorthy *et al*, 2005; Moorthy *et al*, 2006). However, the development of specific skills taxonomies for anaesthetists (Fletcher *et al*, 2003) and surgeons (Yule *et al*, 2006b), although similar do show some significant differences from each other and from those developed in other high-risk industries. They are therefore context-specific and must be used accordingly.

The HPS perioperative scenario course for surgical trainees ran well and proved far more popular than expected. The unanticipated demand for places strongly suggests a perceived need on the part of the trainees for a course to address issues arising from the care of critically ill patients. Furthermore, statements from trainees strongly support the development of future HPS-based courses to address both technical and non-technical aspects of critical care.

VI.5. SUMMARY.

- The use of HPS within a one-day scenario-based critical care course was both feasible and popular with participating trainees, who stated that they found it highly applicable to their everyday practice.

- Assessment of BSTs' performances within HPS-based critical care scenarios demonstrated acceptable reliability, construct validity and a degree of concurrent validity, but this area requires further work.

- Freely made comments by trainees recognised the influence of non-technical skills upon clinical performance.

- The unanticipated demand for places strongly suggests a perceived need on the part of the trainees for such a course addressing technical and non-technical aspects of critical care.

- The incorporation of HPS courses into surgical training should be considered in order to address shortfalls in critical care skills.

Table VI.1

Study Participants' Range of Experience

Time since full registration	n	
≤ 6 months	**5**	Min = 1 month
6 months to 1 year	**8**	
12 to 18 months	**9**	
18 months to 2 years	**10**	
2 to 2½ years	**12**	
2½ to 3 years	**2**	
3 to 3½ years	**6**	Max = 40 months

Table VI.2

Internal Consistency

a. Simulator Assessment	Fields	α
HPS-GA	32	**0.95**

b. Clinical Assessment	Fields	α
EBSTAF I : Communication	9	**0.80**
EBSTAF II : Knowledge	9	**0.83**
EBSTAF III : Teamwork	11	**0.87**
EBSTAF IV : Clinical Skills	22	**0.90**
EBSTAF V : Technical Skills	19	**0.89**
EBSTAF I-V : Overall	72	**0.96**
EBSTAF : PseudoSim	29	**0.92**

HPS-GA : Global assessment in high-fidelity simulator.

EBSTAF : Edinburgh Basic Surgical Trainee Assessment Form.

Pseudosim : virtual assessment series drawn from EBSTAF to mirror fields in high-fidelity simulator global assessment.

α denotes Cronbach's Alpha coefficient: >0.80 is acceptable for high-stakes assessment

Table VI 3a

Construct validity of simulator assessment as measured by global assessment (HPS-GA)

Post	n=	Median	IQR	Mann-Whitney p=
2	4	46.02	42.52 – 68.52	
				0.042
3	7	76.24	60.35 – 91.08	
				0.054
4	8	56.58	45.88 – 68.31	
				ns
5	9	72.72	52.06 – 86.88	
				0.073
6	2	95.58	93.90 – 97.26	
				ns
7	5	83.67	58.50 – 93.95	
	35			

Post : number of six-month clinical posts since full registration.
IQR : inter-quartile range

Kruskal Wallis test p = 0.028

Table VI.3b

Construct validity of simulator assessment of clinical skills & management (VAS CS&M)

Post	n=	Median	IQR	Mann-Whitney p=
2	4	38.86	31.93 – 60.19	
				ns
3	7	56.52	50.00 – 66.30	
				ns
4	8	47.69	38.68 – 55.44	
				ns
5	9	57.07	46.20 – 74.46	
				ns
6	2	81.88	76.81 – 86.96	
				ns
7	5	73.37	50.00 – 77.72	
	35			

Post : number of six-month clinical posts since full registration.
IQR : inter-quartile range

Kruskal Wallis test p = 0.044

Table VI.3c

Construct validity of simulator assessment of communication skills (VAS Comm)

Post	n=	Median	IQR	Mann-Whitney p=
2	4	52.45	46.06 – 75.41	
				ns
3	7	71.74	62.50 – 78.26	
				ns
4	8	58.42	47.46 – 73.78	
				ns
5	9	60.87	45.92 – 77.27	
				ns
6	2	81.79	76.09 – 87.50	
				ns
7	5	73.91	53.80 – 79.89	
	35			

Post : number of six-month clinical posts since full registration.
IQR : inter-quartile range

Kruskal Wallis test p = ns

Table VI.4a.

Determination of Concurrent Validity by Correlation.

Spearman (*rho*) Coefficient (bold) and p value.		Comm[n]	Knowledge	Teamwork	Clinical Skills	Technical Skills	VAS	Overall	PseudoSim	VAS Comm	VAS CS&M	Global
EBSTAF	Knowledge	**0.642** <0.001										
	Teamwork	**0.645** <0.001	**0.693** <0.001									
	Clinical Skills	**0.634** <0.001	**0.70** <0.001	**0.579** <0.001								
	Technical Skills		**0.40** 0.016		**0.597** <0.001							
	VAS	**0.506** 0.001	**0.482** 0.003	**0.643** <0.001	**0.559** <0.001	**0.484** 0.002						
	Overall	**0.455** 0.009	**0.713** <0.001	**0.573** 0.001	**0.772** <0.001	**0.711** <0.001	**0.441** 0.012					
	PseudoSim	**0.737** <0.001	**0.786** <0.001	**0.663** <0.001	**0.896** <0.001	**0.518** 0.001	**0.604** <0.001	**0.776** <0.001				
HPS	VAS Comm				**0.344** 0.043							
	VAS CS&M									**0.786** <0.001		
	HPS-GA				**0.395** 0.019					**0.811** <0.001	**0.945** <0.001	

Comm[n] : Communication. **EBSTAF** : Edinburgh Basic Surgical Trainee Assessment Form. **VAS** : visual analogue scale estimating overall impression of trainee. **Pseudosim** : virtual assessment series drawn from EBSTAF to mirror fields in high-fidelity simulator global assessment. **HPS** : high-fidelity patient simulator. **VASComm** : visual analogue scale assessment of communication. **VASCS&M** : visual analogue scale assessment of clinical skills and management. **HPS-GA** : Global assessment in high-fidelity simulator.

Table VI.4b:

Determination of Concurrent Validity by Concordance.

Kendall's *tau-b* (bold) and p value		Comm[n]	Knowledge	Teamwork	Clinical Skills	Technical Skills	VAS	Overall	oSim	VAS Comm	VAS CS&M	HPS-GA
EBSTAF	Knowledge	**0.490** <0.001										
	Teamwork	**0.516** <0.001	**0.584** <0.001									
	Clinical Skills	**0.524** <0.001	**0.563** <0.001	**0.543** <0.001								
	Technical Skills		**0.323** 0.009		**0.445** <0.001							
	VAS	**0.374** 0.002	**0.374** 0.002	**0.504** <0.001	**0.413** <0.001	**0.353** 0.002						
	Overall	**0.342** 0.008	**0.583** <0.001	**0.438** 0.001	**0.583** <0.001	**0.548** <0.001	**0.341** 0.006					
	PseudoSim	**0.616** <0.001	**0.651** <0.001	**0.527** <0.001	**0.760** <0.001	**0.364** 0.002	**0.451** <0.001	**0.609** <0.001				
HPS	VAS Comm				**0.250** 0.039							
	VAS CS&M									**0.620** <0.001		
	HPS-GA				**0.288** 0.017					**0.618** <0.001	**0.806** <0.001	

Comm[n] : Communication. **EBSTAF** : Edinburgh Basic Surgical Trainee Assessment Form. **VAS** : visual analogue scale estimating overall impression of trainee. **Pseudosim** : virtual assessment series drawn from EBSTAF to mirror fields in high-fidelity simulator global assessment. **HPS** : high-fidelity patient simulator. **VASComm** : visual analogue scale assessment of communication. **VASCS&M** : visual analogue scale assessment of clinical skills and management. **HPS-GA** : Global assessment in high-fidelity simulator.

Table VI.5

Effect of Debriefing.

a.Trainee Self-Assessments

	Pre-Debrief	Post-Debrief	p= *
HPS-GA	68.3	73.8	0.006
VASComm	57.2	59.9	0.030
VASCS&M	50.8	55.4	0.015

b.Faculty vs. Trainee vs. Peer Assessments

	Faculty	Trainee		Peer	
		Score	p= *	Score	p= *
Pre-Debrief					
HPS-GA	66.0	68.3	ns	83.5	<0.001
VASComm	61.5	57.2	0.018	66.9^{+}	0.013
VASCS&M	55.7	50.8	0.008	63.6	0.001
Post-Debrief					
HPS-GA	66.0	73.8	0.005	85.9	<0.001
VASComm	61.5	59.9	ns	69.9^{+}	<0.001
VASCS&M	55.7	55.4	ns	67.0	<0.001

* : Wilcoxon, 2-tailed: + Significant increase seen with debriefing (p= 0.021)

Table VI.6a

Overall evaluation of the high-fidelity patient simulator surgical critical care course.

	Response Rate	I Q R	Median
The course was well organised	51 of 56 (91%)	8 - 10	9
The course was interesting	51 of 56 (91%)	8 – 10	9
The course was enjoyable	51 of 56 (91%)	8 – 10	9
The course met my educational needs	51 of 56 (91%)	8 - 10	9

Responses were graded from 0 (strongly disagree) through 5 (ambivalent) to 10 (strongly agree).

IQR : inter-quartile range

Table VI.6b

Trainee evaluation of their own scenario.

	Response Rate	I Q R	Median
I feel better able to deal with a similar clinical situation	33 of 56 (59%)	7 to 10	8
The debriefing was helpful	36 of 56 (64%)	8 to 10	9
I learned something useful	34 of 56 (61%)	8 – 10	8
I found the scenario threatening	48 of 56 (86%)	5 – 8	6

Responses were graded from 0 (strongly disagree) through 5 (ambivalent) to 10 (strongly agree).

IQR : inter-quartile range

Table VI.6c

Trainee evaluation of observed scenarios.

	Response Rate	I Q R	Median
I feel better able to deal with a similar clinical situation	33 of 56 (59%)	6 to 8	7
The debriefing was helpful	36 of 56 (64%)	8 to 9	8
I learned something useful	34 of 56 (61%)	7 to 9	8
I found the scenario threatening	48 of 56 (86%)	2 to 7	5

Responses were graded from 0 (strongly disagree) through 5 (ambivalent) to 10 (strongly agree).

IQR : inter-quartile range

Table VI.6d

Take-home points resulting from trainees' attendance on the high-fidelity patient simulator critical care course.

Comments were made freely on the evaluation sheet within a box entitled
"What do you think will change in you clinical practice as a result of attending this course"
Duplicates have purposely not been removed

Earlier, better communication.
Re-checking.
More aggressive fluid resuscitation.
Remember to begin with Airway, Breathing & Circulation.
More methodical secondary survey.
Ask for help earlier.
Check machines and rely more on clinical findings.
Think about treating more potential differentials.
Will be more circumspect in diagnosis formation.
Take control and delegate.
Oxygen for the unwell post-op patient.
Gynaecological causes of abdominal pain.
Systematic approach to trauma patient.
Stay calm, think things through.
Methodical approach.
Review findings regularly.
Confidence to ask for advanced investigations, e.g. CT.
Confidence to take charge.
Inform senior earlier.
Structuring differential diagnosis.
Review of ALS procedures.
Increased confidence in own ability.
Has made me more aware of things normally done for me.
Reinforced certain ideas and knowledge.
BP Monitoring.
New arrest protocols.
Will attend ALS course!
Re-emphasise importance of keeping an open mind.
Improved cardiac arrest / ALS.
Better appreciation of sepsis.
Great revision of trauma.
Communication
Epidural considerations
Attempt to be more systematic

Table VI.6d contd.

Reminder of arrest protocols
More logical approach to an ill patient
Will have better understanding of limitations of monitoring equip.
Attempt to be more thorough
Will be confident to stop epidurals
Will have a more structured approach to assessing gynaecological problems.
Clearer delegation of tasks
Better communication
Appropriateness of seeking help EARLY / informing consultants of sick patients.
Importance mental checklist for tasks / observations previously taken as given.
Frequent BP checking, counting respiration rate, checking BM, catheterisation, monitoring urine output
Not to be distracted by tasks being done by others.
Keep up to date on resuscitation protocols
Principles of management, especially HDU
Relationship to seniors, how / who to ask for help.
Appreciation of info sources
Importance of accurate communication.
Think out loud so all involved know thought process.
Sharpen up management of acute problems / emergencies.
Delegation/assessment of abilities / organisation of team.
Understanding patient's perspective
Take a more measured, slow approach to patient
Take a step back to look at the whole picture
Awareness of physiological development
Helped organisation of thoughts
More systematic & methodical approach to problems
Help identify critically ill patients
Ask for help early
Learnt importance of being more organised and being able to justify my actions
Think systematically
Consider calling for help earlier
More thorough history
Better able to deal with trauma patients
Be more cautious with opiates in sick people.
Learnt who to call when things go pear shaped.

Table VI.6e

Comments freely made by trainees at conclusion of the high-fidelity patient simulator critical care course.

Comments were made freely on the evaluation sheet within a box entitled

"Any further comments as to how the course might be improved"

GENERAL COMMENTS

Excellent course, hope to do it again.
Would be useful to all on BST at any stage.
Very useful. Video playback especially useful, allowing self-criticism.
Allowed more objective view of self and own practice.
Testing and arduous experience, especially my own scenario, but very educational and great environment to learn.
Would highly recommend to other SHO's.
Excellent "dummy" run. Promotes anxiety. Very realistic.
Debriefing sessions are excellent - lots of teaching & constructive comments.
Very good variety of scenarios
I have no doubt that at least one life has been saved as a result of my attending this course

SUGGESTIONS ON HOW TO IMPROVE FUTURE COURSES

More scenarios per person.
Scenario 2 does not lend itself well to the simulator – observations are bad but clinical examination is not helpful.
Debriefing should be precise & straight to the point.
At end of each scenario update what was actually wrong and discussion of the ideal management of that situation.

SUGGESTIONS FOR FUTURE SCENARIOS

GI Bleed
Acute Airway
Anaphylaxis

FIGURE VI.1
Schematic Breakdown of Trainee Data Analysis.

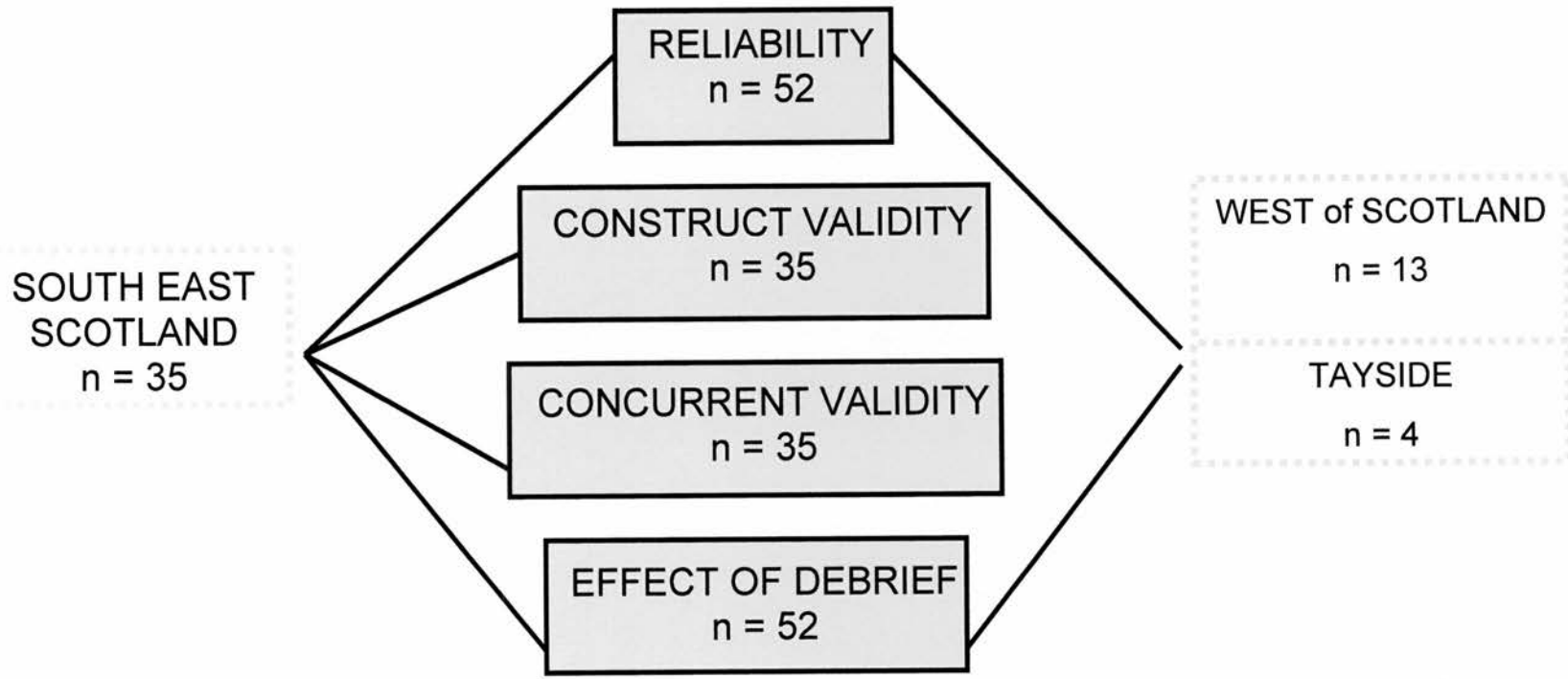
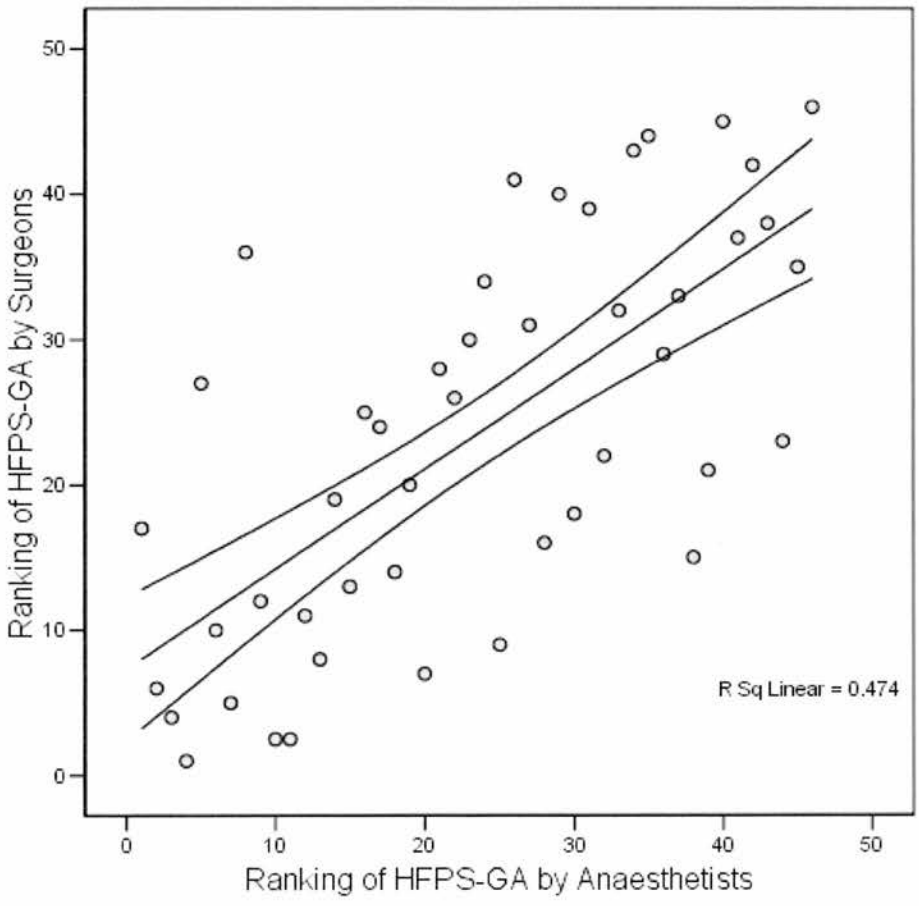


Figure VI.2a

Plot to illustrate agreement between surgeon and anaesthetist rankings of simulator performance as measured by HPS-GA.

Spearman rank correlation *rho* r_s = 0.689, p < 0.001
Kendall's concordance, *tau b* = 0.514, p < 0.001

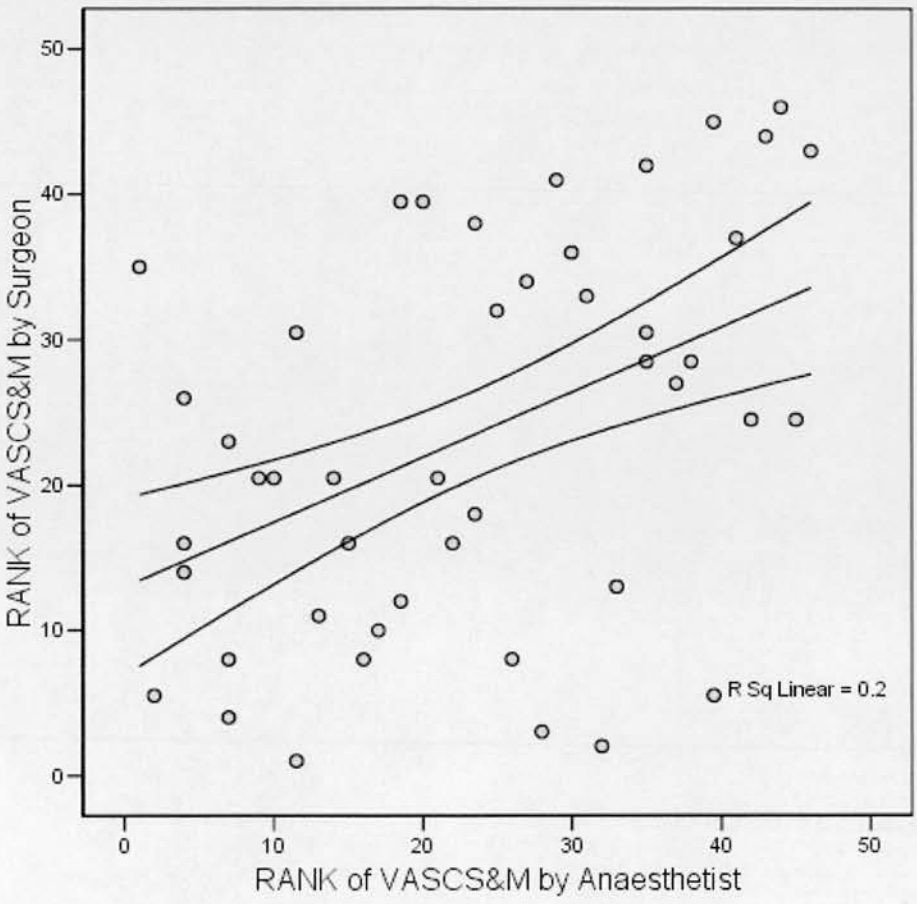


Rank number 1 (i.e. bottom left) = highest score.
Lines denote best fit with 95% confidence limits.

Figure VI.2b

Plot to illustrate agreement between surgeon and anaesthetist rankings of simulator performance as measured by VASCS&M.

Spearman rank correlation *rho* r_s = 0.447, p = 0.002
Kendall's concordance, *tau b* = 0.295, p = 0.004

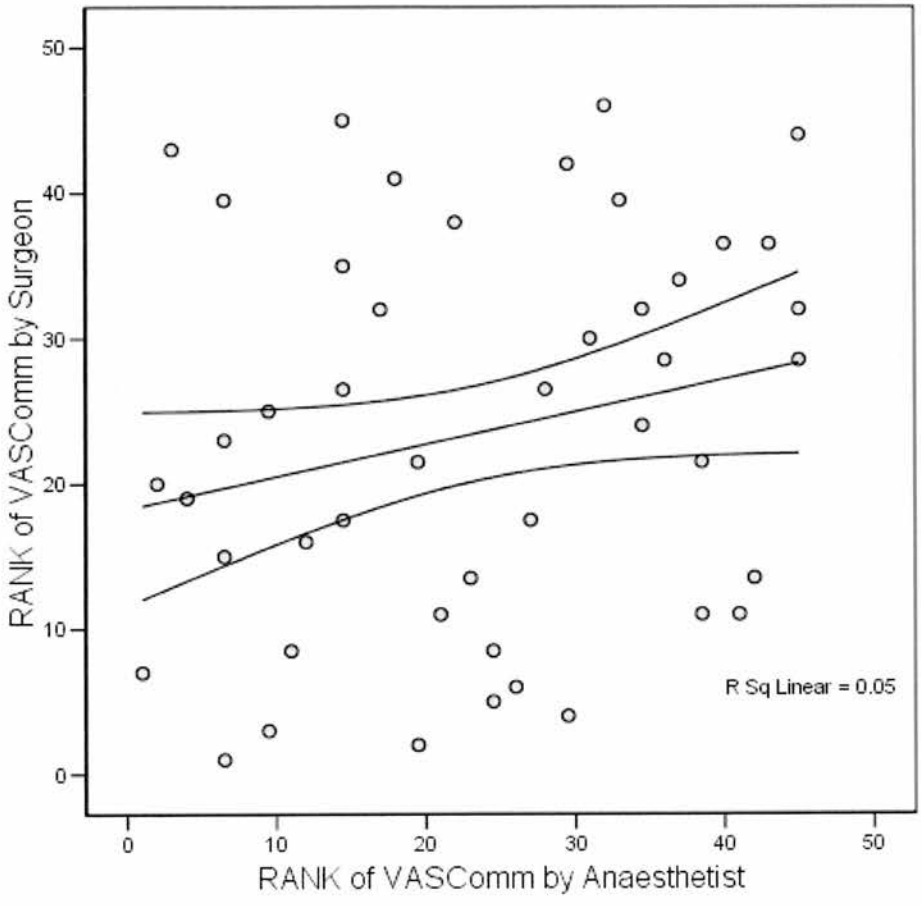


Rank number 1 (i.e. bottom left) = highest score
Lines denote best fit with 95% confidence limits.

Figure VI.2c

Plot to illustrate agreement between surgeon and anaesthetist rankings of simulator performance as measured by VASComm.

Spearman rank correlation *rho* r_s = 0.223, p = ns
Kendall's concordance, *tau b* = 0.155, p = ns



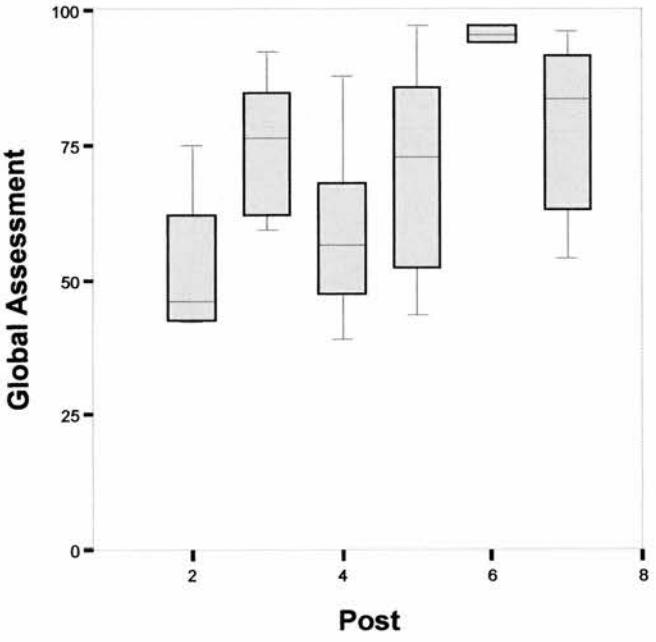
Rank number 1 (i.e. bottom left) = highest score
Lines denote best fit with 95% confidence limits.

Figure VI.3a

Construct validity of simulator assessment measured by HPS-GA.

Boxes represent median, inter-quartile range and range.

Kruskal Wallis, p = 0.028
Mann-Whitney (Post 2 vs. Post 3), p = 0.042



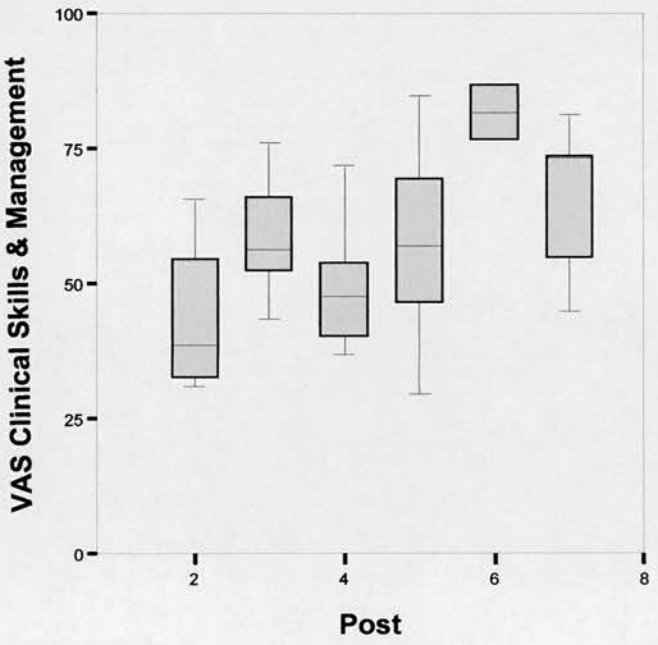
Post: 6-months clinical posts completed since time of full registration

Figure VI.3b

Construct validity of simulator assessment measured by VAS-CS&M

Boxes represent median, inter-quartile range and range.

Kruskal Wallis test p = 0.044



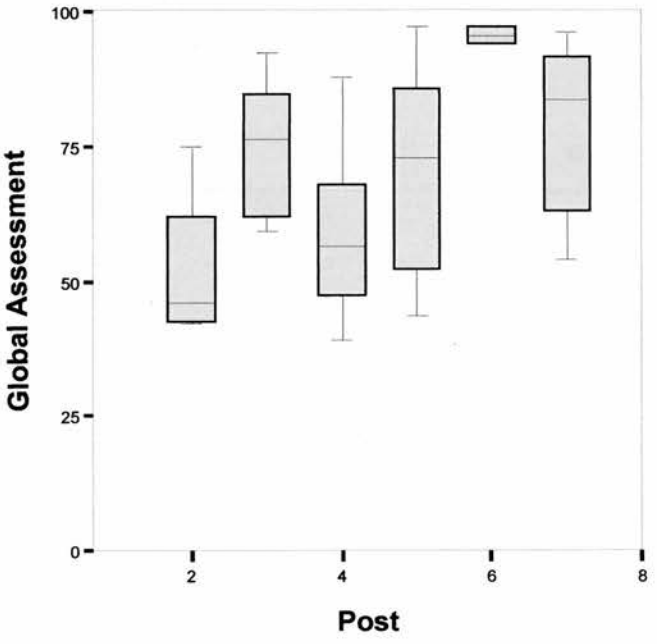
Post: 6-months clinical posts completed since time of full registration

Figure VI.3c

Construct validity of simulator assessment measured by VAS-Comm.

Boxes represent median, inter-quartile range and range.

Kruskal Wallis test p = ns



Post: 6-months clinical posts completed since time of full registration

Section VII.

VIDEO ASSESSMENT OF BASIC SURGICAL TRAINEES' OPERATIVE SKILLS.

VII.1 INTRODUCTION

The frequent incorporation of cameras into operating lights offers the opportunity to record procedures for later review. Using these videos for trainer-trainee feedback of performance or self-review may then promote reflective practice. However, the utility of such methods may be limited by learners' poor self-assessment abilities which result in low correlation with expert ratings (Morton *et al*, 1977; Risucci *et al*, 1989; Gordon, 1991; Das *et al*, 1998; Johnson *et al*, 1998; Ward *et al*, 2002; Ward *et al*, 2003). Improved trainee self-assessment skill may be achieved by benchmarking performance to those of others (Martin *et al*, 1998) but simple self-review of video recorded performance has also been shown to improve a trainee's ability to self-evaluate (Ward *et al*, 2003). Video recorded procedures may also offer the chance to objectively assess operative skill, but before this can be considered the validity and reliability of such methods must be fully understood.

VII.2. AIMS.

- To determine the reliability and validity of assessing BST tissue-handling skills during real-time procedures assessed by a single assessor using EBSTAF.

- To investigate the feasibility, reliability and construct validity of assessing BST tissue-handling skills using edited and anonymised video-recorded procedures assessed by panels of consultants and trainees.

- To examine the concurrent validity of real-time and video assessments by comparison with in-post assessments using the technical skills domain of EBSTAF.

To examine the relationship between trainer and trainee assessment scores in an effort to determine whether the use of EBSTAF has an effect on trainee self-assessment skills.

VII.3. RESULTS.

VII.3.a. Study Participants.

Eleven BSTs were recruited to perform a video-recorded hernia repair after their first six-month post in general surgery. 9 of the same BSTs completed a second hernia repair after their second six-month post, also in general surgery. Shortfall was due to circumstances outside the researchers' control; cancellation of the case and clashes with other clinical commitments. Trainees not recorded on both occasions were excluded leaving 9 trainees (and therefore 18 procedures) for further analysis in combination with two consultant-performed procedures and duplicates at 6 months, 12 months and consultant levels. The result was a total of 23 individual assessment videos in random order for further study.

VII.3.b. Real-Time Assessment (RTA) of Surgical Performance.

VII.3.b.i. Reliability (Table VII.1.)

Trainer ratings of performance demonstrated excellent internal consistency as estimated by Cronbach's Alpha ($\alpha = 0.95$). The internal consistency of trainee assessments was a little reduced but still achieved an acceptable level ($\alpha = 0.85$). Both trainer and trainee assessments therefore

demonstrated acceptable reliability as determined by the estimation of internal consistency.

VII.3.b.ii. Construct Validity (Table VII.2a)

For an assessment of surgical performance to be useful, it must demonstrate improvement as a result of training. Neither trainers' assessments (figure VII.2a) nor trainees' self-assessments (figure VII.2b) demonstrated significant improvement between procedures assessed by RTA at 6 and 12 months. Construct validity was therefore not demonstrated during RTA.

VII.3.b.iii. Concurrent Validity (Tables VII.3a and VII.3c)

Trainer assessments demonstrated a strong relationship between EBSTAF-Tech and RTA-VAS scores (figure VII.3a) but this was not observed in trainees' assessments (figure VII.3b).

No correlation was seen between in-post EBSTAF-TS and either EBSTAF-Tech or RTA-VAS by trainers (figure VII.3e). However, trainee assessments demonstrated a strong relationship between the EBSTAF-TS and EBSTAF-Tech. The correlation between trainee EBSTAF-TS and RTA-VAS was weak and fell just short of significance (figure VII.3f).

VII.3.b.iv. Trainer-Trainee Correlation (Table VII.4a)

No correlation was demonstrated between trainer and trainee assessments (figure VII.4a).

VII.3.b.v. Targeted Suturing Task (Table VII.5)

Targeted suture placement failed to demonstrate a significant improvement between six and twelve month time-points. Neither was there correlation with real-time assessments by either assessor group using EBSTAF-Tech or RTA-VAS, nor with in-post trainer-assessments using EBSTAF-TS.

VII.3.c. Video-Assessment of Surgical Performance.

VII.3.c.i. Feasibility.

Actual procedures took a median total time of 46 minutes (range 26-83), significantly reduced to 17 minutes in the edited videos (range 11-26, $p < 0.001$).

Completed scorebooks were returned by 7 of 9 consultants (78%) and 5 of 7 trainees (71%).

Median time to assess each video was 10 minutes (range 1-25), a significant time saving over the assessment of the actual procedure ($p < 0.001$).

The return rate combined with increased time-efficiency supports the feasibility of the use of edited and anonymised video in the assessment of BST tissue handling skills.

VII.3.c.ii. Reliability (Table VII.1)

Individual assessors showed high levels of consistency in their ratings of individual operators as determined by the estimation of internal consistency by Cronbach's Alpha (EBSTAF-Tech; trainers $\alpha = 0.89$, trainees $\alpha = 0.76$: Toronto; trainers $\alpha = 0.78$, trainees $\alpha = 0.95$).

Individuals were also rated consistently using EBSTAF-Tech, Toronto and V-VAS across the two assessor panels, demonstrating good inter-rater reliability as determined by intra-class correlation (trainers 0.86 to 0.93, trainees 0.69 to 0.84).

Strong and significant correlation and concordance between test and retest scores was obtained for all three scoring systems and supported test-retest reliability for both trainers (figure VII.1a) and trainees (figure VII.1b).

VII.3.c.iii. Construct Validity.

VII.3.c.iii. (a) Assessment by Consultant Panel (Table VII.2b, Figure VII.2c)

Significant score difference was seen between trainees and consultants for all three assessment methods (Mann-Whitney, $p \leq 0.001$). Consultant

assessors were able to distinguish between trainees at 6 and 12 months (Wilcoxon, $p = 0.023$). Consultant panel assessment of basic operative techniques therefore appears to be both construct valid and highly sensitive.

VII.3.c.iii. (b) Assessment by Trainee Panel (Table VII.2c, Figure VII.2d)

Significant score difference was seen between trainees and consultant for all three assessment methods (Mann-Whitney, $p \leq 0.019$). Trainee panel assessment of basic operative techniques may therefore be considered to be construct valid. However, the trainee panel was not as sensitive in its assessments, being unable to distinguish between the two closely related trainee levels.

VII.3.c.iv Concurrent Validity

Excellent correlation and concordance was observed between EBSTAF-Tech, Toronto and V-VAS for both trainer (figure VII.3c) and trainee (figure VII.3d) assessors (Table VII.3.b).

No relationship was demonstrated between video assessment ratings and those obtained in-post using EBSTAF-TS (Table VII.3.d) for either trainers (figures VII.3g) or trainees (figure VII.3.h).

VII.3.c.v. Trainer-Trainee Correlation (Table VII.4b, Figure VII.4b)

Video assessment showed good correlation and concordance between trainer and trainee ratings across all three scoring systems.

VII.3.c.vi. Estimation of Training Stage (Table VII.6)

Only those estimations applying to the 12-month trainees were analysed since those for 6 months and consultant levels would be skewed by the lack of options both above and below the true level. Trainers correctly estimated training level in only 40% of cases, tending to underestimate the remainder. Trainees' estimations were equally distributed and correct in 43% of cases.

VII.4. DISCUSSION.

If assessment procedures are to be used as measures of surgical ability then scores must correlate with surgical experience, improve with training or practice and discriminate between the surgically naïve and experienced operators (construct validity).

The lack of construct validity for trainer RTA ratings is disappointing but perhaps not surprising. It may be due to a number of factors:

First and foremost, EBSTAF was designed (Baldwin *et al*, 1999) and validated (Paisley *et al*, 2001a) for the multisource assessment of observed performance in practice over a period of six months by multi-disciplinary assessors. EBSTAF was never intended, nor validated, for the assessment of a single procedure (or part thereof) by a single assessor but rather the trainee's entire operative experience over that time (termed 'long-loop' feedback).

Secondly, individual real-life cases are highly variable and cannot be duplicated. Operative assessment is therefore *case specific* with the potential for the same trainee to score very differently between cases. As a result, a junior trainee in an easy case may have achieved a similar or even higher score to a superior trainee in a more difficult case where the operator clearly

struggled, thus confounding the assessments. Multiple cases for each trainee are therefore mandatory to minimise case specificity.

Thirdly, the assessment of BST tissue-handling skills using a single procedure scored by a single trainer is also *assessor specific*, influenced by the assessor's personal surgical preferences. Modern trainees are trained by multiple trainers and therefore inevitably develop techniques different to those preferred by individual assessors.

Fourthly, the RTA assessor will have been aware of the training level of the trainee. Although this might be expected to favour reliability and construct validity, it may actually confound things further. An assessor may expect more of a more senior trainee and therefore mark more harshly should the trainee fall short of expectations. Conversely, less may be expected of the junior trainee and the assessor may be more lenient.

Fifth, the inter-personal relationship between assessor and trainee will also have inevitably had an effect on RTA ratings. However, this is impossible to quantify.

Sixth, trainees were assessed during RTA by one of their own trainers. A low score may have been felt by some trainers to adversely reflect on his training and therefore on the assessor himself. This *response bias* may have resulted

in scores that unknowingly approximated to what was expected (or hoped for) rather than honest opinion.

Seventh, trainees may have failed to demonstrate their true abilities due to factors unrelated to the assessment itself (halo effect) or nerves related to what equates to an operative examination in front of an examiner.

Eighth, assessments were made at 6 and 12 months by different trainers. Although this is analogous to the original validation of EBSTAF, the lack of construct validity and correlation with in-post assessments serves to further emphasise the 'snap-shot' nature of assessment procedures when compared to broad-based assessment over a longer time period across multiple cases and multiple procedures.

Finally, trainers could not be blinded to the herniorrhaphy and mesh placement parts of the procedure. There was therefore a considerable procedural component to the assessment of generic tissue-handling skills that would have undoubtedly influenced RTA assessments. This may have been avoided by asking the assessor to leave the procedure at the point of division of external oblique, only to return at the equivalent stage of closure. However, this would have demanded two trainers for each procedure adding a further level of coordination to the procedure.

Also worthy of mention is the relationship between trainees' RTA EBSTAF-Tech and in-post EBSTAF-TS ratings that showed strong and significant correlation and concordance. This suggests that trainee ratings were more influenced by their overall opinion of their technical skills than their performance during the assessment procedure itself; another instance of *halo effect*. Alternatively, trainees may have provided what they felt to be the expected responses rather than honest opinion as a result of *response bias* effecting either one or both assessment methods.

The targeted suturing task failed to demonstrate construct validity. It is unclear why this should be the case, having been shown to be construct valid in previous studies (Seki, 1989; Paisley *et al*, 2001b). It may reflect the lack of reality of such a task, highlighted when carried out immediately after a real life procedure. Alternatively, the trainee may have been preoccupied by their performance in the hernia repair and therefore unable to give the task their full attention. Different results may have been obtained had the task been duplicated at a different time and location, analogous to previous studies.

The use of multiple assessors to reduce the subjectivity of real-time assessments has previously been shown to demonstrate construct validity (Reznick *et al*, 1997; Scott *et al*, 1999). However, the presence of multiple assessors in an operating theatre not only contravenes theatre protocol but also inevitably affects trainee performance, particularly at a junior level. Assessors might alternatively be placed in a nearby video-linked room, but

the logistics of coordinating such an assessor panel is unrealistic in everyday clinical practice. The circulation of video recordings to individuals allows evaluation at the assessor's convenience whilst still maintaining the objectivity of scoring by multiple assessors.

Video assessment (using EBSTAF-Tech, Toronto and V-VAS) of BST tissue-handling skills by consultant assessors using circulated video recordings proved feasible and efficient in terms of the assessors' time, taking approximately one fifth of the time of the whole procedure. Although this finding does not take account of the recording and editing time, the latter may not be required if assessment was to be by consultants unknown to the trainees. An assessor might therefore be able to assess five times as many procedures using video as would be possible in real-time. Although this study examines the application of video assessment to multiple trainees, this technique may be equally applicable to multiple procedures by the same trainee in the form of an operative portfolio.

Overall, video assessment was seen to be reliable, although the examination of the raw data, as shown in Table VII.2.b, clearly shows the consultant group displaying both harsh *hawkish* (assessors 2 and 7) and lenient *dove-like* (assessor 5) assessment behaviours. By the identification of such individuals, the reliability may be further enhanced by their re-training in the assessment process or by their replacement. However, this does illustrate how such a panel of assessors reduces the influence of such extremes.

Video assessment also demonstrated construct validity. This was most likely achieved as a result of the editing and anonymisation of the video-recorded procedures. Ignorance of the operator's identity and training level removed the potential halo effect and response bias while the use of multiple assessors minimised assessor specificity. Furthermore, by blinding assessors to the herniorraphy and mesh placement part of the procedure, assessments became less procedure-specific facilitating the assessment of simple tissue-handling skills in isolation.

Video assessment by the trainer panel was highly discriminatory, separating closely related levels of junior trainee. Previous studies addressing the construct validity of technical skills assessment have addressed widely disparate trainee levels by comparing the performances of various combinations of medical students, SHOs, SpRs and Consultants (Winckel *et al*, 1994; Reznick *et al*, 1997; Martin *et al*, 1997; Regehr *et al*, 1998; Datta *et al*, 2004). Some studies examining more closely related levels of trainee have failed to demonstrate sufficient reliability (Scott *et al*, 2000) while others that do achieve this do so at a higher level of training by employing both procedure-specific checklists and global assessments (Dath *et al*, 2004). Previous assessments of basic technical skills have shown most skills to be generic whilst some are specialty specific (Beard *et al*, 2005). Basic skills assessment must therefore be applied early on in training in order to encourage the good trainee and identify and guide the trainee that may be

failing. However, BSTs are often unable to complete an entire procedure and so technical skills assessments must employ generic rather than procedure-specific methods, achieved in this study by both Toronto and EBSTAF-Tech.

The Global Ratings Scale of Operative Performance (Toronto) has been widely validated within OSATs for summative purposes (Winckel *et al*, 1994; Jansen *et al*, 1995; Reznick *et al*, 1997; Martin *et al*, 1997; Regehr *et al*, 1998; Anastakis *et al*, 1999; Bann *et al*, 2003b; Datta *et al*, 2004) and may be safely regarded as the current gold-standard. Excellent correlation between EBSTAF-Tech and Toronto within video assessment therefore supports the concurrent validity EBSTAF-Tech. However, correlation between EBSTAF-Tech and in-post assessments (EBSTAF-TS) was not demonstrated. This finding may be explained by the differences between the two assessment methods: EBSTAF-Tech assesses a "snap-shot" of generic tissue-handling skills in isolation while EBSTAF-TS examines observed technical performance in practice over six months using multiple assessors. In-post EBSTAF-TS is likely to offer the most valid assessment of overall basic technical ability since it avoids the factors associated with assessment procedures. However, it does not assess detailed performance during specific procedures, as may be required by future surgical curricula. It may therefore prove most useful as part of early-on selection processes rather than as a measure of subsequent progress towards competence.

In contrast to consultant ratings, trainees were unable to separate the two levels of trainee although a trend towards this might be inferred for EBSTAF-Tech, which may have approached significance. Experienced surgeons have developed their own personal standards by which they assess surgical skill. Trainees have repeatedly been shown to be poor assessors whose assessment skills improve with experience. However, self-assessments by trainees, and therefore trainer-trainee correlation, has been shown to be improved by the provision of benchmarks to the trainee (Martin *et al*, 1998). Further, simple self-review improves trainees' self-evaluative abilities (Ward *et al*, 2003) although there is little evidence that videotape replay is effective in isolation (Wanzel *et al*, 2002b). In this study there was good correlation between trainer and trainee ratings suggesting that EBSTAF-Tech and Toronto may provide the necessary cues and benchmarks to facilitate more accurate self-assessment by trainees. Trainees may not yet operate like a consultant but, given the right guide, they nonetheless appear able to recognise good surgical technique. Directed videotape review has been advocated in the fields of sports psychology (Kernodle *et al*, 1992) and surgery (Cauraugh *et al*, 1999) while there is evidence that learners may do so more effectively when they themselves control the feedback that they receive (Janelle *et al*, 1997). Video-feedback of a trainee's operating combined with these assessment tools at a rate determined by the trainee may therefore facilitate the identification and targeting of operative weaknesses with minimal additional input from trainers, improving skills

acquisition and thus prove beneficial to surgical career progression. This is clearly an area worthy of further investigation.

The simple estimation of training level in video assessment proved interesting. Being blinded to the true training level of trainees, trainers' underestimation of training level suggests that they felt that the trainees were not as good as trainers felt they should be. Indeed some expressed written concerns to this effect. Interestingly, trainees' comments mirrored those of consultants and suggested a common basis underlying the ratings; flow of the operation, knowledge of the stages of the procedure, respect for tissues, knot tying and instrument handling. This likely illustrates the objective nature of assessment by anonymised video in the absence of halo effect.

The ideal assessment of real-life operative skills would employ an in-theatre method developed by consensus that has proven to be objective, reliable and valid while exerting a minimal effect on the performance of the operator so as to obtain a true picture of their everyday technical ability. Only by such a method can surgical competence assessment be a fair proposition. This study would suggest that video-assessment might prove useful in attaining this goal, particularly when combined with structured assessment tools such as EBSTAF-Tech or Toronto. Furthermore, such techniques may prove equally valuable to the trainees themselves by allowing accurate self-review of performance, facilitating the acquisition of good technical skills.

VII.5. SUMMARY.

- Assessment of BST operative skills during a single real-time procedure by a single assessor is not valid and showed no correlation with in-post assessments by EBSTAF.

- Assessment of edited and anonymised video-recorded procedures by consultant and trainee panels is feasible and reliable, demonstrating both concurrent and construct validity.

- Assessment by the consultant panel was sufficiently sensitive to identify the effect of just six months' training.

- Assessment by the trainee panel was also far more sensitive than expected, demonstrating a near-significant difference in scores across six months of training.

- Concurrent validity in terms of the correlation between video-assessment and in-post assessment by EBSTAF was not demonstrated.

- Trainee assessment skills appear to be improved by the combination of simple video-review and structured assessment tools.

Table VII.1

Reliability Measures For Real-Time And Video Assessments.

			EBSTAF-Tech	Toronto	VAS
Internal Consistency.	RTA	Consultant	0.95	-	-
(Cronbach α)		Trainee	0.85	-	-
	VIDEO	Consultant	0.89	0.78	-
		Trainee	0.76	0.95	-
Inter-Rater Agreement	VIDEO	Consultant	0.86	0.93	0.90
(Intra-Class Correlation)		Trainee	0.69	0.84	0.78
Test Re-Test Correlation	VIDEO	Consultant	0.759 (p<0.001)	0.787 (p<0.001)	0.784 (p<0.001)
(Spearman Rank – *rho* r_s)		Trainee	0.694 (p= 0.004)	0.945 (p<0.001)	0.815 (p<0.001)
Test Re-test Concordance	VIDEO	Consultant	0.604 (p<0.001)	0.602 (p= 0.001)	0.609 (p<0.001)
(Kendall's *tau-b*)		Trainee	0.591 (p= 0.007)	0.850 (p<0.001)	0.657 (p=0.001)

EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Toronto** : Modified Toronto Global Rating Scale of Operative Performance.
VAS : Visual Analogue Scale rating of overall performance.

Table VII.2a

Construct Validity of Real Time Assessment.

	Assessment by Trainers				Self-Assessment by Trainees			
	EBSTAF-Tech		VAS		EBSTAF-Tech		VAS	
BST No.	**SHO 6**	**SHO 12**	**SHO 6**	**SHO 12**	**SHO 6**	**SHO 12**	**SHO 6**	**SHO 12**
1	77	90	57	64	95	91	59	65
2	91	78	64	57	61	100	57	58
3	80	90	78	80	63	74	56	45
4	66		47		59	98	47	41
5	75	77	53	61	84	69	42	45
6	90	95	68	68	88	95	57	48
7	28	53	31	52	40	63	49	36
8	100	89		78	73	77	57	57
9		92		80	96	100	58	58
Median	**76**	**83**	**57**	**68**	**73**	**85**	**53**	**48**
Wilcoxon SHO 6 vs SHO 12	**ns**		**ns**		**ns**		**ns**	

SHO 6 : trainees at 6 months training. **SHO 12** : trainees at 12 months training.
EBSTAF-Tech : Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Toronto** : Global Ratings Scale of Operative Performance.
V-VAS : Visual Analogue Scale rating of overall performance in Video Assessment. All percentages have been rounded.
Median : Median percentage score (inter-quartile range). **Kruskal-Wallis** : Kruskal-Wallis analysis of variance. **Wilcoxon** : Wilcoxon matched pairs signed rank test.

Table VII.2b

Construct Validity of Video Assessment by Trainers.

EBSTAF-Tech

	Trainer No. 1	2	3	4	5	6	7
BST No.				SHO 6			
1	70	25	28	66	91	57	23
2	78	24	44	48	93	47	41
3	52	30	83	90	98	86	61
4	54	20	38	26	98	81	16
5	100	58	52	55	100	81	33
6	96	33	100	87	100	97	100
7	62	26	23	45	98	60	29
8	49	49	49	82	84	49	30
9	56	30	22	70	89	68	39
	Median (IQR) = 56 (33 - 86)						
BST No.				SHO 12			
1	100	37	100	100	100	100	38
2	68	28	89	89	100	100	42
3	100	49	38	66	100	87	40
4	45	33	33	45	73	65	20
5	58	18	64	62	88	76	30
6	100	78	47	93	100	100	59
7	72	26	34	100	100	74	46
8	41	30	93	68	100	55	38
9	85	33	33	70	100	100	46
	Median (IQR) = 68 (40 - 100)						
Cons No.				Consultant			
C1	100	84	100	100	100	100	100
C2	100	76	91	100	100	88	87
	Median (IQR) = 100 (88 - 100)						
Kruskal-Wallis				< 0.001			
Wilcoxon SHO 6 vs SHO 12				0.023			

Toronto

	Trainer No. 1	2	3	4	5	6	7
BST No.				SHO 6			
1	46	26	20	51	77	54	43
2	54	46	40	46	51	43	31
3	51	37	63	71	57	40	46
4	40	26	26	29	60	0	26
5	63	37	34	57	63	60	37
6	80	51	97	71	100	60	100
7	51	26	34	46	69	43	0
8	51	51	40	74	57	46	23
9	37	23	23	60	46	0	34
	Median (IQR) = 46 (37 - 60)						
BST No.				SHO 12			
1	0	46	100	94	63	77	54
2	60	46	60	69	100	60	60
3	69	51	0	57	0	51	40
4	40	40	29	43	31	43	23
5	40	23	34	57	60	51	54
6	100	80	46	77	100	74	46
7	49	29	37	83	83	43	46
8	40	31	71	54	69	0	29
9	60	37	40	69	63	74	57
	Median (IQR) = 54 (26 - 63)						
Cons No.				Consultant			
C1	77	77	100	97	100	97	100
C2	100	74	57	100	100	77	94
	Median (IQR) = 97 (77 - 100)						
Kruskal-Wallis				< 0.001			
Wilcoxon SHO 6 vs SHO 12				0.005			

V-VAS

	Trainer No. 1	2	3	4	5	6	7
BST No.				SHO 6			
1	51	7	7	39	59	51	14
2	64	24	23	15	40	16	28
3	43	11	60	65	43	15	47
4	35	5	8	3	41	48	3
5	72	22	24	16	49	58	10
6	64	36	88	52	80	63	92
7	56	7	24	23	53	13	25
8	59	31	27	36	31	25	23
9	31	11	1	33	26	13	18
	Median (IQR) = 31 (15 - 51)						
BST No.				SHO 12			
1	92	28	85	80	55	63	42
2	64	28	63	57	98	55	59
3	73	27	28	35	74	58	24
4	36	23	17	16	20	19	7
5	48	7	0	44	34	35	36
6	90	73	24	70	93	68	34
7	56	5	18	77	74	18	43
8	47	11	58	35	58	19	35
9	59	10	27	45	43	73	42
	Median (IQR) = 43 (26 - 63)						
Cons No.				Consultant			
C1	70	64	90	73	93	99	93
C2	98	60	61	82	93	56	84
	Median (IQR) = 83 (63 - 93)						
Kruskal-Wallis				< 0.001			
Wilcoxon SHO 6 vs SHO 12				0.001			

SHO 6 : trainees at 6 months training. **SHO 12** : trainees at 12 months training.
EBSTAF-Tech : Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Toronto** : Global Ratings Scale of Operative Performance.
V-VAS : Visual Analogue Scale rating of overall performance in Video Assessment. All percentages have been rounded.
Median : Median percentage score (inter-quartile range). **Kruskal-Wallis** : Kruskal-Wallis analysis of variance. **Wilcoxon** : Wilcoxon matched pairs signed rank test.

Table VII.2c

Construct Validity of Video Assessment by Trainees.

EBSTAF-Tech

Bst No.	Trainee No. 1	2	3	4	5
SHO 6					
1	91	88	100	94	82
2	76	98	95	98	62
3	81	73	78	87	82
4	59	93	66	82	74
5	63	98	76	71	85
6	68	100	77	92	84
7	57	93	100	100	92
8	85	93	87	84	94
9	40	92	91	49	98
Median (IQR) = 85 (75 - 93)					
SHO 12					
1	78	100	100	100	100
2	100	100	100	100	96
3	53	83	92	82	100
4	56	92	0	76	84
5	72	98	72	90	90
6	100	100	100	82	100
7	68	84	100	64	74
8	47	100	51	100	96
9	87	88	96	91	100
Median (IQR) = 92 (79 - 100)					
Consultant					
C1	100	100	100	100	100
C2	96	100	100	92	100
Median (IQR) = 100 (96 - 100)					
Kruskal-Wallis	<0.001				
Wilcoxon SHO 6 vs SHO 12	0.087				

Toronto

Bst No.	Trainee No. 1	2	3	4	5
SHO 6					
1	77	63	86	71	49
2	40	66	63	69	34
3	66	51	60	63	63
4	40	60	46	0	46
5	54	49	54	49	57
6	43	83	54	71	51
7	37	63	74	80	69
8	57	60	51	66	63
9	29	60	57	43	63
Median (IQR) = 60 (49 - 66)					
SHO 12					
1	60	74	71	77	86
2	63	71	91	77	83
3	34	49	63	63	77
4	34	49	0	51	51
5	43	63	46	57	66
6	91	83	86	0	74
7	40	49	63	49	46
8	37	66	43	77	69
9	57	63	60	69	80
Median (IQR) = 63 (49 - 77)					
Consultant					
C1	97	100	91	91	80
C2	74	69	91	86	89
Median (IQR) = 90 (79 - 93)					
<0.001					
ns					

V.VAS

Bst No.	Trainee No. 1	2	3	4	5
SHO 6					
1	74	60	68	61	53
2	36	68	48	68	30
3	61	25	48	61	41
4	14	55	30	55	40
5	26	47	38	48	55
6	34	81	34	67	45
7	32	45	69	70	60
8	51	53	38	59	58
9	9	49	58	31	64
Median (IQR) = 51 (37 - 61)					
SHO 12					
1	48	77	60	72	66
2	67	72	75	74	68
3	15	25	50	58	64
4	13	36	0	50	52
5	22	50	26	49	57
6	73	82	68	57	63
7	24	19	56	33	49
8	15	64	26	77	61
9	70	56	49	67	61
Median (IQR) = 57 (39 - 68)					
Consultant					
C1	92	97	61	72	72
C2	65	64	82	94	84
Median (IQR) = 77 (65 - 93)					
<0.001					
ns					

SHO 6 : trainees at 6 months training. **SHO 12** : trainees at 12 months training.
EBSTAF-Tech : Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Toronto** : Global Ratings Scale of Operative Performance.
V-VAS : Visual Analogue Scale rating of overall performance in Video Assessment. All percentages have been rounded.
Median : Median percentage score (inter-quartile range). **Kruskal-Wallis** : Kruskal-Wallis analysis of variance. **Wilcoxon** : Wilcoxon matched pairs signed rank test.

Table VII.3a

Relationships Between Scoring Methods : Real Time Assessment and Targeted Suturing Task.

Trainers		RTA-VAS	Suture
EBSTAF-Tech	*rho* r_s	0.823 (p<0.001)	0.111 (p= ns)
	tau-b	*0.677 (p<0.001)*	*0.078 (p= ns)*
RTA-VAS	*rho* r_s		0.033 (p= ns)
	tau-b		*0.010 (p= ns)*

Trainees		RTA-VAS	Suture
EBSTAF-Tech	*rho* r_s	0.418 (p= 0.085)	-0.086 (p= ns)
	tau-b	*0.280 (p= ns)*	*-0.007 (p= ns)*
RTA-VAS	*rho* r_s		-.030 (p= ns)
	tau-b		*0.000 (p= ns)*

EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment. **Suture** : Targeted Suture Task (Total Deviation). **RTA-VAS** : Visual Analogue Scale assessment of overall performance in Real Time Assessment.
rho r_s : Spearman Rank correlation. ***tau-b*** : Kendall's Concordance (italics)

Table VII.3b

Relationships Between Scoring Methods : Video Assessment.

Trainers		**Toronto**	V-VAS
EBSTAF-Tech	*rho* r_s	0.866 (p<0.001)	0.922 (p<0.001)
	tau-b	*0.710 (p<0.001)*	*0.789 (p<0.001)*
Toronto	*rho* r_s		0.872 (p<0.001)
	tau-b		*0.765 (p<0.001)*

Trainees		**Toronto**	V-VAS
EBSTAF-Tech	*rho* r_s	0.904 (p<0.001)	0.937 (p<0.001)
	tau-b	*0.747 (p<0.001)*	*0.802 (p<0.001)*
Toronto	*rho* r_s		0.921 (p<0.001)
	tau-b		*0.816 (p<0.001)*

EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment. **Toronto** : Modified Toronto Global Rating Scale of Operative Performance. **V-VAS** : Visual Analogue Scale assessment of overall performance in Video Assessment.

rho r_s : Spearman Rank correlation. **tau-b** : Kendall's Concordance (italics)

Table VII.3c

Relationship Between Real Time Assessment and In-Post Assessment.

Trainers		EBSTAF-TS Consultant
EBSTAF-Tech	*rho* r_s	0.239 (p= ns)
	tau-b	*0.174 (p= ns)*
RTA-VAS	*rho* r_s	-0.067 (p= ns)
	tau-b	-0.062 (p= ns)

Trainees		EBSTAF-TS Self-assessment
EBSTAF-Tech	*rho* r_s	0.925 (p<0.001)
	tau-b	*0.786 (p<0.001)*
RTA-VAS	*rho* r_s	0.494 (p= 0.052)
	tau-b	*0.494 (p= 0.060)*

EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment. **RTA-VAS** : Visual Analogue Scale assessment of overall performance in Real Time Assessment. **EBSTAF-TS** : Technical Skills portion of EBSTAF as used in Parallel In-Post Assessment assessed by Consultant or Trainee respectively. **rho** r_s : Spearman Rank correlation. **tau-b** : Kendall's Concordance

Table VII.3d

Relationship Between Video Assessment and In-Post Assessment.

Trainers		In-Post EBSTAF-TS Consultant
EBSTAF-Tech	*rho* r_s	-0.048 (p= ns)
	tau-b	*-0.033 (p= ns)*
Toronto	*rho* r_s	0.043 (p= ns)
	tau-b	*0.048 (p= ns)*
V-VAS	*rho* r_s	0.111 (p= ns)
	tau-b	*0.082 (p= ns)*

Trainees		In-Post EBSTAF-TS Self-assessment
EBSTAF-Tech	*rho* r_s	0.317 (p= ns)
	tau-b	*0.232 (p= ns)*
Toronto	*rho* r_s	0.168 (p= ns)
	tau-b	*0.130 (p= ns)*
V-VAS	*rho* r_s	0.235 (p= ns)
	tau-b	*0.146 (p= ns)*

EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Toronto** : Modified Toronto Global Rating Scale of Operative Performance. **V-VAS** : Visual Analogue Scale assessment of overall performance in Video Assessment. **EBSTAF-TS** : Technical Skills portion of EBSTAF as used in Parallel In-Post Assessment assessed by Consultant or Trainee respectively.

rho r_s : Spearman Rank correlation. ***tau-b*** : Kendall's Concordance (italics)

Table VII.4

Relationship between Trainer & Trainee Ratings in Real Time and Video Assessment.

a. Real-Time Assessment

EBSTAF-Tech	*rho* r_s	0.427 (p= 0.099)
	tau-b	*0.318 (p= 0.093)*
RTA-VAS	*rho* r_s	0.181 (p= ns)
	tau-b	*0.118 (p= ns)*

b. Video Assessment

EBSTAF-Tech	*rho* r_s	0.511 (p= 0.021)
	tau-b	*0.353 (p= 0.033)*
Toronto	*rho* r_s	0.586 (p= 0.007)
	tau-b	*0.460 (p= 0.006)*
V-VAS	*rho* r_s	0.615 (p= 0.004)
	tau-b	*0.459 (p= 0.006)*

EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment. **Toronto** : Modified Toronto Global Rating Scale of Operative Performance. **RTA-VAS & V-VAS** : Visual Analogue Scale assessment of overall performance in Real Time and Video Assessments respectively.

rho r_s : Spearman Rank correlation. ***tau-b*** : Kendall's Concordance (italics)

Table VII.5

Targeted Suture Placement.

a. Construct Validity

	Median Deviation	IQR	Wilcoxon p=
SHO 6 Months	13.5mm	10.8 – 16.0	ns
SHO 12 Months	11.0mm	9.0 – 14.3	

b. Correlation with Operative Assessment

Trainers	Suture
EBSTAF-Tech	0.111 (p= ns)
RTA-VAS	-0.225 (p= ns)
Trainees	Suture
EBSTAF-Tech	-0.086 (p= ns)
RTA-VAS	-.030 (p= ns)

c. Correlation with In-Post Assessment by EBSTAF-TS

Trainers	0.38 (p= ns)
Trainees	-0.015 (p= ns)

EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Suture** : Targeted Suture Task (Total Deviation). **RTA-VAS** : Visual Analogue Scale assessment of overall performance in Real Time Assessment.
Wilcoxon : Wilcoxon Signed Rank Matched Pairs Test. Correlations by Spearman Rank (*rho*)

Table VII.6

Estimation of Level of Training.

Trainers were asked to estimate the level of training of the operator having first assessed the procedure using EBSTAF-Tech, Toronto and V-VAS.

Only responses to SHO 12 trainee procedures are shown

	Trainers (n = 62/63)	**Trainees** (n = 44/45)	**Overall** (n = 106/108)
Correct	25 (40%)	19 (43%)	44 (42%)
Underestimate	23 (37%)	12 (27%)	35 (33%)
Overestimate	14 (23%)	13 (30%)	27 (25%)

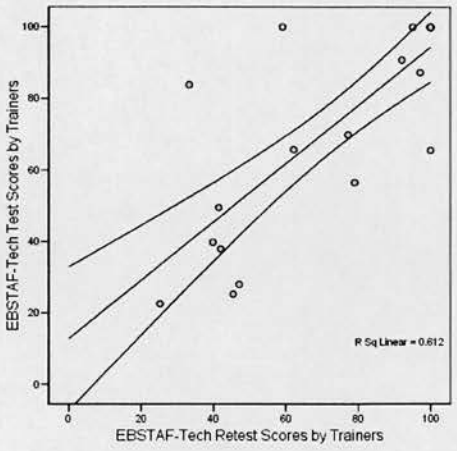
Figure VII.1a

Test re-test reliability of video-assessment by Trainers.

(i) EBSTAF-Tech

rho=0.759, p<0.001

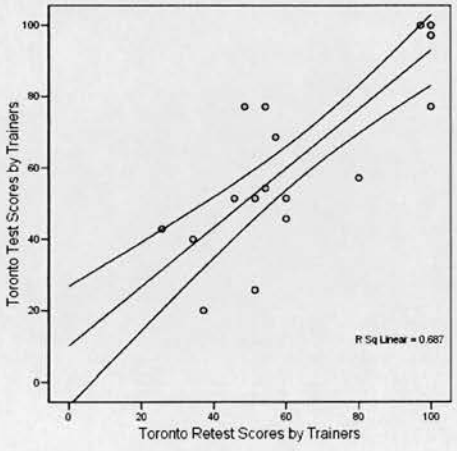
tau-b=0.604, p<0.001



(ii) Toronto

rho=0.787, p<0.001

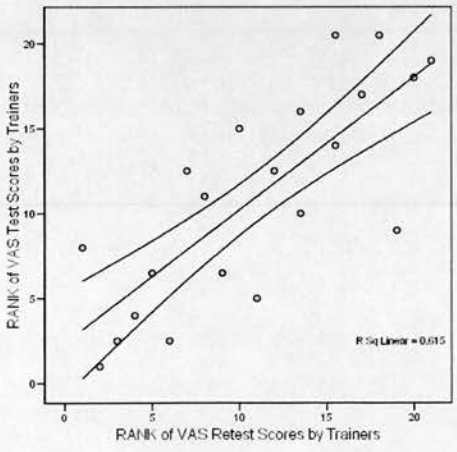
tau-b=0.602, p=0.001



(iii) V-VAS

rho=0.784, p<0.001

tau-b=0.609, p<0.001



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Toronto** : Modified Toronto Global Rating Scale of Operative Performance. **V-VAS** : Visual Analogue Scale rating of overall performancein Video Assessment. ***rho*** : Spearman rank correlation. ***tau-b*** : Kendall's concordance

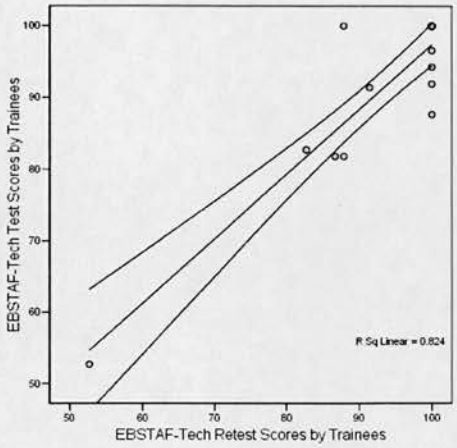
Figure VII.1b

Test re-test reliability of video-assessment by Trainees.

(i) EBSTAF-Tech

rho=0.694, p=0.004

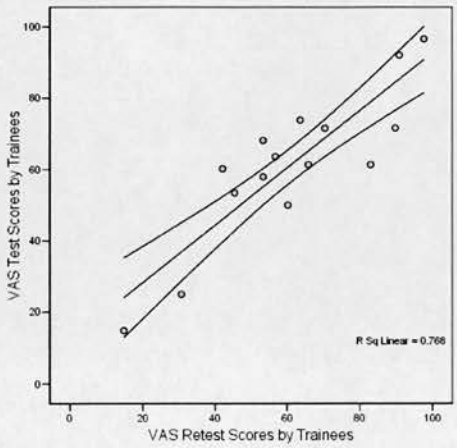
tau-b=0.591, p<0.007



(ii) Toronto

rho=0.945, p<0.001

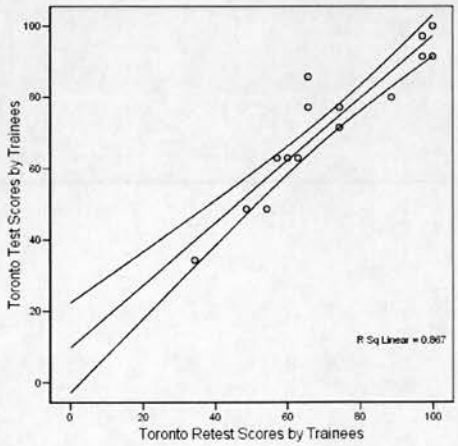
tau-b=0.850, p<0.001



(iii) V-VAS

rho=0.815, p<0.001

tau-b=0.657, p=0.001



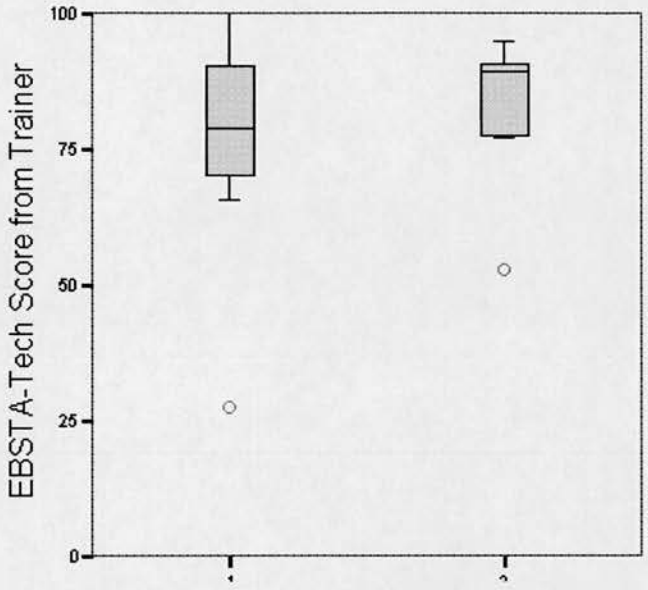
EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Toronto** : Modified Toronto Global Rating Scale of Operative Performance. **V-VAS** : Visual Analogue Scale rating of overall performance in Video Assessment. ***rho*** : Spearman rank correlation. ***tau-b*** : Kendall's concordance

Figure VII.2a

Construct Validity of Real-Time Assessment by Trainers.

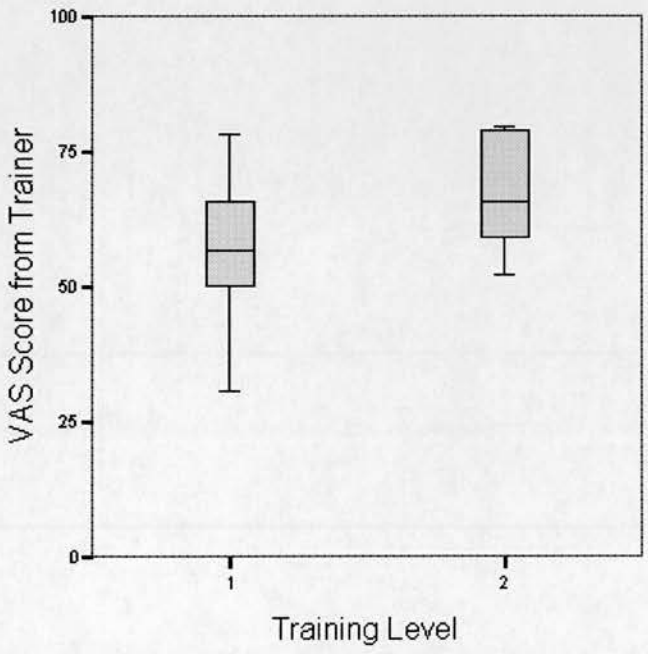
(i) EBSTAF-Tech

Wilcoxon p = ns



(ii) RTA-VAS

Wilcoxon p = ns



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form.

RTA-VAS : Visual Analogue Scale of overall performance in Real Time Assessment.

Training Level : 1 = trainee at 6 months, 2 = trainee at 12 months.

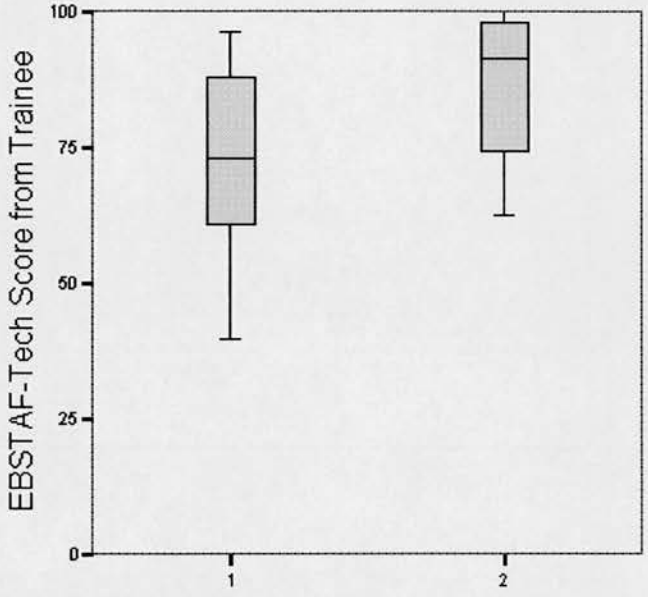
Boxes represent median, inter-quartile range and range with outliers represented by ○.

Figure VII.2b

Construct Validity of Real-Time Assessment by Trainees.

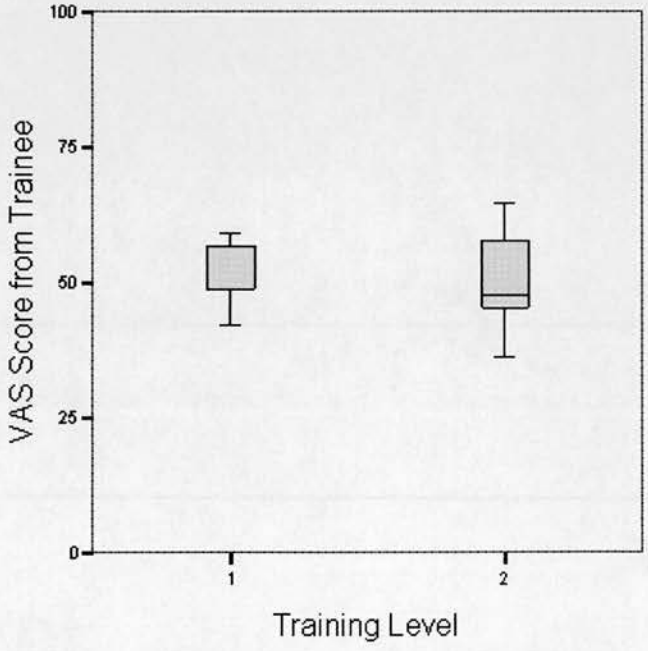
(i) EBSTAF-Tech

Wilcoxon p = ns



(ii) RTA-VAS

Wilcoxon p = ns



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form.

RTA-VAS : Visual Analogue Scale of overall performance in Real Time Assessment.

Training Level : 1 = trainee at 6 months, 2 = trainee at 12 months.

Boxes represent median, inter-quartile range and range with outliers represented by O.

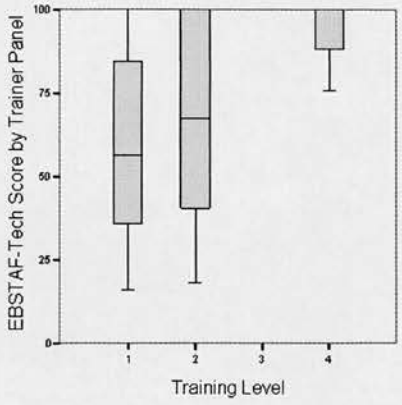
Figure VII.2c

Construct Validity of Video Assessment by Trainers.

(i) EBSTAF-Tech

Kruskal Wallis p <0.001

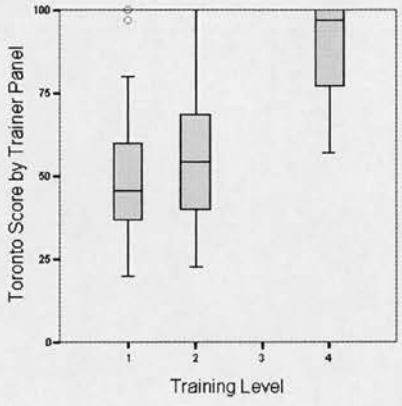
Wilcoxon p = 0.023



(ii) Toronto

Kruskal Wallis p <0.001

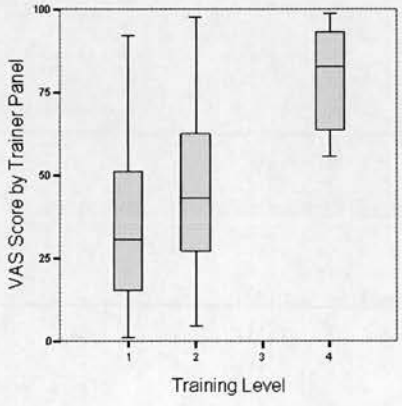
Wilcoxon p =0.005



(iii) V-VAS

Kruskal Wallis p <0.001

Wilcoxon p = 0.001



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Toronto** : Modified Toronto Global Rating Scale of Operative Performance. **V-VAS** : Visual Analogue Scale of overall performance in Video Assessment. **Training Level** : 1 = trainee at 6 months, 2 = trainee at 12 months, 4 = consultant. Boxes represent median, inter-quartile range and range with outliers represented by O.

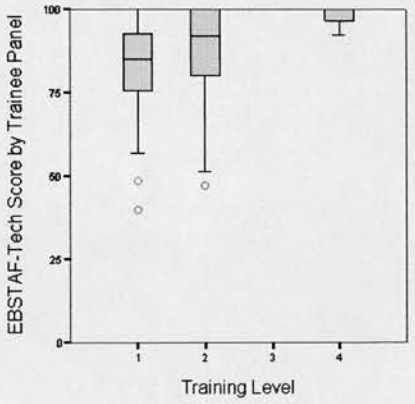
Figure VII.2d

Construct Validity of Video Assessment by Trainees.

(i) EBSTAF-Tech

Kruskal Wallis p <0.001

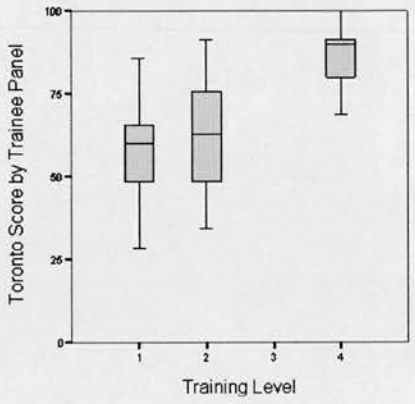
Wilcoxon p =0.087



(ii) Toronto

Kruskal Wallis p<0.001

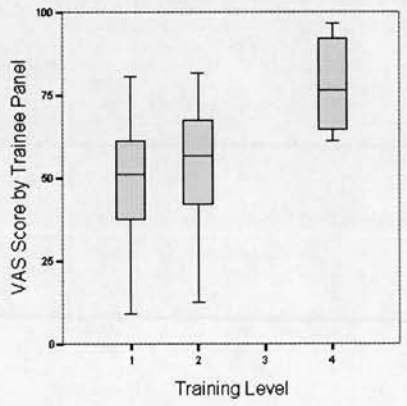
Wilcoxon p = ns



(iii) V-VAS

Kruskal Wallis p <0.001

Wilcoxon p = ns



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form.
Toronto : Modified Toronto Global Rating Scale of Operative Performance. **V-VAS** : Visual Analogue Scale of overall performance in Video Assessment. **Training Level** : 1 = trainee at 6 months, 2 = trainee at 12 months, 4 = consultant.
Boxes represent median, inter-quartile range and range with outliers represented by O.

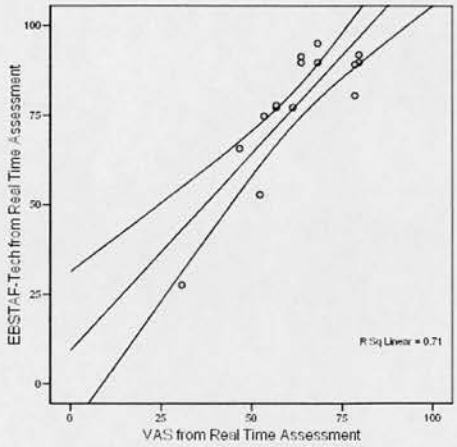
Figure VII.3a

Concurrent Validity of Real Time Assessment by Trainers.

(i) EBSTAF-Tech vs RTA-VAS

rho=0.823, p<0.001

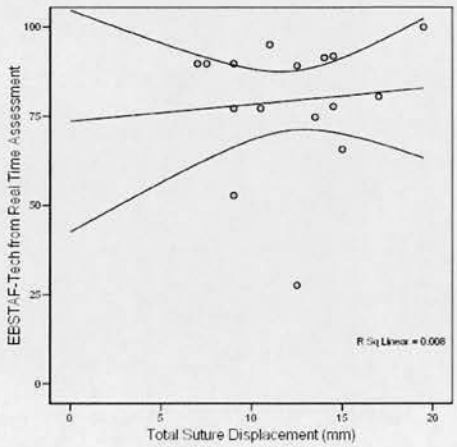
tau-b=0.677,p<0.001



(ii) EBSTAF-Tech vs Suture

rho p = ns

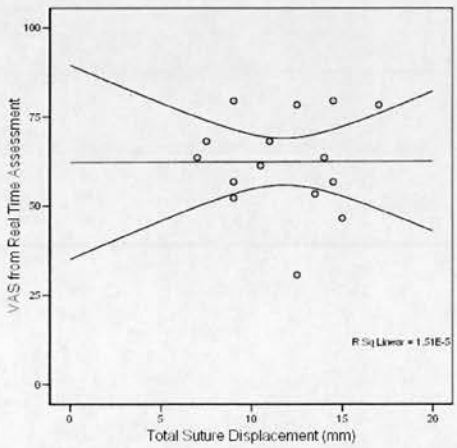
tau-b p = ns



(iii) RTA-VAS vs Suture

rho p = ns

tau-b p = ns



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Suture** : Targeted Suture Task (Total Deviation). **RTA-VAS** : Visual Analogue Scale rating of overall performancein Real Time Assessment. ***rho*** : Spearman rank correlation. ***tau-b*** : Kendall's concordance

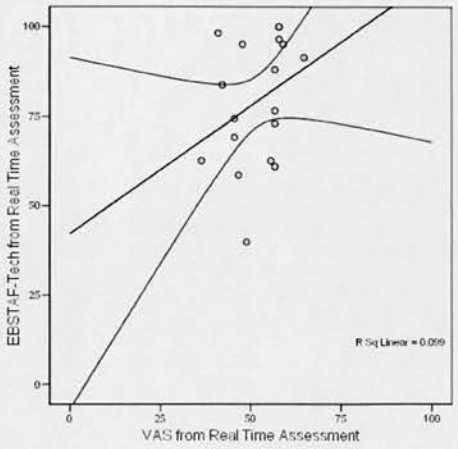
Figure VII.3b

Concurrent Validity of Real Time Assessment by Trainees.

(i) EBSTAF-Tech vs. Real-VAS

rho=0.418, p =0.085

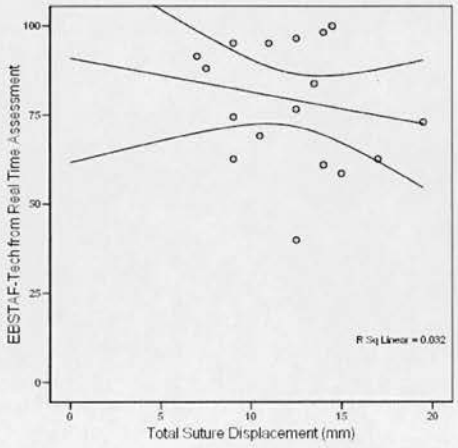
tau-b p = ns



(ii) EBSTAF-Tech vs. Suturing

rho p = ns

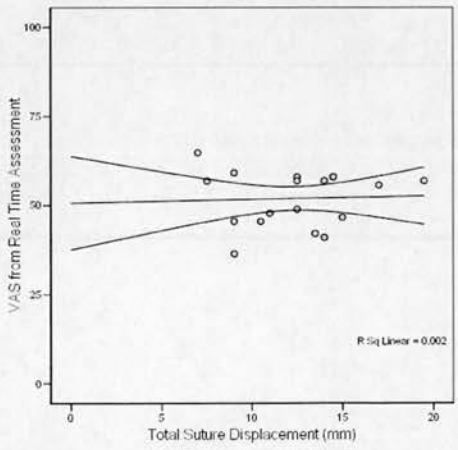
tau-b p = ns



(iii) RTA-VAS vs. Suturing

rho p = ns

tau-b p = ns



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Suture** : Targeted Suture Task (Total Deviation). **RTA-VAS** : Visual Analogue Scale rating of overall performance in Real Time Assessment. ***rho*** : Spearman rank correlation. ***tau-b*** : Kendall's concordance

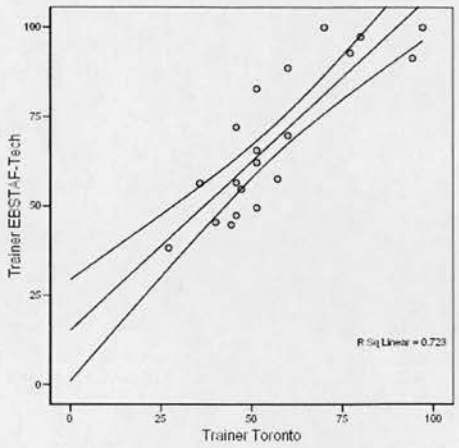
Figure VII.3c

Concurrent Validity of Video Assessment by Trainers.

(i) EBSTAF-Tech vs. Toronto

rho=0.866 p<0.001

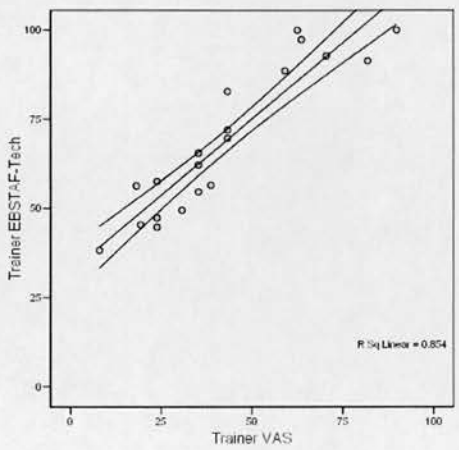
tau-b=0.710 p<0.001



(ii) EBSTAF-Tech vs. V-VAS

rho=0.922 p<0.001

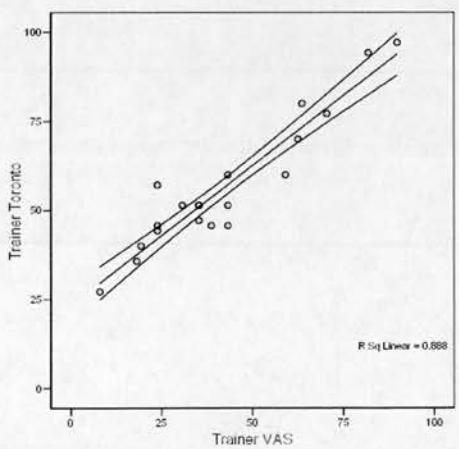
tau-b=0.789 p<0.001



(iii) Toronto vs. V-VAS

rho=0.872 p<0.001

tau-b=0.765 p<0.001



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form.
Toronto : Modified Toronto Global Rating Scale of Operative Performance. **V-VAS** : Visual Analogue Scale rating of overall performance in Video Assessment. ***rho*** : Spearman rank correlation. ***tau-b*** : Kendall's concordance

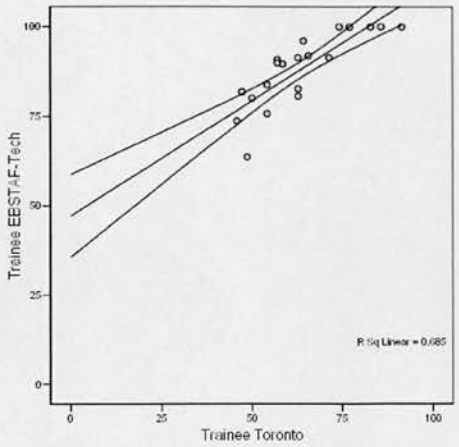
Figure VII.3d

Concurrent Validity of Video Assessment by Trainees.

(i) EBSTAF-Tech vs. Toronto

rho=0.904 p<0.001

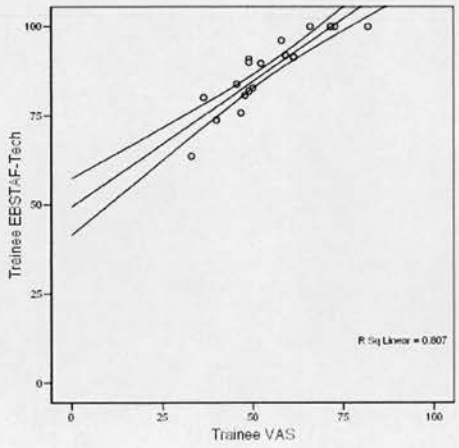
tau-b=0.747 p<0.001



(ii) EBSTAF-Tech vs. V-VAS

rho=0.937 p<0.001

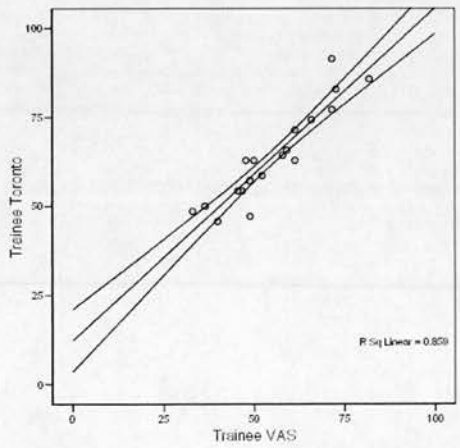
tau-b=0.802 p<0.001



(iii) Toronto vs. V-VAS

rho=0.921 p<0.001

tau-b=0.876 p<0.001



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Toronto** : Modified Toronto Global Rating Scale of Operative Performance. **V-VAS** : Visual Analogue Scale rating of overall performance in Video Assessment. ***rho*** : Spearman rank correlation. ***tau-b*** : Kendall's concordance *tau-b*

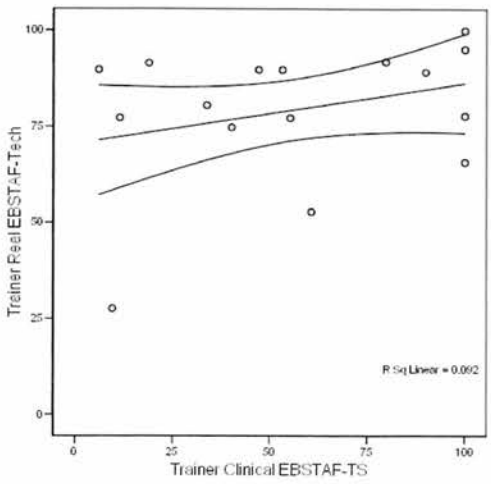
Figure VII.3e

Relationship Between Real Time Assessment and In-Post Assessment by Trainers.

(i) **EBSTAF-Tech vs. EBSTAF-TS**

rho p = ns

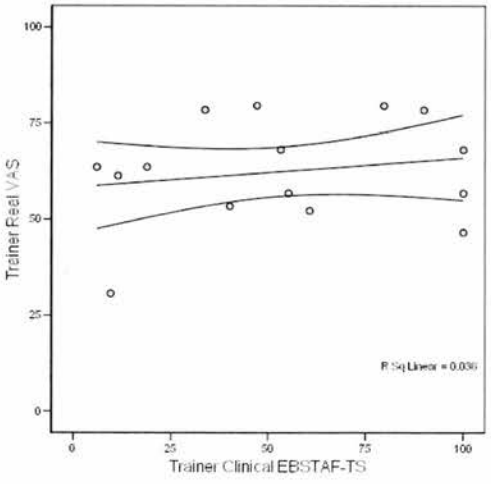
tau-b p = ns



(ii) **RTA-VAS vs. EBSTAF-TS**

rho p = ns

tau-b *p = ns*



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **EBSTAF-TS** : Technical Skills portion of EBSTAF as used in Parallel In-Post Assessment assessed by Consultant or Trainee respectively. **RTA-VAS** : Visual Analogue Scale assessment of overall performance in Real Time Assessment. ***rho*** r_s : Spearman Rank correlation. ***tau-b*** : Kendall's Concordance (italics).

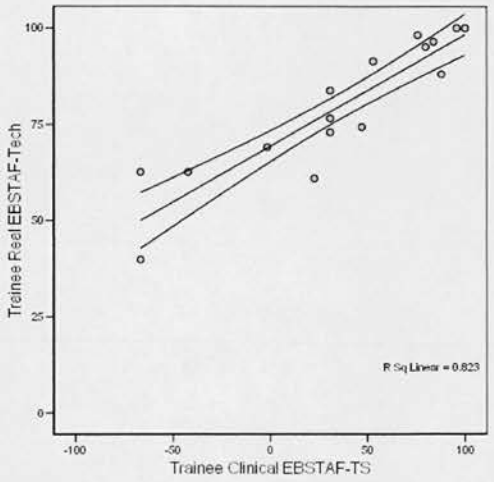
Figure VII.3f

Relationship Between Real Time Assessment and In-Post Assessment by Trainees.

(i) EBSTAF-Tech vs. EBSTAF-TS

rho=0.925 p<0.001

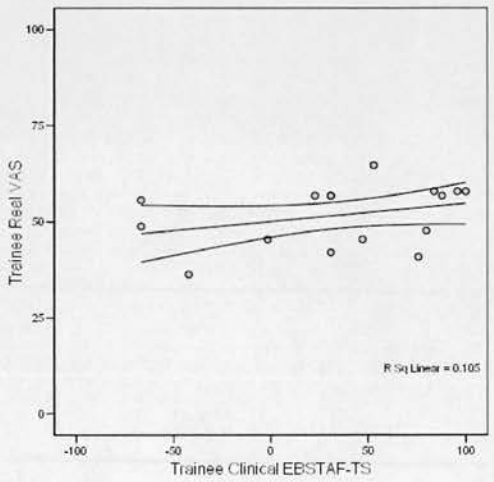
tau-b=0.786 p<0.001



(ii) RTA-VAS vs. EBSTAF-TS

rho=0.494 p =0.052

tau-b=0.494 p=0.060



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form.
EBSTAF-TS : Technical Skills portion of EBSTAF as used in Parallel In-Post Assessment assessed by Consultant or Trainee respectively. **RTA-VAS** : Visual Analogue Scale assessment of overall performance in Real Time Assessment.
rho r_s : Spearman Rank correlation. ***tau-b*** : Kendall's Concordance (italics)

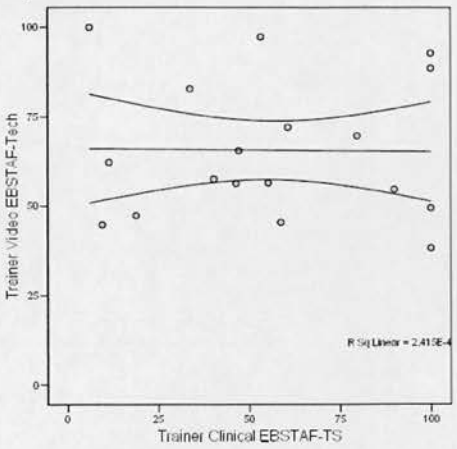
Figure VII.3g

Relationship Between Video Assessment and In-Post Assessment by Trainers.

(i) EBSTAF-Tech vs. EBSTAF-TS

rho p = ns

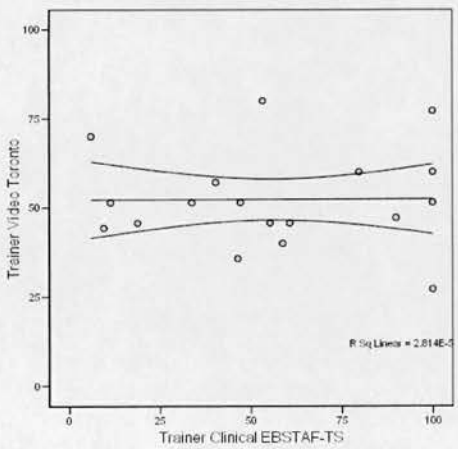
tau-b p = ns



(ii) Toronto vs. EBSTAF-TS

rho p = ns

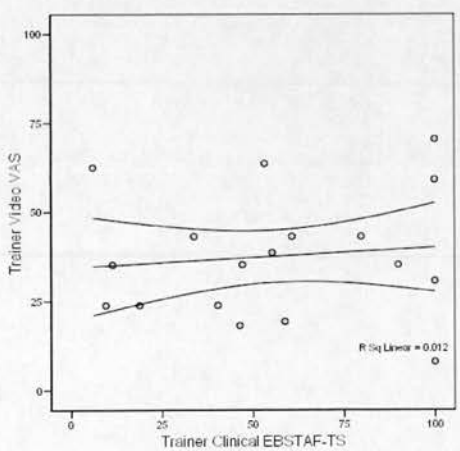
tau-b p = ns



(iii) V-VAS vs. EBSTAF-TS

rho p = ns

tau-b p = ns



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form.

EBSTAF-TS : Technical Skills portion of EBSTAF as used in Parallel In-Post Assessment assessed by Consultant or Trainee respectively. **RTA-VAS** : Visual Analogue Scale assessment of overall performance in Real Time Assessment.

rho r_s : Spearman Rank correlation. ***tau-b*** : Kendall's Concordance (italics)

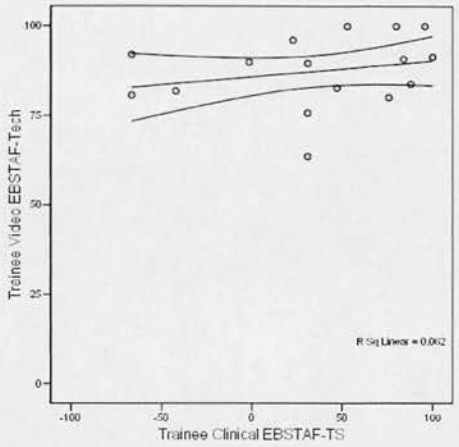
Figure VII.3h

Relationship Between Video Assessment and In-Post Assessment by Trainees.

(i) EBSTAF-Tech vs. EBSTAF-TS

rho p = ns

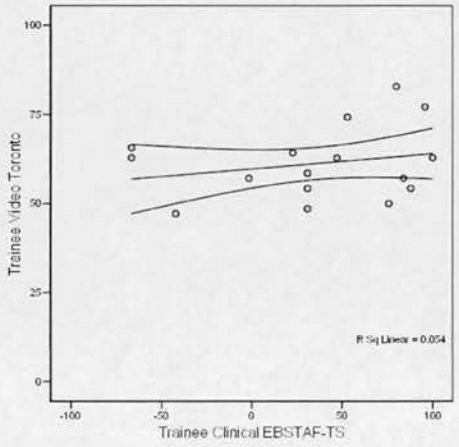
tau-b p = ns



(ii) Toronto vs. EBSTAF-TS

rho p = ns

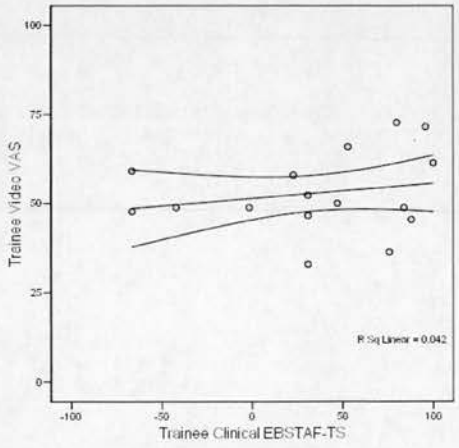
tau-b p = ns



(iii) V-VAS vs. EBSTAF-TS

rho p = ns

tau-b p = ns



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form.

EBSTAF-TS : Technical Skills portion of EBSTAF as used in Parallel In-Post Assessment assessed by Consultant or Trainee respectively. **RTA-VAS** : Visual Analogue Scale assessment of overall performance in Real Time Assessment.

rho r_s : Spearman Rank correlation. ***tau-b*** : Kendall's Concordance (italics)

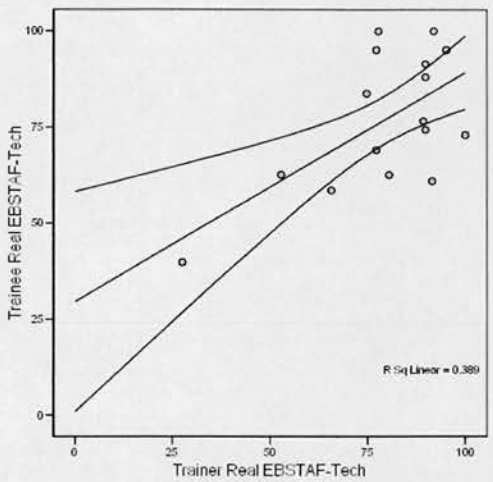
Figure VII.4a

Relationship Between Trainer and Trainee Scores in Real Time Assessment.

(i) EBSTAF-Tech

rho=0.427 p=0.099

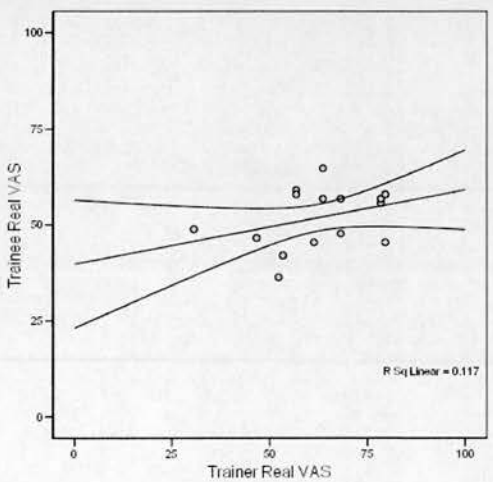
tau-b=0.318 p=0.093



(ii) Toronto

rho p = ns

tau-b p = ns



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form.

Toronto : Modified Toronto Global Rating Scale of Operative Performance.

rho r_s : Spearman Rank correlation. ***tau-b*** : Kendall's Concordance (italics)

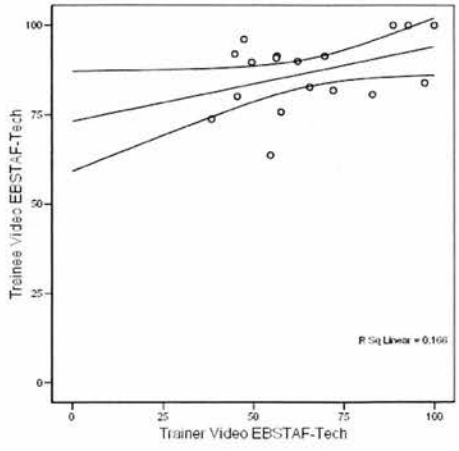
Figure VII.4b

Relationship Between Trainer and Trainee Scores in Video Assessment.

(i) EBSTAF-Tech

rho=0.511 p=0.021

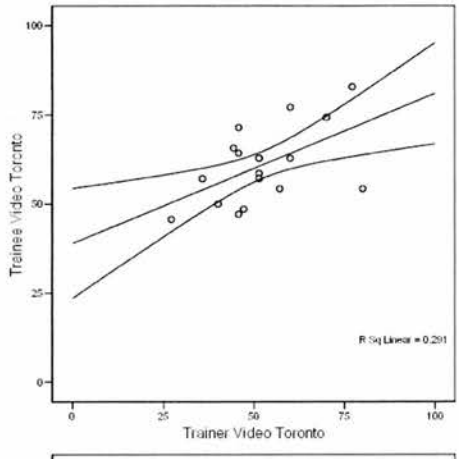
tau-b=0.353 p=0.033



(ii) Toronto

rho=0.586 p=0.007

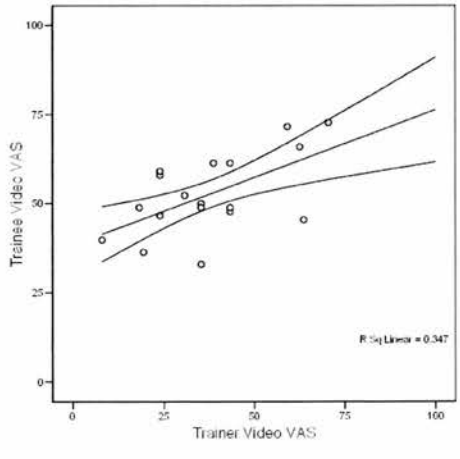
tau-b=0.460 p=0.006



(iii) V-VAS

rho=0.615 p=0.004

tau-b=0.459 p=0.006



EBSTAF-Tech ; Technical Skills portion of Edinburgh Basic Surgical Trainees Assessment Form. **Toronto** : Modified Toronto Global Rating Scale of Operative Performance. **V-VAS** : Visual Analogue Scale of overall performance in Video Assessment. ***rho*** r_s : Spearman Rank correlation. ***tau-b*** : Kendall's Concordance (italics)

Section VIII.

GENERAL DISCUSSION.

The GMC expects a competent doctor to maintain up-to-date professional knowledge and skills while recognising the limits of their professional competence (General Medical Council , 2001). Epstein quotes McPhee and Westberg when he states that "exemplary doctors (and surgeons) seem to have a capacity for critical self-reflection that pervades all aspects of their practice including communication with the patient, problem-solving, eliciting and disseminating information, making evidence-based decisions, performing technical skills and defining their own values" (Epstein, 1999). However, for the trainee to develop what Epstein terms 'mindful practice' requires effective self-evaluation and careful guidance in addition to simple experience. Experience alone may promote confidence but this bears little relation to competence as determined by standardised tests (Jolly et al, 1996; Morgan et al, 2002). Indeed, an inverse relationship between confidence and performance has been demonstrated, albeit in medical students during cardiorespiratory resuscitation (CPR) assessments (Marteau et al, 1991). Poorly performing trainees may genuinely be unaware of their failings due to a lack of professional insight. This may be countered by providing comparisons of trainee performance with the required standards, resulting in improvements in both performance and self-evaluation skills (Hays et al, 2002). This forms the basis of both formative assessment and feedback. However, it is vital that such information is not only relevant to the curriculum but is also seen to be relevant by trainees who, as adult learners, direct a problem-centred approach towards gaining knowledge or skills that they see as necessary for their future practice (Zemke *et al* ,1984). In order

to optimise trainees' learning, they should therefore agree with both the contents of the curriculum and how it is to be assessed. This does not equate to trainees determining the curriculum itself since they are not equipped with the knowledge or skills to do so. Rather, their opinions of the expert-prescribed curriculum should be sought to promote a mutual understanding of what is required and how best to achieve it. At first glance, this concept may appear alien to surgery but outstanding trainers have long applied this subconsciously. By explaining the benefit of a particular action to the patient, the trainer ensures that the trainee learns from the situation and will likely apply what they have learnt to similar situations in their future practice. This concept is perhaps less intuitive when it comes to assessment, but ultimately the same rules apply.

Acceptability of EBSTAF to the Trainees.

It was for this reason that the contents of EBSTAF were put to the trainees for their opinions of the skills and attributes that it examined, themselves prescribed by the expert consensus view of trainers. The results were very positive, with a 100% response rate and 68 of 70 fields (97%) rated by trainees as equal or greater importance compared to consultant ratings. It is therefore likely that trainees will value subsequent guidance based upon these EBSTAF assessments. Assessment of in-post trainee performance using the contents of EBSTAF is therefore acceptable to the trainees, representing the first time this issue has been formally addressed in surgical

training. Subsequent to this study, Ringsted et al describe an examination of the acceptability to trainees of the Danish in-training assessment programme in anaesthesia. However, assessments were not multidisciplinary and trainee response rate was low at only 15 of 27 (56%), with evaluations limited to general impressions rather than specific examination of content (Ringsted *et al*, 2003). No surgical assessment has been similarly tested to date, making EBSTAF the only assessment tool shown to be acceptable to the subjects of the assessment process, namely the trainees themselves.

Just as the acceptability to patients is a vital consideration in the development of clinical interventions, this process must be considered during the development and application of future assessments. The literature would suggest the acceptability of assessments to trainees to be of low priority, but this must change if truly robust assessment *for* training and *of* training is to be achieved.

Having demonstrated the skills examined by EBSTAF to be acceptable to trainees, it was necessary to determine how best to apply the information obtained from such assessments.

Application of EBSTAF to Selection Processes.

The purpose of any assessment is to identify good and poor performers for the purposes of guidance, targeted training or selection. Most studies in this area address concurrent validity, the relationship between one assessment method and another, each with their own strengths and weaknesses. However the goal of any BST, and indeed a basic surgical training programme, is to produce a competent trainee able to attain an SpR post. Therefore, the predictive validity of EBSTAF assessments was legitimately examined for their relationship with subsequent career progression to an SpR post and, perhaps in the future, a consultant position. EBSTAF shows promise in this area by identifying poor performers who appeared to be more likely to struggle in the future. However, the question of whether a slower career progression was to the detriment of the trainees studied (or their patients) remains unclear. Historically, the long training pathway may have allowed the 'slower' trainee to reach competence and the additional advantage that trainers came to know their trainees well, allowing them to make informed judgements as to their suitability for consultant practice, however subjective such judgements inevitably were. However, recent changes in the national training structure for all doctors as a result of Modernising Medical Careers (MMC) (Department of Health 2003), the subject of a recent review (Tooke *et al* 2007), has created a 6-year timeline for the training of all doctors. This will inevitably result in the loss of a few trainees from the surgical profession who have the potential to become

consultants if afforded a little more time. Conversely, robust selection to such a system is of vital importance, irrespective of specialty, since entry to *run-through* training may equate to selection for future consultant practice. It is therefore vital for trainees to be objectively assessed. Those showing an aptitude towards a particular specialty may then be encouraged and selected while those who are more likely to struggle in that specialty may be supported or guided elsewhere. However specialties such as surgery pose a particularly difficult challenge as the assessment and selection of trainees must take place at a stage in their training when they have gained little, if any, technical skill or surgical experience. Assessments of purely operative skills are therefore less discriminatory and may in fact bear little relation to a trainee's future ability. Instead, trainees must be judged on their overall performance according to assessments of generic skills and attributes. Having been developed specifically for the assessment of BSTs, EBSTAF examines 18 generic fields, within the domains of communication and teamworking, which were determined by consensus to be vital for a successful career in surgery. Assessment by EBSTAF therefore provides a detailed and longitudinal insight into a trainee's generic strengths and weaknesses, examined over a six-month period of everyday practice, in addition to the more traditional and specialty-specific technical aspects of surgical practice. The examination of primary EBSTAF assessments completed by medical staff suggests the potential to predict poor career progression post-BST. EBSTAF may therefore be useful in the early identification of trainees who are more likely to struggle later on in their

surgical training. However, the low numbers involved in this study make it difficult to draw firm conclusions and demand a more widespread application and examination of subsequent career progression before application to selection processes that would otherwise provide a self-fulfilling prophecy.

The selection of trainees sets further challenges, however. EBSTAF examines the attainment of *competence*, an all-or-nothing concept with an individual either being competent or not. Although a good trainee may be assessed as competent in more fields than a less able trainee, a ceiling is imposed by the attainment of competence that cannot, given the present structure of EBSTAF, be surpassed. EBSTAF is not therefore able to identify excellence, and its future application to trainee selection would require a restructuring to allow grading above that of merely competent. Furthermore, essential fields requiring competence *prior* to selection for surgical training need to be identified and, for these, competence could become the minimum requirement for selection. Such a major restructuring of EBSTAF would demand full revalidation prior to its application but this must be recommended as a future direction.

Despite changes in career structure as a result of MMC, the same problems challenge the selection of current basic trainees, and the incorporation of components of EBSTAF into selection processes should be considered.

Application of EBSTAF to Trainee Feedback.

Feedback of trainee performance has traditionally involved a brief discussion soon after a particular clinical scenario with the advantage of improved recall of events. However, the feedback of single scenarios by lone assessors is inevitably both case and assessor specific. An inappropriately hawkish assessment of a challenging scenario may demoralise a good trainee while a lenient assessment of a more straightforward one, where the trainee really should have performed better, is unlikely to result in any change in behaviour. In addition, trainees are adult learners who have their own opinions on their performance and a highly subjective opinion that conflicts with their own may be simply discarded rather than accepted.

Longitudinal multi-source assessment of performance across multiple scenarios spread over a period of time provides feedback that is far more robust and less easy for the trainee to discard, however different it may appear to their own opinion. EBSTAF provides just such an assessment that, due to its highly detailed nature, offers less opportunity for individual interpretation. This study shows that the trainees valued highly the information provided by EBSTAF feedback forms. Furthermore, a positive effect on performance was suggested. However the low number of trainees assessed and the lack of sensitivity of the EBSTAF instrument itself, in part as a result of the grading system as discussed earlier, may have resulted in a failure to *demonstrate* significance in the domains examined rather than it not

occurring. The study is likely to have been further confounded by the multifactorial nature of training and the lack of limitations that could be placed on informal feedback between trainer and trainee in either study cohort. Indeed, the very act of completing assessments on trainees may have improved the extent of this informal feedback. However, the formal and structured feedback of everyday performance assessments from multidisciplinary assessors should have, if nothing else, improved trainees' self-assessment abilities. This was not specifically examined in this study but must be considered as part of future research, along with an examination of the effect of structured feedback. This is likely to demand a large multi-centre (or possibly national) study randomising trainees to informal feedback or formal structured feedback to determine its true effect upon performance and trainee self-assessment skills. This is further addressed in Section IX.

Application of EBSTAF to Hi-Fidelity Human Patient Simulation.

The application of EBSTAF to hi-fidelity human patient simulation of critical care scenarios for BSTs was particularly interesting. As might have been expected from its high face validity and demonstrated validity from studies in other specialties, scenarios were able to distinguish trainees of increasing experience in the care of critically ill patients. However, both the debriefing session and comments from the trainees themselves clearly recognised the importance of non-technical skills in what has always been considered to be a technical area of practice. White quotes Collins, the then chairman of the

RCSEng Patient Liason Group, who suggested that the nature of surgical decision-making made surgeons more self-reliant and therefore less able to work as part of a team (White, 2002). Yet it was clear during the scenarios that the good trainees better utilised the skill mix of their team and displayed a higher degree of situation awareness during the scenarios. The poor performers on the most part knew how to manage a situation, but failed to recognise it within the scenario for a variety of non-technical reasons. Once highlighted, trainees recognised the benefits of improved non-technical skills but were keen to point out that they had never been trained with reference to this area of practice. There is, as previously described, a growing interest in the area of non-technical skills in medicine and in particular surgery. However the inability, until recently, to robustly define or assess non-technical skills and therefore give feedback on non-technical performance has limited their introduction to training and incorporation into everyday surgical practice. The "quick fix" of adopting skills taxonomies from other industries, or even allied specialties within medicine such as anaesthesia, is flawed and should be discouraged (Yule *et al*, 2006a). As a result of observations made during the HPS course combined with techniques previously applied in the development of a taxonomy non-technical skills in aviation and anaesthetics, the Non Technical Skills in Surgery (NOTSS) taxonomy, with an appropriate behavioural marker system, has been developed (Yule *et al*, 2006b). This has allowed the assessment of surgeons' non-technical skills and has subsequently demonstrated that they can be improved by training (as yet unpublished data).

The ability to assess both clinical and non-technical skills in combination using high-fidelity patient simulators suggests great potential in the assessment of trainee critical care skills. Their incorporation into courses such as CCrISP has been considered by the Colleges of Edinburgh and Glasgow, but was felt that the substantial financial and manpower costs were prohibitive at the present time. However, this must not be allowed to mask what has been learned from the HPS environment. Such adjuncts to traditional training and assessment must be considered for both training and future consultant revalidation processes.

Application of EBSTAF to Video Assessment of Tissue-Handling Skill.

The application of EBSTAF-Tech to the video-assessment of BSTs tissue-handling skills also proved useful. Many studies have been described which address validity derived from improved scores between widely disparate levels of surgeon. The development of assessment methods in this context will, no doubt, prove useful in the assessment of trainees' progress as they pass through run-through surgical training. However, little work has been carried out at the most junior end of surgical training. As detailed earlier, selection processes are likely to prove difficult at best and, if incorrectly applied, run the very real risk of selecting individuals not suited for surgery who subsequently fail to attain competence. It would thus be valuable to assess basic surgical skills prior to the selection process. Both EBSTAF-

Tech and Toronto, when applied to multi-assessor video-assessment of basic tissue handling, offer extremely sensitive measures of ability by discriminating between very close levels of the most junior surgical trainees. Clearly the predictive validity of such assessments requires further attention before their incorporation into selection processes, but it may be that courses similar to the Royal Colleges' Basic Surgical Skills course could be made a pre-requisite for application to surgery with an in-built grading of ability based upon such measures. This does however raise issues concerning the availability of such courses, which are already oversubscribed and frequently only available later in surgical training.

As discussed in Section VII, the fact that EBSTAF-Tech failed to demonstrate concurrent validity with clinical assessment of technical skills should not be allowed to detract from what appears to be a very robust assessment tool. Indeed, it may be considered reassuring that the attempt failed to show significance, illustrating the difference between single short-loop assessment procedures and the longitudinal multi-source assessment of performance-in-practice provided by EBSTAF. For assessment processes to prove their worth in the selection of future surgeons, it is important to recognise what it is that needs to be assessed at each training stage. It is also essential to ensure that what is observed is not a surrogate measure of ability; a high level of non-technical skill may mask more technical failings in the short term, but longitudinal multi-source assessment of an individual's everyday practice will provide a more detailed picture of weak areas to guide further training.

The assessment of everyday practice is unavoidably complicated and labour intensive, but it is essential if those individuals who are likely to struggle to reach competence are to be identified. EBSTAF appears to provide just such an assessment tool, applicable at an early stage of training, and its incorporation into existing assessment processes should be recommended. However, the issues highlighted above must be addressed in its wider application before formative EBSTAF assessments *for* training are applied to summative assessments *of* training for the purposes of selection. As always, the right tool for the right job.

Section IX.

CONCLUSIONS.

- EBSTAF identifies those trainees that may struggle to progress in their subsequent surgical career.

- The skills and attributes examined by EBSTAF are acceptable to basic surgical trainees.

- The use of EBSTAF assessments for detailed structured feedback appears to have a positive effect on subsequent in-post performance.

- Trainees greatly valued detailed and structured feedback of in-post performance based upon EBSTAF assessments.

- The use of EBSTAF in HPS-based critical care scenarios demonstrated reliability and construct validity. The course also highlighted the importance of non-technical skills as an important factor in poor performances within the scenarios.

- The use of EBSTAF in video-based assessment of basic tissue-handling skills of BSTs demonstrated reliability and validity. It may also have improved trainee self-assessment skills.

Section X.

AN UPDATE OF SUBSEQUENT DEVELOPMENTS

AND

RECOMMENDATIONS FOR FUTURE WORK.

X.1. SUBSEQUENT DEVELOPMENTS AND POTENTIAL PROBLEMS.

The work described in this thesis was carried out between February 2001 and February 2003, during what will likely come to be recognised as the most turbulent decade in the history of surgical training. As a result, there have been many changes to both its structure and processes for assessment and selection. It is therefore necessary to put the above-described work into context with the present system and suggest how it may be applied to guide and improve future developments.

X.1.a. Modernising Medical Careers.

Broadly, the recommendations of MMC (Department of Health 2003)have been sequentially implemented. Having graduated from medical school, doctors enter a mandatory two-year foundation programme in clinical practice (FY1 and FY2), gaining a broad range of experience across clinical specialties whilst remaining under close supervision. During this period, those trainees intent upon a career in surgery must gain evidence to support their application to Run-Through Specialist Training Programmes, in the form of research and audit publications, whilst regularly undergoing formative assessments in the form of Mini-Patient Assessment Tools (Mini-PAT), Mini-Clinical Evaluation Exercises (Mini-CEX), Direct Observations of Procedural Skills (DOPS), Case-Based

Discussions (CBD)) and Procedure-Based Assessments (PBA). These are all coordinated on-line by PMETB and the Intercollegiate Surgical Curriculum Project (ISCP) across an interactive curriculum using competency assessment with explicit standards. However, these are all simply assessment tools and it is worthwhile examining them in detail to highlight their strengths and weaknesses.

X.1.a.i. Mini Peer Assessment Tool (Mini-PAT).

The mini-PAT was derived from the Sheffield Peer Review Assessment Tool (SPRAT), an established multisource 360 degree feedback instrument designed for more senior doctors (Archer and Davies 2003; Archer *et al*, 2005). It comprises multi-source assessment across 16 highly generic fields graded using a likert scale (1 to 10) and cross-referenced to the MMC curriculum, itself based upon Good Medical Practice (General Medical Council , 2001). These fields are broadly similar to generic fields examined by EBSTAF [see Appendix: Section XII Part 6]. However, there are many fields within EBSTAF, more specifically related to surgery and determined as necessary for surgical success (Baldwin *et al*, 1999), that are not examined by Mini-PAT. Mini-Pat has been validated using standard criteria (Archer *et al*, 2008) but its predictive validity remains unexamined.

X.1.a.ii. Mini Clinical Evaluation Exercise (Mini-CEX).

The Mini-CEX was originally developed by Norcini (Norcini *et al*, 1995) [see Appendix: Section XII Part 6]. It comprises a 15-minute assessment of a doctor-patient encounter roughly equivalent to an observed clinical long case examining demonstrated clinical skills, attitudes and behaviour. Four to six such encounters are to be completed during each foundation year, each by a different assessor who then provides immediate feedback of the trainee's strengths and weaknesses (lasting just five minutes) along with a subsequent action plan to address issues that may be forthcoming. Assessment covers seven criteria detailed within a single-page Assessor Written Training document; history taking, physical examination skills, communication, clinical judgement, professionalism, organisation/efficiency and overall clinical care. Grading (1 to 6) is determined in relation to what the assessors would expect at the specific training level. Here, assessments do have two gradings with the descriptors of "above expectations for FY1/FY2 completion", allowing the potential identification of individuals that excel. However, each assessment episode will be inevitably highly subjective, using a single assessor over a single case (albeit weighted within the assessment form). The selection of individuals on this basis may therefore easily be criticised. Similarly, feedback that the trainee receives will also be highly subjective, based upon the assessor's likes and dislikes. In comparison, EBSTAF provides a multisource longitudinal assessment of performance

based upon specific fields derived from consensus opinion. Individual assessments are therefore less subjective while the multisource nature of each assessment episode results in far greater objectivity and therefore superior feedback.

X.1.a.iii. Direct Observation of Procedural Skills (DOPS).

DOPS is equivalent to a mini-CEX of practical skills determined as essential for good clinical care by the Royal College of Physicians rather than the surgical Colleges [see Appendix: Section XII Part 6]. Procedures include venepuncture, cannulation, peripheral and central blood culture, intravenous infusions, the obtaining of an ECG, arterial blood sampling, subcutaneous / intradermal / intramuscular / intravenous injection, urethral catheterisation, airway care, insertion of a nasogastric tube and tracheal intubation. The wide range of procedures is entirely non-surgical and can be seen to include procedures rarely performed by doctors at FY1/FY2 level, such as tracheal intubation. The choice of timing, procedure and observer are at the discretion of the trainee and assessment is graded 1 to 6 across 11 criteria, again with relation to what is expected at FY1/FY2 level. DOPS seeks to identify competence rather than excellence and although the same two grades identify a standard above expectations for FY1/FY2, selection on the basis of DOPS will be difficult. Data is heterogeneous with trainees excelling in relatively simple tasks, such as cannulation, being compared to trainees who perhaps

have failed in their attempts at more challenging assessment procedures, such as tracheal intubation. As a result, data will be highly procedure specific as well as assessor specific, again using single assessors. EBSTAF does not address trainee performance across specific procedures, having never been designed to do so. However, it does assess many of the same criteria as DOPS over a more prolonged period using multiple assessors. It is therefore less procedure specific and more objective in its assessment of trainees' procedural skills.

X.1.a.iv. Case-Based Discussions (CBD).

CBDs revolve around a trainee's involvement in a particular case [see Appendix: Section XII Part 6]. They are designed to assess clinical decision-making and the application of medical knowledge in the care of a trainee's own patients. This is not a new concept in medical/surgical practice; cases have long been discussed between trainer and trainee at all levels. The difference lies in the assessment and documentation of discussions across 7 criteria graded (1 to 6) relative to what is expected of a FY1/FY2 trainee. Examined criteria comprise medical record-keeping, clinical assessment, investigations and referrals, treatment, follow-up and future planning, professionalism and overall clinical judgement.

Yet again, discussions are case and assessor specific. However, of particular relevance to CBD are changes in trainee working practices to meet working time regulations. As a result, clinical decisions are rarely taken at junior level while discontinuity of care means that a trainee may have had limited involvement in the majority of suitable cases, increasing the hypothetical nature of the discussion.

X.1.a.v. Procedure-Based Assessment (PBA).

This assesses trainees' technical and professional skills across a range of specialty-specific procedures and is made up of 6 competency domains (consent, pre-operative planning, pre-operative preparation, exposure and closure, intra-operative technique and post-operative management) and a global summary [see Appendix: Section XII Part 6]. PBA is again a snap-shot of trainee performance assessed by a single assessor which, in sufficient numbers, is hoped to result in more objective assessment of trainees' skills. However, the procedure and timing of the assessment remains at the discretion of the trainee and although it is planned that the assessor will be the respective educational supervisor, this may not be practical on the basis of working patterns and workload.

EBSTAF examines everyday performance across multiple cases using long-loop multisource assessment that may be applied to both feedback and selection. It therefore avoids much of the case-specificity and

assessor-specificity that is inherent to the above short-loop assessment tools whilst potentially assessing the uppermost level of Miller's competence pyramid, that of everyday practice. EBSTAF should be regarded as a useful, if not essential, adjunct to current methods and its inclusion into present assessment and selection processes should be recommended.

During FY2, trainees apply to Specialty Training (ST) posts. This was initially via the NHS Medical Training Application Service (MTAS) but following emotional criticism of the on-line system, this has more recently been by MMC selection processes that vary widely according to region and specialty but employ nationally agreed person-specifications for each specialty and generic application and reference forms. Short-listed candidates then enter a region-specific series of interviews ranging from discussion of application form and curriculum vitae to clinical scenarios and even technical exercises.

The number of posts for surgery in particular is severely limited and the logistics of countrywide application processes results in application deadlines in the first half of FY2 with some trainees yet to work in their intended specialty being forced to base their application for their future career on experiences as a medical student. Furthermore, the application form has been criticised for being as much a test of creative writing as an

assessment of cognitive surgical achievement. Initial failures in the MTAS on-line applications system led to extensive and emotional criticism and along with the intensive and unrewarded clinician input essential for short-listing and interview processes, this has led to considerable scepticism of current MMC selection methodology. However, there are other issues that have not been highlighted and that may deeply impact upon the future surgical profession.

X.1.a.vi. MMC Selection is a Self-Fulfilling Prophecy.

Not one of the tools employed in the assessment or selection of trainees at foundation level have been demonstrated to have predictive validity in terms of subsequent career progression. This is an area that appears to have been simply overlooked during their development. Despite constant reference in the literature to the importance of the *demonstration* of robust assessment prior to application to selection, what is perhaps the single most important piece of supporting evidence is lacking. This is understandable since it is also the most difficult to demonstrate, but that does not remove the fact that trainees are currently being selected (and excluded) on the basis of only partially valid assessment criteria. Trainees who score poorly during the application process are excluded from subsequent surgical training at even the most junior level, despite the fact that in the past it was recognised that not all surgeons knew their true vocation until surgical training at SHO or even SpR level. Conversely,

only those trainees who are selected are offered the opportunity of a career in surgery. Irrespective of the subsequent dropout rate, MMC will be seen as a successful selection system since there is no way of determining how many trainees were incorrectly excluded. As a result the *sensitivity* and *specificity* of the process cannot be determined because the number of *false negatives* is unknown. This would be totally unacceptable for a clinical intervention and yet it has been rolled out nationally for the selection of trainees across all specialties. Many of the assessment methods have been individually applied for a number of years in their originating institutions during their development. This would therefore facilitate the rapid retrospective examination of career progression relative to previous assessment scores and this must be urgently addressed in an effort to confirm their validity and modify the process as necessary.

EBSTAF is one of very few assessment tools where the predictive validity has been examined prior to its application for high-stakes assessment. However, the low numbers available to this study mean that although the findings are promising, the predictive validity of EBSTAF is far from clear-cut. A wider application of EBSTAF is required to further evaluate this issue and to determine whether EBSTAF offers a more predictive multisource assessment of everyday trainee performance than current methods. Furthermore, changes in career structure mean that the clinical level of previous SHOs equate to the first or second years of specialist

training (ST1/ST2), at which such a study should be aimed since a further selection process will soon be put in place between ST2 and ST3 training grades. Such an investigation would require national coordination and funding, however, and this might best be undertaken by existing coordinating bodies, such as the ISCP themselves, in parallel with existing systems to allow direct comparison between the two.

X.1.b. Feedback of Trainee Performance in Surgery.

The effects of detailed structured feedback on trainees' subsequent performance, and ultimately their career progression, continues to be ignored in the literature despite its' potential to optimise skills acquisition, improve self-assessment skills and thus promote reflective practice. This is an area that must be addressed.

X.1.c. The Rise of Non-Technical Skills in Surgery.

As a direct result of observations made during the HPS Critical Care Course described within this thesis, the surgery-specific taxonomy of non-technical skills, NOTSS, has been developed (Yule *et al*, 2006b) and applied to non-technical skills training (Yule *et al*, 2007). Subsequently, the system's reliability has been demonstrated in the assessment of consultant surgeon non-technical performance during simulated operative scenarios (Yule *et al*, 2008). It should now be examined for reliability and

validity during the assessment of both consultant and trainee behaviours in the real operating environment. It should then be applied to trainees prior to and following non-technical skills courses to demonstrate that such skills can be improved by training.

In the meantime, it is essential that the development of assessment and training in non-technical skills does not become not marred by literature that applies inadequately developed tools. The practice of applying non-technical skills taxonomies developed from other industries must be abandoned following the demonstration of significant differences between medical specialties as closely aligned as surgery and anaesthetics (Fletcher *et al*, 2003; Yule *et al*, 2006b). Instead, the NOTSS system should be adopted and applied more widely to confirm its validity.

X.1.d. The Role of EBSTAF in Video Assessment of Trainees' Tissue-Handling Skills.

Video assessment of operative skills appears to have changed little from the time of this study and although EBSTAF demonstrated validity in the video assessment of trainees' tissue-handling skills, it was never designed for this purpose. Although it is very encouraging that the operative criteria examined within the Technical Skills domain of EBSTAF were so sensitive in their discrimination of trainee levels, the use of EBSTAF in this context adds little to the wealth of literature to support the

use of other global ratings scales, such as the Objective Structured Assessment of Technical Skill (OSATS) tool (Martin *et al*, 1997), that were specifically designed for this purpose. However, the literature continues to ignore the need to discriminate between very junior levels of surgeons using the assessment of generic rather than procedural operative skills for the purposes of selection for MMC at both FY2/ST1 and ST2/ST3 levels. Even the most recent publications, which employ video-based motion-tracking for the assessment of surgical skill, continue to compare novice and expert operators (Aggarwal *et al*, 2007; Aggarwal *et al*, 2008). Although it is clearly essential to highlight poorly performing surgeons on the basis of results and operative skills, the selection of the right trainees in the first instance is paramount. This study shows that such discrimination is possible using the correct methods and this is an area that must be urgently examined further using well-validated assessment instruments.

Possibly the most promising finding from the video-assessment study was the strength of correlation between trainee and trainer assessments. This would suggest that video-review in combination with structured assessment tools may improve trainee self-assessment skills and therefore provide better insight into their own performance and promote reflective practice. The only robust means of determining whether this is a true effect would be by randomisation of trainee groups to receive video-based feedback of operative skills and independently assessing their

operative skill whilst blinded to whether such feedback had been provided. This is certainly a study that should be completed in the future although the large numbers required would demand multi-centre involvement and considerable ethical considerations in order to avoid actual or perceived disadvantage to either trainee group.

X.2. RECOMMENDATIONS FOR FUTURE WORK

- EBSTAF has been shown to be a robust long-loop assessment tool that should be considered for inclusion within future trainee assessment and selection processes as it provides unique insight into trainees' everyday practice. This should be in combination with other reliable and validated methods, recognising that no one single assessment method can provide sufficient information for high-stakes assessment. However, assessment methods employed for the purposes of selection must first have demonstrated predictive validity to optimise the selection of the right trainees. This is an aspect of current selection processes that is lacking and must be urgently addressed.

- Multi-assessor video assessment of basic tissue-handling skills using validated tools, such as EBSTAF or Toronto, should incorporated into existing specialty selection and ongoing assessment processes. In particular, this must be applied at the most junior level in order to select the right trainees for the specialty, an area that has been previously neglected but must now be addressed.

- The value of video-feedback of trainee operative performance must be further examined as it offers the potential to improve trainee self-assessment skills and reflective operative practice whilst optimising

skills acquisition. A randomised and multi-centre trial to examine the effect providing video-based feedback of performance to trainees is required.

- High-fidelity human patient simulation of critical care scenarios should be considered a safe and valid adjunct to existing training methods and should be incorporated into surgical training. Its potential contribution to courses such as the Royal College of Surgeons of England CCrISP course should be recognised and if possible realised by the application of HPS to the courses themselves and subsequent assessment of critical care skills during surgical training.

- Non-technical skills should be applied, taught and assessed both in specific courses and in the workplace in order to improve both surgical practice and patient safety. They should also be incorporated into existing courses, such as CCrISP, so that the benefits of good non-technical skills become more widely recognised while specific courses addressing non-technical skills should also be developed. However, the incorporation of non-technical skills assessments into selection processes should be approached with caution.

- Further work to follow the BSTs in these studies should be carried out to confirm whether there is a relationship between poor assessment scores at the beginning of basic surgical training and the failure to attain a consultant post.

Section XI.

REFERENCES.

1. Abrahamson S, Denson JS, Wolf RM. 1969. Effectiveness of a simulator in training anesthesiology residents. J Med Educ 44(6):515-9.

2. Aggarwal R, Grantcharov TP, Moorthy K. 2007. An evaluation of the feasibility, validity and reliability of laparoscopic skills assessment in the operating room. Ann Surg 245:992-9.

3. Aggarwal R, Grantcharov TP, Moorthy K, Milland T, Darzi A. 2008. Toward feasibile, valid and reliable video-based assessmernts of technical skills in the operating room. Ann Surg 247(2):372-9.

4. Albo D, Taylor CW, Page B, Chang FC, Moody FG. 1976. Multifactor evaluations of surgical trainees and teaching services. Surgery 80:115-21.

5. Allen PI. 1990. Anastomosis: a craft workshop for surgical trainees. Aust N Z J Surg 60:943-5.

6. Altman DG. 1991. Practical Statistics for Medical Research. London: Chapman & Hall.

7. Anastakis DJ, Regehr G, Reznick RK, Cusimano M, Murnaghan J, Brown M, Hutchison C. 1999. Assessment of technical skills transfer from the bench training model to the human model. Am J Surg 177(2):167-70.

8. Ansell JS, Boughton R, Cullen T. 1979. Lack of agreement between subjective ratings of instructors and objective testing of knowledge acquisition in a urological continuing medical education course. J Urol 122:721-3.

9. Anthoney TR. 1986. A discrepancy in objective and subjective measures of knowledge: do some medical students with learning problems delude themselves? Med Educ 20(1):17-22.

10. Anwar RAH, Bosk C, Greenburg AG. 1981. Resident evaluation: is it, can it, should it be objective? J Surg Res 30:27-41.

11. Apley AG. 1980. Fixation of fractures: organising a course. Ann R Coll Surg Engl 62:219-22.

12. Archer J, Norcini J, Southgate L, Heard S, Davies H. 2008. mini-PAT (Peer Assessment Tool): a valid component of a national assessment programme in the UK? Advances in Health Sciences Education 13(2):181-92.

13. Archer, J. C. and Davies, H. A. Sheffield peer review assessment tool for consultants (SPRAT): screening for poorly performing doctors. Bern, Switzerland: Association of Medical Education of Europe (AMEE); 2003.

14. Archer JC, Norcini JJ, Davies HA. 2005. Use of SPRAT for peer review of paediatricians in training. BMJ 330:1251-3.

15. Arnold L, Willoughby TL, Calkins EV. 1985. Self-evaluation in undergraduate medical education: a longitudinal perspective. J Med Educ 60:21-8.

16. Atkinson P, Pugsley L. 2005. Making sense of ethnography and medical education. Med Educ 39:228-34.

17. Avermaete J, Kruijsen E. 1998. NOTECHS. The evaluation of non-technical skills of multi-pilot aircrew in relation to the JAR-FCL requirements. Final Report NLR-CR-98443. Amsterdam, Netherlands.: National Aerospace Laboratory.;

18. Bailey RW, Imbembo AL, Zucker KA. 1991. Establishment of a laparoscopic cholecystectomy training program. American Surgeon 57(4):231-6.

19. Baker R. 1990. Development of a questionnaire to assess patients' satisfaction with consultations in general practice. Br J Gen Pract 40:487-90.

20. Baker R. 2004. Implications of Harold Shipman for general practice. Postgrad Med J 80(944):303-6.

21. Baldwin PJ, Paisley AM, Paterson-Brown S. 1999. Consultant surgeons' opinion of the skills required of basic surgical trainees. Br J Surg 86(8):1078-82.

22. Bann S, Datta V, Khan M, Darzi A. 2003a. The surgical error examination is a novel method for objective technical knowledge assessment. Am J Surg 185(6):507-11.

23. Bann S, Khan M, Datta V, Darzi A. 2005. Surgical skill is predicted by the ability to detect errors. Am J Surg 189(4):412-5.

24. Bann S, Kwok KF, Lo CY, Darzi A, Wong J. 2003b. Objective assessment of technical skills of surgical trainees in Hong Kong. Br J Surg 90(10):1294-9.

25. Barnes RW. 1987. Surgical handicraft: teaching and learning surgical skills. Am J Surg 153:422-7.

26. Barnes RW, Lang NP, Whiteside MF. 1989. Halstedian technique revisited: innovations in teaching surgical skills. Ann Surg 210:118-21.

27. Bauchner H, Vinci R. 2001. What have we learnt from the Alder Hey affair? That monitoring physicians' performance is necessary to ensure good practice. BMJ 322(7282):309-10.

28. Beard JD, Jolly BC, Southgate LJ, Newble DI, Thomas EG, Rochester J. 2005. Developing assessments of surgical skills for the GMC Performance Procedures. Ann R Coll Surg Engl 87(4):242-7.

29. Beckman HB, Markakis KM, Suchman AL, Frankel RM. 1994. The doctor-patient relationship and malpractice. Lessons from plaintiff depositions. Arch Intern Med 154(12):1365-70.

30. Beecham L. 1996. New Scottish CMO criticises training reforms. BMJ 313:947.

31. Bell JA. 1998. Royal Air Force selection procedures. Ann R Coll Surg Engl 70:270-5.

32. Ben David MF. 2003. Life beyond OSCE. Medical Teacher 25(3):239-40.

33. Bergen PC, Littlefield JH, O'Keefe GE, Rege RV, Anthony TA, Kim LT, Turnage RH. 2000. Identification of high-risk residents. J Surg Res 92:239-44.

34. Bernadin H, Dahmus S, Redmon G. 1993. Attitudes of first-line supervisors toward subordinate appraisals. Human Resource Management 32(2-3):315-24.

35. Bevan PG. 1986. Craft workshops in surgery. Br J Surg 73:1-2.

36. Beyea SC. 2004. Human patient simulation: a teaching strategy. AORN Journal 80(4):738-2.

37. Black P, William D. 1998. Assessment and Classroom Learning. Assessment in Education 5(1):7-71.

38. Bligh J. 2001. Assessment: the gap between theory and practice. Med Educ 35(4):312.

39. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. 1956. Taxonomy of Educational Objectives: the Classification of Educational Goals. New York: David McKay Company, Inc.

40. Blum RH, Raemer DB, Carroll JS, Sunder N, Felstein DM, Cooper JB. 2004. Crisis resource management training for an anaesthesia faculty: a new approach to continuing education. Medical Education 38(1):45-55.

41. Bottomley V. 1992. The National Health Service; Where are we now? Where are we going? ASIT Yearbook. 1992-3 ed. London: Stroudgate plc..

42. Boulet JR, Murray D, Kras J, Woodhouse J, McAllister J, Ziv A. 2003. Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. Anesthesiology 99(6):1270-80.

43. Bourne MC, Paterson-Brown S. 1999. Calman and the new deal--compromising doctor training and patient care. Scott Med J 44(5):147-8.

44. Boyse TD, Patterson SK, Cohan RH, Korobkin M, Fitzgerald JT, Oh MS, Gross BH, Quint DJ. 2002. Does medical school performance predict radiology resident performance? Acad Radiol 9(4):437-45.

45. Brearley S. 1994. Health care delivery systems: effects on surgical education in the United Kingdom. World J Surg 18(5):725-9.

46. Bulstrode C, Bell Y, Gray M. 1993. Senior house officers: the lost tribes. Br J Hosp Med 50(10):572-3.

47. Bulstrode C, Holsgrove G. 1996. Education for educating surgeons. BMJ 312(7027):326-7.

48. Bulstrode CJ, Hunt V. 2001. Examining in Practice. Oxford: Skills Unit (Oxford).

49. Butterfield OS, Mazzaferri EL. 1991. A new rating form for use by nurses in assessing residents' humanistic behaviour. J Gen Intern Med 6:155-61.

50. Butterfield OS, Pearson JA. 1990. Nurses in resident evaluation: a qualitative study of the participants' perpectives. Evaluation and the Health Professions 13:453-73.

51. Byrne AJ, Jones JG. 1997. Responses to simulated anaesthetic emergencies by anaesthetists with different durations of clinical experience. Br J Anaesth 78(5):553-6.

52. Calman KC, Donaldson M. 1991. The pre-registration house officer year: a critical incident study. Med Educ 25(1):51-9.

53. Campbell LM, Murray TS. 1996. Summative assessment of vocational trainees: results of a 3-year study. Br J Gen Pract 46(408):411-4.

54. Case SM, Swanson DB, Stillman PL. 1988. Evaluating diagnostic pattern recognition: the psychometric characteristics of a new item format. Proc Annu Conf Res Med Educ 27:3-8.

55. Catto GR. 2000. Interface between university and medical school: the way ahead? BMJ 320(7235):633-6.

56. Cauraugh JH, Martin M, Martin KK. 1999. Modeling surgical expertise for motor skill acquisition. Am J Surg 177(4):331-6.

57. Chaudhry A, Sutton C, Wood J, Stone R, McCloy R. 1999. Learning rate for laparoscopic surgical skills on MIST VR, a virtual reality simulator: quality of human-computer interface. Ann R Coll Surg Engl 81(4):281-6.

58. Chopra V, Gesink BJ, de Jong J, Bovill JG, Spierdijk J, Brand R. 1994. Does training on an anaesthesia simulator lead to improvement in performance? Br J Anaesth 73(3):293-7.

59. Church AH. 2000. Do higher performing managers actually receive better ratings? a validation of multirater assessment methodology. Consulting Psychology Journal: Practice and Research(52):-99.

60. Clark CI, Snooks S. 1993. Objectives of basic surgical training. Br J Hosp Med 50(8):477-9.

61. Cohen JA. 1960. A coefficient of agreement for nominal scales. Educ Psych Research 10:37-47.

62. Cohen R, Reznick RK, Taylor BR, Provan J, Rothman A. 1990. Reliability and validity of the objective structured clinical examination in assessing surgical residents. Am J Surg 160(3):302-5.

63. Coleman J, Nduka CC, Darzi A. 1994. Virtual reality and laparoscopic surgery. Br J Surg 81(12):1709-11.

64. Collins RE. 1995. Surgeons and the new deal--good deal or raw deal? Ann R Coll Surg Engl 77(6 Suppl):297-8.

65. Cooper GE, White MD, Lauber JK. 1980. Resource Management on the Flight Deck: Proceedings of a NASA/Industry Workshop (NASA CP-2455). Moffett Field, CA: NASA - Ames Research Center.

66. Cronbach LJ. 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16:297-334.

67. Crosbie SJ, Gilberstadt H. 1961. Contrast between several means of appraising physicians. J Med Educ 36:1310.

68. Crossley J, Davies H, Humphris G, Jolly B. 2002a. Generalisability: a key to unlock professional assessment. Med Educ 36(10):972-8.

69. Crossley J, Humphris G, Jolly B. 2002b. Assessing health professionals. Med Educ 36(9):800-4.

70. Cundiff GW. 1997. Analysis of the effectiveness of an endoscopy training program in improving residents' laparoscopic skills. Obstet Gynecol 90:854-9.

71. Cuschieri A. 1992. The dust has settled - let's sweep it clean: training in minimal access surgery. J R Coll Surg Edinb 37:213-4.

72. Cuschieri A, Wilson RG, Sunderland G, McIntyre IM, Youngson GG, Cash JD, Mackay N, Shields SR. 1997. Training initiative list scheme (TILS) for minimal access therapy: the MATTUS experience. J R Coll Surg Edinb 42(5):295-302.

73. Cusimano MD, Cohen R, Tucker W, Murnaghan J, Kodama R, Reznick R. 1994. A comparative analysis of the costs of administration of an OSCE (objective structured clinical examination). Acad Med 69(7):571-6.

74. Darzi A, Mackay S. 2001. Assessment of surgical competence. Qual Health Care 10 Suppl 2:II64-II69.

75. Darzi A, Smith S, Taffinder N. 1999. Assessing operative skill. Needs to become more objective. BMJ 318(7188):887-8.

76. Das M, Mpofu D, Dunn E, Lanphear JH. 1998. Self and tutor evaluations in problem-based learning tutorials: is there a relationship? Med Educ 32:411-8.

77. Dashfield AK, Lambert A, Campbell W. 2001. Correlation between psychometric test scores and learning tying of surgical reef knots. Ann R Coll Surg Engl 83:139-43.

78. Dath D, Regehr G, Birch D, Schlachta E, Poulin E, Mamazza J, Reznick R, MacRae H. 2004. Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. Surg Endosc 18(12):1800-4.

79. Datta V, Bann S, Beard J, Mandalia M, Darzi A. 2004. Comparison of bench test evaluations of surgical skill with live operating performance assessments. J Am Coll Surg 199(4):603-6.

80. Datta V, Chang A, Mackay S, Darzi A. 2002a. The relationship between motion analysis and surgical technical assessments. Am J Surg 184(1):70-3.

81. Datta V, Mackay S, Mandalia M, Darzi A. 2001. The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. J Am Coll Surg 193(5):479-85.

82. Datta V, Mandalia M, Mackay S, Darzi A. 2002b. The PreOp flexible sigmoidoscopy trainer. Validation and early evaluation of a virtual reality based system. Surg Endosc 16(10):1459-63.

83. Davidge AM, Hull AL. 1980. A system for the evaluation of medical students' clinical competence. J Med Educ 55:65-7.

84. Davies HT, Shields AV. 1999. Public trust and accountability for clinical performance: lessons from the national press reportage of the Bristol hearing. J Eval Clin Pract 5(3):335-42.

85. Dawson JL. 1998. Are basic surgical trainees being 'short changed'? Ann R Coll Surg Engl (suppl) 80:s163-s165.

86. de Cossart I, Fish D. 2005a. Assessment and its role in education for clinical practice: an overview. Cultivating a thinking surgeon: new perspectives on clinical teaching, learning and assessment. Shrewsbury: tfm Publishing; p 93-118.

87. de Cossart I, Fish D. 2005b. Professional values and the traditions of practice in surgery. Cultivating a thinking surgeon: new perspectives on clinical teaching, learning and assessment. Shrewsbury: tfm Publishing; p 19-36.

88. de Cossart I, Fish D. 2005c. The significance of clinical practice in the education of a surgeon. Cultivating a thinking surgeon: new perspectives on clinical teaching, learning and assessment. Shrewsbury: tfm Publishing; p 3-18.

89. de Leval MR, Carthey J, Wright DJ, Farewell VT, Reason JT. 2000. Human factors and cardiac surgery: a multicenter study. J Thorac Cardiovasc Surg 119(4 Pt 1):661-72.

90. Deary IJ, Graham KS, Maran AG. 1992. Relationships between surgical ability ratings and spatial abilities and personality. J R Coll Surg Edinb 37(2):74-9.

91. Denson JS, Abrahamson S. 1969. A computer-controlled patient simulator. JAMA 208(3):504-8.

92. Dent TL. 1991. Training, credentialling and granting of clinical priveleges for laparoscopic general surgery. Am J Surg 161:399-403.

93. Department of Health. 1998. A Guide to Specialist Registrar Training. London: HMSO.

94. Department of Health. 1998a. Record of In-training Assessment (RITA). A Guide to Specialist Registrar Training. p 139-54.

95. Department of Health. 1998b. Review and appeal procedures. p 155-63.

96. Department of Health. 2003. Modernising Medical Careers. The response of the four UK Health Ministers to the consultation on *Unfinished Business*: Proposals for reform of the Senior House Officer grade. London:

97. Derossis AM, Fried GM, Abrahamowicz M, Sigman HH, Barkun JS, Meakins JL. 1998. Development of a model for training and evaluation of laparoscopic skills. Am J Surg 175:482-7.

98. Devitt JH, Kurrek MM, Cohen MM, Cleave-Hogg D. 2001. The validity of performance assessments using simulation. Anesthesiology 95(1):36-42.

99. Devitt JH, Kurrek MM, Cohen MM, Fish K, Fish P, Murphy PM, Szalai JP. 1997. Testing the raters: inter-rater reliability of standardized anaesthesia simulator performance. Can J Anaesth 44(9):924-8.

100. Devitt JH, Kurrek MM, Cohen MM, Fish K, Fish P, Noel AG, Szalai JP. 1998. Testing internal consistency and construct validity during evaluation of performance in a patient simulator. Anesth Analg 86(6):1160-4.

101. Dielman TE, Hull AL, Davis WK. 1980. Psychometric properties of clinical performance ratings. Evaluation and the Health Professions(3):-103.

102. Donabedian A. 1980. Explorations in Quality Assessment and Monitoring. The Definition of Quality and Approaches to its Assessment. Ann Arbor, MI: Health Administration Press.

103. Donaldson L. 2002. Unfinished Business ; Proposals for reform of the Senior House Officer Grade. Department of Health;

104. Dosis A, Aggarwal R, Bello F, Moorthy K, Munz Y, Gillies D, Darzi A. 2005. Synchronized video and motion analysis for the assessment of procedures in the operating theater. Arch Surg 140(3):293-9.

105. Downing SM, Downing SM. 2004. Reliability: on the reproducibility of assessment data. Med Educ 38(9):1006-12.

106. Downing SM, Downing SM. 2003. Validity: on meaningful interpretation of assessment data. Med Educ 37(9):830-7.

107. Dunnington GL, DaRosa DA. 1994. Changing surgical education strategies in an environment of changing health care delivery systems. World J Surg 18(5):734-7.

108. Durning SJ, Cation LJ, Markert RJ, Pangaro LN. 2002. Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. Acad Med 77(9):900-4.

109. Ebel RL. 1951. Estimation of the reliability of ratings. Psychometrika 16:407-24.

110. Ecke U, Klimek L, Muller W, Ziegler R, Mann W. 1998. Virtual reality: preparation and execution of sinus surgery. Comput Aided Surg 3(1):45-50.

111. Edwards CC, Bailey RW. 2000. Laparoscopic hernia repair: the learning curve. Surgical Laparoscopy, Endoscopy & Percutaneous Techniques 10(3):149-53.

112. Epstein RM. 1999. Mindful practice. JAMA 282(9):833-9.

113. Erlandson EE, Calhoun JG, Barrack FM, Hull AL, Youmans LC, Davis WK, Bartlett RH. 1982. Resident selection: applicant selection criteria compared with performance. Surgery 92(2):270-5.

114. European Council. 1993. Directive 93/104/EC. Official Journal of the European Communities No L(307):18-24.

115. Eva KW, Reiter HI, Rosenfeld J, Norman GR. 2004. The ability of the multiple mini-interview to predict preclerkship performance in medical school. Acad Med 79(10 Suppl):S40-S42.

116. Evans AW, Aghabeigi B, Leeson R, O'Sullivan C, Eliahoo J. 2002. Are we really as good as we think we are? Ann R Coll Surg Engl 84(1):54-6.

117. Ewy GA, Felner JM, Juul D, Mayer JW, Sajid AW, Waugh RA. 1987. Test of a cardiology patient simulator with students in fourth-year electives. J Med Educ 62(9):738-43.

118. Faulkner H, Regehr G, Martin J, Reznick R. 1996. Validation of an objective structured assessment of technical skill for surgical residents. Acad Med 71(12):1363-5.

119. Fedor D, Bettenhausen K. 1989. The impact of purpose, participant preconceptions and rating level on the acceptance of peer evaluations. Group and Organisational Studies 14(2):182-97.

120. Fletcher C. 1999. The implications of research on gender differences in self-assessment and 360 degree appraisal. Human Resource Management 9(1):39-46.

121. Fletcher G, Flin R, McGeorge P, Glavin RJ, Maran NJ, Patey R. 2001. Final Report: Development of a Behavioural Marker System for Anaesthetists Non-Technical Skills (ANTS). University of Aberdeen Grant Report for SCPMDE project reference: RDNES/991/C.

122. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. 2003. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. Br J Anaesth 90(5):580-8.

123. Flin R, O.Connor P, Mearns K. 2002. Crew resource management: improving safety in high reliability industries. Team Performance Management 8:68-78.

124. Fonseka C. 1996. To err was fatal. BMJ 313(7072):1640-2.

125. Fowell S, Jolly B. 2000. Combining marks, scores and grades. Reviewing common practices reveals some bad habits. Med Educ 34(10):785-6.

126. Fowell SL, Southgate LJ, Bligh JG. 1999. Evaluating assessment: the missing link? Med Educ 33(4):276-81.

127. Francis NK, Hanna GB, Cresswell AB, Carter FJ, Cuschieri A. 2001a. The performance of master surgeons on standard aptitude testing. Am J Surg 182(1):30-3.

128. Francis NK, Hanna GB, Cuschieri A. 2001b. Reliability of the Advanced Dundee Endoscopic Psychomotor Tester for bimanual tasks. Arch Surg 136(1):40-3.

129. Gaba DM. 1992. Improving anesthesiologists' performance by simulating reality. Anesthesiology 76(4):491-4.

130. Gaba DM. 2004. The future vision of simulation in health care. Quality & Safety in Health Care 13 Suppl 1:i2-10.

131. Gaba DM, DeAnda A. 1988. A comprehensive anesthesia simulation environment: re-creating the operating room for research and training. Anesthesiology 69(3):387-94.

132. Gaba DM, Fish SK, Howard SK. 1994. Crisis Management in Anaesthesiology. New York: Churchill Livingstone.

133. Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R. 1998. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. Anesthesiology 89(1):8-18.

134. Galasko C, Mackay N. 1997. Unsupervised surgical training. BMJ 315:1306-7.

135. Galasko CS. 2000. Competencies required to be a competent surgeon. Ann R Coll Surg Engl 82(3 Suppl):89-90.

136. Gardiner Q, Oluwole M, Tan L, White PS. 1996. An animal model for training in endoscopic nasal and sinus surgery. J Laryngol Otol 110:425-8.

137. Gawande AA, Zinner MJ, Studdert DM, Brennan TA. 2003. Analysis of errors reported by surgeons at three teaching hospitals. Surgery 133:614-21.

138. General Medical Council. 1997. The New Doctor. The General Medical Council;

139. General Medical Council. 1998a. Good Medical Practice. 2 ed. London: General Medical Council.

140. General Medical Council. 1998b. Revalidating Doctors. London: General Medical Council.

141. General Medical Council. 2000a. Revalidating doctors: ensuring standards, securing the future. London: General Medical Council.

142. General Medical Council. 2000b. Teamworking in medicine. General Medical Council Website (www.gmc-uk.org/med_ed/teamwork): General Medical Council.

143. General Medical Council. 2001. Good Medical Practice. 2 ed. London: General Medical Council.

144. General Medical Council. 2003. A Licence to Practice & Revalidation. London: General Medical Council.

145. Gibbons RD, Baker RJ, Skinner DB. 1986. Field articulation testing: a predictor of technical skills in surgical residents. J Surg Res 41(1):53-7.

146. Giddings AEB. 2001. Human Factors in Surgery. Discussion Document. London: Association of Surgeons of Great Britain and Ireland.;

147. Giddings AEB, Mansfield A. 2000. A Handbook on Consultant Surgeons - Team Working in Surgical Practice. Edinburgh: Royal College of Surgeons of Edinburgh.

148. Gilbart MK, Cusimano MD, Regehr G. 2001. Evaluating surgical resident selection procedures. Am J Surg 181(3):221-5.

149. Gillard JH, Dent TH, Smyth-Pigott PJ, Eaton J. 2000. Recent changes in the workload and clinical experience of pre-registration house officers: experiences over four years in south-west England. Med Educ 34(5):371-3.

150. Gilligan JH, Welsh FK, Watts C, Treasure T. 1999. Square pegs in round holes: has psychometric testing a place in choosing a surgical career? A preliminary report of work in progress. Ann R Coll Surg Engl 81(2):73-9.

151. Gipps C. 1994. Beyond Testing: towards a theory of educational assessment. London: Falmer Press.

152. Good ML. 1990. Anaesthesia simulators and training devices. Anaesthesia 45(7):525-6.

153. Good ML, Gravenstein JS. 1989. Anesthesia simulators and training devices. Int Anesthesiol Clin 27(3):161-8.

154. Gordon MJ. 1991. A review of the validity and accuracy of self-assessments in health professions training. Acad Med 66:762-9.

155. Gordon MS, Ewy GA, Felner JM, Forker AD, Gessner I, McGuire C, Mayer JW, Patterson D, Sajid A, Waugh RA. 1980. Teaching bedside cardiologic examination skills using "Harvey", the cardiology patient simulator. Med Clin North Am 64(2):305-13.

156. Gough MH. 1988. How should we select surgical trainees? Aust Clin Rev 8(31):163-9.

157. Gough MH. 1993. Aptitude testing for specialization. Br J Hosp Med 50(4):161-2.

158. Gough MH, Holdsworth R, Bell JA, Keeman JA, Lagaay MB, Van de Loo PJM, Droog A. 1988. Personality assessment techniques and ability testing as aids to the selection of surgical trainees. Ann R Coll Surg Engl 70:265-79.

159. Greenburg AG, McClure DK, Penn NE. 1982. Personality traits of surgical house officers: faculty and resident views. Surgery 92(2):368-72.

160. Greenhalgh RM, Eastcott HH, Mansfield AO, Taylor DE. 1987. Aneurysm jig for anastamosis technique. Ann R Coll Surg Engl 69:199-200.

161. Griffiths S. 2006. The role of the postgraduate medical education and training board. Arch Dis Child 91(2):195-7.

162. Grossman RS, Fincher RM, Layne RD, Seelig CB, Berkowitz LR, Levine MA. 1992. Validity of the in-training examination for predicting American Board of Internal Medicine certifying examination scores. J Gen Intern Med 7(1):63-7.

163. Guadagnoli M, Holcomb W, Davis M. 2002. The efficacy of video feedback for learning the golf swing. J Sports Sci 20(8):615-22.

164. Guest CB, Regehr G, Tiberius RG. 2001. The life long challenge of expertise. Med Educ 35(1):78-81.

165. Hall W, Violato C, Lewkonia R, Lockyer J, Fidler H, Toews J, Jennett P, Donoff M, Moores D. 1999. Assessment of physician performance in Alberta: the physician achievement review. Can Med Assoc J 161:52-7.

166. Hamdorf JM, Hall JC. 2000. Acquiring surgical skills. Br J Surg 87(1):28-37.

167. Hamdorf JM, Hall JC. 2001. The development of undergraduate curricula in surgery: III. Assessment. ANZ J Surg 71(3):178-83.

168. Hamdy H, Prasad K, Williams R, Salih FA. 2003. Reliability and validity of the direct observation clinical encounter examination (DOCEE). Med Educ 37(3):205-12.

169. Handfield-Jones RS, Mann KV, Challis ME, Hobma SO, Klass DJ, McManus IC, Paget NS, Parboosingh IJ, Wade WB, Wilkinson TJ. 2002. Linking assessment to learning: a new route to quality assurance in medical practice. Medical Education 36(10):949-58.

170. Hanna GB, Drew T, Cuschieri A. 1997. Technology for psychomotor skills testing in endoscopic surgery. Seminars in Laparoscopic Surgery 4:120-4.

171. Harden RM, Gleeson FA, Harden RM, Gleeson FA. 1979. Assessment of clinical competence using an objective structured clinical examination (OSCE).[see comment]. Medical Education 13(1):41-54.

172. Hardy KJ, Demos LL, McNeil JJ. 1998. Undergraduate surgical examinations: an appraisal of the clinical orals. Med Educ 32(6):582-9.

173. Hargreaves DH. 1996. A training culture in surgery. BMJ 313(7072):1635-9.

174. Harris CJ, Herbert M, Steele RJ. 1994. Psychomotor skills of surgical trainees compared with those of different medical specialists. Br J Surg 81(3):382-3.

175. Hasan A, Pozzi M, Hamilton JR. 2000. New surgical procedures: can we minimise the learning curve?. BMJ 320(7228):171-3.

176. Hays RB, Jolly BC, Caldon LJ, McCrorie P, McAvoy PA, McManus IC, Rethans JJ. 2002. Is insight important? measuring capacity to change performance. Med Educ 36(10):965-71.

177. Healey AN, Undre S, Vincent CA. 2004. Developing observational measures of performance in surgical teams. Quality & Safety in Health Care 13 Suppl 1:i33-i40.

178. Helmreich RL, Merritt AC, Wilhelm JA. 1999. The evolution of Crew Resource Management training in commercial aviation. Int J Aviat Psychol 9(1):19-32.

179. Hilton JR, Shiralka SP, Samsudin A, Wheeler JMB, Saad R, Galland RB, Lewis MH. 2002. Disruption of the on-call surgical team. Ann R Coll Surg Engl (suppl) 84:50-3.

180. Hodges B. 2003. Validity and the OSCE. Medical Teacher 25(3):250-4.

181. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. 1999. OSCE checklists do not capture increasing levels of expertise. Acad Med 74(10):1129-34.

182. Holzman RS, Cooper JB, Gaba DM, Philip JH, Small SD, Feinstein D. 1995. Anesthesia crisis resource management: real-life simulation training in operating room crises. J Clin Anesth 7(8):675-87.

183. Howard SK, Gaba DM, Fish KJ, Yang G, Sarnquist FH. 1992. Anesthesia crisis resource management training: teaching anesthesiologists to handle critical incidents. Aviat Space Environ Med 63(9):763-70.

184. Howie J, Heaney D, Maxwell M, Walker J. 1998. A comparison of a patient enablement instrument against two established scales as an outcome measure of primary care consultations. Fam Pract 15:165-71.

185. Howie J, Heaney D, Maxwell M, Walker J, Freeman G, Rai H. 1999. Quality at general practice consultations: cross sectional survey. BMJ 319:738-43.

186. Hurley PA, Paterson-Brown S. 1999. Senior House Officer training: some myths exposed. J R Coll Surg Edinb 44(5):324-7.

187. Intercollegiate Surgical Curriculum Project. 2005. Syllabus. http://www.iscp.ac.uk/Syllabus/Structure.aspx#

188. Intercollegiate Surgical Curriculum Project. 2006. New Surgical Curriculum. http://www.rcseng.ac.uk/curriculum/index.html

189. Isbister WH. 2002. Some thoughts on 'operative' training in general surgery. ANZ J Surg 72(4):307-8.

190. Issenberg SB, Gordon MS, Gordon DL. 2001. Simulation and new learning technologies. Medical Teacher 23:16-23.

191. Jackson BT. 1998. An Introduction to the history of surgery. In: Burnand KG, Young AE, editors. The New Aird's Companion in Surgical Studies. Second ed. Edinburgh: Churchill Livingstone; p 1-22.

192. Janelle CM, Barba DA, Frehlich SG, Tennant LK, Cauraugh JH. 1997. Maximizing performance feedback effectiveness through videotape replay and a self-controlled learning environment. Res Q Exerc Sport 68(4):269-79.

193. Jansen JJ, Tan LH, van der Vleuten CP, van Luijk SJ, Rethans JJ, Grol RP. 1995. Assessment of competence in technical clinical skills of general practitioners. Med Educ 29(3):247-53.

194. Johnson D, Cujec B. 1998. Comparison of self, nurse and physician assessment of residents rotating through an intensive care unit. Crit Care Med 26:1811-6.

195. Jolly B, Grant J. 1997. The Good Assessment Guide: A practical guide to assessment and appraisal for Higher Specialist Training. 1 ed. London: Joint Centre for Education in Medicine.

196. Jolly B. 2001. Square pegs in round holes. Med Educ 35(6):522-3.

197. Jolly BC, Joes A, Dacre JA, Elzubeir M, Kopelman P. 1996. Relationships between students' clinical experiences in introductory clinical courses and their performances on an objective structured clinical examination (OSCE). Acad Med 71:909-16.

198. Jones JS, Hunt SJ, Carlson SA, Seamon JP. 1997. Assessing bedside cardiologic examination skills using "Harvey," a cardiology patient simulator. Acad Emerg Med 4(10):980-5.

199. Jones PF. 1993. Surgical training. Training is inconsistent. BMJ 307(6913):1210.

200. Jonsen AR. 1990. The New Medicine and The Old Ethics. Cambridge, Massetusets: Harvard University Press.

201. Kassebaum DG, Eaglen RH. 1999. Shortcomings in the evaluation of students' clinical skills and behaviors in medical school. Acad Med 74(7):842-9.

202. Kassirer JP, Gorry GA. 1978. Clinical problem solving: a behavioural analysis. Ann Intern Med 89:245-55.

203. Keck JW, Arnold L, Willoughby L, Calkins V. 1979. Efficacy of cognitive/noncognitive measures in predicting resident-physician performance. J Med Educ 54(10):759-65.

204. Kennedy I. 2001. Learning from Bristol. The report of the public inquiry into children's heart surgery at the Bristol Royal Infirmary 1984 to 1995. Norwich: The Stationery Office Limited;

205. Kernodle MW, Carlton LG. 1992. Information feedback and the learning of multiple-degree-of-freedom activities. J Motor Behaviour 24:187-96.

206. Klampfer P, Flin R, Helmreich RL, Hausler R, Sexton B, Fletcher G, Field P, Staender S, Lauche K, Dieckmann P, Amacher A. 2001. Enhancing Performance in High Risk Environments: Recommendations for the use of Behavioural Markers. Position paper from the Behavioural Markers Workshop, Swissair Training Centre, Zurich, 5-6 July 2001. Ladenburg.: Daimler-Benz Shiftung;

207. Kneebone R, ApSimon D. 2001. Surgical skills training: simulation and multimedia combined. Med Educ 35:909-15.

208. Kolb D, Fry R. 1975. Towards an applied theory of experiential learning. In: Cooper C, editor. Theories of Group Processes. London: Wiley.

209. Kron IL, Kaiser DL, Nolan SP, Rudolf LE, Muller WH, Jr., Jones RS. 1985. Can success in the surgical residency be predicted from preresidency evaluation? Ann Surg 202(6):694-5.

210. Krummel TM. 1998. Surgical simulation and virtual reality: the coming revolution. Ann Surg 228:635-7.

211. Lambert TW, Goldacre MJ, Evans J. 2000. Views of junior doctors about their work: survey of qualifiers of 1993 and 1996 from United Kingdom medical schools. Med Educ 34(5):348-54.

212. Lazar HL, DeLand EC, Tompkins RK. 1980. Clinical performance versus in-training examinations as measures of surgical competence. Surgery 87(4):357-62.

213. Lee SK, Pardo M, Gaba D, Sowb Y, Dicker R, Straus EM, Khaw L, Morabito D, Krummel TM, Knudson MM. 2003. Trauma assessment training with a patient simulator: a prospective, randomized study. J Trauma 55(4):651-7.

214. Levinson W, Roter DL, Mullooly JP, Dull VT, Frankel RM. 1997. Physician-patient communication. The relationship with malpractice claims among primary care physicians and surgeons. JAMA 277(7):553-9.

215. Lighthall GK, Barr J, Howard SK, Gellar E, Sowb Y, Bertacini E, Gaba D. 2003. Use of a fully simulated intensive care unit environment for critical event management training for internal medicine residents.[see comment]. Crit Care Med 31(10):2437-43.

216. Linn BS, Arostegui M, Zeppa R. 1975. Performance self assessment. Br J Med Educ 9:98-101.

217. Linn BS, Zeppa Z. 1984. Does surgery atract students who are more resistant to stress? Ann Surg 200:638-43.

218. Livingstone JI, Thomas JM. 1996. Surgery. Training or derailment? Lancet 348 Suppl 2:sII25.

219. Lounsbery M, Sharpe T. 1996. Plotting the effects of corrective self-analysis on elite volleyball skill performance. Applied Res Coach Athletics Ann:31-42.

220. Mabe P, West S. 1982. Validity of self-evaluation of ability: a review and meta-analysis. Applied Psychology 67:280-96.

221. MacDonald J, Williams RG, Rogers DA. 2003. Self-assessment in simulation-based surgical skills training. Am J Surg 185(4):319-22.

222. MacIntyre IM. 1996. UK Surgical Training: current problems and possible solutions. J R Coll Surg Edinb 41:209-12.

223. MacIntyre IM, Munro A. 1990. Simulation in surgical training. BMJ 300(6732):1088-9.

224. MacLaren IF. 1988. Quality control in surgical training--do we need higher examinations? J R Coll Surg Edinb 33(2):98-102.

225. Macmillan AI, Cuschieri A. 1999. Assessment of innate ability and skills for endoscopic manipulations by the Advanced Dundee Endoscopic Psychomotor Tester: predictive and concurrent validity. Am J Surg 177(3):274-7.

226. MacRae H, Regehr G, Leadbetter W, Reznick RK. 2000. A comprehensive examination for senior surgical residents. Am J Surg 179(3):190-3.

227. MacRae HM, Cohen R, Regehr G, Reznick R, Burnstein M. 1997. A new assessment tool: the patient assessment and management examination. Surgery 122(2):335-43.

228. Malik SL, Manchanda SK, Deepak KK, Sunderam KR. 1988. The attitudes of medical students to the objective structured practical examination. Med Educ 22(1):40-6.

229. Manogue M, Brown G, Foster H. 2001. Clinical assessment of dental students: values and practices of teachers in restorative dentistry. Med Educ 35(4):364-70.

230. Maran N, Glavin RJ. 2001. Medical errors. Courses on crisis avoidance and resource management are available. BMJ 322(7299):1422-3.

231. Marteau TM, Humphrey C, Matoon G, Kidd J, Lloyd M, Horder J. 1991. Factors influencing the communication skills of first-year clinical medical students. Med Educ 25:127-34.

232. Martella AT. 1995. Priorities in general surgical training. Am J Surg 169:271-2.

233. Martin D, Regehr G, Hodges B, McNaughton N. 1998. Using videotaped benchmarks to improve the self-assessment ability of family practice residents. Acad Med 73(11):1201-6.

234. Martin IG, Jolly B. 2002. Predictive validity and estimated cut score of an objective structured clinical examination (OSCE) used as an assessment of clinical skills at the end of the first clinical year. Med Educ 36(5):418-25.

235. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M. 1997. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 84(2):273-8.

236. Maxim BR, Dielman TE. 1987. Dimensionality, internal consistency and interrater reliability of clinical performance ratings. Med Educ 21:130-7.

237. McCarthy AM, Garavan TN. 2001. 360 degree feedback processes: performance improvement and employee career development. Journal of European Industrial Training 25(1):3-32.

238. McDonald P. 1998. Training for surgeons after the year 2000. J R Soc Med 91(8):401.

239. McEvoy G, Buller P. 1987. User acceptance of peer appraisals in an industrial setting. Personnel Psychology 40:785-97.

240. McKinstry B, Walker J, Blaney D, Heaney D, Begg D. 2004. Do patients and expert doctors agree on the assessment of consultation skills? A comparison of two patient consultation assessment scales with the video component of the MRCGP. Fam Pract 21(1):75-80.

241. McMahon DJ, Chen S, MacLellan DG. 1995. Formal teaching of basic surgical skills. ANZ J Surg 65(8):607-9.

242. McManus IC, Smithers E, Partridge P, Keeling A, Fleming PR. 2003. A levels and intelligence as predictors of medical careers in UK doctors: 20 year prospective study. BMJ 327(7407):139-42.

243. Medina M. 1993. The Borinquen ring: introduction of a new laparoscopic simulation surgery training instrument. J Laparoendosc Surg 3(6):593-7.

244. Medina M. 2002. The laparoscopic-ring simulation trainer. JSLS 6(1):69-75.

245. Meier AH, Rawn CL, Krummel TM. 2001. Virtual reality: surgical application--challenge for the new millennium. J Am Coll Surg 192(3):372-84.

246. Miller GE. 1990. The assessment of clinical skills / competence / performance. Acad Med 65(9 Suppl):S63-S67.

247. Moorthy K, Munz Y, Adams S, Pandey V, Darzi A. 2005. A human factors analysis of technical and team skills among surgical trainees during procedural simulations in a simulated operating theatre. Ann Surg 242(5):631-9.

248. Moorthy K, Munz Y, Adams S, Pandey V, Darzi A. 2006. Self-assessment of performance among surgical trainees during simulated procedures in a simulated operating theater. Am J Surg 192:114-8.

249. Moorthy K, Munz Y, Jiwanji M, Bann S, Chang A, Darzi A. 2004. Validity and reliability of a virtual reality upper gastrointestinal simulator and cross validation using structured assessment of individual performance with video playback. Surg Endosc 18(2):328-33.

250. Moorthy K, Smith S, Brown T, Darzi A. 2003. Evaluation of virtual reality bronchoscopy as a learning and assessment tool. Respiration 70(2):195-9.

251. Morgan PJ, Cleave-Hogg D. 2002. Comparison between medical students' experience, confidence and competence. Med Educ 36(6):534-9.

252. Morton JB, Macbeth WA. 1977. Correlations between staff, peer and self assessments of fiurth year students in surgery. Med Educ 11:167-70.

253. Muller WK, Ziegler R, Bauer A, Soldner EH. 1995. Virtual reality in surgical arthroscopic training. J Image Guid Surg 1(5):288-94.

254. Munro A, Park KG, Atkinson D, Day RP, Capperauld I. 1994a. A laparoscopic surgical simulator. J R Coll Surg Edinb 39:174-6.

255. Munro A, Park KG, Atkinson D, Day RP, Capperauld I. 1994b. Skin simulation for minor surgical procedures. J R Coll Surg Edinb 39(3):174-6.

256. Munz Y, Moorthy K, Bann S, Shah J, Ivanova S, Darzi A. 2004. Ceiling effect in technical skills of surgical residents. Am J Surg 188(3):294-300.

257. Murdoch JR, Bainbridge LC, Fisher SG, Webster MH. 1994. Can a simple test of visual-motor skill predict the performance of microsurgeons? J R Coll Surg Edinb 39(3):150-2.

258. Murray D, Boulet J, Ziv A, Woodhouse J, Kras J, McAllister J. 2002. An acute care skills evaluation for graduating medical students: a pilot study using clinical simulation. Med Educ 36(9):833-41.

259. Newble DI, Dawson B. 1994. Guidelines for assessing clinical competence. Teach Learn Med 6(3):213-20.

260. Newble DI. 1983. The critical incident technique: a new approach to the assessment of clinical competence. Med Educ 17:401-3.

261. Newble DI, Hoare J, Sheldrake PF. 1980. The selection and training of examiners for clinical examinations. Med Educ 14(5):345-9.

262. Newble DI, Jaeger K. 1983. The effect of assessments and examinations on the learning of medical students. Med Educ 17(3):165-71.

263. NHS Management Executive. 1991. Junior Doctors, the 'New Deal'. Working Arrangements for Hospital Doctors and Dentists in Training. London: NHS Management Executive.

264. NHS Management Executive. 2001. Appraisal for consultants working in the NHS: guidance. http://www.dh.gov.uk/assetRoot/04/01/46/07/04014607.pdf: accessed 4-9-2006.

265. Nichols, D. P. 1998. Choosing an Intraclass Correlation Coefficient. SPSS Keywords 67: http://www.ats.ucla.edu/stat/spss/library/whichicc.htm: accessed 28-5-2004.

266. Norcini JJ. 2001. The validity of long cases. Med Educ 35(8):720-1.

267. Norcini JJ. 2002. The death of the long case?. BMJ 324(7334):408-9.

268. Norcini JJ, Blank LL, Arnold GK, Kimball HR. 1995. The Mini CEX (Clinical Evaluation Exercise). A preliminary investigation. Ann Intern Med 1995:795-9.

269. Norman G. 2002. The long case versus objective structured clinical examinations. BMJ 324:748-9.

270. Oxford University Press. 2006. The Oxford English Dictionary. Oxford: Oxford University Press.

271. Paget NS, Newble DI, Saunders NA, Du J. 1996. Physician assessment pilot study for the Royal Australasian College of Physicians. J Contin Educ Health Prof 16:103-11.

272. Paisley AM. 2002. Prospective evaluation of assessment tools for use with surgical trainees. [dissertation]. Oxford University.

273. Paisley AM, Baldwin PJ, Paterson-Brown S. 2001a. Feasibility, reliability and validity of a new assessment form for use with basic surgical trainees. Am J Surg 182(1):24-9.

274. Paisley AM, Baldwin PJ, Paterson-Brown S. 2001b. Validity of surgical simulation for the assessment of operative skill. Br J Surg 88(11):1525-32.

275. Papp KK, Polk HC, Jr., Richardson JD. 1997. The relationship between criteria used to select residents and performance during residency. Am J Surg 173(4):326-9.

276. Pendleton D, Schofield T, Tate P. 1984. A method for giving feedback. The consultation: an approach to learning and teaching. Oxford: Oxford University Press; p 68-71.

277. Pietroni M. 1993a. Guidelines and standards in surgical training. Ann R Coll Surg Engl 75(5):305-7.

278. Pietroni M. 1993b. The assessment of competence in surgical trainees. Ann R Coll Surg Engl 75(6 Suppl):200-2.

279. Playter R, Raibert M. 1997. A virtual surgery simulator using advanced haptic feedback. Minimal Invasive Therapy and Advanced Technology 6:117-21.

280. Postgraduate Medical Education and Training Board. 2004a. PMETB: the first three years. PMETB.

281. Postgraduate Medical Education and Training Board. 2004b. Principles for an assessment system for postgraduate medical training. www.pmetb.org/pmetb/publications/principles.pdf

282. Prescott LE, Norcini JJ, McKinlay P, Rennie JS. 2002. Facing the challenges of competency-based assessment of postgraduate dental training: Longitudinal Evaluation of Performance (LEP). Med Educ 36(1):92-7.

283. Prystowsky JB, Regehr G, Rogers DA, Loan JP, Hiemenz LL, Smith KM. 1999. A virtual reality module for intravenous catheter placement. Am J Surg 177(2):171-5.

284. Rafiq A, Moore JA, Zhao X, Doarn CR, Merrell RC. 2004. Digital video capture and synchronous consultation in open surgery. Ann Surg 239(4):567-73.

285. Raibert M, Playter R, Krummel TM. 1998. The use of a virtual reality haptic device in surgical training. Acad Med 73(5):596-7.

286. Ramsey PG, Carline JD, Blank LL, Wenrich MD. 1996. Feasibility of hospital-based use of peer ratings to evaluate the performances of practicing physicians. Acad Med 71(4):364-70.

287. Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. 1993. Use of peer ratings to evaluate physician performance. JAMA 269:1655-60.

288. Razaboni RM, Ballantyne DL, Harper AD, Shaw WW. 1980. The microvascular technique of vein grafting in rats as a training and experimental model. J Microsurgery 2:148-50.

289. Reason J. 1990. Human Error. New York: Cambridge University Press.

290. Reed MWR. 1993. Evaluation of surgical training - Urgent improvement needed. Ann R Coll Surg Engl 75:198-9.

291. Regehr G, MacRae H, Reznick RK, Szalay D. 1998. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med 73(9):993-7.

292. Rethans JJ, Sturmans F, Drop R, van d, V, Hobus P. 1991. Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. BMJ 303(6814):1377-80.

293. Rethans JJ, van Leeuwen Y, Drop R, van d, V, Sturmans F. 1990. Competence and performance: two different concepts in the assessment of quality of medical care. Fam Pract 7(3):168-74.

294. Reynolds N. 1999. Calman and the new deal--compromising doctor training and patient care. Scott Med J 44(5):131.

295. Reznek M, Smith-Coggins R, Howard S, Kiran K, Harter P, Sowb Y, Gaba D, Krummel T. 2003. Emergency Medicine Crisis Resource Management (EMCRM): Pilot Study of a Simulation-based Crisis Management Course for Emergency Medicine. Acad Emerg Med 10(4):386-9.

296. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. 1997. Testing technical skill via an innovative "bench station" examination. Am J Surg 173(3):226-30.

297. Reznick RK. 1993. Teaching and testing technical skills. Am J Surg 165(3):358-61.

298. Reznick RK, Blackmore D, Cohen R, Baumber J, Rothman A, Smee S, Chalmers A, Poldre P, Birtwhistle R, Walsh P, . 1993. An objective structured clinical examination for the licentiate of the Medical Council of Canada: from research to reality. Acad Med 68(10 Suppl):S4-S6.

299. Reznick RK, Blackmore D, Dauphinee WD, Rothman AI, Smee S. 1996. Large-scale high-stakes testing with an OSCE: report from the Medical Council of Canada. Acad Med 71(1 Suppl):S19-S21.

300. Richards P. 1992a. Educational improvement of the preregistration period of general clinical training. Council of Deans of United Kingdom Medical Schools and Faculties. BMJ 304(6827):625-7.

301. Richards P. 1992b. Improving preregistration training. BMJ 304(6840):1510.

302. Ringsted C, Ostergaard D, van der Vleuten CP. 2003. Implementation of a formal in-training assessment programme in anaesthesiology and preliminary results of acceptability. Acta Anaesthesiol Scand 47(10):1196-203.

303. Risucci DA. 2002. Visual spatial perception and surgical competence. Am J Surg 184(3):291-5.

304. Risucci DA, Tortolani AJ, Ward RJ. 1989. Ratings of surgical residents by self, supervisors and peers. Surg Gynecol Obstet 169(6):519-26.

305. Ritchie I. 2001. Career advice - the role of appraisal. J R Coll Surg Edinb 46(4):213-5.

306. Ro CY, Toumpoulis IK, Ashton RCJ, Jebara T, Schulman C, Todd GJ, Deroase JJJ, McGinty JJ. 2005. The LapSim: a learning encironment for both experts and novices. Stud Health Technol Inform 111:414-7.

307. Roethlisberger FJ, Dickson WJ. 1939. Management and the Worker: An Account of a Research Program Conducted by Western Electric Company, Hawthorne Works, Chicago. Cambridge, Massachusetts: Harvard University Press.

308. Rogers DA, Elstein AS, Bordage G. 2001. Improving continuing medical education for surgical techniques: applying the lessons learned in the first decade of minimal access surgery. Ann Surg 233(2):159-66.

309. Rogers DW, Blackman J, Lanzafame RJ, Hinshaw JR. 1986. Two simple models for teaching fiberoptic choledochoscopy techniques. Surg Gynecol Obstet 162:584-6.

310. Rolfe I, McPherson J. 1995. Formative assessment: how am I doing? Lancet 345(8953):837-9.

311. Rosenbaum DA. 1992. Reaching and grasping. Human Motor Control. San Diego: Academic Press Inc.; p 197-225.

312. Rosser JC, Rosser LE, Savalgi RS. 1997. Skill acquisition and assessment for laparoscopic surgery. Arch Surg 132(2):200-4.

313. Royal College of General Practitioners. 2006. Video Assessment of Consulting Skills. http://www.rcgp.org.uk/exam/videoworkbook/doc/WBOOK06.pdf: http://rcgp.org.uk

314. Sackier JM, Berci G, Paz-Partlow M. 1991. A new training device for laparoscopic cholecystectomy. Surgical Endoscopy 5:158-9.

315. Sadler DR. 1989. Formative assessment and the design of instructional systems. Instructional Science 18:119-44.

316. Saifi J, Chang BB, Paty PSK, Kaufman J, Leather RP, Shah DM. 1990. An animal model for instructing and the study of *in situ* arterial bypass. J Vasc Surg 12:538-40.

317. Satish U, Streufert S, Marshall R, Smith JS, Powers S, Gorman P, Krummel T. 2001. Strategic management simulations is a novel way to measure resident competencies. Am J Surg 181(6):557-61.

318. Schon DA. 1987. Educating the reflective practitioner: toward a new design for teaching and learning in the professions. San Francisco: Jossey-Bass.

319. Schueneman AL, Carley JP, Baker WH. 1994. Residency evaluations. Are they worth the effort? Arch Surg 129(10):1067-73.

320. Schueneman AL, Pickleman J, Freeark RJ. 1985. Age, gender, lateral dominance, and prediction of operative skill among general surgery residents. Surgery 98(3):506-15.

321. Schueneman AL, Pickleman J, Hesslein R, Freeark RJ. 1984. Neuropsychologic predictors of operative skill among general surgery residents. Surgery 96(2):288-95.

322. Schuwirth LW, van der Vleuten CP. 2003. ABC of learning and teaching in medicine: Written assessment. BMJ 326:643-5.

323. Schwartz GF, Gonnella JS. 1973. Measurement of clinical competence in the surgical clerkship. J Med Educ 48(8):762-3.

324. Schwartz RW, Barclay JR, Harrell PL. 1994a. Defining the surgical personality. Surgery 115:62-8.

325. Schwartz RW, Donnelly MB, Sloan DA, Johnson SB, Strodel WE. 1994b. Assessing senior residents' knowledge and performance: an integrated evaluation program. Surgery 116(4):634-7.

326. Schwid HA, Rooke GA, Carline J, Steadman RH, Murray WB, Olympio M, Tarver S, Steckner K, Wetstone S, Anesthesia Simulator Research Consortium. 2002. Evaluation of anesthesia residents using mannequin-based simulation: a multiinstitutional study. Anesthesiology 97(6):1434-44.

327. Scott DJ, Bergen PC, Euhus DM. 1999. Intensive laproscopic skills training improves operative performance of surgical residents. Am Coll Surg Surg Forum 50:670-1.

328. Scott DJ, Rege RV, Bergen PC, Guo WA, Laycock R, Tesfay ST, Valentine RJ, Jones DB. 2000. Measuring operative performance after laparoscopic skills training: edited videotape versus direct observation. J Laparoendosc Adv Surg Tech A 10(4):183-90.

329. Sedlack RE, Kolars JC. 2002. Colonoscopy curriculum development and performance-based assessment criteria on a computer-based endoscopy simulator. Acad Med 77(7):750-1.

330. Seki S. 1987. Accuracy of suture placement. Br J Surg 74(3):195-7.

331. Seki S. 1989. Suturing techniques of individual surgeons--differences in accuracy and mechanics. Jpn J Surg 19(4):425-31.

332. Seymour NE, Gallagher AG, Roman SA, O'Brien MK, Bansal VK, Andersen DK, Satava RM. 2002. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 236(4):458-63.

333. Shah J, Buckley D, Frisby J, Darzi A. 2003. Reaction time does not predict surgical skill. Br J Surg 90(10):1285-6.

334. Shah J, Darzi A. 2002. Virtual reality flexible cystoscopy: a validation study. BJU Int 90(9):828-32.

335. Small SD, Wuerz RC, Simon R, Shapiro N, Conn A, Setnik G. 1999. Demonstration of high-fidelity simulation team training for emergency medicine. Acad Emerg Med 6(4):312-23.

336. Smith S, Wan A, Taffinder N, Read S, Emery R, Darzi A. 1999a. Early experience and validation work with Procedicus VA--the Prosolvia virtual reality shoulder arthroscopy trainer. Stud Health Technol Inform 62:337-43.

337. Smith SG, Torkington J, Darzi A. 1999b. Objective assessment of surgical dexterity using simulators. Hosp Med 60(9):672-5.

338. Spence R, Cole D, Brown A, Camishon R, Pello M. 1987. Training for professional competence in general surgery. Curr Surg 44(4):273-8.

339. Squire D, Giachino AA, Profitt AW, Heaney C. 1989. Objective comparison of manual dexterity in physicians and surgeons. Can J Surg 32(6):467-70.

340. Steele RJ, Walder C, Herbert M. 1992. Psychomotor testing and the ability to perform an anastomosis in junior surgical trainees. Br J Surg 79(10):1065-7.

341. Stewart MA. 1995. Effective physician-patient communication and health outcomes: a review. CMAJ 152(9):1423-33.

342. Stotter AT, Becket AJ, Hansen JP, Capperauld I, Dudley HA. 1986. Simulation in surgical training using freeze dried material. Br J Surg 73(1):52-4.

343. Suchman AL, Roter D, Green M, Lipkin M, Jr. 1993. Physician satisfaction with primary care office visits. Collaborative Study Group of the American Academy on Physician and Patient. Med Care 31(12):1083-92.

344. Swanson DB, Norman GR, Linn R. 1995. Performance-based assessment: lessons learnt from health professionals. Educ Res 24:5-11.

345. Szalay D, MacRae H, Regehr G, Reznick R. 2000. Using operative outcome to assess technical skill. Am J Surg 180(3):234-7.

346. Taffinder N, Smith S, Mair J, Russell RC, Darzi A. 1999a. Can a computer measure surgical precision? Reliability, validity and feasibility of the ICSAD. Surg Endosc 13(suppl 1):81.

347. Taffinder N, Smith SG, Huber J, Russell RC, Darzi A. 1999b. The effect of a second-generation 3D endoscope on the laparoscopic precision of novices and experienced surgeons. Surg Endosc 13(11):1087-92.

348. Taffinder N, Sutton C, Fishwick RJ, McManus IC, Darzi A. 1998a. Validation of virtual reality to teach and assess psychomotor skills in laparoscopic surgery: results from randomised controlled studies using the MIST VR laparoscopic simulator. Stud Health Technol Inform 50:124-30.

349. Taffinder NJ, McManus IC, Gul Y, Russell RC, Darzi A. 1998b. Effect of sleep deprivation on surgeons' dexterity on laparoscopy simulator. Lancet 352(9135):1191.

350. The Department of Health. 1993. Hospital Doctors. Training for the future; the Report of the Working Group on Specialist Medical Training. The Department of Health.

351. Thomas RE. 1992. Teaching medicine with cases: student and teacher opinion. Med Educ 26(3):200-7.

352. Thomas SD, Hathaway DK, Arheart KL. 1992. Face validity. West J Nurs Res 14(1):109-12.

353. Thomas WE. 2000. Core skills, courses and competency. Ann R Coll Surg Engl 82(1 Suppl):18-20.

354. Thomas WE, Lee PW, Sunderland GT, Day RP. 1996. A preliminary evaluation of an innovative synthetic soft tissue simulation module ('Skilltray') for use in basic surgical skills workshops. Ann R Coll Surg Engl 78(6 Suppl):268-71.

355. Thorndike EL. 1920. A Constant Error on Psychological Rating. J Appl Psych IV:25-9.

356. Tombleson P, Fox RA, Dacre JA. 2000. Defining the content for the objective structured clinical examination component of the professional and linguistic assessments board examination: development of a blueprint. Med Educ 34(7):566-72.

357. Toogood GJ, Stableforth CF, O'Brien TS. 1996. Surgical Skills of Pre-Registration House Surgeons. Ann R Coll Surg Engl 78(S):114-5.

358. Tooke J, Ashtiany S, Carter DC, Cole A, Michael J, Rashid A, Smith PC, Tomlinson S. 2007 Jul. Aspiring to Excellence: Findings and Recommendations of the Independent Inquiry into Modernising Medical Careers. London: Universities UK;

359. Treasure T. 1998. Lessons from the Bristol case. More openness--on risks and on individual surgeons' performance. BMJ 316(7146):1685-6.

360. Tweed M, Cookson J. 2001a. The face validity of a final professional clinical examination. Med Educ 35:465-73.

361. Tweed M, Miola J. 2001b. Legal vulnerability of assessment tools. Medical Teacher 23:312-4.

362. Van der Heijden BI, Nijhof AH. 2004. The value of subjectivity: problems and prospects for 360 degree appraisal systems. Int J Resource Management 15(3):493-511.

363. van der Vleuten CP, Norman GR, Graaf E. 1991. Pitfalls in the pursuit or objectivity: issues of reliability. Med Educ 25:110-8.

364. van der Vleuten CP, Swanson DB. 1990. Assessment of clinical skills with standardised patients: state of the art. Teach Learn Med 2:58-76.

365. Van Rij AM, McDonald JR, Pettigrew RA, Putterill MJ, Reddy CK, Wright JJ. 1995. Cusum as an aid to early assessment of the surgical trainee. Br J Surg 82(11):1500-3.

366. Vanchieri C. 1999. Virtual reality: will practice make perfect? J Natl Cancer Inst 91(3):207-9.

367. Vickers MD, Reeve PE. 1990. Selection methods in medicine: a case for replacement surgery? J R Soc Med 83(9):541-3.

368. Vozenilek J, Huff JS, Reznek M, Gordon JA. 2004. See one, do one, teach one: advanced technology in medical education. Acad Emerg Med 11(11):1149-54.

369. Wakeford R, Southgate L. 1992. Postgraduate medical education: Modifiying trainees' study approaches by changing the examination. Teach Learn Med 4(210):213.

370. Wanzel KR, Hamstra SJ, Anastakis DJ, Matsumoto ED, Cusimano MD. 2002a. Effect of visual-spatial ability on learning of spatially-complex surgical skills. Lancet 359(9302):230-1.

371. Wanzel KR, Hamstra SJ, Caminiti MF, Anastakis DJ, Grober ED, Reznick RK. 2003. Visual-spatial ability correlates with efficiency of hand motion and successful surgical performance. Surgery 134(5):750-7.

372. Wanzel KR, Ward M, Reznick RK. 2002b. Teaching the surgical craft: From selection to certification. Curr Probl Surg 39(6):573-659.

373. Ward M, Gruppen L, Regehr G. 2002. Measuring self-assessment: current state of the art. Adv Health Sci Educ Theory Pract 7(1):63-80.

374. Ward M, MacRae H, Schlachta C, Mamazza J, Poulin E, Reznick R, Regehr G. 2003. Resident self-assessment of operative performance. Am J Surg 185(6):521-4.

375. Wass V, Van der Vleuten C. 2004. The long case. Med Educ 38(11):1176-80.

376. Wass V, Van der Vleuten C, Shatzer J, Jones R. 2001. Assessment of clinical competence. Lancet 357(9260):945-9.

377. Weller JM, Bloch M, Young S, Maze M, Oyesola S, Wyner J, Dob D, Haire K, Durbridge J, Walker T, Newble D. 2003. Evaluation of high fidelity patient

simulator in assessment of performance of anaesthetists. Br J Anaesth 90(1):43-7.

378. Weller JM, Robinson BJ, Jolly B, Watterson LM, Joseph M, Bajenov S, Haughton AJ, Larsen PD. 2005. Psychometric characteristics of simulation-based assessment in anaesthesia and accuracy of self-assessed scores. Anaesthesia 60(3):245-50.

379. Wenrich MD, Carline JD, Giles LM, Ramsay P. 1993. Ratings of the performances of practising internists by hospital based registered nurses. Acad Med 68:680-7.

380. White C. 2002. Surgeons top the number of referrals to assessment authority. BMJ 325(7358):235.

381. Wilkinson TJ, Challis M, Hobma SO, Newble DI, Parboosingh JT, Sibbald RG, Wakeford R. 2002. The use of portfolios for assessment of the competence and performance of doctors in practice. Med Educ 36(10):918-24.

382. Williamson JA, Webb RK, Sellen A, Runciman WB, van der Walt JH. 1993. Human failure: an analysis of 2000 incident reports. Anaesth Intensive Care 21:678-83.

383. Wilson MS, Middlebrook A, Sutton C, Stone R, McCloy RF. 1997. MIST VR: a virtual reality trainer for laparoscopic surgery assesses performance. Ann R Coll Surg Engl 79(6):403-4.

384. Wimer S, Nowack K. 1998. Thirteen common mistakes using 360 degree feedback. Training & Development 52(5):69-79.

385. Winchell SW, Safar P. 1966. Teaching and testing lay and paramedical personnel in cardiopulmonary resuscitation. Anesth Analg 45:441-9.

386. Winckel CP, Reznick RK, Cohen R, Taylor B. 1994. Reliability and construct validity of a structured technical skills assessment form. Am J Surg 167(4):423-7.

387. Winfrey M, Weeks D. 1996. Effects of self-modeling on self-efficacy and balance beam performance. Percept Mot Skills 77:907-13.

388. Wingard JR, Williamson JR. 1973. Grades as predictors of physicians' career performance: an evaluative literature review. J Med Educ 48:311-20.

389. Wood L, Hassell A, Whitehouse A, Bullock A, Wall D. 2006. A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design. Medical Teacher 28(7):e185-e191.

390. Wood L, O'Donnell E. 2000. Assessment of competence and performance at interview. BMJ (Classified):2-3.

391. Woods JR, Ansbacher R, Castro RJ, Marshall W, Trabal JF. 1980. Animal surgery: an adjunct to training in obstetrics and gynaecology. Obstetrics & Gynecology 56:373-6.

392. Woolliscroft JO, Howell JD, Patel BP, Swanson DB. 1994. Resident-patient interactions: the humanistic qualities of internal medicine residients assessed by patients, attending physicians, program supervisors, and nurses. Acad Med 69(3):216-24.

393. Woolliscroft JO, TenHaken J, Smith J, Calhoun JG. 1993. Medical students' clinical self-assessments: comparisons with external measures of performance and the students' self-assessments of overall performance and effort. Acad Med 68:285-94.

394. Wragg A, Wade W, Fuller G, Cowan G, Mills P. 2003. Assessing the performance of specialist registrars. Clin Med 3(2):131-4.

395. Yuki G, Lepsinger R. 1995. How to get the most of 360 degree feedback. 32 12(45):50.

396. Yule S, Flin R, Maran N, Rowley D, Youngson G, Paterson-Brown S. 2008. Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behaviour rating system. World Journal of Surgery 32(4):548-56.

397. Yule S, Flin R, Paterson-Brown S, Maran N. 2006a. Non-technical skills for surgeons in the operating room: a review of the literature. Surgery 139(2):140-9.

398. Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. 2006b. Development of a rating system for surgeons' non-technical skills. Medical Education 40:1098-104.

399. Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D, Youngson G. 2007. Teaching surgeons about non-technical skills. Surgeon 5(2):86-9.

400. Zemke, R. and Zemke, s. 1984. 30 things we know for sure about adult learning. Innovation Abstracts VI(8): http://www.hawaii.edu/intranet/committees/FacDevCom/guidebk/teachtip/adults-3.htm: accessed 25-10-2001.

401. Ziegler R, Fischer G, Muller W, Gobel M. 1995. Virtual reality arthroscopy training simulator. Comput Biol Med 25(2):193-203.

Section XII.

APPENDIX.

Part 1: EBSTAF.

Part 2: Trainee Questionnaire to Determine Acceptability of the Contents of EBSTAF.

Part 3: Structured Feedback Document.
Trainee Feedback Evaluation Form.

Part 4: HPS Scenarios.
HPS Assessment Form (HPS-GA).

Part 5: Video Assessment Form.

Part 6: Mini-PAT (Peer Assessment Tool)
Mini-CEX (Clinical Evaluation Exercise)
Surgical DOPS (Direct Observation of Procedural Skills.
CBD (Case-Based Discussion).
General Surgical PBA (Procedure-Based Assessment) – Appendicectomy.

Part 1:

EBSTAF.

Consultant / SpR

CONFIDENTIAL

The Edinburgh Basic Surgical Trainee Assessment Form.

GUIDELINES FOR COMPLETION.

1. The form should be filled out and returned to me within **TWO WEEKS** of the end of the post in order to optimise the accuracy of the assessment.
2. The assessment takes the form of 6 domains, each skill therein being scored by shading the appropriate circle. There may be some areas where you do not feel able to directly assess the SHO; there are boxes denoted 'not observed / assessed by me' which should shaded as appropriate in such cases. Please try not to leave any lines blank.
3. The assessment **should be completed without discussion with either the SHO or the other assessors**, in direct comparison with the intercollegiate assessment form where it is encouraged that discussion should take place with the SHO. This is vital to ensure that feedback on the form may be formalised and thus objectively analysed.
4. If you do not feel able to provide a fair assessment of the trainee named below, please pass this form on to a colleague who can do so, but please ask them to fill in their details overleaf for future reference. Otherwise please contact me as soon as possible to give me time to arrange another assessment.

SHO No
Specialty
Start Date
End Date

Assessor Name
Assessor Post
Assessor Location
Assessor Hospital

DATE OF COMPLETION

ASSESSMENT SHOULD BE BASED UPON YOUR OWN OBSERVATIONS AND NOT THOSE OF COLLEAGUES.

Observed Skill	ASSESSED / OBSERVED BY ME			NOT OBSERVED / ASSESSED BY ME				COMMENTS
	Competent	More Practice	Unable	Performed	No Evidence	NoOpportunity	CannotComment	
1. Communication with Patients and Relatives :								
Establishes a rapport with patients	○	○	○	○	○	○	○	
Sensitive and empathic towards patients	○	○	○	○	○	○	○	
Explains any potential risks in treatment	○	○	○	○	○	○	○	
Able to explain management in layman's terms	○	○	○	○	○	○	○	
Able to explain diagnosis in layman's terms	○	○	○	○	○	○	○	
Able to allay anxiety	○	○	○	○	○	○	○	
Able to diffuse anger and hostility	○	○	○	○	○	○	○	
Relates management to individual patient's needs	○	○	○	○	○	○	○	
Aware of patient's social history	○	○	○	○	○	○	○	
2. Application of Knowledge :								
Knows the natural history of disease	○	○	○	○	○	○	○	
Actively seeks out further information	○	○	○	○	○	○	○	
Knows the relative merits of different management	○	○	○	○	○	○	○	
Can co-ordinate available information on a case	○	○	○	○	○	○	○	
Can present material clearly	○	○	○	○	○	○	○	
Critically evaluates published work	○	○	○	○	○	○	○	
Can teach or explain with enthusiasm	○	○	○	○	○	○	○	
Can complete research	○	○	○	○	○	○	○	
Can initiate research	○	○	○	○	○	○	○	
3. Teamwork :								
Seeks advice when beyond limits of competence	○	○	○	○	○	○	○	
Can be trusted to carry out instructions	○	○	○	○	○	○	○	
Able to communicate clearly with other staff members	○	○	○	○	○	○	○	
Accepts feedback on own performance	○	○	○	○	○	○	○	
Can keep to time	○	○	○	○	○	○	○	
Keeps GP informed	○	○	○	○	○	○	○	
Understands other staff members' points of view	○	○	○	○	○	○	○	
Delegates when appropriate	○	○	○	○	○	○	○	
Aware of the role of other specialties	○	○	○	○	○	○	○	
Able to offer constructive criticism to others	○	○	○	○	○	○	○	
Can cope with unreasonable colleagues	○	○	○	○	○	○	○	

Observed Skill	ASSESSED / OBSERVED BY ME			NOT OBSERVED / ASSESSED BY ME				COMMENTS
	Competent	More Practice	Unable	Performed	No Evidence	NoOpportunity	CannotComment	
4. Clinical Skills :								
Can identify the acutely ill	○	○	○	○	○	○	○	
Carries out thorough clinical examination	○	○	○	○	○	○	○	
Takes full history	○	○	○	○	○	○	○	
Extracts relevant information from history &	○	○	○	○	○	○	○	
Conscientious in postoperative care	○	○	○	○	○	○	○	
Keeps accurate notes	○	○	○	○	○	○	○	
Pays attention to changes in clinical picture	○	○	○	○	○	○	○	
Listens to additional information from relatives	○	○	○	○	○	○	○	
Reviews diagnosis and management regularly	○	○	○	○	○	○	○	
Uses information in referral letter	○	○	○	○	○	○	○	
Adapts quickly if problems in management arise	○	○	○	○	○	○	○	
Knows when NOT to intervene	○	○	○	○	○	○	○	
Remains calm in an emergency	○	○	○	○	○	○	○	
Can formulate a working diagnosis & give rationale	○	○	○	○	○	○	○	
Interprets results with reference to other information	○	○	○	○	○	○	○	
Generates & ranks appropriate differential diagnosis	○	○	○	○	○	○	○	
Initiates investigations promptly	○	○	○	○	○	○	○	
Decides quickly in an emergency	○	○	○	○	○	○	○	
Knows when follow up is appropriate	○	○	○	○	○	○	○	
Knows when discharge is appropriate	○	○	○	○	○	○	○	
Can improvise where necessary	○	○	○	○	○	○	○	
Aware of cost & clinical value of investigations	○	○	○	○	○	○	○	
5. Technical Skills :								
Handles tissue gently	○	○	○	○	○	○	○	
Handles dangerous instruments safely	○	○	○	○	○	○	○	
Demonstrates sound knowledge of anatomy	○	○	○	○	○	○	○	
Competent in tying all knots	○	○	○	○	○	○	○	
Can distinguish normal from abnormal	○	○	○	○	○	○	○	
Makes incisions appropriately	○	○	○	○	○	○	○	
Can identify and expose tissue planes	○	○	○	○	○	○	○	
Demonstrates manual dexterity	○	○	○	○	○	○	○	
Able to position patient on operating table	○	○	○	○	○	○	○	
Able to control bleeding by swab, sucker & clips	○	○	○	○	○	○	○	
Able to close skin neatly	○	○	○	○	○	○	○	
Can use diathermy techniques	○	○	○	○	○	○	○	
Has good hand eye co-ordination	○	○	○	○	○	○	○	
Able to control bleeding by suturing	○	○	○	○	○	○	○	
Has 3-dimensional spatial awareness	○	○	○	○	○	○	○	
Selects correct instruments	○	○	○	○	○	○	○	
Considers the aesthetic appearance of wound	○	○	○	○	○	○	○	
Anticipates movements during assistance	○	○	○	○	○	○	○	
Is economical in movements	○	○	○	○	○	○	○	

Observed Skill	ASSESSED / OBSERVED BY ME				NOT OBSERVED / ASSESSED BY ME			
	Performed Supervised.			Assisted Only.	Evidence trainee has performed procedure alone	Evidence trainee has performed procedure supervised	Opportunity present. No evidence that trainee has performed	No opportunity
	Competent. Could Perform Alone.	Needs More Practice.	Unable to Perform Procedure					
6. Operative Experience : GENERAL / BREAST								
Incision and drainage of abscess	○	○	○	○	○	○	○	○
Excision of skin lesion	○	○	○	○	○	○	○	○
Excision of ingrowing toenail	○	○	○	○	○	○	○	○
Sigmoidoscopy	○	○	○	○	○	○	○	○
Appendicectomy	○	○	○	○	○	○	○	○
Banding/ injection of haemorrhoids	○	○	○	○	○	○	○	○
Primary varicose vein ligation and strip	○	○	○	○	○	○	○	○
Open and close abdomen	○	○	○	○	○	○	○	○
Inguinal hernia repair	○	○	○	○	○	○	○	○
Circumcision	○	○	○	○	○	○	○	○
Peritoneal lavage	○	○	○	○	○	○	○	○
Central line insertion	○	○	○	○	○	○	○	○
Breast biopsy	○	○	○	○	○	○	○	○
Haemorrhoidectomy	○	○	○	○	○	○	○	○
Upper GI endoscopy	○	○	○	○	○	○	○	○
Axillary node sampling	○	○	○	○	○	○	○	○
Intestinal anastomosis	○	○	○	○	○	○	○	○
Mastectomy	○	○	○	○	○	○	○	○
Able to use laparoscope	○	○	○	○	○	○	○	○

Please complete the following visual analogue scales by placing a vertical line at the appropriate point.

Overall Impression of Trainee.	*Poor*	*Excellent*
Working Relationship with Trainee	*Poor – difficult to get on with*	*Excellent*

Additional Comments :

(Please identify specific strengths or weaknesses that should be addressed in the next post.)

THANK YOU FOR YOUR TIME IN COMPLETING THIS FORM.

Please return completed forms to

Peter Driscoll, University Department of Surgery, Royal Infirmary of Edinburgh, Lauriston Pl., Edinburgh EH35 5NJ

A SELF-ADDRESSED ENVELOPE WAS ENCLOSED WITH THIS FORM TO AID ITS RETURN.

Part 2:

Trainee Questionnaire to Determine

Acceptability of the Contents of EBSTAF.

«SHO_No»

QUALITIES VALUED IN A CONSULTANT ...

ACCORDING TO CONSULTANTS !

The following qualities were considered to be of variable value in trainee surgeons by a panel of consultants. Please rank these qualities **in your own opinion**.

GENERAL SURGICAL SKILLS.	Essential	Important	Useful	Irrelevant
Able to position the patient on operating table				
Handles dangerous instruments safely				
Demonstrates a sound knowledge of anatomy				
Makes incisions appropriately				
Can distinguish normal from abnormal				
Can identify and expose tissue planes				
Handles tissues gently				
Demonstrates manual dexterity				
Chooses instruments correctly				
Competent in tying all knots				
Able to control bleeding by swab, sucker & clips				
Able to close skin neatly				
Can use diathermy techniques				
Has good hand eye co-ordination				
Able to control bleeding by suturing				
Has 3-dimensional spatial awareness				
Considers aesthetic appearance of wound				
Anticipates movements during assistance				
Is economical with movements				

PATIENTS AND RELATIVES	Essential	Important	Useful	Irrelevant
Establishes a rapport with patients				
Able to allay anxiety				
Able to diffuse anger and hostility				
Able to explain diagnosis in layman's terms				
Able to explain management in layman's terms				
Explains any potential risks in treatment				
Sensitive and empathic towards patients				
Aware of patient's social history				
Relates management to individual patient's needs				

APPLICATION OF KNOWLEDGE	Essential	Important	Useful	Irrelevant
Knows the natural history of disease				
Knows relative merits of different management plans				
Actively seeks out further information				
Critically evaluates published work				
Can coordinate available information on a case				
Can teach / explain with enthusiasm				
Can initiate research				
Can complete research				
Can present material clearly				

CLINICAL SKILLS	Essential	Important	Useful	Irrelevant
Takes a full history				
Carries out a thorough clinical examination				
Extracts relevant information from history & examination				
Uses information in referral letter				
Listens to additional information from patient / relatives				
Can identify the acutely ill				
Generates and ranks alternative hypotheses for diagnosis				
Can formulate a working diagnosis and give rationale				
Aware of cost and clinical value of investigations				
Initiates investigations promptly				
Interprets results with reference to other information				
Keeps accurate notes				
Pays attention to any change in clinical picture				
Adapts quickly if problems in management arise				
Conscientious in post-operative care				
Reviews diagnosis and management regularly				
Remains calm in an emergency				
Decides quickly in an emergency				
Can improvise where necessary				
Knows when NOT to intervene				
Knows when follow-up is appropriate				
Knows when discharge is appropriate				

TEAMWORK	Essential	Important	Useful	Irrelevant
Accepts feedback on own performance				
Can offer constructive criticism to others				
Seeks advice when beyond limits of competence				
Delegates to others when appropriate				
Able to communicate clearly with other staff				
Can be trusted to carry out instructions				
Can cope with unreasonable colleagues				
Understands other staff members' point of view				
Can keep to time				
Keeps G.P. informed				
Aware of role of other surgical / non-surgical specialties				

Thank you for your time in completing this questionnaire.

Please return to Peter Driscoll in the Department of Surgery.

A self-addressed envelope has been provided.

Part 3:

Structured Feedback Document.

(An anonymised copy of an actual feedback form is included by way of example).

Trainee Feedback Evaluation Form.

CONFIDENTIAL

BASIC SURGICAL TRAINEE ASSESSMENT FEEDBACK FORM.

MARCH 2001

X = Assessor Mark

X = Self-Assessment

TRAINEE NAME	:	**????**
SPECIALTY	:	**General / ITU – Post 1**

SECTION 1 PATIENTS AND RELATIVES	Observed by Assessor.			Not Observed by Assessor			
	Trainee demonstrated skill competently	Trainee demonstrated skill but more practice needed	Opportunity present but trainee unable to demonstrate skill	Evidence that trainee performed skill but not observed by me	Opportunity present but no evidence that trainee performed skill	No opportunity/ not appropriate in this unit	Unable to comment on opportunity
Establishes a rapport with patients	XXXX	XX		X			
Sensitive and empathic towards patients	XXX	XX	X	X			
Explains any potential risks in treatment	XXX	XX		X			X
Able to explain management in layman's terms	XXXX	XX					X
Able to explain diagnosis in layman's terms	XXXXX	X					X
Able to allay anxiety	XXXX	XX		X			
Able to diffuse anger and hostility	X	XX					XXXX
Relates management to individual patient's needs	XX	XX					XXX
Aware of patient's social history	XX	XX	X				XX

SECTION 2 APPLICATION OF KNOWLEDGE	Observed by Assessor.			Not Observed by Assessor			
	Trainee demonstrated skill competently	Trainee demonstrated skill but more practice needed	Opportunity present but trainee unable to demonstrate skill	Evidence that trainee performed skill but not observed by me	Opportunity present but no evidence that trainee performed skill	No opportunity/ not appropriate in this unit	Unable to comment on opportunity
Knows the natural history of disease	XXX	XX	X				X
Actively seeks out further information	XXXX	XX	X				
Knows the relative merits of different management plans	XX	XXX		X			X
Can co-ordinate available information on a case	XXX	X	X	X			X
Can present material clearly	XX	XX	X				XX
Critically evaluates published work		XX					XXXXX
Can teach or explain with enthusiasm	X	XX					XXXX
Can complete research		X	X				XXXXX
Can initiate research			X			X	XXXXX

SECTION 3 TEAMWORK **Relationship with all colleagues e.g. medical, nursing and secretarial staff**	Observed by Assessor.			Not Observed by Assessor			
	Trainee demonstrated skill competently	Trainee demonstrated skill but more practice needed	Opportunity present but trainee unable to demonstrate skill	Evidence that trainee performed skill but not observed by me	Opportunity present but no evidence that trainee performed skill	No opportunity/not appropriate in this unit	Unable to comment on opportunity
Seeks advice when beyond limits of competence	XXXX	XXX					
Can be trusted to carry out instructions	XXXXX	X					X
Able to communicate clearly with other staff members	XXXX	XXX					
Accepts feedback on own performance	XXXX	X				X	X
Can keep to time	XX	XXX	X				XX
Keeps GP informed	XX	X			X		XX
Understands other staff members' points of view	XXX	XXX			X		
Delegates when appropriate	XXX	X	X		X		X
Aware of the role of other specialties	XXXX	XX	X				
Able to offer constructive criticism to others	XX	XX			X	X	X
Can cope with unreasonable colleagues	XXX	X					XXX

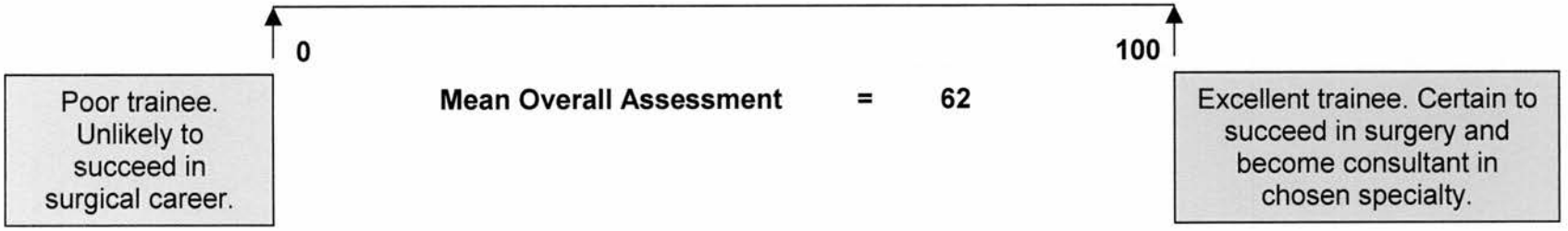
SECTION 4 CLINICAL SKILL	Observed by Assessor.			Not Observed by Assessor			
	Trainee demonstrated skill competently	Trainee demonstrated skill but more practice needed	Opportunity present but trainee unable to demonstrate skill	Evidence that trainee performed skill but not observed by me	Opportunity present but no evidence that trainee performed skill	No opportunity/ not appropriate in this unit	Unable to comment on opportunity
Can identify the acutely ill	XX	XXXX				X	
Carries out thorough clinical examination	XX	XXX		X			X
Takes full history	XX	XXX		X			X
Extracts relevant information from history & examination	XX	XXX		X			X
Conscientious in postoperative care	XXX	XXX					X
Keeps accurate notes	XXX	XX	X				
Pays attention to changes in clinical picture	XXX	XXXX					
Listens to additional information from relatives	XXX	X		X	X		X
Reviews diagnosis and management regularly	XXX	XX			X		X
Uses information in referral letter	XXX	X	X				XX
Adapts quickly if problems in management arise	XXXX	XX					X
Knows when NOT to intervene	XXX	XX		X			X
Remains calm in an emergency	X	XX		X		X	XX
Can formulate a working diagnosis & give rationale	XXXX	X	X	X			
Interprets results with reference to other information	XXX	XXX					
Generates & ranks appropriate differential diagnosis	XX	XX	X				XX
Initiates investigations promptly	XXX	XXX		X			
Decides quickly in an emergency	X	XX		X		X	XX
Knows when follow up is appropriate	X	XXXX		X			X
Knows when discharge is appropriate	XX	XXX		X			X
Can improvise where necessary	XX	XX					XXX
Aware of cost & clinical value of investigations	XX	X					XXXX

SECTION 5 TECHNICAL SKILL	Observed by Assessor.			Not Observed by Assessor			
	Trainee demonstrated skill competently	Trainee demonstrated skill but more practice needed	Opportunity present but trainee unable to demonstrate skill	Evidence that trainee performed skill but not observed by me	Opportunity present but no evidence that trainee performed skill	No opportunity/not appropriate in this unit	Unable to comment on opportunity
Handles tissue gently	X	XX		X		XX	X
Handles dangerous instruments safely	XX	X		X		X	XX
Demonstrates sound knowledge of anatomy	XX	XX		X		X	X
Competent in tying all knots	XX	X		X		XX	X
Can distinguish normal from abnormal	XX	X		X		X	XX
Makes incisions appropriately	XX	X		X		XX	X
Can identify and expose tissue planes	X	XX		X		XX	X
Demonstrates manual dexterity	XX	XX		X		X	X
Able to position patient on operating table		XX	X	X		XX	X
Able to control bleeding by swab, sucker & clips	X	XX		X		XX	X
Able to close skin neatly	XX	X		X		XX	X
Can use diathermy techniques	XX	X		X		XX	X
Has good hand eye co-ordination	XX	X		X		XX	X
Able to control bleeding by suturing	X	X		X	X	XX	X
Has 3-dimensional spatial awareness	XX	X		X		X	XX
Selects correct instruments	X	XX		X		X	X
Considers the aesthetic appearance of wound	X	XX		X		XX	X
Anticipates movements during assistance	X	XX		X		XX	X
Is economical in movements	X	XX		X		X	XX

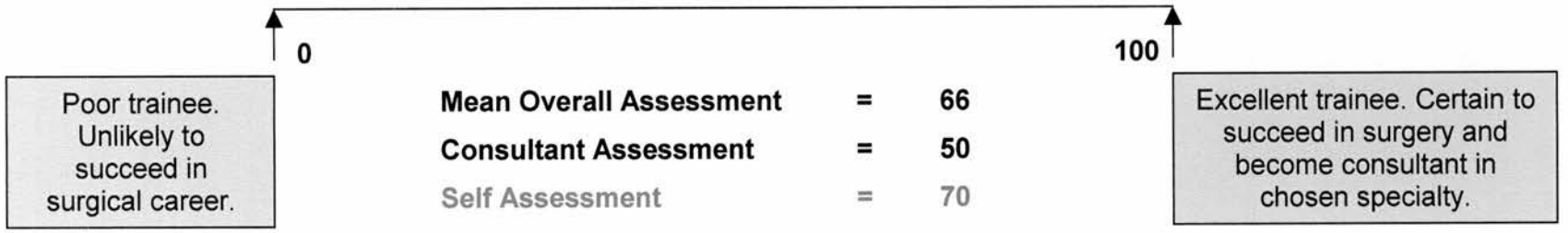
SECTION 6 OPERATIVE EXPERIENCE GENERAL SURGERY	Observed by Assessor.					Not Observed by Assessor			
	Performed unsupervised		Performed supervised		Assisted only	Evidence that trainee has assisted/performed procedure but not observed by me	Opportunity present but no evidence that trainee has assisted/performed procedure	No opportunity/ not appropriate in this unit	Unable to comment on opportunity
	Competent	More practice needed	Competent	More practice needed					
Incision and drainage of abscess	XX					XX		XX	X
Excision of skin lesion	XX		X			X		XX	X
Excision of ingrowing toenail	X			X		X		XX	XX
Sigmoidoscopy	XX	X	X			X			XX
Appendicectomy	X		X	X		X		XX	X
Banding/ injection of haemorrhoids				X		X		XXX	XXXX
Primary varicose vein ligation and stripping				XX		X		XX	XX
Open and close abdomen				XXX		X		XX	X
Inguinal hernia repair				XX		X		XX	XX
Circumcision				X		X	X	XX	XX
Peritoneal lavage				XX		X	X	XX	X
Central line insertion			X	X		XX		XX	X
Breast biopsy								XXXXX	XX
Haemorrhoidectomy								XXXXX	XX
Upper GI endoscopy				XX		X	X	XX	X
Axillary node sampling								XXXXX	XX
Intestinal anastomosis				XX		X	X	XX	X
Mastectomy								XXXXX	XX
Able to use laparoscope				XXX		X		XX	X

Sections 7 and 8 provide an overall global rating of your performance.

SECTION 7 Overall impression of trainee



SECTION 8 Working relationship with trainee



SECTION 9 Additional comments

Any specific strengths or weaknesses that should be addressed in next post.

ASSESSORS' COMMENTS :

Limited opportunity to assess but trainee appeared very competent and efficient during time observed by me

This man is not focussed

Despite direct advice to see patients and assess them prior to theatre, the very next case 2/7 later I asked him to operate upon with me, he had not seen the patient and was not particularly familiar with the anatomy. This trainee seems reluctant to take on board the most basic training advice

Can sometimes be dismissive towards staff when asked to do certain tasks. However, his manner has improved throughout his placement.

I have maintained a good professional working relationship with ????. However, some of the junior staff found him to be a little off-hand with both themselves and occasionally patients

Did not seem interested in theatre, demonstrated no enthusiasm and made comments to support this

EMERGENCY THEATRE UNABLE TO ASSESS

OWN COMMENTS :

I feel I am good with patients but a little disorganised

I have improved my operating skills steadily throughout the job but am still not economical enough with my movements

Knowledge has improved but still requires more work

Points of improvement for the next post ;

- Increase my knowledge base
- Become more dextrous
- Try to overcome my inherent disorganisation

ASSESSMENT OF THE FEEDBACK PROCESS.

Please place a vertical mark along the line for each domain at the point that corresponds to your opinion.

Did you find the formal feedback useful to your training ?

Not at all _____ *Very helpful*

Did you find the feedback given to be fair ?

Unfair _____ *Very fair*

How did you find the level of feedback ?

Not enough *Just Right* *Too much*

_____

Did you find being assessed by your colleagues threatening ?

Not at all _____ *Very threatening*

Did you find being assessed by nursing staff threatening ?

Not at all _____ *Very threatening*

PLEASE TURN OVER

Which part of the feedback did you find most useful / informative ?

Tick-box assessments

Visual-analogue scales

Comments

ALL OF THE ABOVE

NONE OF THE ABOVE

How might the assessment / feedback be improved ?

Part 4:

HPS Scenarios.

HPS Assessment Form (HPS-GA).

HPS Scenarios.

1: Accident & Emergency – Multiply Injured Patient.

Scenario:

23yo male, smelling of alcohol, found alongside a car that had left the road and hit a tree. May have been unrestrained and ejected. GCS 15 at scene but uncooperative with ambulance crew. Cardio-vascularly stable but open fracture right femoral shaft with possible arterial bleeding. Laceration to scalp and bruising to chest.

Events:

Little history from patient who continually shouts about his leg, demanding pain relief.

Ambulance crewman who is currently applying pressure to bleeding point overlying femoral fracture repeatedly draws Dr's attention to leg as he is keen to get away at the end of his shift.

Initial survey as reported but patient slowly develops pneumothorax (then tension pneumothorax if unnoticed) requiring needle thoracocentesis then formal chest drain at an appropriate time.

Patient gradually begins to drop GCS requiring anaesthetic input, CT scan and discussion with neurosurgery. Patient is accepted for transfer once stabilised.

2: Accident & Emergency – Ruptured Appendix Abscess.

Scenario.

29yo female with a 5 day flu-like illness with lower abdominal pains, nausea, vomiting and some diarrhoea. Pain suddenly worse today in lower abdomen associated with rigors and high temperatures. Unable to move with the pain. Patient is drowsy and able to answer questions. Observations, once obtained, show her to be in shock and arterial blood gases, if obtained, demonstrate severe metabolic acidosis demanding rapid fluid resuscitation and anaesthetic input with a view to theatre having first discussed the case with the on-call; consultant.

Events.

If resuscitation does not occur in a timely fashion, the patient collapses and may arrest. Issues addressed are the obtaining of frequent observations and

appropriate blood tests with rapid interpretation and decision-making. Communication with the on-call consultant was also addressed by having a consultant who was willing to come in but seemed to think the trainee was happy to carry on without him.

3: High-Dependency Unit.

Scenario.

73yo lady, 8 hours following 'routine' anterior resection for rectal carcinoma, now in the HDU. Over the last 2 hour the urine output has reduced despite fluid challenges from the houseman. The patient has an effective epidural running, is prescribed her usual beta-blockers, which she received pre-operatively, and is currently mildly hypotensive. THE PATIENT IS BLEEDING.

Events.

Operative notes, charts and drug chart should be reviewed to get a full picture of events. The epidural should be stopped while the situation is assessed, bloods should be taken for urgent full blood count and clotting and fluid resuscitation should be given despite a normal pulse rate (beta-blocker). The scenario addresses the danger of epidural and beta-blockers in a bleeding patient by masking reflex tachycardia and explaining away the hypotension. The case should be discussed with the consultant on-call immediately the haemoglobin concentration is phoned back by the lab at 4g/dl with a view to resuscitation with cross-matched blood and return to theatre. Anaesthetic colleagues should also be called and informed appropriately. IF THESE MEASURES ARE NOT PUT IN PLACE THE PATIENT CAN BE MADE TO ARREST, REQUIRING APPROPRIATE RUNNING OF CPR.

4: High-Dependency Unit.

Scenario.

64yo male with a history of angina awaiting aneurysm repair the next morning who presents with chest pain and hypotension secondary to an ischaemic event. The patient is tachycardic, hypotensive and peripherally shutdown, consistent with either hypovolaemic or cardiogenic shock.

Events.

The case requires a full history to be taken and a diagnosis of ischaemia to be made, confirmed by ECG. However, the nurse looking after the patient has been on the vascular ward for a number of years and is convinced that the patient has a leaking aneurysm and should be taken immediately to theatre. Management of this scenario therefore requires an objective evaluation of the situation and subsequent discussion with both the patient's consultant and cardiology for transfer to the coronary care unit.

STIRLING CLINICAL SIMULATOR.
Surgeons Course.
Score Sheet.

SHO No :

Assessor :

Observed Skill	Competent	Needs More Practice	Unable to Complete	Not Applicable
Takes command of the situation	○	○	○	○
Can identify the acutely ill	○	○	○	○
Able to allay anxiety	○	○	○	○
Carries out thorough clinical examination	○	○	○	○
Uses information in referral letter	○	○	○	○
Actively seeks out further information	○	○	○	○
Extracts relevant information from history & examination	○	○	○	○
Can co-ordinate available information on a case	○	○	○	○
Can formulate a working diagnosis & give rationale	○	○	○	○
Generates & ranks appropriate differential diagnosis	○	○	○	○
Knows the natural history of disease	○	○	○	○
Initiates investigations promptly	○	○	○	○
Interprets results with reference to other information	○	○	○	○
Relates management to individual patient's needs	○	○	○	○
Knows the relative merits of different management plans	○	○	○	○
Reviews diagnosis and management regularly	○	○	○	○
Pays attention to changes in clinical picture	○	○	○	○
Adapts quickly if problems in management arise	○	○	○	○
Remains calm in an emergency	○	○	○	○
Decides quickly in an emergency	○	○	○	○
Seeks advice when beyond limits of competence	○	○	○	○
Knows when NOT to intervene	○	○	○	○
Delegates when appropriate	○	○	○	○
Aware of the role of other specialties	○	○	○	○
Can present material clearly	○	○	○	○
Can be trusted to carry out instructions	○	○	○	○
Able to communicate clearly with other staff members	○	○	○	○
Understands other staff members' points of view	○	○	○	○
Can improvise where necessary	○	○	○	○
Identified problems appropriately	○	○	○	○
Managed problems appropriately	○	○	○	○
Accepts feedback on own performance	○	○	○	○

Overall Score in COMMUNICATION — *Poor* _____ *Excellent*

Overall Score in CLINICAL SKILLS & MANAGEMENT — *Poor* _____ *Excellent*

PLEASE MAKE ANY ADDITIONAL COMMENTS OVERLEAF.

Part 5.

Video Assessment Form.

VIDEO-ASSESSMENT
of
OPERATIVE SKILL LEVEL.

Dear Assessor,

Thank you participating in this study into the use of video as a means of assessment of Basic Surgical Trainees.

PROCEDURES
Each surgeon (SHO to Consultant) has been videotaped whilst performing a standard Lichtenstein open hernia repair. Because a trainee who was performing the procedure for the first time may not have been able to proceed from start to finish, the videos have then been edited to show the incision and dissection to the level of the external oblique aponeurosis and subsequent closure from this level back to skin. Thus we are not asking you to assess how good the trainee is at a Lichtenstein repair, but rather to determine their level of skill in general tissue and instrument handling.

SCORING
Please take a moment to look at the score sheets. You will see that we are employing two forms of structured scoring system. First, the technical skills section of the Edinburgh Basic Surgical Assessment Form has been adapted to include the fields that may be scored during such a procedure. Second, we have obtained permission to utilise the Toronto Global Rating Scale of Operative Performance. This has been previously studied with whole-procedure video-assessments of Anterior Resection and Laparoscopic Nissen Fundoplication in pigs, but has not, to date, been used on more basic generic technical skills in clinical practice.
For each procedure please give a mark for each field without leaving any gaps – if the particular skill was not demonstrated then the 'did not occur' box should be ticked, while if the video does not allow assessment of a particular field, please place a line through the field in question.
In addition we have asked you to estimate the training level of the surgeon in each procedure and to briefly describe on what basis this conclusion was drawn. We would also be interested as to whether you recognise the surgeon in question and how long it takes you to score each procedure.
Finally, we would ask you to make any comments as to your impression of the use of video to objectively assess tissue handling skills.

In total there are 24 (A-X) procedures to score, each lasting between 10 and 20 minutes. Previous work in this area has suggested that once you have seen sufficient of a particular procedure, you may fast-forward on to the next relevant part of the same procedure or the next surgeon. Scoring may be carried out in one or more sittings, and we have included some wine to help pass the time.

VIDEO-ASSESSMENT
of
OPERATIVE SKILL LEVEL.

Dear Trainee,

Thank you for participating in this study into the use of video as a means of assessment of Basic Surgical Trainees.

PROCEDURES
Each surgeon (SHO to Consultant) has been videotaped whilst performing a standard Lichtenstein open hernia repair. Because a trainee who was performing the procedure for the first time may not have been able to proceed from start to finish, the videos have then been edited to show the incision and dissection to the level of the external oblique aponeurosisosis and subsequent closure from this level back to skin. Thus we are not asking you to assess how good the trainee is at a Lichtenstein repair, but rather to determine their level of skill in general tissue and instrument handling.

SCORING
Please take a moment to look at the score sheets. You will see that we are employing two forms of structured scoring system. First, the technical skills section of the Edinburgh Basic Surgical Assessment Form has been adapted to include the fields that may be scored during such a procedure. Second, we have obtained permission to utilise the Toronto Global Rating Scale of Operative Performance. This has been previously studied with whole-procedure video-assessments of Anterior Resection and Laparoscopic Nissen Fundoplication in pigs, but has not, to date, been used on more basic generic technical skills in clinical practice.
For each procedure please give a mark for each field without leaving any gaps – if the particular skill was not demonstrated then the 'did not occur' box should be ticked, while if the video does not allow assessment of a particular field, please place a line through the field in question.
In addition we have asked you to estimate the training level of the surgeon in each procedure and to briefly describe on what basis this conclusion was drawn. We would also be interested as to whether you recognise the surgeon in question along how long it takes you to score each procedure.
Finally, we would ask that you rate all of the training modalities that you may have been exposed to during your surgical training to date.

In total there are 24 (A-X) procedures to score, each lasting between 10 and 20 minutes. Previous work in this area has suggested that once you have seen sufficient of a particular procedure, you may fast-forward on to the next relevant part of the same procedure or the next surgeon. Scoring may be carried out in one or more sittings, and we have included some wine to help pass the time.

VIDEO ASSESSMENT SCORE SHEET

TECHNICAL SKILL	COMPETENT **Could perform procedure unsupervised**	NOT COMPETENT **More practise needed.**	NOT COMPETENT **Trainee unable to perform skill.**	NOT APPLICABLE **Situation did not occur.**
Able to position patient on operating table				
Makes incisions appropriately				
Handles dangerous instruments safely				
Selects correct instruments				
Can identify and expose tissue planes				
Demonstrates sound knowledge of anatomy				
Can distinguish normal from abnormal				
Handles tissue gently				
Competent in tying appropriate knots				
Can use diathermy techniques				
Able to control bleeding by swab, sucker & clips				
Able to control bleeding by suturing				
Able to close skin neatly				
Considers the aesthetic appearance of wound				
Demonstrates manual dexterity				
Is economical in movements				
Has good hand eye co-ordination				
Has 3-dimensional spatial awareness				
Demonstrates knowledge of the procedure				

PLACE A VERTICAL LINE TO INDICATE YOUR OVERALL IMPRESSION

Overall Score Poor Excellent

Should not Perform unsupervised

Could perform unsupervised

GLOBAL RATING SCALE OF OPERATIVE PERFORMANCE
Developed by and used with permission of
University of Toronto Academic Surgical Unit.

Please circle the number corresponding to the candidate's performance.

Respect for Tissue				
1 Frequently used unnecessary force on tissue or caused damage by inappropriate instrument use.	2	3 Careful handling of tissue but occasionally caused inadvertent damage.	4	5 Consistently handles tissue appropriately with minimal damage to tissue.
Time and Motion				
1 Many unnecessary moves.	2	3 Efficient use of time / motion but some unnecessary moves.	4	5 Clear economy of movement and maximum efficiency.
Instrument Handling				
1 Repeatedly made tentative or awkward moves with instruments through inappropriate use.	2	3 Competent use of instruments but occasionally appeared stiff or awkward.	4	5 Fluid movements with instruments and no stiffness or awkwardness.
Flow of Operation				
1 Frequently stopped operating and seemed unsure of next move.	2	3 Demonstrated some forward planning and reasonable progression of procedure.	4	5 Obviously planned course of operation with effortless flow from one move to the next.
Use of Assistant				
1 Consistently placed assistant poorly or failed to use them.	2	3 Appropriate use of assistant most of the time.	4	5 Strategically used assistant to the best advantage.
OVERALL PERFORMANCE				
1 Very Poor.	2	3 Competent.	4	5 Clearly Superior
QUALITY OF FINAL PRODUCT				
1 Very Poor	2	3 Competent.	4	5 Clearly Superior

At what level of training would you estimate this surgeon to be ?

1st Post SHO

2nd Post SHO

Specialist Registrar

Consultant

Please briefly state on what basis you base this estimate? (e.g. good/bad techniques, good/bad habits)

Do you recognise the surgeon ?

YES / NO

If YES, then who do you think it is?

.....

How long did your assessment take? ..... minutes?

Part 6.

Mini-PAT (Peer Assessment Tool)

Mini-CEX (Clinical Evaluation Exercise)

Surgical DOPS (Direct Observation of Procedural Skills.

CBD (Case-Based Discussion).

General Surgical PBA (Procedure-Based Assessment).

(PBA for appendicectomy is included by way of example)

ISCP INTERCOLLEGIATE SURGICAL CURRICULUM PROGRAMME

Mini-PAT (Peer Assessment Tool)

Please complete the questions using a cross: ☒ Please use black ink and CAPITAL LETTERS

Trainee's surname:

Trainee's forename:

Trainee's GMC No. Hospital:

Trainee level: ST1☐ ST2☐ ST3☐ ST4☐ ST5☐ ST6☐ ST7☐ ST8☐ Other ____

Specialty: ☐ Cardio ☐ General ☐ Neuro ☐ O&M ☐ Otol ☐ Paed ☐ Plast ☐ T&O ☐ Urology

How do you rate this trainee in their:	Standard: The assessment should be judged against the standard expected at completion of this level of training. Levels of training are defined in the syllabus						
	Below expectations		Borderline	Meets expectations	Above expectations		U/C[1]
	1	2	3	4	5	6	
Good Clinical Care							
1. Ability to diagnose patient problems							
2. Ability to formulate appropriate management plans							
3. Awareness of own limitations							
4. Ability to respond to psychosocial aspects of illness							
5. Appropriate utilisation of resources e.g. ordering investigations							
Maintaining good medical practice							
6. Ability to manage time effectively/ prioritise							
7. Technical skills (appropriate to current practice)							
Teaching and Training, Appraising and Assessi							
8. Willingness and effectiveness when teaching/training colleagues							
Relationship with Patients							
9. Communication with patients							
10. Communication with carers and/or family							
11. Respect for patients and their right to confidentiality							
Working with colleagues							
12. Verbal communication with colleagues							
13. Written communication with colleagues							
14. Ability to recognise and value the contribution of others							
15. Accessibility/Reliability							
16. Overall, how do you rate this doctor compared to a doctor ready to complete this level of training?							

[1] Please mark this if you have not observed the behaviour and therefore feel unable to comment.

PTO:

Acknowledgements: Mini PAT is derived from SPRAT (Sheffield Peer Review Assessment Tool)

08.07

Anything especially good?	Please describe any behaviour that has raised concerns or should be a particular focus for development. Include an explanation of any rating below 'Meets expectations':

Sample for guidance only

Please record your Mini-Pat Online

Do you have any concerns about this doctor's probity or health? ☐ Yes ☐ No
If yes, please state your concerns

Environment observed
(please choose one answer only)

- ☐ Inpatients
- ☐ Outpatients
- ☐ Both In and Out-patients
- ☐ A&E/Admissions
- ☐ Intensive care
- ☐ Theatre
- ☐ Other (please specify)

Your position:

- ☐ Consultant
- ☐ SASG
- ☐ SpR/StR
- ☐ Foundation/PRHO
- ☐ Nurse
- ☐ SHO
- ☐ Allied Health Professional
- ☐ Other (please specify)

Have you read the mini-PAT guidance notes? ☐ Yes ☐ No

How long has it taken you to complete this form in minutes? ☐

Assessor satisfaction with mini-PAT: Not at all 1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐ 8☐ 9☐ Highly 10☐

Assessor's signature: ..... GMC Number ☐☐☐☐☐☐☐
(or other medical registration number as applicable)

Date: / /

Assessor's name: ☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

Assessor's institutional e-mail address: _____

Acknowledgements: Mini PAT is derived from SPRAT (Sheffield Peer Review Assessment Tool)

ISCP INTERCOLLEGIATE SURGICAL CURRICULUM PROGRAMME

Mini-Clinical Evaluation Exercise (CEX)

Please complete the questions using a cross: ☒ Please use black ink and CAPITAL LETTERS

Trainee's surname:

Trainee's forename:

GMC Number: Hospital:

Trainee level: ST1☐ ST2☐ ST3☐ ST4☐ ST5☐ ST6☐ ST7☐ ST8☐ Other ____

Specialty: ☐ Cardio ☐ General ☐ Neuro ☐ O&M ☐ Otol ☐ Paed ☐ Plast ☐ T&O ☐ Urology

Clinical setting: e.g. Outpatients

Clinical problem: e.g. Inguinal Hernia

New ☐ or Follow up ☐ Focus of clinical encounter: History ☐ Diagnosis ☐ Management ☐ Explanation ☐

Complexity of case: Low ☐ Average ☐ High ☐ Assessor's position: Consultant ☐ SASG ☐ SpR ☐ Other HCP ☐ Self ☐

Please grade the areas below using the scale 1-6:	Standard: The assessment should be judged against the standard expected at completion of this stage of training (e.g. initial stage ST1/ST2). Stages of training are defined in the syllabus. Some specialties have also indicated standards associated with each training level (e.g. ST1, ST2 etc) which can also be applied.						
	Below expectations		Borderline	Meets expectations	Above expectations		U/C[1]
	1	2	3	4	5	6	
1. History taking							
2. Physical Examination Skills							
3. Communication Skills							
4. Clinical Judgement							
5. Professionalism							
6. Organisation/Efficiency							
7. Overall Clinical Care[2]							

[1] Please mark this if you have not observed the behaviour and therefore feel unable to comment.
[2] Do not complete the overall rating unless you have observed the entire procedure.

Anything especially good?	Suggestions for development: Please include an explanation of any rating below 'Meets expectations'

Agreed action:

Trainee satisfaction with Mini-CEX: Not at all 1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐ 8☐ 9☐ Highly 10☐

Assessor satisfaction with Mini-CEX: 1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐ 8☐ 9☐ 10☐

Have you had training in the use of this assessment tool? ☐No ☐Yes: Written Training ☐Yes: Face to face ☐Yes: Web/CD Rom

Time taken for observation (in minutes): ☐

Time taken for feedback (in minutes): ☐

Assessor's signature: ..... GMC Number: Date: / /

Assessor's name:

Assessor's institutional e-mail address:

Acknowledgements: Adapted with the permission of the American Board of Internal Medicine

ISCP INTERCOLLEGIATE SURGICAL CURRICULUM PROGRAMME

Direct Observation of Procedural Skills in Surgery (Surgical DOPS)

Please complete the questions using a cross: ☒ Please use black ink and CAPITAL LETTERS

Trainee's surname:

Trainee's forename:

GMC number: Hospital:

Specialty: ☐ Cardio ☐ General ☐ Neuro ☐ O&M ☐ Otol ☐ Paed ☐ Plast ☐ T&O ☐ Urology

Trainee level: ST1 ☐ ST2 ☐ Other (please state level): _____

Name of procedure:

Number of times procedure performed by trainee:

Difficulty of procedure: Easier than usual ☐ Average difficulty ☐ More difficult than usual ☐

Please grade the areas below using the scale 1-6:	Standard: The assessment should be judged against the standard expected at completion of this stage of training (e.g. initial stage ST1/ST2). Stages of training are defined in the syllabus. Some specialties have also indicated standards associated with each training level (e.g. ST1, ST2 etc) which can also be applied.						
	Below expectations		Borderline	Meets expectations	Above expectations		U/O[1]
	1	2	3	4	5	6	
1. Describes indications, relevant anatomy, & details of procedure							
2. Obtains informed consent, after explaining procedure & comps							
3. Prepares for procedure according to an agreed protocol							
4. Administers effective analgesia or safe sedation (if no anaesthetist)							
5. Demonstrates good asepsis and safe use of instruments/sharps							
6. Performs the technical aspects in line with the guidance notes							
7. Deals with any unexpected event or seeks help when appropriate							
8. Completes required documentation (written or dictated)							
9. Issues clear post-procedure instructions to patient and/or staff							
10. Communicates with patient & staff in a professional manner							
11. Overall ability to perform whole procedure[2]							

[1] Please mark this if you have not observed this aspect and therefore feel unable to comment.
[2] Do not complete the overall rating unless you have observed the entire procedure.

Please use this space to record areas of strength or any suggestions for development.

Assessor training? No ☐ Written ☐ Web/CD ☐ Workshop ☐ Time taken for observation (mins): _____
Time taken for feedback (mins): _____

	Not at all									Highly
Trainee satisfaction with Surgical DOPS	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	10 ☐
Assessor satisfaction with Surgical DOPS	1 ☐	2 ☐	3 ☐	4 ☐	5 ☐	6 ☐	7 ☐	8 ☐	9 ☐	10 ☐

Assessor's name:

Assessor's position: Consultant ☐ SAS/G ☐ SpR ☐ Nurse ☐ Other HCP ☐ Self ☐

Assessor's signature: ..... Assessor's institutional e-mail address: _____

GMC Number _____ Date: / /

Acknowledgements: Adapted with the permission of the UK Royal Colleges of Physicians 08.07

ISCP INTERCOLLEGIATE SURGICAL CURRICULUM PROGRAMME

Case-based Discussion (CBD)

Please complete the questions using a cross: ☒ Please use black ink and CAPITAL LETTERS

Trainee's surname:

Trainee's forename:

GMC Number: Hospital:

Trainee level: ST1☐ ST2☐ ST3☐ ST4☐ ST5☐ ST6☐ ST7☐ ST8☐ Other ____

Specialty: ☐ Cardio ☐ General ☐ Neuro ☐ O&M ☐ Otol ☐ Paed ☐ Plast ☐ T&O ☐ Urology

Clinical setting: e.g. Outpatients

Clinical problem: e.g. Inguinal Hernia

Focus of clinical encounter: Medical record keeping☐ Clinical Assessment☐ Management☐ Professionalism☐

Complexity of case: Low☐ Average☐ High☐ Assessor's position: _____

CBD used for reflective practice ☐

Please grade the areas below using the scale 1-6:	Standard: The assessment should be judged against the standard expected at completion of this stage of training (e.g. initial stage ST1/ST2). Stages of training are defined in the syllabus. Some specialties have also indicated standards associated with each training level (e.g. ST1, ST2 etc) which can also be applied.						
	Below expectations		Borderline	Meets expectations	Above expectations		U/C[1]
	1	2	3	4	5	6	
1. Medical Record Keeping							
2. Clinical Assessment							
3. Investigation and Referrals							
4. Treatment							
5. Follow-up and Future Planning							
6. Professionalism							
7. **Overall Clinical Judgement**							

[1] U/C Please mark this if you have not observed the behaviour and therefore feel unable to comment.

Anything especially good?	Suggestions for development: Please include an explanation of any rating below 'Meets expectations'
Agreed action:	

Trainee satisfaction with CbD (Not at all 1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐ 8☐ 9☐ Highly 10☐)

Assessor satisfaction with CbD (1☐ 2☐ 3☐ 4☐ 5☐ 6☐ 7☐ 8☐ 9☐ 10☐)

Have you had training in the use of this assessment tool?
☐No ☐Yes: Face to face
☐Yes: Read Guidelines ☐Yes: Web/CD Rom

Time taken for observation (in minutes): ☐

Time taken for feedback (in minutes): ☐

Assessor's signature: ..... GMC Number ☐☐☐☐☐☐☐ Date: / /

Assessor's name:

Assessor's institutional e-mail address: _____

General Surgery PBA: Appendicectomy APPROVED SURGICAL TEMPLATE Jun 08 updated 10.07

Trainee:	Assessor:	Date:
Assessor's Position*:	Email (institutional):	GMC No:
Duration of procedure (mins):	Duration of assessment period (mins):	Hospital:
Operation more difficult than usual? Yes / No (If yes, state reason)		

* Assessors are normally consultants (senior trainees may be assessors depending upon their training level and the complexity of the procedure)

IMPORTANT: 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 procedure to the patient / relatives / 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		
PL2	Demonstrates ability to make reasoned choice of appropriate equipment, materials or devices (if any) taking into account appropriate investigations e.g. x-rays		
PL3	Checks materials, equipment and device requirements with operating room staff		
PL4	Ensures the operation site is marked where 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		
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. catheter, diathermy)		
PR7	Ensures appropriate drugs administered		
PR8	Arranges for and deploys specialist supporting equipment (e.g. laparoscopic stack, image intensifiers) effectively		

Competencies and Definitions		Rating N/D/S	Comments
IV.	**Exposure and closure**		
E1	Demonstrates knowledge of optimum skin incision / portal / access		
E2	Achieves an adequate exposure through purposeful dissection in correct tissue planes and identifies all structures correctly		
E3	Completes a sound wound repair where appropriate		
E4	Protects the wound with dressings, splints and drains where appropriate		
V.	**Intra operative technique: global (G) and task- specific items (T)**		
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		
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)	Performs exploration of the right iliac fossa in a logical fashion		
IT13 (T)	Mobilises appendix safely		
IT14 (T)	Achieves secure haemostasis of mesoappendix then divides this safely		
IT15 (T)	Divides the appendix safely with appendix stump secured		
IT16 (T)	Examines the omentum, terminal ileum and pelvic organs when the appendix is found to be macroscopically normal		
IT17 (T)	Manages intraperitoneal contamination at end of procedure appropriately		
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		

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 complications that arose)	