British Journal of Anaesthesia **90 (5)**: 580–8 (2003)
DOI: 10.1093/bja/aeg112

BJA

CLINICAL INVESTIGATIONS

Anaesthetists' Non-Technical Skills (ANTS): evaluation of a behavioural marker system†

G. Fletcher[1], R. Flin[1]*, P. McGeorge[1], R. Glavin[2], N. Maran[2] and R. Patey[3]

[1]*Department of Psychology, University of Aberdeen, King's College, Aberdeen AB24 2UB, UK.*
[2]*Scottish Clinical Simulation Centre, Stirling Royal Infirmary, Livilands Gate, Stirling FK8 2AU, UK*
[3]*Department of Anaesthesia, Aberdeen Royal Infirmary, Foresterhill, Aberdeen AB25 2ZN, UK*

**Corresponding author. E-mail: r.flin@abdn.ac.uk*

Background. Non-technical skills are critical for good anaesthetic practice but are not addressed explicitly in normal training. Realization of the need to train and assess these skills is growing, but these activities must be based on properly developed skills frameworks and validated measurement tools. A prototype behavioural marker system was developed using human factors research techniques. The aim of this study was to conduct an experimental evaluation to establish its basic psychometric properties and usability.

Method. The Anaesthetists' Non-Technical Skills (ANTS) system prototype comprises four skill categories (task management, team working, situation awareness, and decision making) divided into 15 elements, each with example behaviours. To investigate its experimental validity, reliably and usability, 50 consultant anaesthetists were trained to use the ANTS system. They were asked to rate the behaviour of a target anaesthetist using the prototype system in eight videos of simulated anaesthetic scenarios. Data were collected from the ratings forms and an evaluation questionnaire.

Results. The results showed that the system is complete, and that the skills are observable and can be rated with acceptable levels of agreement and accuracy. The internal consistency of the system appeared sound, and responses regarding usability were very positive.

Conclusions. The findings of the evaluation indicated that the ANTS system has a satisfactory level of validity, reliability and usability in an experimental setting, provided users receive adequate training. It is now ready to be tested in real training environments, so that full guidelines can be developed for its integration into the anaesthetic curriculum.

Br J Anaesth 2003; **90**: 580–8

Keywords: anaesthetists; education, training

Accepted for publication: January 17, 2003

Non-technical skills have a vital role in anaesthetic practice but have not traditionally been addressed in anaesthetic training.[1] This situation is not unique to anaesthesia. By the early 1980s, the aviation industry had recognized that high technical proficiency in pilots was not enough to guarantee safety[2] and responded by introducing Crew Resource Management (CRM) training.[3] This was designed to enhance the performance of non-technical skills in everyday operations, providing 'a set of error countermeasures'.[4] In common with other high-reliability industries,[5] similar training programmes are now emerging in medicine.[6–10] However, CRM-style training in medicine must be underpinned by properly developed skills frameworks.

†*Declaration of interest*: The ANTS system was developed under research funding from the Scottish Council for Postgraduate Medical and Dental Education, now part of NHS Education for Scotland, through grants to the University of Aberdeen from September 1999 to August 2003. 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 body.

It is first necessary to identify the skills required for a specific job (and operational environment) using appropriate task analysis techniques.[11–13] It is also essential to be able to assess these skills explicitly to provide structured feedback about performance[14] and to allow training effectiveness to be evaluated.[15] To meet a similar need for objective and transparent methods assessing CRM (non-technical) skills, the aviation industry developed behavioural marker systems.[16 17] Behavioural markers are 'observable, non-technical behaviours that contribute to superior or substandard performance within a work environment'.[17] Derived from empirical data, they are usually developed into structured skill taxonomies and combined with a rating scale to allow the skills, which are demonstrated through behaviour, to be assessed by trained, calibrated raters. In addition to providing a tool for assessing aspects of performance traditionally judged on gut feeling, behavioural marker systems supply a common language for discussing non-technical skills and can function as frameworks to structure teaching and debriefing.

Behavioural markers are already being used in the medical domain.[10 18 19] However, these tools have mainly been developed outside the UK and from existing aviation systems, e.g. the Line/Line Operational Simulation Checklist (LLC),[20] for specific purposes, e.g. to investigate team performance,[20] and to measure particular aspects of performance, e.g. crisis management.[19] Cultural differences at the organizational, professional or national level have been found to have a considerable impact on crew resource management attitudes and behaviour,[20] and so should be taken into account when developing a behavioural marker system. Until now, there have been very few attempts to design a marker system for anaesthetists' non-technical skills from first principles, based on a systematic analysis of their task requirements, and none of these have been conducted in the UK. Fewer studies still have sought to evaluate empirically the measurement properties of such behavioural marker systems, yet unless the behavioural marker system is valid and reliable it has little value as an assessment tool.[17 21] The aim of this study was to investigate the experimental validity, reliability and usability of the Anaesthetists' Non-Technical Skills (ANTS) system.

Method

Prototype behavioural marker system

The ANTS system prototype was developed using psychological research techniques to identify the skills and structure them into a meaningful hierarchy.[22 23] The results of a literature review[2] identified six existing behavioural marker systems currently being used in anaesthesia (and emergency medicine).[7 10 18 19 24 25] These did not fit the requirements for this project (e.g. they were team scales rather than individual ones), but they were examined to establish their structure and content. Their individual skills were extracted and grouped by common themes to guide subsequent data collection activities. Cognitive task analysis interviews[13] were conducted with 29 consultant anaesthetists, who were asked to recall and describe their management of a particularly challenging case or critical incident[26 27] and to list the skills they thought were important for good practice in anaesthesia. The interview data were analysed using a grounded theory approach,[28] to identify the non-technical skills. A prototype taxonomy was developed during workshop discussions with the project team. The hierarchical structure of the prototype was based on the framework of the European aviation marker system NOTECHS,[29] and practical criteria established by the project team (three psychologists and three consultant anaesthetists). The initial prototype was then refined by re-coding a sample of interviews, reviewing anaesthesia incident reports, and from observations in theatre. Examples of good and poor behaviour were included for each skill, and were written as action statements. This process is summarized in Fig. 1. The results from a study of UK anaesthetists' attitudes to safety and teamwork[30] were also taken into account.

The main structure of the ANTS system prototype is shown in Fig. 2. Some skills identified from the interviews as being important, e.g. stress management and level of self-composure, were excluded from the taxonomy because they would not be easily identifiable through observation of behaviour.[22 23] Furthermore, the reliance of the ANTS system on communication to detect the skill elements means that the ANTS system does not contain 'communication' as a separate category or element. A preliminary evaluation of the prototype by 11 consultant anaesthetist simulator instructors (pilot study) confirmed that the prototype appeared to be complete and the skills could be identified through observation as required. The prototype was therefore ready for a formal evaluation study to establish experimentally its basic validity, reliability and usability.

Design

An experimental method adapted from an earlier behavioural marker system evaluation was used.[31] This had been tested previously for anaesthetists in the pilot study. A number of specific experimental hypotheses were developed and used to drive the data collection and analysis process (Table 1). The design of the study required trained participants to watch videos of scripted anaesthetic situations and to rate the non-technical skills of the main anaesthetist in each scenario using the ANTS system.

The video scenarios were written in advance to ensure all the elements were portrayed at varying levels of performance, and were filmed in a high-fidelity patient simulator with practising anaesthetists, surgeons and theatre staff acting the main roles. Ten scenarios were produced (eight

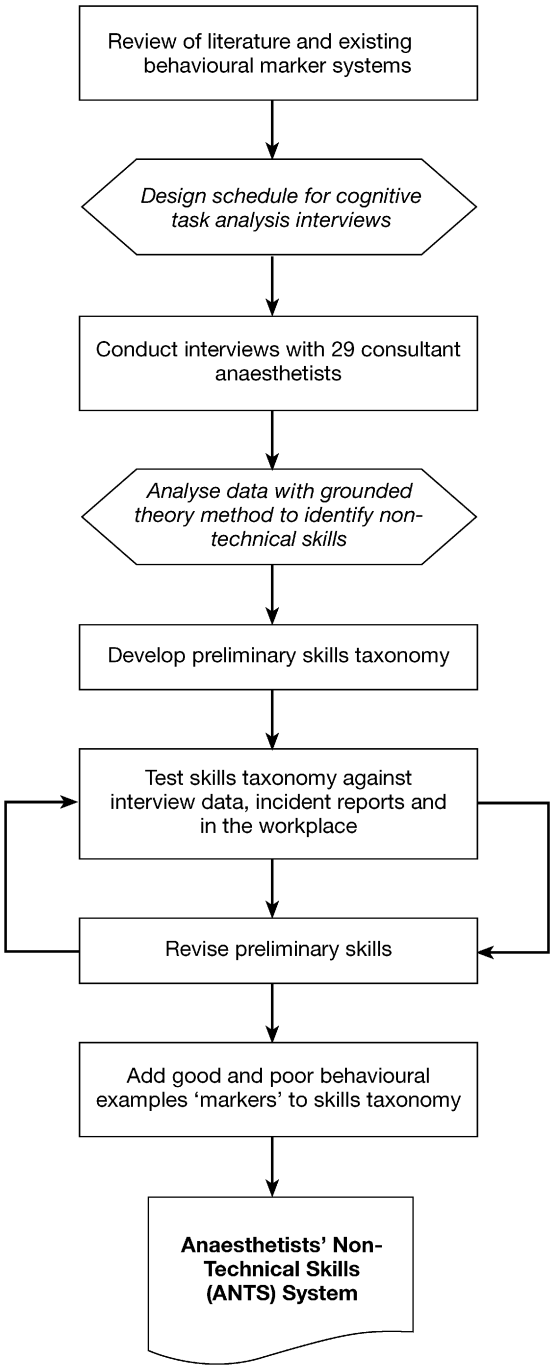

Fig 1 The ANTS system development process.

test and two practice), ranging from 4 to 21 min. The scenarios showed a variety of anaesthetic activities in different circumstances being undertaken by both trainees and consultants. A definitive rating for each of the non-technical skills demonstrated in the scenario was obtained from the three project anaesthetists, who rated the scenarios individually and then discussed their ratings to produce an agreed reference rating. These ratings were used as a benchmark in subsequent data analysis.

It is widely recognized that raters need to be trained in order to assess non-technical skills.[17 36 37] This is particularly important for the ANTS system, where users do not have knowledge of the system and are not experienced in making explicit assessments of non-technical skills. While necessarily constrained because of time availability, the training package was developed to address components previously established as being effective with behavioural performance measures.[36] Training was provided by a psychology researcher with assistance from a consultant anaesthetist. The course consisted of: (i) background on human factors and non-technical skills, including information about human error, threat management and crew resource management training; (ii) an introduction to the ANTS system and how to make behavioural assessments, which included detailed descriptions of the categories, elements and behavioural markers, supported by showing video snippets of examples; and (iii) instructions for rating non-technical skills, possible biases to avoid, and practice in scoring two scenarios with the full system and rating scale. Importantly, participants were told that, while the layout of the categories and elements in the table may suggest a temporal sequence, this is not necessarily meant to reflect an ordering priority when making their observations. Participants were sent a booklet describing the full ANTS system in advance and were able to use this for reference throughout the evaluation. No attempt was made to calibrate the raters to a standard scoring. Not only would this have taken a considerable amount of time, but it would also have resulted in an evaluation of the calibration process and the ANTS system, not just of the ANTS system. Hence a calibration phase was excluded to prevent compromising the data.

Materials for data collection were a set of rating forms and an evaluation questionnaire. The rating forms showed the ANTS system elements and categories with a four-point scale (Fig. 3). Each point of the scale had a descriptor to provide guidance on when it should be used. An additional rating option of 'not observed' was provided for when the skills could not be identified in a particular situation, because either they did not need to be used or they could not be detected from behaviour. Separate element and category rating forms were supplied for each scenario. The evaluation questionnaire was divided into five parts plus an 'other comments' section: (i) 10 general questions about the completeness and design of the system and the observability of the non-technical skills; (ii) four questions asking about the rating scale; (iii) five questions about the training; (iv) three questions about the video scenarios; and (v) five questions about the role of the ANTS system. A separate background information questionnaire was used to collect data on experience as a consultant anaesthetist, involvement in training, and assessment and knowledge of human factors.

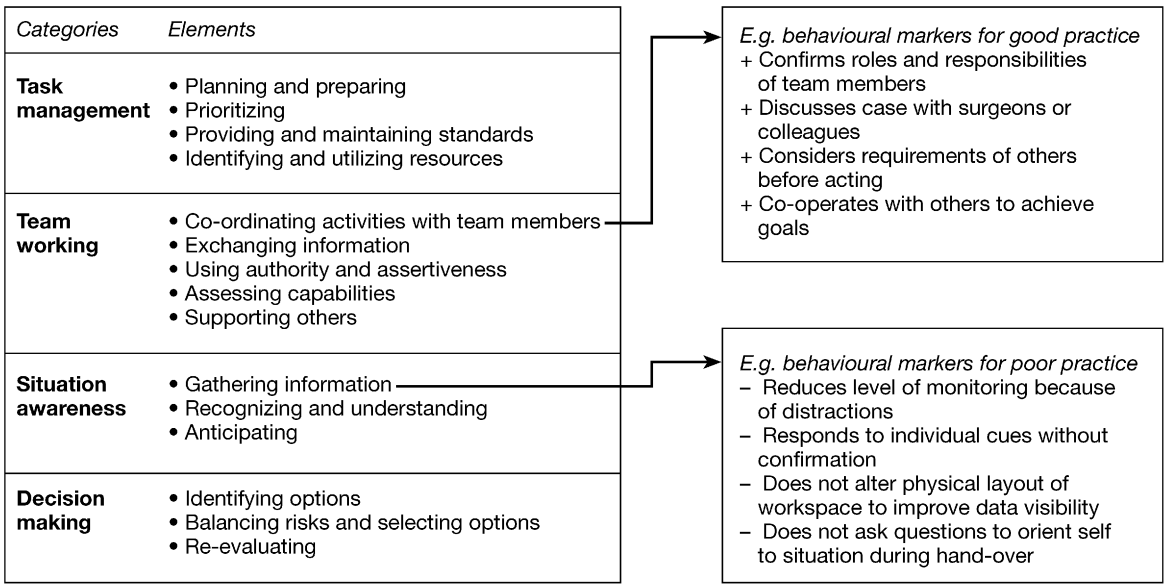

Fig 2 The ANTS system prototype.

Table 1 Evaluation hypotheses, data sources and analysis techniques

Evaluation criterion	Hypothesis	Data source and analysis
Validity		
Completeness	The ANTS system provides a suitably comprehensive set of categories and elements to describe anaesthetists' non-technical skills	Questionnaire data: basic frequency analysis and content review to identify any superfluous or missing elements
Observability	Anaesthetists' non-technical skills can be identified by observation of behaviour using the ANTS system	Ratings data: basic descriptive statistics and χ^2 tests to establish the extent to which non-technical skills were observed *vs* not observed. Questionnaire data: frequency analysis, content review and *t* tests where appropriate
Reliability		
Inter-rater agreement	Using the ANTS system to rate non-technical skills, participants will achieve inter-rater agreement at (a) category level and (b) element level consistent with recognised criteria for acceptance	Ratings data: within-group inter-rater agreement statistic[32 33] to show the level of rater consensus (i.e. whether they rate performances the same): $r_{wg}=1-(S\chi^2/\sigma_E^2)$ where $S\chi^2$=variance of observed ratings and σ_E^2=population variance for a discrete rectangular distribution of ratings (i.e. it represents a random response where each scale point would have an equal number ratings). This is calculated as $\sigma_E^2=(A^2-1)/12$, where A is the number of points on the scale
Accuracy/sensitivity	Category and element ratings given by participants will be consistent with reference ratings agreed by a panel of experts	Ratings data: mean absolute deviation (MAD) from the reference ratings[34 35] and basic difference from reference ratings to establish the level of accuracy or error for ratings
Internal consistency	The ANTS system has an acceptable level of internal consistency between the categories and their elements.	Ratings data: Cronbach α coefficient for correlation between elements within a category and Pearson reliability coefficient for mapping of elements to categories
Usability		
Acceptability	The ANTS system is an acceptable tool for (a) training and (b) assessing non-technical skills in anaesthesia	Questionnaire data: basic descriptive statistics and content review to establish the level of acceptance for different uses of the system
Usability	The ANTS system is straightforward for anaesthetists to use to rate non-technical skills	Questionnaire data: basic descriptive statistics and content review. Ratings data: overall indication of effective use of the system

Participants

Participants were 50 consultant anaesthetists involved in training and assessment from 17 hospitals across Scotland, who attended one of the eight 1-day sessions held for the study. Numbers of participants at each session varied from two to 10.

Procedure

Each session consisted of approximately 4 h of training, as described above, followed by 3 h for rating the eight experimental video scenarios. As practised in the training phase, ratings were made of the non-technical skills of the main anaesthetist in each scenario using the ANTS system

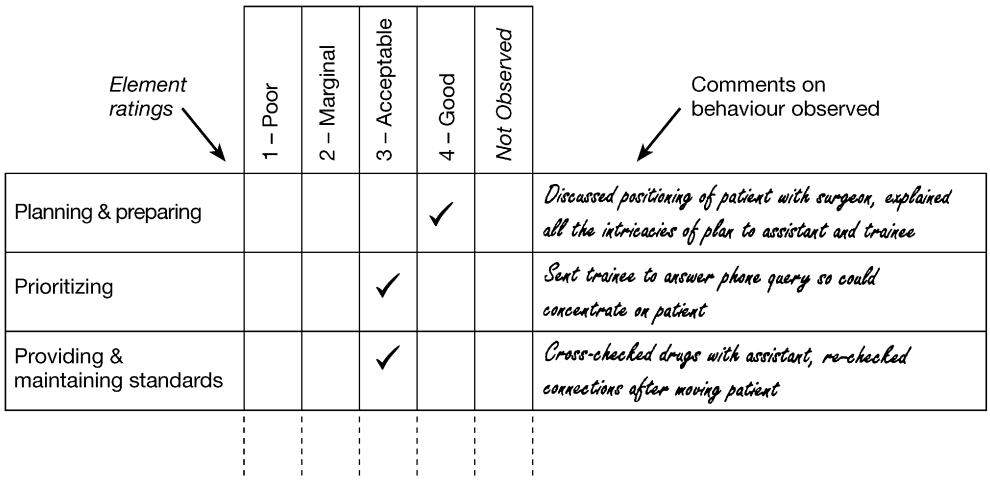

Fig 3 Example of ANTS system rating form.

rating forms. Participants were instructed to watch the scenario first, if necessary making notes of key behaviours observed, and, once the scenario was over, to rate the observed elements on the rating form. Having scored the element, participants rated the higher-level ANTS categories. All ratings were made individually and participants were not permitted to discuss their scores with others. At the end of the session, participants completed the evaluation questionnaire.

Data from the rating forms were transferred into SPSS (Statistical Package for Social Sciences; SPSS, Chicago, IL, USA) and data from the questionnaires into Microsoft Access and Excel. A number of analyses were conducted on the ratings data according to the hypothesis being tested (Table 1). The nature of the ratings data was such that for most of the analyses each scenario was examined separately (scores were expected to vary across the different scenarios and so averaging them would render them meaningless). To provide an overall result to test the hypotheses, an average was taken of the results from each of the eight scenarios.

Results

The consultancy experience of participants ranged from 1 to more than 25 yr (mean 8 yr). Their involvement in anaesthetic training varied from general supervision of trainees to specific duties such as being college tutor and a simulator instructor. A total of 67% reported some involvement in assessment, but only 42% of these had received special training for this task. Some participants had been involved in research at earlier stages of this project (e.g. being interviewed), but for 72% of participants the study was their first exposure to the ANTS system and its concepts.

As a considerable amount of data was analysed,[38] only the key findings for each evaluation criteria (validity, reliability and usability) are described. These are shown in Tables 2–4.

Discussion

The ANTS system[22 23] was designed to describe the main non-technical skills that are important for good anaesthetic practice. It was therefore necessary to test whether the system was suitably comprehensive and whether the skills are observable. The results from the evaluation suggest that the system does capture the most important non-technical skills (Table 2). Concerns about complexity were felt to be due to limited familiarity with the system and will be addressed through further training. To indicate that 'personal factors' (e.g. stress management and self-presentation) are recognized as being important, behavioural markers illustrating how they can affect various elements will be added. All 15 elements could be observed at various levels of performance, as could their overarching categories. Some elements were more difficult to recognize than others, but this may also have been due to lack of familiarity and also to the design of the scenarios. Generally it is anticipated that, if non-technical skills are being observed in real training settings, there will be more opportunities to see if appropriate behaviour is being demonstrated. However, it is important to accept that in some circumstances it will not be possible to observe certain skill elements, either because the situation does not require them or because the associated behaviour is so subtle it goes unnoticed. Other aspects of validity are dependent on reliability, as discussed below, and ultimately on use in real training environments. Only when data are available on predicative power and concordance with other measures of performance can the full operational validity of the ANTS system be established.

The reliability of the prototype was investigated from a number of perspectives. In the ANTS system, it was expected that the elements within each category would be closely related to each other (internal consistency) and that the individual elements would be related to their own categories better than to other categories. The results from

Table 2 Summary of results for validity

Evaluation criteria	Results
Completeness of ANTS system	(1) Did it address the key NTS behaviours displayed? *n*=50 **Yes=100%** (2) Do you think any elements or categories are missing? *n*=50 **No=84%** Yes=8% Comment only=8% (3) Do you think any elements or categories are superfluous? *n*=46 **No=81%** Yes=17% Comment only=2%
Observability of NTS	(1a) Averaged across all scenarios NTS observability was good, ranging from 100% (gathering information, recognizing and understanding) to 66% (assessing capabilities). Overall, 13 elements observable >80%, and all categories were observable >95% (Appendix 1) (1b) Across all scenarios, only 5% of ratings (8 out of 152) showed no difference between use of 'observed' *vs* 'not observed', i.e. χ^2 was not significant (2a) How easy was it to relate behaviours to elements? *n*=50 **Average to Easy=78%** Difficult to Very difficult=22% (2b) How easy was it to relate behaviours to categories? *n*=50 **Very easy to Average=82%** Difficult to Very difficult=18% Categories were significantly easier to relate to behaviours than elements (t=0–3.06, P<0.05)

Table 3 Summary of results for reliability

Evaluation criteria	Results
Inter-rater agreement	(1) At element level r_{wg}=**0.55–0.67**, 'recognising and understanding' being lowest and 'identifying and utilising resources' being highest (2) At category level r_{wg}=**0.56–0.65**, Situation Awareness being lowest and Task Management and Team Working being joint highest (Appendix 1) Overall Situation Awareness and its individual elements were associated with the lowest levels of agreement. Values of r_{wg} varied considerably across scenarios
Accuracy	(1) Accuracy good as measured by rater-reference rater difference: **>88% accuracy to 1 scale point** (2) Mean absolute deviation from the reference i.e. error score **0.49–0.84**, which, while showing significant difference between elements, suggests only minor differences arising across boundaries (Appendix 1)
Internal consistency	(1) Correlations were strongest, indicating the **'best fit', for existing category-element grouping for 13 of the 15 elements**; for the remaining two, mapping was not the highest on 2–3 scenarios (2) Consistency between elements in each category using Cronbach α ranged from **0.79 to 0.86**, which, while reasonably high, suggests commonality but not duplication

the study support this and show that the structure of the ANTS system appears sound.

Inter-rater reliability is of particular concern amongst practitioners. In spite of their limited familiarity with the ANTS system, participants were still able to use the system with a reasonable level of agreement. Indeed, the levels reached were higher than might have been expected, especially with such a large sample. Across the whole system, Situation Awareness showed a slightly lower level of agreement than other categories. This is not surprising as it is a cognitive skill that makes observable behaviours more difficult to detect, and it is not a concept currently described in UK anaesthetic training. The levels of rater agreement for the ANTS system shown in this preliminary test exercise are obviously not as high as recommended for trained non-technical skill assessors.[37] Nevertheless, they provide a good indication of the basic reliability afforded by the system with minimal training (half a day). When users become more familiar with both the system and the rating task, inter-rater reliability can be expected to improve. Research into behavioural marker systems has shown that, with comprehensive training and calibration (2–3 days for people already familiar with human factors concepts), inter-rater agreement can be increased to above 0.7.[37] For the ANTS system, it will be important to establish the amount of training that is practical whilst allowing users to develop the skills to use the tool at the necessary level of inter-rater reliability.

The last psychometric property investigated in this evaluation is rater accuracy. This is the degree of raters' agreement with the baseline reference. The levels of accuracy achieved by the consultants using the ANTS system were acceptable; that is, averaged across scenarios, 88–97% of raters matched the reference rating to within 1 scale point (Table 3). The participants suggested that limitations in accuracy occurred as a consequence of not

Table 4 Summary of results for usability

Evaluation criteria	Results
Acceptability	(1) Was the ANTS system useful for structuring observations? *n*=50 **Yes=100%** (2) Would the ANTS system be helpful for consultants giving training to junior anaesthetists? *n*=50 **Yes=94%** (*describes skills/gives framework, useful for problem trainees*) Comment only=6% *(more evidence and familiarity needed*) (3) Would the ANTS system be helpful for consultants assessing junior anaesthetists? *n*=50 **Yes=78%** (*gives it structure/puts it into words; some with caveats, e.g. training, validation*) No=8% (*informal feedback/self-assessment; not good enough at assessment*) Comment only=7% (*validation, more familiarity practice*) (4) Do you think the ANTS system could be used to support in-theatre teaching? *n*=50 **Yes=94%** (*highlights important skills, to help with observation/ giving feedback*) No=4% Comment only=2%
Design of ANTS system	Was the wording used for category and element labels meaningful? *n*=48 **Yes=98%** No=2% (2) Were the descriptions clear? *n*=48 **Yes=96%** No=4% (3) Were the 'good' behavioural markers helpful? *n*=50 **Yes=96%** Comment only=4% (4) Were the 'poor' behavioural markers helpful? *n*=48 **Yes=92%** Comment only=8% (5) Do you think the rating scale gave you enough flexibility to rate the performance levels seen? *n*=50 **Yes=94%** No=6% *Comments varied across both groups, some preferring a longer scale, some a shorter, some wanting a mid-point and others not*

knowing where to set the boundaries for each scale point and should therefore be resolvable with training and calibration. A previous study identified difficulties for intra-rater reliability when raters had to average performance across longer periods.[19] This could yet be encountered when the ANTS system is used in real training situations.

The final test of the ANTS system relates to its usability. This includes its acceptability to anaesthetists, which is very important, given that the ANTS system is being developed primarily as a tool for them to use during training. The results from the evaluation questionnaire were extremely encouraging (Table 4). Anaesthetists reported that the process of rating both elements and categories was useful, and the behavioural markers were helpful for this process. There were no major problems with the layout and language of the prototype. One of the most important findings from the evaluation questionnaire was that the participants clearly thought the ANTS system could be used to support regular training of non-technical skills, both at the simulator and in routine hospital-based teaching. They also thought that it could be used for assessing junior anaesthetists but with the requirement that adequate training should be given to the assessors. Overall, the respondents recognized that the ANTS system addressed an important area of anaesthetic practice that is currently not well explained, and they did not appear to have any major problems.

From the results of this evaluation, the ANTS system appears not only to have a high level of acceptability but also to provide a reasonable level of reliability and accuracy when used by anaesthetists in an experimental setting to rate non-technical skills demonstrated in simulator scenarios. There were some limitations in the study. The first is the use of scripted videos viewed in a controlled setting rather than live anaesthetic situations. Hence the current results refer to experimental evaluation and not to real-world testing. The second is that the participants in the study could only be given limited training and were not permitted to calibrate their ratings. This was reflected by their feelings of unfamiliarity with the system, the presence of boundary errors, and the level of rater agreement. Nonetheless, taken together, these experimental results show that the ANTS system has a satisfactory basic level of validity, reliability and usability. Once field testing has been undertaken and proper guidelines for its implementation have been produced, the ANTS system can become an important

Table 5 Results for observability, inter-rater agreement and accuracy averaged across all scenarios. [a]High percentages indicate a good level of observability; [b]r_{wg}=1 represents perfect agreement, r_{wg}=0 represents no agreement; [c]low numbers indicate a low error rate and good accuracy

ANTS	Observability (% of 'observed' ratings)[a]	Inter-rater agreement (r_{wg} scores)[b]	Accuracy (% ratings accurate ±1 scale point)	Accuracy (mean absolute difference)[c]
Elements				
Planning and preparing	83	0.64	96	0.49
Prioritizing	95	0.59	90	0.79
Providing and maintaining standards	90	0.58	91	0.84
Identifying and utilizing resources	85	0.67	94	0.50
Coordinating with team	98	0.62	94	0.62
Exchanging information	98	0.58	93	0.54
Using authority and assertiveness	94	0.63	96	0.48
Assessing capabilities	66	0.61	88	0.73
Supporting others	70	0.66	95	0.55
Gathering information	100	0.58	94	0.53
Recognizing and understanding	100	0.55	93	0.58
Anticipating	95	0.56	92	0.71
Identifying options	91	0.61	97	0.55
Balancing risks and selecting options	87	0.62	96	0.55
Re-evaluating	93	0.57	93	0.60
Categories				
Task management	99	0.65	94	0.60
Team-working	99	0.65	92	0.57
Situation awareness	100	0.56	92	0.58
Decision-making	98	0.61	95	0.61

tool for non-technical skills training in anaesthesia, supporting non-technical skills training and simulator-based human factors courses. It could also be used as a measure to allow the effectiveness of such training to be evaluated.[39] Long-term feedback on the use of the system will allow broader conclusions to be drawn about its operational validity, and from these it will be possible to make recommendations about more advanced use.

Appendix 1

The results for observability, inter-rater agreement and accuracy averaged across all scenarios are given in Table 5.

References

1 Fletcher GCL, McGeorge P, Flin RH, Glavin RJ, Maran N. The identification and measurement of anaesthetists' non-technical skills: a review of current literature. *Br J Anaesth* 2002; **88**: 418–29

2 Helmreich R. Managing human error in aviation. *Sci Am* 1997; **276**: 40–5

3 Weiner E, Kanki B, Helmreich R, eds. *Cockpit Resource Management*. San Diego, California: Academic Press, 1993

4 Helmreich RL. The evolution of Crew Resource Management. Paper presented at the IATA Human Factors Seminar, October 31, 1996, Warsaw, Poland; 5

5 Flin R, O'Connor P, Mearns K. Crew Resource Management: enhancing team performance in high reliability industries. *Team Perform Manag* 2002; **8**: 68–78

6 Pizzi L, Goldfarb N, Nash D. Crew Resource Management and its applications in medicine. In: Shojania K, Duncan B, McDonald K, Wachter R, eds. *Making Health Care Safer: A Critical Analysis of Patient Safety Practices*. Washington, DC: Agency for Healthcare Research and Quality, 2001; Chapter 44: 511–19

7 Howard SK, Gaba DM, Fish KJ, Yang GS, Sarnquist FH. Anesthesia crisis resource management training: teaching anesthesiologists to handle critical incidents. *Aviat Space Environ Med* 1992; **63**: 763–70

8 Maran N, Glavin R, Fletcher GCL. Training in human factors for anaesthetists in Scotland: identifying key skills and developing a training programme. *Proceedings of the 7th European Forum on Quality Improvement in Health Care, Mar 21–23 2002, Edinburgh, Scotland*. London: BMJ Publishing Group, 2002

9 Sexton B, Marsch S, Helmreich R, Betzendoerfer D, Kocher T, Scheidegger D. Participant evaluation of Team Oriented Medical Simulation, In: Henson LC, Lee AC, eds. *Simulators in Anesthesiology Education*. New York: Plenum, 1998: 107–8

10 Small SD, Wuerz RC, Simon R, Shapiro N, Conn A, Setnik G. Demonstration of high-fidelity simulation team training for emergency medicine. *Acad Emerg Med* 1999; **6**: 312–23

11 Kirwan B, Ainsworth LK, eds. *A Guide to Task Analysis*. London: Taylor & Francis, 1992

12 Seamster TL, Kaempf GL. Identifying resource management skills for airline pilots. In: Salas E, Bowers C, Edens E, eds. *Improving Teamwork in Organisations: Applications of Resource Management Training*. Mahwah, New Jersey: Laurence Erlbaum, 2001; 9–30

13 Seamster TL, Redding RE, Kaempf GL. *Applied Cognitive Task Analysis in Aviation*. Aldershot: Avebury Aviation, 1997

14 Greaves JD, Grant J. Watching anaesthetists work: using the professional judgement of consultants to assess the developing clinical competence of trainees. *Br J Anaesth* 2000; **84**: 525–33

15 Murray E, Gruppen L, Catton P, Hays R, Woolliscroft JO. The accountability of clinical education: its definition and assessment. *Med Educ* 2000; **43**: 871–9

16 Flin R, Martin L. Behavioural markers for Crew Resource Management: a survey of current practice. *Int J Aviat Psychol* 2001; **11**: 95–118

17 Klampfer B, Flin R, Helmreich RL, *et al*. Enhancing performance in high risk environments: recommendations for the use of

behavioural markers. Ladenburg: Daimler-Benz Shiftung, 2001. Downloadable version available from: www.psyc.abdn.ac.uk/serv02: 10

18 Helmreich RL, Schaefer H-G, Sexton JB. Operating room checklist. Aerospace Crew Resource Project Technical Report 95–4. Austin, Texas: University of Texas, 1995

19 Gaba DM, Howard SK, Flanagan B, Smith BE, Fish KJ, Botney R. Assessment of clinical performance during simulated crises using both technical and behavioral ratings. *Anesthesiology* 1998; **89**: 123–47

20 Helmreich RL, Merritt AC. *Culture at Work in Aviation and Medicine*. Aldershot: Ashgate, 1998

21 Holt RW, Boehm-Davis DA, Beaubien JM. Evaluating resource management training. In: Salas E, Bowers C, Edens E, eds. *Improving Teamwork in Organisations: Applications of Resource Management Training*. Mahwah, New Jersey: Laurence Erlbaum, 2001; 165–88

22 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. *Cognition Technology and Work*. (Submitted)

23 Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Final report: the identification and measurement of anaesthetists' non-technical skills. University of Aberdeen Grant Report for SCPMDE, 2001. Aberdeen: University of Aberdeen, 2001

24 Altmaier EM, From RP, Pearson KS, Gorbatenko-Roth KG, Ugolini KA. A prospective study to select and evaluate anesthesiology residents: phase 1, the critical incident technique. *J Clin Anesth* 1997; **9**: 629–36

25 Department of Anaesthesia, University of Basel-Kantonspital. Kommunikations-Status (KOMSTAT), Operationssal-Beobachtungen, Ver. 2.0, 07/97. Personal communication, 2000; available from Swiss anaesthesia server: http://www.medana.unibas.ch [in German]

26 Flanagan JC. The Critical Incident Technique. *Psychol Bull* 1954; **51**: 327–58

27 Klein GA, Calderwood R, MacGregor D. Critical decision method for eliciting knowledge. *IEEE Trans Syst Man Cybern* 1989; **19**: 462–72

28 Strauss A, Corbin J. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Newbury Park, California: Sage Publications, 1990

29 Flin R, Goeters KM, Hormann J, Martin L. A generic structure of non-technical skills for training and assessment. *Proceedings of 23rd Conference of the European Association for Aviation Psychology, Sept 14–18, 1998, Vienna, Austria*. Vienna, 1998

30 Flin R, Fletcher G, McGeorge P, Sutherland A, Patey R. Anaesthesists' attitudes to teamwork and safety. *Anaesthesia* 2003; **58**: 233–42

31 O'Connor P, Hörmann H-J, Flin R, Lodge M, Goeters K-M, JARTEL Group. Developing a method for evaluating Crew Resource Management skills: a European perspective. *Int J Aviat Psychol* 2002; **12**: 263–85

32 James LR, Demaree RG, Wolf G. Estimating within-group interrater reliability with and without response bias. *J Appl Psychol* 1984; **69**: 85–98

33 James LR, Demaree RG, Wolf G. r_{wg}. An assessment of within-group interrater agreement. *J Appl Psychol* 1993; **78**: 306–9

34 Johnson PJ, Goldsmith TE. The importance of quality data in evaluating aircrew performance. US Federal Aviation Authority Technical Report, 1998. Available from Federal Aviation Authority website: www.faa.gov/avr/afs/aqphome

35 Goldsmith TE, Johnson PJ. Assessing and improving evaluation of aircrew performance. *Int J Aviat Psychol* 2002; **12**: 223–40

36 Baker DP, Mulqueen C, Dismukes RK. Training raters to assess resource management skills. In: Salas E, Bowers C, Edens E, eds. *Improving Teamwork in Organisations: Applications of Resource Management Training*. Mahwah, New Jersey: Laurence Erlbaum, 2001; 131–45

37 Williams DM, Holt RW, Boehm-Davis DA. Training for inter-rater reliability: baselines and benchmarks. *Proceedings of the 9th International Symposium on Aviation Psychology, Apr 27–May 1, 1997, Columbus, Ohio*. Columbus, Ohio: Ohio State University, 1997: 514–20

38 Fletcher G, Flin R, McGeorge P, Glavin R, Maran N, Patey R. Evaluation of the prototype Anaesthesists Non-Technical Skills (ANTS) behavioural marker system: WP7 Experimental Report. University of Aberdeen Technical Report for NHS Education for Scotland, 2002. Aberdeen: University of Aberdeen, 2002

39 Byrne AJ, Greaves JD. Assessment instruments used during anaesthetic simulation: review of published studies. *Br J Anaesth* 2001; **86**: 445–50