ANZJSurg.com

Experience matters: comparing novice and expert ratings of non-technical skills using the NOTSS system

Steven Yule,* David Rowley,† Rhona Flin,* Nikki Maran,‡ George Youngson,§ John Duncan¶ and Simon Paterson-Brown**

*School of Psychology, University of Aberdeen, and
§Royal Aberdeen Children's Hospital, Aberdeen, and
†Royal College of Surgeons of Edinburgh, and
Departments of ‡Anaesthesia and
**Surgery, Royal Infirmary, Edinburgh, and
¶Department of Surgery, Raigmore Hospital, Inverness, UK

Key words
behaviour rating, non-technical skill, patient safety, surgical skill, workplace assessment.

Abbreviations
ANTS, Anaesthetists' Non-Technical Skills; NOTSS, non-technical skills for surgeons.

Correspondence
Dr Steven Yule, School of Psychology, University of Aberdeen, Aberdeen AB24 2UB, Scotland, UK.
Email: s.j.yule@abdn.ac.uk

S. Yule MA, MSc, PhD; **D. Rowley** BMed Bio I, MD, FRCS; **R. Flin** BSc, PhD; **N. Maran** MBChB, FRCA; **G. Youngson** MBChB, PhD,FRCS; **J. Duncan** ChM, FRCEd, FACS; **S. Paterson-Brown** MBBS, MS, FRCSEd.

Accepted for publication 6 July 2008.

doi: 10.1111/j.1445-2197.2008.04833.x

Abstract

There is growing evidence that non-technical skills (NTS) are related to surgical outcomes and patient safety. The aim of this study was to further evaluate a behaviour rating system (NOTSS: Non-Technical Skills for Surgeons) which can be used for workplace assessment of the cognitive and social skills which are essential components of NTS. A novice group composed of consultant surgeons (n = 44) from five Scottish hospitals attended one of six experimental sessions and were trained to use the NOTSS system. They then used NOTSS to rate surgeons' behaviors in six simulated scenarios filmed in the operating room. The behaviours demonstrated in each scenario were compared to expert ratings to determine accuracy. The mode rating from the novice group (who received a short training session in behaviour assessment) was the same as the expert group in 50% of ratings. Where there was disagreement, novice raters tended to provide lower ratings than the experts. Novice raters require significant training in this emerging area of competence in order to accurately rate non-technical skills.

Introduction

The surgical profession is rapidly changing to cope with internal and external pressures such as the European Working Time Directive, which restricts the working week to 48 hours; the challenges of new professional roles such as nurse practitioners, the modernization of training and education, and new technology.[1] Technological developments and innovations have wide-reaching implications for surgery which are not necessarily matched by advances in training and systems analysis. Current professional development expects operating theatre personnel to gain a range of skills for the intraoperative management of patients that are increasingly complex, yet under-specified. The focus of surgical training still heavily favours technical skill acquisition. Yet surgeons increasingly operate using teams with which they may be unfamiliar, especially in an emergency setting and use recent technology with which the team may not yet be accustomed. This changes the nature of surgery for all operating theatre personnel and impacts on the behavioural and cognitive demands of their work.

Surgeons have acknowledged that cognitive skills contribute to skilful surgery but little attention has been paid to the cognitive

processes that underpin surgical competence.[2] Similarly, the literature on surgical decision making is restricted almost entirely to the pre-operative phase of surgery[3] with little emphasis placed on intraoperative judgement and decision making which is a critical part of surgical performance. According to many consultant surgeons,[4] teamwork is also highly relevant for a successful surgical career, and surgeons' leadership behaviours in the operating theatre have been shown to be important, especially when new technology is being adopted.[5] Despite this, cognitive and interpersonal skills are currently only trained and assessed on a rather tacit and discretionary basis and are not part of the surgical curriculum in the UK.

Considerable money is spent on ensuring that technology affords certain innovative surgical actions, but comparatively little investment is made on ensuring the whole system operates effectively to support such action safely for the patient. In such cases, favouring technical training and assessment may be suitable for highly procedural and predictable routine surgery, but does not help surgeons be flexible or adaptive if the system of surgery suddenly becomes unstable or unfamiliar.

Changes in surgical training and assessment

Changes to the configuration of surgical training and education are currently under way in the UK to attempt to streamline development of competent doctors who are skilled at communicating and working as effective members of a team. The adoption of this approach recommends that progress through and completion of surgical training be based on competence and has moved the emphasis of assessment away from set-piece examinations of knowledge towards learning and assessment of skills in the workplace. Selection of trainees into surgical specialties has also been radically altered and provides an opportunity to formalize the role of non-technical skills in surgical education and assessment.

The main methods of workplace-based assessment of trainees in the UK are observational tools, which cover skills such as ability to work in a multiprofessional team (Mini-PAT: Peer Assessment Tool) and communication (Mini-CEX: Clinical Evaluation Exercise). However, these tools are for the assessment of perioperative skills, often using interactions with patients as a basis for assessment. This is to be encouraged, but the skills assessed do not necessarily relate to those required for working with other professionals during a surgical procedure, commonly with an anaesthetized patient. The systems that are used to assess trainees' intraoperative competence such as surgical Direct Observation of Procedural Skills and Procedure Based Assessment are focused almost entirely on technical ability and do not cover non-technical skills or systems aspects of surgery.

The cognitive and social skills, which underpin clinical and technical proficiency are recognized as requirements for a competent surgeon[4] and rank highly as core competencies within organizations such as CanMeds, the General Medical Council, and the Royal Colleges of Surgery in the UK but until recently there were no tools to reliably assess these skills in the workplace.[6]

Development of NOTSS

Non-technical skills for surgeons are defined as 'behavioural aspects of performance in the operating theatre which underpin medical expertise, use of equipment and drugs'.[7] They are the cognitive and interpersonal skills which underpin clinical and technical skills and are requirements for a competent surgeon. The NOTSS system was developed and tested under funding from the Royal College of Surgeons of Edinburgh and NHS Education for Scotland, from 2003–2007. The project was run by the University of Aberdeen, with a steering group of surgeons, psychologists and an anaesthetist. The research drew on previous work in Scotland on surgical competence, professionalism, and the skills surgeons required to operate safely[4,8] and followed on from a similar project which developed a behaviour rating system for anaesthetists – the Anaesthetists' Non-Technical Skills (ANTS) system.[9] The aim of the NOTSS project was to develop and test an educational system for assessment and training based on observed skills in the intraoperative phase of surgery. The system was developed from the bottom up with subject matter experts (consultant surgeons), instead of adapting existing frameworks used in other industries. It was considered important to recognize and understand the unique aspects of non-technical skills in surgery, and not to assume that those non-technical skills identified for pilots, nuclear power controllers or anaesthetists would be exactly mirrored in, or be relevant to surgery. The NOTSS system is in surgical language for suitably trained surgeons to observe, rate and feedback on non-technical skills in a structured manner. An adapted model of systems design[10] was used to guide the design of the system. The three phases in the model relate to the three objectives set by the NOTSS steering group in 2003: to identify the relevant non-technical skills required by surgeons, to develop a system to allow surgeons to rate these skills and to test the system for reliability and usability.

In phase 1 we used task analysis to identify the relevant skills. Methods used included field notes of observation sessions in the operating theatre, analysis of surgical mortality reports, a review of published research,[7] attitude survey of operating theatre personnel[11] and cognitive interviews with subject matter experts.[12] This generated a list of 150 non-technical skills.

In phase 2, four independent groups of consultant surgeons used an iterative process to develop a skills taxonomy from these skills which formed the basis of the system. The NOTSS system follows the same hierarchical structure of categories, elements and behaviours as behaviour rating systems in other professions such as ANTS (anaesthetists) and NOTECHS,[13] which is used to assess pilots' non-technical skills by several European airlines. Exemplar good and poor behaviours were written by a further $n = 16$ consultant surgeons to complete the prototype system. The NOTSS skills taxonomy (Table 1) was revised after psychometric evaluation of the prototype

Table 1 Non-technical skills for surgeons skills taxonomy v1.2

Category	Element
Situation awareness	Gathering information
	Understanding information
	Projecting and anticipating future state
Decision making	Considering options
	Selecting and communicating option
	Implementing and reviewing decisions
Communication and teamwork	Exchanging information
	Establishing a shared understanding
	Co-ordinating team
Leadership	Setting and maintaining standards
	Supporting others
	Coping with pressure

system in phase 3, where $n = 44$ surgeons rated standardized video scenarios of surgeons' behaviours in the operating theatre.[14] The skills taxonomy comprises two cognitive skills (Situation Awareness and Decision Making) and two interpersonal skills (Communication and Teamwork and Leadership), broken down into constituent elements.

Translating NOTSS into practice

A user handbook was then written which included advice for using NOTSS, definitions and behavioural examples of the NOTSS categories and elements, and a set of rating forms for users. An initial usability trial in which the system was used to observe skills and debrief trainees after 43 operations revealed that observing and rating non-technical skills was feasible for surgeons and that the system was viewed as a positive adjunct to available methods for assessing trainees.[15] As part of the evaluation, it emerged that the training given in using the system was not sufficient for many users as they did not have background knowledge in human performance from psychology and human factors. To address this, a one-day training course was developed for surgeons focusing on awareness of human factors and non-technical skills (level 1 training) and practice in the basics of workplace assessment using NOTSS (level 2 training). This developed into a two-day course on safer operative surgery. These courses were designed for higher trainee and consultant surgeons only and were based on task analysis of surgeons' non-technical skills, the NOTSS behaviour rating system, principles of safety science, and underlying psychology.[16] Current plans are to develop these courses for a multi-disciplinary audience and embed non-technical skills training and workplace assessment of non-technical skills in the undergraduate and postgraduate education of surgeons. Before this is possible, it is important to have valid and reliable tools to structure observations of behaviour, provide feedback and make formative or summative workplace assessments. The NOTSS system has already been subject to a degree of psychometric testing, which focused on agreement of raters within groups on ratings.[14]

The present study

Previous analysis focused on the psychometric properties of the scale, especially the reliability and sensitivity of the system. This paper focuses on how the system was used focusing on absolute ratings provided by novice raters and 'reference ratings' provided by a group of experts.

Methods

Participants

Adverts were placed on the University and Royal College of Surgeons of Edinburgh websites and letters were sent to surgeons who had taken part in the development of the NOTSS taxonomy to invite them to take part in a further study to evaluate the system. Forty-four consultant surgeons from five Scottish hospitals opted in to one of six experimental sessions. Participants were from general surgery ($n = 18$, 41%), orthopaedic surgery ($n = 11$, 25%), paediatric surgery ($n = 3$, 7%), plus two urologists, one breast surgeon and one cardiothoracic surgeon. Eight participants (18%) did not disclose their speciality. Mean experience at consultant level was 8.9 years (standard deviation 7.5 years); 95% ($n = 42$) were men. This group of 44 surgeons constituted the 'novice' group.

Procedure

In other high consequence industries, it is understood that raters must receive specific training in order to assess non-technical skills.[17,18] As this basic training is currently lacking from surgeons' formal education, the NOTSS system training course was developed. Lasting two and a half hours, the course was based on guidance from aviation on behaviour rating and used precompiled surgical video examples to train raters.[17] It comprised: (i) background on human factors and non-technical skills; (ii) an introduction to the NOTSS system and how to make behavioural assessments, and (iii) practice in observing and rating behaviors with NOTSS in three training scenarios. All scenarios used in the observation session were simulations, and this was made clear to participants. Although some discussion followed each of the training scenarios, raters were not formally calibrated. After the training, participants rated the consultant surgeons' behaviors in six video scenarios. Participants were instructed to watch each scenario and to rate the observed skills using the NOTSS rating form. All ratings were made individually, using the NOTSS rating form to record scores for NOTSS categories and elements on a 4-point rating scale: 4 good, 3 acceptable, 2 marginal, 1 poor, and N/A not applicable (skill not required or expected for given clinical situation).

Design of video scenarios

For this evaluation, 6 video scenarios were selected from 11 filmed, illustrating surgeons' non-technical skills (good to poor) in a range of realistic simulated situations. The scenarios were filmed in operating theatres, using a patient simulator and practising surgeons, anaesthetists and nurses acting the main roles. The scenarios were designed by two surgeons, an anaesthesiologist experienced in non-technical skills training, and two psychologists. To minimise typecasting, five consultant surgeons played the lead roles across scenarios, which ranged in length from 2 min 30 s to 5 min 40 s in length. Three further scenarios were selected for practice during pre-experiment training. See Yule *et al.* (2008) for further details of the specific content of the scenarios and Figure 1 for a sample of stills showing the high fidelity of these simulated scenarios.[14]

Reference ratings

A set of 'reference ratings' were collected for comparison with the participants' ratings. The reference ratings were provided by the scenario designers who were also practising surgical team members with up to 10 years expertise in behaviour rating and assessment of technical and non-technical skills. In comparison with their peers, they were some of the most experienced clinicians available to provide 'expert' opinion. They provided a judgement of the level of each non-technical skill shown in each scenario, expressed as an agreed set of category and element ratings for each scenario.

Results

The analysis focused on a comparison of participants' ratings against reference ratings. Figure 2 shows a comparison of the mode rating from participants with the reference rating provided by the experts on each NOTSS skill category for each scenario. The mode rating was selected because for this analysis we were interested in comparing the raters' most commonly used absolute rating with the reference rating. The results show that the mode score is generally close to reference ratings but as a group, the majority of novice raters do not always agree with the experts. Expert ratings across the six scenarios ranged from ceiling to floor. This validates the scenarios as showing the full range of non-technical skills from good to poor across scenarios.

Of the 24 category rating opportunities in this study (4 categories × 6 scenarios), in 12 cases (50%) the rater mode was not the same as the reference rating. Of those 12 cases, on 5 occasions (42.5%) the experts rated higher (to indicate better observed non-technical performance) than the novices, and on 7 occasions (57.5%) novices rated higher. There was only one occasion where the difference between the mode rating provided by participants was more than one scale point different from the reference rating: for the leadership category in scenario two, the mode rating was 4 and the reference rating 2.

The mode score is useful for comparing the general view of participants with the reference ratings, but it does not tell us how many raters agreed with reference ratings. With four rating points and a 'not applicable' option on the NOTSS scale, there is the potential for much disagreement. In relation to how the NOTSS categories were used across scenarios, Fig. 3 presents a breakdown of the percentage of raters who agreed with the reference rating, the percentage of raters who rated lower and the percentage of raters who rated higher than the reference rating for each category. The percentage of raters who were two points different from reference ratings and percentage who rated a category as N/A (not applicable) are also presented. Figure 3 shows that highest number of participants agreed with the reference ratings for each category. Where they disagreed, participants were more likely to rate lower than the reference ratings for decision-making, communication and teamwork and leadership. The widest range in ratings was for the social skills with 15% of ratings of communication and teamwork, and 13% of leadership ratings were two points from the reference. A small proportion of N/A ratings were given to decision-making, communication and teamwork and leadership but 23% of situation awareness ratings were N/A. It is of note that the experts felt able to rate this behaviour in all cases.

Discussion

This study was designed to look at levels of agreement between expert and novice raters using the NOTSS system to rate surgeons'

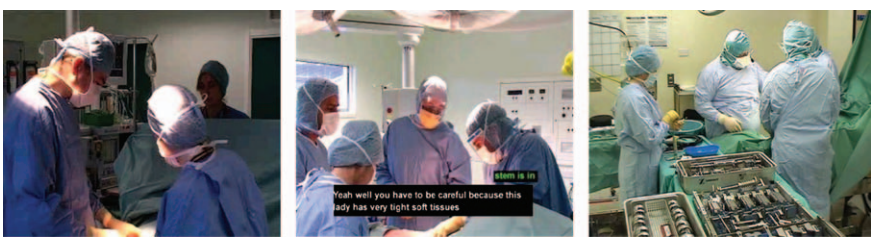

Fig. 1. Still shots from a selection of the non-technical skills for surgeons video scenarios. (a) A general surgeon enters theatre and does not appear to know which patient he is operating on. He struggles to make appropriate decisions during the operation (scenario 1). (b) An orthopaedic surgeon loses his situation awareness, blames the rest of the surgical team for breaking the patient's femur and displays poor leadership (scenario 4). (c) An orthopaedic surgeon has poor communication skills and expects the scrub nurse to know what he wants next without having to ask for it (scenario 6).

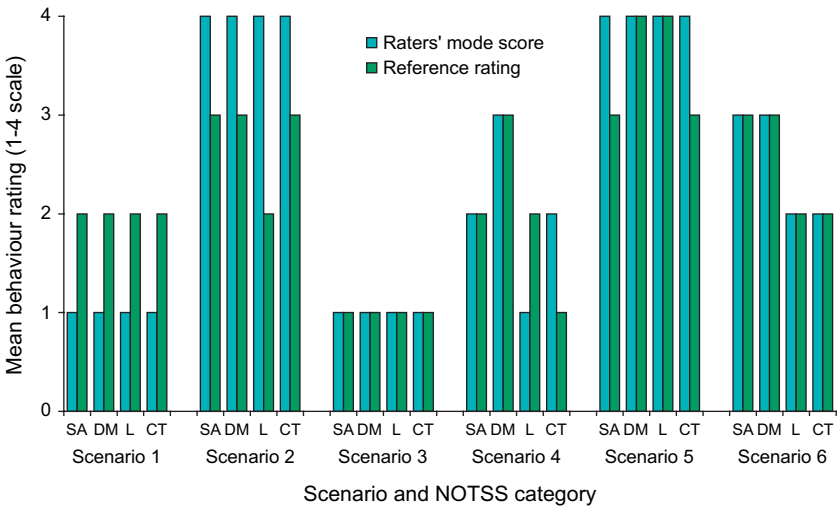

Fig. 2. Comparing expert and novice ratings within each scenario. CT, communication and teamwork; DM, decision-making; L, Leadership; SA, situation awareness.

non-technical skills in six simulated scenarios filmed in the operating theatre. The results indicate that although half the novice surgeons agreed with the expert rating, significant numbers of participants did not. When they did not agree, novice raters tended to be harsher in their ratings than the expert group. This study has raised several issues regarding the reliability of workplace assessment of non-technical skills and has several practical implications.

The majority view of untrained raters tends to agree with the experts, that is, there is evidence of a regression to the mode with an increasing numbers of raters. However, the system is likely to be used by one rater so even with training, it is unlikely that they will rate skills in line with the experts on every occasion.

Accuracy within scenarios

It is interesting to note from Figure 2 that the range of ratings across scenarios show a range of behaviours, from floor (scenario 3) to almost ceiling (scenario 5), to mid-range (scenario 6), mid-low (scenario 1), mid-high (scenario 2). There was more consistency around ceiling or floor ratings – raters found it easier to rate skills that were obviously good or poor (i.e. scenarios 1 and 3). Although it is reassuring that rating with the NOTSS system appears reliable in identifying clearly inadequate performance, this sort of behaviour will be observed infrequently. Assuming that such extremes of poor behaviour are rare in the real workplace, it may be important to focus the training of future NOTSS observers to discriminating between behaviours in the mid range (i.e. from marginal to acceptable).

Most disagreement was for scenarios depicting skills in the mid-range (as judged by the experts). Scenario 4 generated the most mid-range NOTSS ratings and was also the scenario with least agreement between raters and experts. This scenario had a number of ambiguous behaviours and the consultant surgeon who was the target of ratings also displayed good and poor behaviours for each category throughout the scenario. For example, he briefed the trainee surgeon on the risks of the operation and showed good situation awareness immediately before the start of the operation but then clearly lost his awareness of the operation and was distracted during a period when the trainee broke the patient's femur. Immediately after this event, the consultant surgeon regained good situation awareness to manage the incident. Raters clearly struggled to provide one rating for these periods of behaviour as they were forced to decide whether the loss of awareness meant that only a low rating could be provided as the surgeons' behaviours endangered patient safety (and by definition triggered the lowest rating possible of 1 – poor), or whether the other periods of good awareness should be reflected in the overall global rating. This led to a high degree of variance in ratings. In practice, it is suggested that individual institutions need to provide guidance on how to rate cases like this with reference to their own performance standards so NOTSS can be used consistently.

Accuracy within categories

The widest range in ratings was for the social skills with 15% of ratings of communication and teamwork and 13% of leadership ratings two points from the reference rating. This was notable because the previous psychometric evaluation found that the social skills were most reliably rated among the group.[14]

Strengths and limitations

One benefit of testing the reliability of the NOTSS system using simulated surgical scenarios was the ability to video-record specified behaviors in a stable context. For the purpose of this study, it was important to measure non-technical skills across a number of different clinical encounters with different surgeons being rated, and showing the full range of non-technical behaviour. For this reason, we chose to film simulated scenarios because this gave us control over the clinical contexts and the behaviours to be demonstrated within each. A glance at the range of category ratings across scenarios in Figure 2 demonstrates that this was achieved. All scenarios were clinically appropriate and looked realistic.

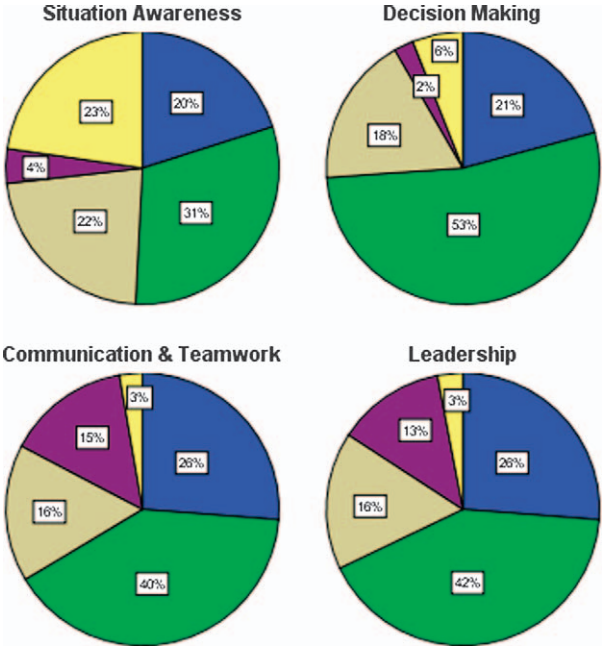

Fig. 3. Distribution of participants' ratings against reference rating for each category across scenarios. ■, one scale point below reference rating; ■, agree with reference rating; ■, one scale point above reference rating; ■, two points from reference rating; ■ rated as N/A, not applicable.

The main limitation was the inadequate amount of training raters received. The lack of experience of the rating group is a likely explanation for the range in performance compared with the reference ratings. The surgeons who participated as raters had no previous experience of behaviour rating, and their only experience of the underlying concepts was received during the short NOTSS training session prior to the evaluation. Other high reliability domains where non-technical skills training and assessment is commonplace recommend that a minimum of 2 days training is provided for using this type of rating system. Part of this process will involve calibration and checking of raters, which has been shown to increase inter-rater agreement. For the purpose of this study the short training delivered was the maximum feasible for the sample of consultant surgeons (n = 44) who volunteered to participate.

Conclusion

There is increasing evidence that non-technical skills are important in the safe care of patients in the operating theatre and beyond. The NOTSS taxonomy provides surgeons with a structure and language to observe, rate, and provide feedback on behaviours within the operating room following routine elective or emergency surgical cases and is well received.[15] However, as the assessment of non-technical skills becomes more commonplace,[14,19,20] and begins to be integrated into surgical training curricula, it is essential that the systems for rating these skills in surgery have acceptable validity and reliability.[16] In common with other high reliability domains, our study shows that this degree of reliability cannot be achieved without more in-depth training and calibration of raters. Further work to establish the level of training required to ensure reliability is required.

Acknowledgements

The NOTSS system was developed under funding from the Royal College of Surgeons of Edinburgh and NHS Education Scotland. The views presented in this paper are those of the authors and should not be taken to represent the position or policy of the funding bodies. We would like to thank the surgeons who took part in this study. Further information on the NOTSS system can be found at the Royal College of Surgeons of Edinburgh Patient Safety Board website: www.patientsafetyboard.org and at the University of Aberdeen: www.abdn.ac.uk/notss

References

1. Kneebone R, Darzi A. New professional roles in surgery. *Br. Med. J.* 2005; **330**: 803–4.
2. Hall JC, Ellis C, Hamdorf J. Surgeons and cognitive processes. *Br. J. Surg*. 2002; **90**: 10–16.
3. Flin R, Youngson GG, Yule S. How do surgeons make intraoperative decisions? *Qual. Saf. Health Care* 2007; **16**: 235–239.
4. Baldwin PJ, Paisley AM, Paterson-Brown S. Consultant surgeons' opinions of the skills required of basic surgical trainees. *Br. J. Surg*. 1999; **86**: 1078–82.
5. Edmondson AC. Speaking Up in the Operating Room: How Team Leaders Promote Learning in Interdisciplinary Action Teams. *J. Manag. Stud.* 2003; **40**: 1419–52.
6. GMC. *Good Medical Practice*. London: General Medical Council, 2001.
7. Yule S, Flin R, Paterson-Brown S, Maran N. Non-technical skills for surgeons: a review of the literature. *Surgery* 2006; **139**: 140–49.
8. Paisley AM, Baldwin PJ, Paterson-Brown S. Feasibility, reliability and validity of a new assessment form for use with basic surgical trainees. *Am. J. Surg*. 2001; **182**: 24–9.
9. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Rating non-technical skills: Developing a behavioural marker system for use in anaesthesia. *Cogn. Tech. Work* 2004; **6**: 165–71.
10. Gordon SE. *Systematic Training Programme Design: Maximising Effectiveness and Minimizing Liability.* Englewood Cliffs, NJ: Prentice Hall, 1993.
11. Flin R, Yule S, McKenzie L, Paterson-Brown S, Maran N. Attitudes to teamwork and safety in the operating theatre. *Surgeon* 2006; **4**: 145–51.
12. Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons' non-technical skills. *Med. Educ*. 2006; **40**: 1098–104.
13. Flin R, Goeters K, Amalberti R *et al.* The development of the NOTECHS system for evaluating pilots' CRM skills. *Hum. Factors Aerospace Saf.* 2003; **3**: 95–117.
14. Yule S, Flin R, Maran N, Rowley DR, Youngson GG, Paterson-Brown S. Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behaviour rating system. *World J. Surg*. 2008; **32**: 548–56.

15. Yule S, Flin R, Rowley D, Mitchell A, Youngson GG, Paterson-Brown S. Debriefing surgeons on non-technical skills (NOTSS). *Cogn. Tech. Work*. 2008; **10**: 265–274.
16. Flin R, Yule S, Paterson-Brown S, Maran N, Rowley D, Youngson G. Teaching surgeons about non-technical skills. *Surgeon* 2007; **5**: 86–9.
17. Baker D, Mulqueen C, Dismukes R. Training raters to assess resource management skills. In: Salas E, Bowers C, Edens E (eds). *Improving Teamwork in Organizations*. New Jersey: LEA, 2001; 131–45.
18. Goldsmith T, Johnson P. Assessing and improving evaluation of aircrew performance. *Int. J. Aviation Psychol*. 2002; **12**: 223–40.
19. Mishra A, Catchpole K, Dale T, McCulloch P. The influence of non-technical performance on technical performance in laparoscopic cholecystectomy. *Surg. Endosc*. 2008; **22**: 68–73.
20. Sevdalis N, Davis R, Koutantji M, Undre S, Darzi A, Vincent C. Reliability of a revised NOTECHS scale for use in surgical teams. *Am. J. Surg*. 2008; **196**: 184–90.