World J Surg (2008) 32:548–556
DOI 10.1007/s00268-007-9320-z

World Journal of Surgery

Surgeons' Non-technical Skills in the Operating Room: Reliability Testing of the NOTSS Behavior Rating System

Steven Yule · Rhona Flin · Nicola Maran · David Rowley · George Youngson · Simon Paterson-Brown

Published online: 8 February 2008

Abstract
Background Previous research has shown that surgeons' intraoperative non-technical skills are related to surgical outcomes. The aim of this study was to evaluate the reliability of the NOTSS (Non-technical Skills for Surgeons) behavior rating system. Based on task analysis, the system incorporates five categories of skills for safe surgical practice (Situation Awareness, Decision Making, Task Management, Communication & Teamwork, and Leadership).
Methods Consultant (attending) surgeons ($n = 44$) from five Scottish hospitals attended one of six experimental sessions and were trained to use the NOTSS system. They then used the system to rate consultant surgeons' behaviors in six simulated operating room scenarios that were presented using video. Surgeons' ratings of the behaviors demonstrated in each scenario were compared to expert ratings ("accuracy"), and assessed for inter-rater reliability and internal consistency.
Results The NOTSS system had a consistent internal structure. Although raters had minimal training, rating "accuracy" for acceptable/unacceptable behavior was above 60% for all categories, with mean of 0.67 scale points difference from reference (expert) ratings (on 4-point scale). For inter-rater reliability, the mean values of within-group agreement (r_{wg}) were acceptable for the categories Communication & Teamwork (.70), and Leadership (.72), but below a priori criteria for other categories. Intra-class correlation coefficients (ICC) indicated high agreement using average measures (values were .95–.99).
Conclusions With the requisite training, the prototype NOTSS system could be used reliably by surgeons to observe and rate surgeons' behaviors. The instrument should now be tested for usability in the operating room.

S. Yule (✉) · R. Flin
School of Psychology, University of Aberdeen,
Aberdeen AB24 2UB, Scotland, United Kingdom
e-mail: s.j.yule@abdn.ac.uk

N. Maran
Department of Anaesthesia, Royal Infirmary of Edinburgh,
51 Little France Crescent, Old Dalkeith Road, Edinburgh,
Scotland EH16 4SA, United Kingdom

D. Rowley
Department of Orthopaedic and Trauma Surgery, University of Dundee, Ninewells Hospital, Dundee, Scotland DD2 9SY,
United Kingdom

G. Youngson
Royal Aberdeen Children's Hospital, Cornhill Road, Aberdeen,
Scotland AB25 2ZD, United Kingdom

S. Paterson-Brown
Department of Surgery, Royal Infirmary Edinburgh, 51 Little France Crescent, Old Dalkeith Road, Edinburgh, Scotland EH16 4SA, United Kingdom

Introduction

Surgical patients may be involved in up to 60% of adverse medical events [1, 2]. Studies of behavior in the operating room show that breakdowns in non-technical skills such as teamwork, leadership, communication, situation awareness, and poor decision making are not uncommon [2] and can lead to errors [3], poor outcomes [4], and higher compensation payouts [5]. Conversely, some expert surgeons demonstrate good non-technical skills as an integral part of their operating behavior [6, 7]. Giddings from the Royal College of Surgeons of England recently warned that there are still avoidable deaths in surgery because surgeons do not learn from past failures and in some cases have a

"seriously flawed opinion of their own capabilities," which can turn into arrogance [8]. Similarly, Davidson argues that future surgical training will need to encompass more than just clinical and technical skills [9]. High-hazard industries have also recognized that technical expertise does not guarantee safe operations and they introduced non-technical skills training, often called Crew Resource Management (CRM), designed to enhance performance of these skills [10]. Non-technical skills training in surgery is one method of enhancing surgeons' performance in the operating room, in order to improve patient safety.

Training of surgeons' non-technical skills needs to be preceded by task analysis to identify the critical skill set. In addition, a performance measure is required so that demonstration of the skills can be assessed for feedback or for formal competence assurance. Normally these performance evaluations are made from observations of behavior on task or in a simulator, using behavior rating systems, such as NOTECHS for airline pilots [11] and ANTS (Anaesthetists' Non-technical Skills) for anesthetists [12]. These tools assess "observable, non-technical behaviors that contribute to superior or substandard performance within a work environment" (p. 10) [13]. All behaviour-rating systems must be evaluated according to psychometric criteria regarding validity and reliability before they can be used in the workplace [14].

We developed a taxonomy and a behavior rating system of Non-technical Skills for Surgeons (NOTSS) [15] using several methods of task analysis with consultant surgeons [16]. This skill set consists of five categories (Situation Awareness, Decision Making, Task Management, Communication & Teamwork, and Leadership), divided into 14 elements, each with example good and poor behaviors. The content validity of the NOTSS system was derived from its systematic development process with subject matter experts. Before this system can be used to rate behaviors, it must be evaluated according to a set of a priori criteria. The aim was to test the reliability and initial usability of the NOTSS framework using standardized video scenarios of surgeons' behavior in the operating room. Ethical clearance was granted by Grampian Research Ethics Committee (GREC).

Materials and methods

The system was tested in relation to the following components:

- *Sensitivity*: the level of accuracy between participants' and "reference" (expert) ratings.
- *Inter-rater reliability*: the degree of agreement within six groups of surgeons ($n = 44$) who rated six standardized video scenarios filmed in the operating room.
- *Internal structure:* the relationship between category and element ratings.

Design of video scenarios

Video scenarios ($n = 11$) illustrating surgeons' non-technical skills (good to poor) in a range of realistic simulated situations were filmed in operating rooms, using a patient simulator and practicing surgeons, anesthetists, and nurses acting the main roles. This ensured that the reliability of the NOTSS system could be tested across different clinical encounters. The scenarios were designed by two surgeons, an anesthesiologist experienced in non-technical skills training, and two psychologists. To minimize typecasting, five consultant surgeons played the lead roles across scenarios. For the evaluation, six scenarios were selected, ranging from 2.30 to 5.40 min in length (see Appendix 1). A further three scenarios were selected for practice during pre-experiment training.

Reference ratings

A set of "reference ratings" was used to test the sensitivity of participants' ratings (i.e., to what extent participants could detect the non-technical skills and discriminate between good and poor performance). The reference ratings were provided by the scenario designers who were also practicing surgical team members with up to 10 years expertise in behavior 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 judgment of the level of each non-technical skill demonstrated in each scenario, expressed as an agreed set of category and element ratings for each scenario.

Participants

Advertisements 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. As a result, 44 consultant surgeons from five Scottish hospitals attended one of six experimental sessions: 18 (41%) from general surgery, 11 (25%) from orthopedic surgery, three (7%) from pediatric surgery, 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 (s.d. 7.5 yr); 95% ($n = 42$) were male. Most surgeons ($n = 37$, 84%) reported some involvement in assessing trainees, but of these, only 17 (39%) had received training in performance assessment. Some surgeons ($n = 8$) had been involved in earlier stages of the

NOTSS project, but the most were unfamiliar with the system.

Procedure

The data-collection materials used were NOTSS rating forms, a participant demographics form, and an evaluation questionnaire. The rating form was used to record scores for NOTSS categories and elements on a four-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). It is well recognized that raters must receive specific training to assess non-technical skills [14, 17]. The NOTSS system training course (2.5 h) was developed to give the surgeons who participated in the study a basic background in human factors, and in observing and rating behaviors, which is lacking in their current professional training. The course was based on guidance from aviation on behavior rating [14] and used surgical examples to train raters. It comprised (1) background on human factors and non-technical skills; (2) an introduction to the NOTSS system and how to make behavioral assessments, and (3) practice in observing and rating behaviors with NOTSS in three training scenarios. Raters were not formally calibrated but they did discuss their observations and ratings after each of the three training scenarios.

After the training, the experimental evaluation (to rate the consultant surgeon's behaviors in the six evaluation video scenarios) took 1 h. Participants were instructed to watch each scenario and to rate the observed skills using the NOTSS rating form. Participants were informed of the simulated nature of the scenarios. All ratings were made individually. At the end of the session, participants completed an evaluation questionnaire.

Results

System sensitivity

Non-technical Skills for Surgeons sensitivity (accuracy) at category level was assessed by calculating the mean absolute difference for category ratings and the corresponding reference rating within each scenario. The mean of these differences across all six experimental scenarios was then calculated. Lower scores indicate increasing sensitivity. The average NOTSS sensitivity at category level was .67 scale points and a breakdown of sensitivity by category is presented in Figure 1. The Task Management category had the highest sensitivity, and ratings for the two other interpersonal skills categories (Leadership and Communication & Teamwork) were roughly equivalent at a mean difference from reference rating of around 0.8 scale points. Of the two cognitive skills categories, Decision Making had more sensitivity than Situation Awareness, which was the least sensitively rated category of all.

The accuracy of ratings when distinguishing between behaviors which might be deemed to be "acceptable" and "unacceptable" was examined by comparing collapsed ratings of 1 and 2 (unacceptable) with ratings of 3 and 4 (acceptable). With this method, accuracy was assessed by examining the percentage of raters agreeing with the reference rating (unacceptable or acceptable) for each category, averaged across the six scenarios (see Fig. 2). The response N/A was not included in this analysis. From Figure 2, a mean of $n = 36$ (82%) of participants agreed with reference ratings of unacceptable or acceptable behaviors within the Decision Making and Communication & Teamwork categories, which were the most accurately rated. A mean of $n = 30$ (69%) agreed with the reference rating for Leadership and a mean of $n = 28$ (63%) agreed with ratings for Situation Awareness and Task Management.

Inter-rater reliability (IRR)

Mean within-group agreement (r_{wg}) [18] was calculated for the NOTSS categories and elements ratings for each of the six experimental groups. The average across groups was then taken to represent the overall reliability of the prototype system. The within group agreement statistic lies between 0 (no agreement) and 1 (perfect agreement) and represents the degree to which a number of raters agree on the absolute ratings given to targets (in this study, targets are the behaviors in the scenarios that reflect the NOTSS categories and elements). The calculation is based on ratings of behaviors for each scenario, which are then averaged across the six scenarios to give a mean r_{wg} for each of the NOTSS categories and elements. The generally accepted criterion for an acceptable level of agreement is $r_{wg} > 0.7$–0.8 [19].

Table 1 shows that the mean r_{wg} across experimental groups only exceeded the criterion of >.7 for two categories: Leadership and Communication & Teamwork. Within-group agreement for Decision Making approached the criterion, but the values of r_{wg} for Situation Awareness and Task Management were .51 and .66, respectively. In analysis of the NOTSS elements, only ratings for "supporting others," one of the Leadership elements, exceeded the criterion.

The intraclass correlation coefficient [ICC(2)] was also calculated to provide a complementary reliability measure to the r_{wg}. ICC(2) provides an estimate of the reliability of

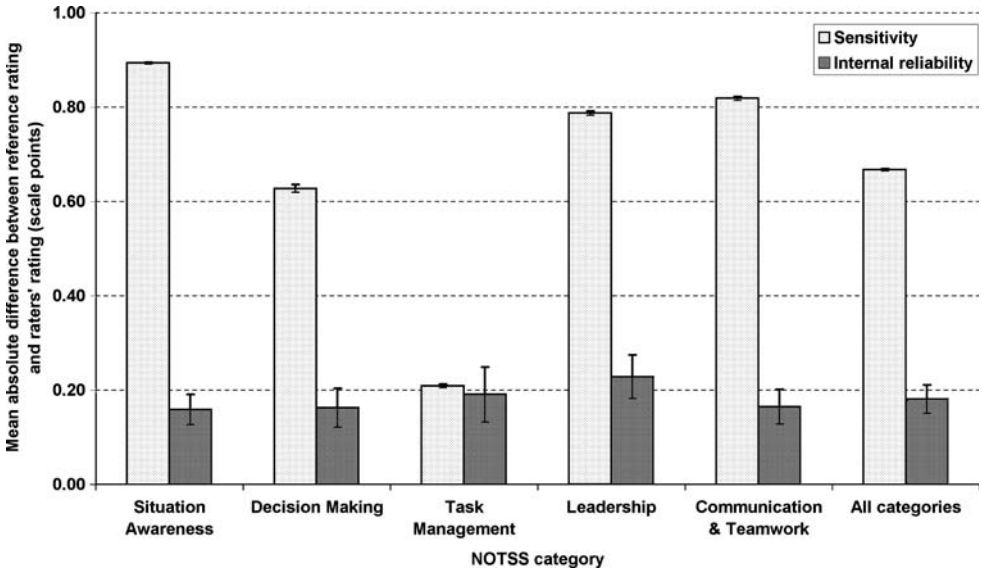

Fig. 1 NOTSS system sensitivity and internal reliability

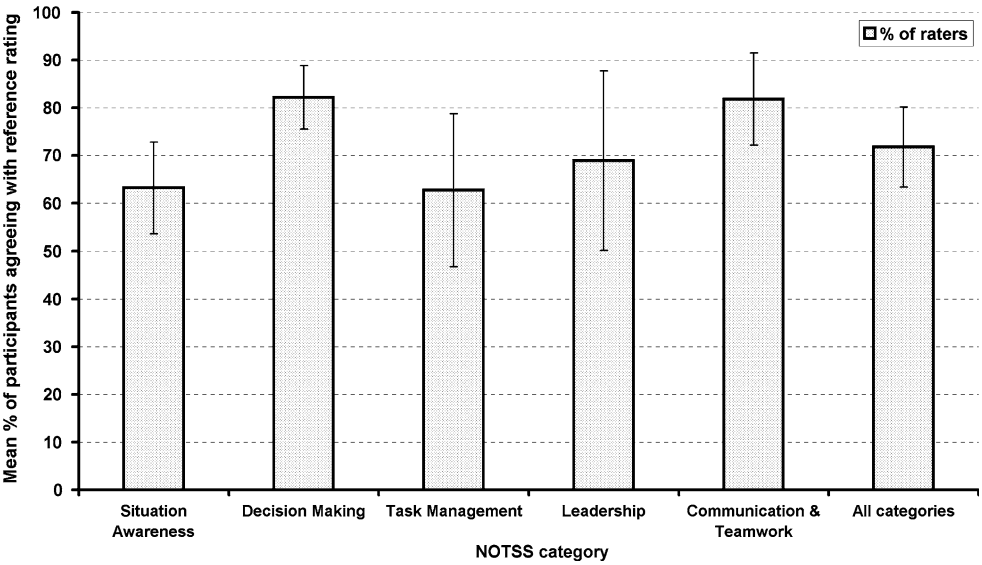

Fig. 2 Mean percentage of raters agreeing with category reference ratings (collapsed scale: unacceptable versus acceptable)

group means and is estimated using mean squares from a one-way random-effects ANOVA [20]. Intraclass correlation coefficient (2), based on absolute agreement, was selected as the most appropriate model [21], and the coefficients for single-rater and average-ratings were examined, as displayed in Table 1. The convention is that ICC $>$ .07 is required for acceptable reliability. For ICC(2) average measures, this was exceeded in all categories (scores from .95 to .99).

The coefficients for ICC(2) single measure did not reach the criterion for any category, but coefficients for Decision Making, Leadership, and Communication & Teamwork were all $>$.6. The ICC(2) single measure coefficients were lowest for Situation Awareness (.29) and Task Management (.39).

Between-scenario differences in IRR

Within-group agreement was higher within scenarios 1, 2, and 3 than for scenarios 4, 5, and 6 (see Table 1). To investigate this further, the mean within-group agreement across categories for each scenario was calculated and is shown in Figure 3. There is substantial variation in surgeons' agreement between scenarios, possibly because behaviors were more extreme and therefore "easier" to

Table 1 Within-group agreement (r_{wg}) and intraclass correlation coefficient (ICC) across six experimental scenarios

NOTSS skills	Experimental group						Mean r_{wg}	ICC(2) Single	ICC(2) Average
	1	2	3	4	5	6			
Categories									
Situation Awareness (SA)	0.65	0.58	0.51	0.52	0.22	0.56	0.51	0.29	0.95
Decision Making (DM)	0.57	0.78	0.70	0.73	0.63	0.66	0.68	0.60	0.99
Task Management (TM)	0.50	0.61	0.76	0.75	0.68	0.65	0.66	0.39	0.97
Leadership (L)	0.60	0.67	0.84	0.78	0.69	0.72	0.72	0.66	0.99
Communication & Teamwork (C&T)	0.68	0.70	0.84	0.82	0.47	0.68	0.70	0.63	0.99
Elements[a]									
Gathering information (SA)	0.59	0.50	0.82	0.48	0.28	0.62	0.55	–	–
Understanding Information (SA)	0.63	0.61	0.60	0.66	0.04	0.29	0.46	–	–
Projecting and anticipating future state (SA)	0.49	0.44	0.58	0.55	0.21	0.60	0.48	–	–
Considering options (DM)	0.61	0.64	0.63	0.87	0.50	0.70	0.66	–	–
Selecting and communicating options (DM)	0.55	0.63	0.82	0.72	0.48	0.65	0.64	–	–
Implementing and reviewing decisions (DM)	0.61	0.56	0.71	0.62	0.66	0.61	0.63	–	–
Planning and preparation (TM)	0.45	0.43	0.85	0.72	0.73	0.45	0.60	–	–
Flexibility/responding to change (TM)	0.51	0.58	0.74	0.67	0.64	0.61	0.63	–	–
Setting and maintaining standards (L)	0.49	0.55	0.74	0.50	0.46	0.40	0.52	–	–
Supporting others (L)	0.71	0.66	0.78	0.78	0.79	0.72	0.74	–	–
Coping with pressure (L)	0.58	0.63	0.78	0.75	0.71	0.62	0.68	–	–
Exchanging information (C&T)	0.53	0.50	0.71	0.65	0.58	0.72	0.61	–	–
Establishing a shared understanding (C&T)	0.59	0.58	0.79	0.77	0.29	0.48	0.58	–	–
Co-ordinating team activities (C&T)	0.55	0.68	0.73	0.65	0.57	0.60	0.63	–	–

[a] Corresponding categories are in parentheses

identify and rate in some scenarios (i.e., scenarios 1 and 3), than others (i.e., scenarios 5 and 6). As behaviors become more extreme and/or raters are better trained, within-group agreement would be expected to improve.

Between-specialization differences in IRR

The type of surgery being depicted in the scenarios may also have an impact on the degree of agreement. Surgeons from several specialties participated in the evaluation, but all scenarios were based either in general or orthopedic surgery. It is possible that surgeons' ratings might be more homogeneous when they are rating scenarios based in their own specialty than when rating other specialties. This possibility was tested by comparing the within-group agreement for raters who were either general surgeons or orthopedic surgeons for each of the scenarios. The results presented in Figure 3 show that orthopedic surgeons' ratings were significantly more in agreement than general surgeons' ratings for all scenarios, irrespective of the type of surgery being depicted in the scenario ($t = 10.17$, $p < 0.00$).

Internal reliability of the system

The NOTSS system allows raters to rate skills at both the category and element level. Because of conceptual overlap, there should be a high degree of consistency between the category rating and the ratings for the two or three underlying elements. This possibility was assessed by calculating the mean absolute difference between raters' element ratings and their rating for the corresponding category [22], displayed in Figure 1 (internal consistency scores). Lower scores indicate closer agreement. Consistency between category and element was very high for all categories ($M < 0.25$ of a scale point between element and category on a four-point scale).

Implications for system development

Following a review of the results, the Task Management category was dropped, and relevant behaviors were incorporated into the other categories where appropriate. This was done because (1) the results of IRR for this category were relatively poor, (2) feedback from participants

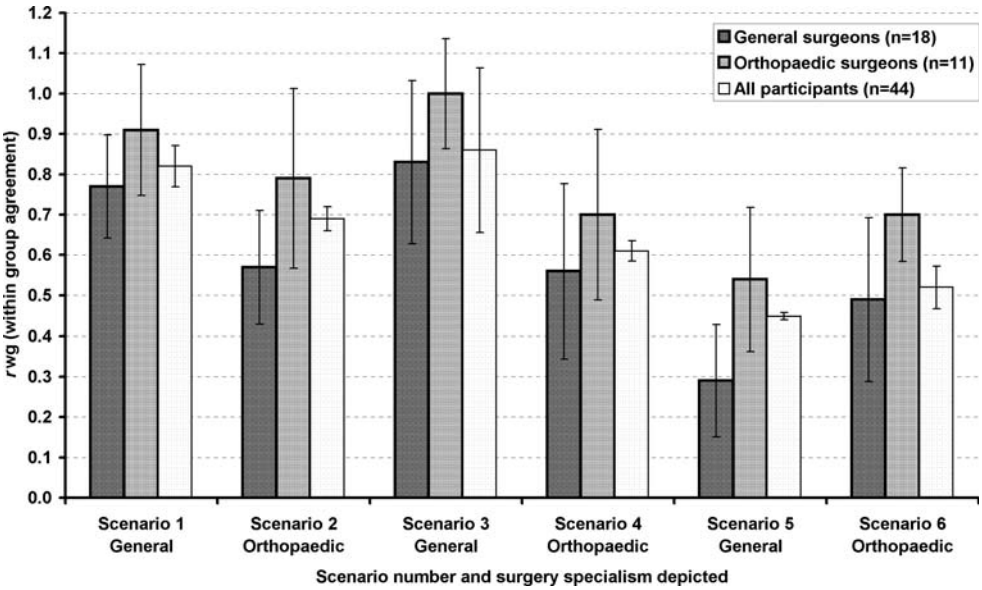

Fig. 3 Within-group agreement by scenario: comparing general and orthopedic surgeons

indicated that surgeons did not engage in task management behaviors intraoperatively because they delegated the majority of these tasks during preoperative planning, (3) many of the Task Management behaviors reflected the underlying concepts of gaining and updating Situation Awareness, and (4) because of feedback regarding complexity, reducing the number of categories made the rating system more parsimonious and reduced cognitive workload for users [23]. The revised NOTSS taxonomy (v1.2) is presented in Table 2.

Discussion

The results indicate that the NOTSS system has a consistent internal structure and acceptable sensitivity when surgeons' ratings are compared with reference ratings of acceptable/unacceptable behavior. Intraclass correlation coefficients exceeded criteria for the scale to be used by multiple raters (ICC average-ratings), and acceptable within-group reliability existed for the interpersonal skills categories (Leadership and Communication & Teamwork). However, within-group reliability did not meet *a priori* criteria for the cognitive skills categories of Situation Awareness and Decision Making. It should be emphasized that the surgeons who participated as raters had no previous experience with this type of behavior rating, and they had not been educated in the underlying concepts as part of their medical training. Also, they had received only 2.5 h of training in use of the NOTSS system. This was the maximum amount of training that we could arrange for groups of consultant surgeons who participated on a discretionary basis and were not compensated for their time. Also, the intention of the 2.5 h training was not to improve the accuracy or reliability of raters per se, but to give them basic training in observing and rating behaviors, as well as some initial practice at doing this on video scenarios. For these reasons, it was remarkable that acceptable reliability was achieved for some categories. Moreover, no attempt was made to calibrate the raters by using feedback and discussion to establish common standards, and this may account for a significant amount of the within-group variability. The lack of agreed "gold standards" in surgery means that several opinions on what is good or desirable treatment in a given situation can co-exist. The fact that some categories did not reach acceptable reliability according to strict *a priori* criteria does not negate the importance of the underlying constructs. Interest in rating the behaviors relating to cognitive skills and assessing their impact on surgical performance should not be abandoned on the basis of this initial reliability study. There may

Table 2 Non-Technical Skills for Surgeons (NOTSS) 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 & Teamwork	Exchanging information
	Establishing a shared understanding
	Co-ordinating team
Leadership	Setting and maintaining standards
	Supporting others
	Coping with pressure

always be more diverse interpretations in surgery compared with industries that are highly proceduralized and regulated, such as civil aviation and nuclear power.

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. Six different scenarios were used to test the system because it was deemed important to establish reliability across a number of different clinical encounters, and the within-group agreement statistic was calculated for each category and element by considering the agreement of ratings for those categories across the six scenarios. By testing the system in this manner—across six scenarios depicting different clinical encounters with different surgeons acting the lead roles—the results were less likely to be skewed, and the applicability of the system to several different scenarios was established. Despite the fact that all scenarios were clinically appropriate and looked realistic, they were relatively brief and did not represent a situation where a rater could spend several hours observing a surgeon's behavior during a case. Although information was provided about the patient in each scenario, some raters said they would want additional contextual information (e.g., the level of the trainee's competence) before they could judge the consultant's behavior. The main limitation, as mentioned above, was the inadequate amount of training raters received. It is recommended that a minimum of 2 days' training be provided for using this type of rating system. However, the 2.5 h training session that was delivered to the 44 surgeons who participated has been viewed as a good investment in an emerging area of surgical training and assessment, one that could affect the future behavior of those surgeons in the operating room. We believe that with increasing familiarity and training, agreement within surgeons about how to rate non-technical skills will increase. There is no requirement for the suggested 2 days of training to be completed back-to-back in a single course; in fact, it may be more sustainable if training is delivered over a period of time. Ongoing research on the usability of NOTSS in the operating room will continue to provide the surgeons who participate with training that will build on the foundations provided in this study. In spite of limitations, participants were able to use the NOTSS system with a level of sensitivity and agreement that approached an acceptable standard. Although this was not as high as recommended for trained non-technical skills assessors, evaluation of behavioral rating systems in other industries has shown that high levels of sensitivity and inter-rater reliability are not achieved unless individuals have proper training and calibration.

Implementing NOTSS into surgical practice

The NOTSS system was designed primarily as an educational tool, to provide surgeons with a structure and with the language to observe, rate, and provide feedback on behaviors during routine cases. To focus on the educational aspect, we are now running a trial to establish the impact of regular debriefings on surgical performance. The NOTSS system was not designed specifically for quality control, but there have been several emergent uses in this area since its release. For example, surgeon-trainers have used NOTSS to deal with underperforming surgeons; to analyze surgical adverse events [24] and to structure non-technical skills training. The ultimate goal is for NOTSS to be adopted into the curriculum of all surgical specialties. To achieve this without over-burdening surgeons with additional training requirements, NOTSS may need to be adapted so that it is in the same format as existing tools for evaluating technical skills. The results of this study show that, with minimal training, surgeons can use NOTSS to provide formative assessment of non-technical skills at the acceptable/unacceptable level, based on observation of behaviors during the performance of surgical procedures. This may be one practical way of implementing the system to increase exposure and familiarity, before surgeons are asked to assess using the whole scale.

A recent systematic review found that clinicians have only limited ability to self-assess their competence [25]. Current proposals to increase medical regulation in the UK [26] may mean that objective assessment of non-technical competencies could become routinely used in healthcare, as it is in other industries [27]. To develop appropriate training and feedback procedures, behavior rating tools such as NOTSS need to be carefully designed and subjected to iterative testing with domain experts. Therefore, with the proviso that adequate training be given to the raters, the NOTSS system should now be subjected to a formal usability evaluation in the operating room, which will also allow trainers and trainees to gain familiarity and confidence in a method of introducing non-technical skills into the dialogs of training.

Acknowledgments The NOTSS system was developed under funding from the Royal College of Surgeons of Edinburgh and NHS Education Scotland. The views presented are those of the authors and should not be taken to represent the position or policy of the funding bodies. The authors thank the surgeons who took part in this study.

Appendix 1. Description of video scenarios

Scenario 1. General surgery

A 62-year-old obese man with symptomatic gallstones is about to undergo emergency laparoscopic cholecystectomy. The patient is in the operating room and has been anesthetised. The surgeon arrives late and does not appear to know the patient. There are no long ports in the operating room or day surgery, so the surgeon decides to operate with short ports instead, exposing the patient to risk.

Scenario 2. Orthopaedic surgery

A 69-year-old woman is undergoing a foot operation. The surgeon appears to be in control but does not communicate with the rest of the team (radiographer, trainee surgeon). He experiences technical difficulties with a drill.

Scenario 3. General surgery

A 51-year-old man is undergoing elective mesh repair of an inguinal hernia. The consultant surgeon is bad-tempered, breaks operating room protocol, and openly questions the competence of operating room team members in a hostile manner.

Scenario 4. Orthopedic surgery

An 86-year-old woman is undergoing a trial reduction of femur. The consultant surgeon tells the trainee to be careful but is chatting with another team member when the trainee breaks the patient's femur. The consultant surgeon blames the trainee and is initially aggressive, but rapidly regains control of the situation.

Scenario 5. General surgery

A 40-year-old man has been anesthetized and is about to undergo repair of an inguinal hernia. A junior member of staff has taken the patient's consent for the operation, and there is confusion over which side should be operated on. The trainee surgeon is keen to proceed and is sure that it is the left side, but the consultant surgeon decides to stop the operation and wake the patient up to make sure, involving other team members in the decision-making process.

Scenario 6. Orthopedic surgery

A 68-year-old woman is undergoing a knee replacement. The consultant surgeon treats the trainee and scrub nurse differently—he is friendly and encouraging with the trainee and not very communicative with the nurse. She makes a minor error as a direct result of the surgeons' ambiguous communication style.

References

1. Gawande AA, Zinner MJ, Studdert DM, et al. (2003) Analysis of errors reported by surgeons at three teaching hospitals. Surgery 133:614–621
2. Yule S, Flin R, Maran N, et al. (2006) Non-technical skills for surgeons in the operating room. A review of the literature. Surgery 39:140–149
3. Stevenson KS, Gibson SC, Rogers PN, et al. Process of care in acute surgical admissions: room for improvement. Br J Surg in press
4. Christian C, Gustafon M, Roth E, et al. (2006) A prospective study of patient safety in the operating room. Surgery 139: 159–173
5. Studdert DM, Mello MM, Gawande AA, et al. (2006) Claims, errors, and compensation payments in medical malpractice litigation. N Engl J Med 354:2024–2033
6. Carthey J, de Leval MR, Wright DJ, et al. (2003) Behavioral markers of surgical excellence. Safety Sci 41:409–425
7. Moorthy K, Munz Y, Adams S, et al. (2005) A human factors analysis of technical and team skills among surgical trainees during procedural simulations in a simulated operating theatre. Ann Surg 242:631–641
8. Templeton S, Feinmann J (2006) Arrogant surgeons 'risk another Bristol babies scandal'. The Sunday Times London September 3
9. Davidson P (2002) The surgeon of the future and implications for training. Aust N Z J Surg 72:822–828
10. Flin R, Maran N (2004) Identifying and training non-technical skills in acute medicine. Qual Safety Healthcare i180–i184
11. Flin R, Martin L, Goeters K, et al. (2003) Pilots' non-technical skills: NOTECHS. Hum Factors Aerospace Safety 3:95–117
12. Fletcher G, Flin R, McGeorge P, et al. (2003) Anaesthetists' non-technical skills (ANTS): evaluation of a behavioral marker system. Br J Anaesthesia 90:580–588
13. Klampfer B, Flin R, Helmreich RL, et al. (2001) Enhancing performance in high risk environments: recommendations for the use of behavioral markers. Berlin: GIHRE
14. Baker D, Mulqueen C, Dismukes R (2001) Training raters to assess resource management skills. In Salas E, Bowers C, Edens E, editors, Improving Teamwork in Organizations. Mahwah, NJ, Lawrence Earlbaum, 131–145
15. Flin R, Yule S, Paterson-Brown S, et al. (2005) Surgeons' non-technical skills. Surg News 4:83–85
16. Yule S, Flin R, Paterson-Brown S, et al. (2006) Development of a rating system for surgeons' non-technical skills. Med Ed 40:1098–1104
17. Goldsmith T, Johnson P (2002) Assessing and improving evaluation of aircrew performance. Int J Aviation Psychology 12: 223–240
18. James L, Demaree R, Wolf G (1993) rwg: an assessment of within-group inter-rater agreement. J Appl Psychology 78: 306–309

19. Nunnally J, Bernstein I (1993) Psychometric Theory. New York, McGraw Hill
20. Bliese P (2000) Within-group agreement, non-independence, and reliability. Implications for data aggregation and analysis. In Klein K, Kozlowski S, editors, Multilevel theory, research, and methods in organizations. San Francisco, Jossey-Bass
21. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86:420–428
22. O'Connor P, Hormann HJ, Flin R, et al. (2002) Developing a method for evaluating Crew Resource Management skills: a European perspective. Int J Aviation Psychology 12:265–288
23. Holt R, Boehm-Davis D, Beaubien J (2001) Evaluating resource management training. In Salas E, Bowers C, Edens E, editors. Improving Teamwork in Organizations. Applications of Resource Management Training. Mahwah, NJ, Lawrence Earlbaum Associates
24. Yule S, Flin R, Rowley D, et al. Debriefing surgeons on non-technical skills (NOTSS). Cognition, Technology & Work (in press)
25. Davis DA, Mazamanian PE, Fordis M, et al. (2006) Accuracy of physician self-assessment compared with observed measures of competence. JAMA 296:1094–1102
26. Donaldson LJ (2006) Good doctors, safer patients: proposals to strengthen the system to assure and improve the performance of doctors and to protect the safety of patients. London, Department of Health
27. Flin R, O'Connor P, Crichton M (2007) Safety at the Sharp End. A Guide to Non-Technical Skills. Aldershot, UK, Ashgate, in press