Journal of the Royal Society of Medicine Open;
0(0) 1–6
DOI: 10.1177/2054270415611834

Lessons from surgery and anaesthesia: evaluation of non-technical skills in interventional radiology

Chun L Pang[1], **Salil B Patel**[2] **and Nicola Pilkington**[3]
[1]Peninsula Radiology Academy, Plymouth NHS Hospitals Trust, Plymouth, UK
[2]Peninsula College of Medicine and Dentistry, Royal Devon and Exeter Hospital, Exeter, UK
[3]Department of Anaesthesia, Derriford Hospital, Plymouth, UK
Corresponding author: Chun L Pang. Email: chunlappang@nhs.net

Summary

In the medical profession, surgery and anaesthesia are leading the way in identifying human errors that negatively affect patient safety. Evidence suggests that the implementation of non-technical skills assessments reduces such errors. Interventional Radiology is a procedural based speciality and therefore may also benefit from formal assessment of non-technical skills. This literature review supports the use of standardised assessment tools used in surgery and anaesthesia. Using the Downing framework of internal validity, the tools demonstrated good internal consistency but a spectrum of inter-rater variability, which can be partially improved with training. At present, a formal Interventional Radiology non-technical skills assessment tool is probably not suitable to be a stand-alone 'high stakes' assessment, but may be a useful adjunct to the existing array of workplace-based assessments.

Keywords

anaesthesia, radiology (diagnostics), surgery

Introduction

There has been plenty of interest in human performances in the medical profession over the last decade. Anaesthesia and surgery are probably the leading medical specialties in addressing issues relating to human factors regarding patient safety initiatives. The surgeon and anaesthetist are jointly responsible for the patient, with a supporting network of clinical staff. Each has unique responsibilities, but there is considerable overlap. Within hospitals, operating theatres are reportedly the most common site for adverse events to occur, most likely due to the complex environment.[1] The aviation industry has identified a correlation between potential adverse events and deficiencies in teamwork behaviour, which can translate to the healthcare environment.

Interventional Radiology (IR) is a relatively new subspeciality within Radiology, and the working environment is similar to the operating theatre. Vascular procedures are often complex, and patients may suffer due to a number of technical and human factors. The main difference between theatre and IR is that a radiologist replaces a surgeon as an operator. Radiologists are not immune to making mistakes; Reason[2] pointed out that errors are inevitable, but identifying system-wide issues might minimise future adverse events. Several methods have been developed for measuring teamwork and cognitive skills in the operating theatre. The ability to measure non-technical skills (NTS) is believed to influence the quality and safety of healthcare. This article aims to review the literature surrounding assessment of NTS in anaesthesia and surgery and evaluate the transferability of such assessments to develop a new workplace-based assessment (WBA) in IR. The construct validity of the assessment will be compared against the Downing[3] framework. Emphasis will be placed on internal consistency and inter-rater variability.

Method

A review of the existing literature was undertaken to look at validated tools used to assess NTS in anaesthetics and surgery. Emphasis was placed on examining internal consistency and inter-rater variability of multisource feedback (MSFs), Mini-clinical Evaluation exercises (mini-CEXs), Observational Teamwork Assessment for Surgery tool (OTAS & OTAS-D), Non-technical Skills for Surgeons (NOTSS), the Observational Skill-based Clinical Assessment tool for Resuscitation (OSCAR), The Oxford Non-technical Skills Scale (NOTECHS) and related WBAs for NTS evaluation in anaesthesia and surgery.[4–11] The relevant literature was identified using online databases including MEDLINE, Google Scholar and Web of Knowledge.

MSF and NTS

The existing MSF or 360-degree assessment arguably does assess NTS and has reasonable validity evidence, showing at least a 'medium' size effect based on an evaluation of 35 studies in a recent meta-analysis.[12] MSF assesses aspects of basic clinical skills competence and non-cognitive behaviour by using multiple questionnaires and different types of rating systems. It provides a more comprehensive evaluation of clinical practice than could typically be provided by one source.[13] Taking into account the heterogeneity of the questionnaires used in the MSF literature, it broadly examines professionalism, clinical competence, communication, management and interprofessional relationships. A study by Ahmed et al.[14] suggested that patient safety in IR relied on many elements including the competence of non-technical skills. The authors proposed NTS domains, which included knowledge, communication skills, cognitive skills, decision-making, mental readiness, ability to cope with stress and teamwork. Research that evaluates WBAs in IR is scarce, therefore surgery and anaesthesia are the closest that IR can build upon. There was a weak, statistically significant negative correlation between technical errors and NTS scores.[9] In surgery, one study investigated a tool designed to assess NTS and teamwork that demonstrated promising validity evidence.[8] A study by Gale et al.[15] introduced a scoring system exploring the relationship between performance at interviews for anaesthesia trainee selections and subsequent performance during training. It showed that NTS rating at appointment interview was a reasonable predictor of workplace performance during the first year of appointment. A study by Crossingham et al.[10] subsequently adopted this scoring system to score anaesthesia trainee NTS in theatre. The descriptors of the score domains observed during in-theatre assessments specifically include situational awareness and working under pressure, which were similar to those described in the study by Ahmed et al. but not traditionally included in MSF. Developing a specific method to capture those parameters would help correlate patient safety and competency, leading to safer procedures.

Internal consistency and inter-rater variability of assessment domains

A questionnaire-based feedback assessment, similar to an MSF, is proposed to assess NTS in IR. To ensure content validity, the intended format will be based on the one outlined by Crossingham et al.[10] Table 1 shows the proposed NTS assessment domains and their statistical characteristics. No data were available for the 'Empathy and Sensitivity' and 'Working under Pressure' sections. Cronbach's alpha (α) is a measure of the internal consistency of a test or scale, with a minimum of 0 and a maximum of 1 (representing strong consistency). It cannot be interpreted only with its numerical values. Detailed discussion of its use is beyond the scope of this article. A study by Graham et al.[11] examined the internal consistency of descriptors used to assess NTS in Australia (a similar model to the United Kingdom), which demonstrated α values between 0.78 and 0.87 to be consistent with good reliability.

In an ideal assessment, the difference in scores among candidates would depend on trainee ability. However, there are a number of 'unwanted' factors that make substantial contributions to scores. Some assessors (doves) consistently use the high end of the scale and some (hawks) consistently use the low end.[16] The estimated variance component and the percentage of total variance are reflections of the intraclass correlation coefficient (ICC), which describes inter-rater agreement.[17] A study by Vassiliou et al.[18] showed an ICC of more than 0.8 among scores could be attributed to true variance among subjects. However, the ICC values could be affected by ranges and slopes of the data and differ according to the different models, types and

Table 1. Non-technical skills assessment in Interventional Radiology: test domains and their reliability.

Domains[10]	Internal consistency (expressed in values of Cronbach's α)[11,16]	Inter-rater agreement (represented by values of intraclass correlation coefficient)[16]
Communication (including verbal and non-verbal communication)	NA	0.77
Organisation and planning/task management	0.85	0.73
Situation awareness and decision-making	0.78–0.87	0.64
Team working	0.85	0.71

measures.[19] A study by Hull et al.[20] assessed the reliability of similar domains to Crossingham et al. The author found that ICC values were between 0.71 and 0.77 for communication, leadership and cooperation and an ICC of 0.64 for situational awareness. The ICC values were below the 0.8 threshold described by Vassiliou et al.[19]

Studies that focus on emergency resuscitation reported more promising results in the assessment of NTS.[8] When compared with results reported by Graham et al.[11] and Hull et al.,[20] Walker et al. claimed values of Cronbach's α between 0.74 and 0.97, suggesting better consistency, with ICC values between 0.65 and 0.91.

To summarise, it is important to acknowledge that an assessment tool cannot be valid unless it has good internal consistency and low inter-rater disagreement. However, the reliability of an instrument does not depend on its validity.[21] In the coming sections, other factors potentially affecting validity are considered before meaningful interpretation of the described statistical measurement.

Inter-rater variability of the assessment items

In the surgical literature, an NTS assessment tool known as OTAS has been reported to demonstrate validity evidence.[4] Sevdalis et al.[6] investigated the correlations between expert and novice raters' score for selected OTAS items and overall score, using Pearson correlations (r). Notable correlations existed between expert and expert raters' scores for 12 of 15 assessed items (r between 0.51 and 0.94) during different phases of an operation. In contrast, only 3 of 15 items have notable correlation between expert and novice raters' scores ($r = 0.52$ and 0.60). Similar patterns were demonstrated in overall scores. Some individual items appear to receive a wider variation of score than others. It led to the belief that fine tuning of these items may be required to increase reliability. Hull et al.[16] investigated which specific items are associated with higher variation, using a Cohen's kappa (k) value of 0.4 as an absolute cut off. The author described reasonable observer agreement (k at least 0.41) for 84% (109 of 130 items), and 16% (21 of 130 items) were deemed problematic ($k < 4$). As a result, the items of concern were either removed or modified. Examples of items considered to have a poor level of agreement between assessors include the following:[16]

- Communication domain – Nurse: Scrub nurse repeats surgeon's request, confirming requirements;
- Cooperation domain – Anaesthetist: Operating department practitioner provides assistance to anaesthetist;
- Monitoring domain – Surgeon: Check table positioning and positions of team members.

In order for the future IR specific questionnaire to be valid, it needs to be scrutinised and improved systematically. The questionnaire should be applied during a preliminary run to test the initial reliability. The results from the study would then be used to refine the questionnaire.

Assessor eligibility and training

Russ et al.[22] looked at the correlation between inexperienced and expert assessors at each training stage. Four of the OTAS domains (communication, cooperation, leadership and monitoring) showed a positive linear trend in standardised scores across all training stages, highlighting a correlation between reduced inter-rater variability and increasing number of observed cases.[22] Standardised scores were compared between psychologist and surgeon assessors across training stages and also overall. There was no significant difference between the two groups from different backgrounds. Training probably improved the reliability of scoring among both novice and expert assessors. Although using psychologists/non-clinical staff may be a feasible option in assessing NTS after training, there are other areas to consider before replacing Radiologists with behavioural scientists.

An NTS score invented by a team in Oxford demonstrated that training did make a difference in scoring ($t = -3.02$, $p = 0.005$).[9] How long does it take to improve assessor ability to accurately score candidates? A study by Baker et al.[23] reported an adequate level of inter-rater agreement for a behavioural marker system after eight hours of training, suggesting that an end ICC > 0.7 would be satisfactory.[23] The University of Aberdeen,[24] which developed NTS assessment in anaesthesia, achieved ICCs between 0.56 and 0.65 after four hours of training. Using a pre-recorded video as a standardised assessment, Graham et al. reported ICCs ranging between 0.11 and 0.62, despite eight hours of structured training, which included feedback and calibration. Non-clinical assessors, for example, psychologists, have demonstrated expert behaviours when they had observed a minimum of 40 cases.[6] Inexperienced non-clinical assessors who observed six cases were unreliable in their scoring. Within the NTS assessment literature among surgery and anaesthesia, a single day of training has been shown not to be adequate for inexperienced clinical assessors.[7,11]

A number of factors probably contributed to these low ICC scores. Misclassification might be a problem when assessors score 'doves' or 'hawks'. Graham et al.[11] recommended an average score could be released to assessors to offer better guidance. Mishra et al.[9] hinted at using observable and classifiable examples of behaviour, such as shouting at colleagues, as a negative modifier, to prompt a lower score. Further work is needed to justify the use of average scores. Another cause of low ICCs might be due to a lack of agreement on safety standards, for example, 'test ventilation'. Some anaesthetists believed it to be mandatory practice, whereas others considered it dangerous. This problem is directly relevant to the future of WBAs, as variation in safety beliefs within the IR community is common and must be considered.[25]

In order to improve OTAS reliability, assessors received 18 hours of comprehensive training, which included observation, calibration, real and video cases and debriefing.[22] At the early stages of training, ICCs ranged mostly from 0.5 to 0.6. Mean ICCs improved steadily through the middle stages of training, reaching levels of 0.6 to 0.7. At the end stage of training, mean ICCs were all above 0.7, except cooperation for which $\text{ICC} = 0.68$. This particular learning curve did not fit all the descriptor domains.[22] Communication, cooperation and leadership improved most rapidly after the middle stages of training. For situational awareness, calibration improved relatively steadily. Coordination, being a technical skill, rather than NTS, was also observed, but interestingly it did not improve ICCs over time.

Short-term training (more than a day) and coaching might be a way to improve reliability of the assessment, provided there is adequate training; both clinicians and non-clinicians would make appropriate assessors. Considering the inter-rater agreement variation (overall $\text{ICC} < 0.8$), NTS assessment in IR should be considered as an addition to existing assessments, rather than an individual 'high stakes' test.

Translating the surgery and anaesthesia literature

A study by Passauer-Baierl et al.[5] demonstrated that it was possible to translate OTAS for a different setting, in this case translating an English tool for German speakers. Inter-observer agreement was good for the majority of the items, and the problematic ones were changed and refined. Substantial agreement was found for 67.1% of the items ($k > 0.6$) and an acceptable level of reliability ($\text{ICC} > 0.72$).[5] To make these domains and items IR specific, modifications will be made and the IR subspecialty curriculum endorsed by the General Medical Council and the Royal College of Radiologists. An expert panel including radiologists, radiographers and radiology nurses, preferably involved in national training, will form a focus group to design a questionnaire. To increase sample size, any pilot study should involve both existing IR consultants as well as trainees, due to the relatively small number of IR trainees. Consultation should be held regarding the design of any questionnaire. Collated feedback will improve and refine the validity of assessments. To develop an IR-specific NTS assessment, the existing MSF and NTS assessment tool within other medical specialities will contribute valid methodology and help to reduce costs.

Scoring system

Using the framework proposed in a study by Crossingham et al.,[10] the scoring system would comprise a 4-point standardised anchored rating scale ($1 = $ poor; $2 = $ borderline; $3 = $ satisfactory; $4 = $ good, excellent) for each domain. To pass, a candidate needs a score of 3 or 4, and a score below this constitutes failure. This should be a system easily understood by both assessor and candidate. A system of specific and generic positive and negative modifiers may also guide assessors to mark any candidate as objectively as possible, which will reduce misclassification bias.[9,11] Training of assessors has been discussed in the previous section. Two assessors will rate a candidate independently consistently of the consultant interventional radiologist in charge and either the radiology scrub nurse or the fluoroscopy radiographer that assisted during the case.

Mini-CEXs offer poor discrimination between candidates because of wide inter-rater variability and the face-to-face nature of the assessment.[17] Mirroring the anaesthesia assessment design model, a paper-based assessment system would be used with candidates being allocated pre-prepared assessment packs.[26] These sealed envelopes would contain the questionnaires for the consultant, nurse or radiographer assessing the case, accompanied by instructions and a return envelope. Once a suitable case is identified, the trainee will hand the assessor the assessment pack before the procedure starts. To ensure standardisation of assessment, an index case, specific to training grade, would be used to assess candidates. Appropriate and feasible cases will be identified through discussions within the expert panel. To achieve IR subspecialty status, trainees are currently encouraged to undertake IR training between ST4 and ST6. Table 2 shows examples of

Table 2. Example cases for each trainee grade.

Grade	Index case
ST4	Performance of insertion of nasogastric tube under fluoroscopy guidance on a cooperative patient with no oesophageal stricture or airway concern
ST5	Performance of insertion of an eight-French locking abdominal drain in a cooperative patient with simple ascites and international normalised ratio < 1.4, under ultrasound guidance
ST6	Performance of insertion of a tunnelled right-sided internal jugular line using ultrasound and fluoroscopic guidance on an elective patient needing chemotherapy

potential cases and corresponding training grades. Vascular procedures generally involve more advanced skills, compared with non-vascular cases.

Conclusion

According to medical and non-medical literature, human performance measured by NTS assessment can be used to improve safety and quality of care. There is some suggestion that a lack of NTS may be associated with adverse outcomes. Accordingly, the IR community should learn from surgery and anaesthesia to minimise the deficiency of teamwork behaviour. The appraised assessment tools demonstrate good internal consistency but moderate interrater variability. Further discussion to identify different sources and threats to validity would help to establish NTS assessment in addition to other existing assessments in IR.

Declarations

Competing interests: None declared

Funding: None declared

Ethical approval: Not Required

Guarantor: CLP

Contributorship: CLP drafted the manuscript. All authors contributed to the manuscript.

Acknowledgements: The idea was originally developed while studying the CLI712 module as part of the Clinical Education Certificate at Plymouth University.

Provenance: Not commissioned; peer-reviewed by Elizabeth Chapman

References

1. Gaba D, Fish KJ and Howard SK. *Crisis management in anesthesiology*. New York: Churchill Livingstone, 1994.
2. Reason J. Human error: models and management. *BMJ* 2000; 320: 760–770.
3. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ* 2003; 37: 830–837.
4. Undre S, Healey AN, Darzi A and Vincent CA. Observational assessment of surgical teamwork: a feasibility study. *World J Surg* 2006; 30: 1774–1783.
5. Passauer-Baierl S, Hull L, Miskovic D, Russ S, Sevdalis N and Weigl M. Re-validating the Observational Teamwork Assessment for Surgery tool (OTAS-D): cultural adaptation, refinement, and psychometric evaluation. *World J Surg* 2014; 38: 305–313.
6. Sevdalis N, Lyons M, Healey AN, Undre S, Darzi A and Vincent CA. Observational teamwork assessment for surgery: construct validation with expert versus novice raters. *Ann Surg* 2009; 249: 1047–1051.
7. Yule S, Rowley D, Flin R, et al. Experience matters: comparing novice and expert ratings of non-technical skills using the NOTSS system. *ANZ J Surg* 2009; 79: 154–160.
8. Walker S, Brett S, McKay A, Lambden S, Vincent C and Sevdalis N. Observational Skill-based Clinical Assessment tool for Resuscitation (OSCAR): development and validation. *Resuscitation* 2011; 82: 835–844.
9. Mishra A, Catchpole K and McCulloch P. The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. *Qual Saf Health Care* 2009; 18: 104–108.
10. Crossingham GV, Sice PJ, Roberts MJ, Lam WH and Gale TC. Development of workplace-based assessments of non-technical skills in anaesthesia. *Anaesthesia* 2012; 67: 158–164.
11. Graham J, Hocking G and Giles E. Anaesthesia non-technical skills: can anaesthetists be trained to reliably use this behavioural marker system in 1 day? *Br J Anaesth* 2010; 104: 440–445.
12. Al Ansari A, Donnon T, Al Khalifa K, Darwish A and Violato C. The construct and criterion validity of the multi-source feedback process to assess physician performance: a meta-analysis. *Adv Med Educ Pract* 2014; 5: 39–51.
13. Bracken D, Timmreck CW and Churck AH. *Introduction: a multisource feedback process model*. San Francisco: Jossey-Bass, 2001.
14. Ahmed K, Keeling AN, Khan RS, et al. What does competence entail in interventional radiology? *Cardiovasc Intervent Radiol* 2010; 33: 3–10.
15. Gale TC, Roberts MJ, Sice PJ, et al. Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. *Br J Anaesth* 2010; 105: 603–609.
16. Weller JM, Jolly B, Misur MP, et al. Mini-clinical evaluation exercise in anaesthesia training. *Br J Anaesth* 2009; 102: 633–641.

17. Bould MD, Crabtree NA and Naik VN. Assessment of procedural skills in anaesthesia. *Br J Anaesth* 2009; 103: 472–483.
18. Vassiliou MC, Feldman LS, Andrew CG, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg* 2005; 190: 107–113.
19. Lee KM, Lee J, Chung CY, et al. Pitfalls and important issues in testing reliability using intraclass correlation coefficients in orthopaedic research. *Clin Orthop Surg* 2012; 4: 149–155.
20. Hull L, Arora S, Kassab E, Kneebone R and Sevdalis N. Observational teamwork assessment for surgery: content validation and tool refinement. *J Am Coll Surg* 2011; 212: 234–243 e1–e5.
21. Tavakol M and Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ* 2011; 2: 53–55.
22. Russ S, Hull L, Rout S, Vincent C, Darzi A and Sevdalis N. Observational teamwork assessment for surgery: feasibility of clinical and nonclinical assessor calibration with short-term training. *Ann Surg* 2012; 255: 804–809.
23. Baker D, Mulqueen C and Dismukes R. *Training raters to assess resource management skills*. Mahwah, NJ: Lawrence Erlbaum Associates, 2001.
24. Fletcher G, Flin R, McGeorge P, Glavin R, Maran N and Patey R. Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system. *Br J Anaesth* 2003; 90: 580–588.
25. Calder I and Yentis SM. Could 'safe practice' be compromising safe practice? Should anaesthetists have to demonstrate that face mask ventilation is possible before giving a neuromuscular blocker? *Anaesthesia* 2008; 63: 113–115.
26. Wilkinson JR, Crossley JG, Wragg A, Mills P, Cowan G and Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Med Educ* 2008; 42: 364–373.