Original article

Prospective observational study to evaluate NOTSS (Non-Technical Skills for Surgeons) for assessing trainees' non-technical performance in the operating theatre

J. Crossley[1], J. Marriott[2], H. Purdie[3] and J. D. Beard[4]

[1]Academic Unit of Medical Education and [2]Department of Reproductive and Developmental Medicine, University of Sheffield, and [3]Clinical Research Facility and [4]Sheffield Vascular Institute, Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK
Correspondence to: Professor J. D. Beard, Sheffield Vascular Institute, Northern General Hospital, Sheffield S5 7AU, UK
(e-mail: Jonathan.D.Beard@sth.nhs.uk)

Background: **Most surgical assessment has been aimed at technical proficiency. However, non-technical skills also affect patient safety and clinical effectiveness. The NOTSS (Non-Technical Skills for Surgeons) assessment instrument was developed specifically to assess the non-technical skills of individual surgeons in the operating theatre. This study evaluated NOTSS as a real-world assessment, with a mix of minimally trained assessors. The evaluation criteria were feasibility, validity and psychometric reliability.**

Methods: **In a standard evaluation of NOTSS, 56 anaesthetists, 39 scrub nurses, two surgical care practitioners and three independent assessors provided 715 assessments of 404 surgical cases of 15 index procedures across six specialties performed by 85 surgical trainees.**

Results: **The assessment was feasible, but important implementation challenges were highlighted. Most respondents considered the method valid, but with reservations about assessing cognition. The factor structure of scores, and their positive relationships with other measures of experience and performance, supported validity. Trainees' non-technical skill scores were relatively procedure-independent and achieved good reliability (generalizability coefficient 0·8 or more) when six to eight assessors observed one case each.**

Conclusion: **Minimally trained assessors, who are typically present in operating theatres, were sufficiently discriminating and consistent in their judgements of trainee surgeons' non-technical skills to provide reliable scores based on an achievable number of observations.**

Paper accepted 26 January 2011
Published online in Wiley Online Library (www.bjs.co.uk). **DOI:** 10.1002/bjs.7478

Introduction

The past decade has seen a major expansion in postgraduate assessment within the medical profession. Assessment continues to be the main means of ensuring that trainee doctors have achieved competence. This means that good clinical practice in the National Health Service will become dependent to some extent upon good assessment practice[1].

Educational research has provided a number of important observations. First, clinical performance is context-specific; a good performance in one case doesn't necessarily predict a good performance in another[2]. Consequently clinicians should be assessed on a sample of cases. Second, doctors judging their peers and trainees largely agree on who is performing well and poorly, but they display some individual differences. Consequently clinicians should be assessed by a sample of suitably experienced judges[1]. Third, attempts to standardize assessment by taking doctors out of their real workplace and into a controlled environment are futile. It is quite possible to assess a doctor or a surgeon in a controlled environment, but competence in such a setting does not predict real workplace performance[3,4]. Therefore, to know how they perform in the workplace, clinicians should be assessed regularly there by a mix of assessors on their day to day work.

In the UK, the medical and surgical Royal Colleges have submitted their postgraduate assessment programmes to the regulatory body (the Postgraduate Medical Education and Training Board, now the General Medical Council). Most of these programmes reflect the findings above[5].

All of them rely heavily on workplace-based assessment (WBA).

The Intercollegiate Surgical Curriculum Programme (ISCP) assessment framework includes five established workplace-based methods: the mini-Peer Assessment Tool, a multisource feedback tool; mini-Clinical Evaluation Exercise, a clinical encounter assessment; case-based discussion; direct observation of procedural skills (DOPS); and procedure-based assessment (PBA).

Surgical trainee performance in the operating theatre is assessed using the latter two methods: DOPS and PBA. DOPS is used for basic clinical procedures and PBA for more complex procedures. Recent evaluation of PBA found it to be a psychometrically reliable instrument for assessing trainee surgeons' technical skills in the operating theatre[6].

Both DOPS and PBA are designed to assess technical aspects of performance. The ISCP framework currently includes no method to assess non-technical skills in the operating theatre. In other high-hazard industries, such as aviation, non-technical performance is recognized as fundamental to safety and effectiveness. There is now a growing body of evidence that the same is true in surgical practice[7–10].

A number of observation-based instruments have been developed to study non-technical performance in the operating theatre. Important examples include Surgical NOTECHS (developed from an earlier NOn-TECHnical Skills aviation instrument)[11] and OTAS (Observational Teamwork Assessment for Surgery)[12]. Both of these instruments have demonstrated some measure of reliability, with inter-item correlation and inter-rater agreement respectively. Team performance correlated when measured by NOTECHS and OTAS, providing evidence for the validity of both; however, both of these instruments focus on the performance of the entire surgical team rather than that of the individual surgeon (trainee or otherwise), and both usually require specialist assessors. Consequently, neither instrument provides a comfortable platform for real-world WBA.

The NOTSS (Non-Technical Skills for Surgeons) instrument was designed to assess individual surgeon non-technical skills. Defining non-technical performance as 'the critical cognitive and interpersonal skills that underpin technical proficiency', a multidisciplinary group of surgeons, psychologists and an anaesthetist used task analysis to develop this four-category behavioural rating system[13]. The categories are: situation awareness, decision-making, communication and teamwork, and leadership. In one study of NOTSS, trained consultant surgeons assessing six video-recorded simulated operations provided scores with good internal reliability within the four behavioural categories, and their scores showed moderate agreement with expert scores[14]. In a separate study comparing novice and expert assessors, novices were less likely to provide ratings of the cognitive category 'situation awareness', and were harsher than experts, especially when rating communication and teamwork, and leadership[15]. Finally, a feasibility study of NOTSS in the operating theatre provided mixed responses[16]. Respondents indicated that NOTSS provided a structure and language to rate trainee surgeons and provide feedback on their non-technical behaviours, but it was difficult to understand some behavioural descriptors and difficult to rate the cognitive categories. In addition, many routine cases presented too few decisions for the decision-making category and the consultant was likely to undermine the trainee's leadership.

This study was conducted as part of an evaluation of three instruments designed to assess trainee surgeons in the operating theatre. The evaluation of NOTSS, PBA and OSATS (Objective Structured Assessment of Technical Skills – currently used by the Royal College of Obstetricians and Gynaecologists) was run in parallel[17]. The aim was to extend the existing evaluation of NOTSS as a real-world working assessment method, especially with regard to the following: to determine whether it was feasible to implement NOTSS assessment in the workplace with minimal support; to find out whether minimally trained assessors (including non-surgeons) in real workplace situations could provide reliable scores of non-technical performance; to collect further evidence about the validity of NOTSS as a measure of non-technical performance; and to compare the reliability of NOTSS with the reliability of technical skills assessments (PBA and OSATS).

Methods

The development and design of the NOTSS system has been described previously[13,18]. It is a behavioural rating system based on a skill taxonomy, with examples of good and poor behavioural markers, and is used to identify observable, non-technical behaviours that contribute to superior, satisfactory or substandard performance (*Fig. 1*). NOTSS has been designed to provide surgeons with explicit ratings, and with feedback on their non-technical performance.

The NOTSS system comprises a three-level hierarchy. The top level identifies four skill categories (situation awareness, decision-making, communication and teamwork, and leadership). The middle level identifies 12 elements (3 per category). The lowest level identifies numerous behavioural markers that typify each category

Hospital ..... Trainer name ..... Date .....

Trainee name ..... Operation .....

Category	Category rating*	Element	Element rating*	Feedback on performance and debriefing notes
Situation awareness		Gathering information		
		Understanding information		
		Projecting and anticipating future state		
Decision-making		Considering options		
		Selecting and communicating option		
		Implementing and reviewing decisions		
Communication and teamwork		Exchanging information		
		Establishing a shared understanding		
		Coordinating team activities		
Leadership		Setting and maintaining standards		
		Supporting others		
		Coping with pressure		

* **1** Poor; **2** Marginal; **3** Acceptable; **4** Good; **NA** Not applicable

1 Poor	Performance endangered or potentially endangered patient safety; serious remediation is required
2 Marginal	Performance indicated cause for concern; considerable improvement is needed
3 Acceptable	Performance was of a satisfactory standard but could be improved
4 Good	Performance was of a consistently high standard, enhancing patient safety; it could be used as a positive example for others
NA	Not applicable

Fig. 1 NOTSS (Non-Technical Skills for Surgeons) summary rating form

and element. These are intended to be indicative rather than comprehensive. For example, the category 'situation awareness' includes the element 'understanding information', which is typified by the positive behaviour 'reflects and discusses significance of information' and by the negative behaviour 'overlooks or ignores important results'.

The assessor observes the surgeon during the 'gloves on, scrubbed up' phase of surgery and then scores each category and element according to the rating scale: 'good' (4), 'acceptable' (3), 'marginal' (2), 'poor' (1). 'Not applicable' is used if a skill is not required or relevant to the procedure. The output thus includes 16 scores: 12 element scores and four category scores. An assessment may be accompanied by feedback at the end of the operation.

To simplify presentation, some of the present analyses combined these scores to provide four 'domain' scores, derived by taking the mean of the category score and the three element scores for each category. These four scores were the 'situation awareness domain score' (SDS), 'decision-making domain score' (DDS), 'communication/teamwork domain score' (CDS) and 'leadership domain score' (LDS). A global score (GS) was also calculated as the mean of the four category scores to provide a reliability coefficient for the tool as a whole.

Study design

This was an ethically approved, prospective observational study conducted within the operating theatres of three teaching hospitals in Sheffield, UK, between April 2007 and June 2009. Trainees were observed directly and assessed using NOTSS by one or more independent assessors (IAs) from the research team, as well as one or more of the following members of the theatre team: anaesthetist, scrub nurse and surgical care practitioner (SCP). The consultant surgeons who were supervising the trainees did not complete the NOTSS form, as they were occupied completing OSATS or PBA forms. The aim was to assess every trainee performing each of their specialty-specific procedures on at least two occasions. Index procedures were selected from six specialties: cardiac (coronary artery bypass, aortic valve

replacement), colorectal (right hemicolectomy, anterior resection), gastrointestinal (laparoscopic cholecystectomy, open inguinal hernia), orthopaedics (primary hip and knee replacement), vascular (varicose vein, aortic aneurysm, carotid endarterectomy), and obstetrics and gynaecology (urgent and elective caesarean section, evacuation of uterus, diagnostic laparoscopy). These reflected the breadth of surgery in each specialty (for example open and laparoscopic procedures) and a range of procedural complexity; they were also performed regularly by trainees.

All patients gave written consent, and all trainees and assessors gave verbal consent to participate. Recruitment was concentrated upon trainees within specialty training (ST3–ST7), although there was no exclusion of those in core training (ST1 or ST2) or in non-training posts.

Training

The investigators briefed the theatre team assessors and trainees on the use of NOTSS and provided them with a copy of the booklet that included explanatory notes[19]. Trainees were instructed to perform the procedure as they usually would, seeking guidance or assistance as required. All IAs were practising in surgery, and had received training in non-technical skills assessment by the developers of the NOTSS tool. They were therefore considered 'experts' in the use of NOTSS compared with the theatre team assessors. The IAs had also completed recognized educational training at the Royal College of Surgeons of England[20].

Data collection

All assessors completed the NOTSS independently of one another to avoid rating bias. Laminated copies of NOTSS behavioural markers were provided in theatre to assist assessors. Trainees were not provided with NOTSS feedback during the study because the consultant surgeons' feedback on their PBA or OSATS assessment was given priority. Trainees and assessors provided initial demographic data. Assessors completed a final questionnaire asking about the feasibility, validity and acceptability of NOTSS once they had completed assessments within the study. Trainee questionnaire evaluations were confined to PBA and OSATS, and were not available for NOTSS because trainees had no direct experience of NOTSS during the study.

Sample size and sampling

There is no equivalent of a power calculation for an exploratory variance component analysis. However, wide and representative sampling of each relevant factor in the assessment process (trainees, operations and assessors) is necessary to produce dependable estimates of reliability. The aim was to assess 450 cases, sampling as many trainees and assessors as possible. This optimized the estimates of reliability.

Study implementation

Implementation of the study is illustrated in *Fig. 2*.

Statistical analysis

All analyses were conducted using SPSS® version 14.0 (SPSS, Chicago, Illinois, USA). Descriptive data on trainee and assessor demographics, assessor mix, and score distributions were presented as text or simple tables (numbers and percentages) and were not subjected to statistical tests. The feasibility evaluation was based on questionnaire responses, which were presented in the same way. Field notes made by the research team were also cited.

The validity evaluation used three strands. First, it used questionnaire responses as above. Second, it explored the internal structure (intercorrelation) of the NOTSS element and category scores by applying an exploratory factor analysis (principal axis factoring with varimax rotation). Third, it examined the correlation between the domain scores (SDS, DDS, CDS and LDS) and the GS, and other measures of training, experience and performance. NOTSS scores were compared with two measures of technical performance (OSATS and PBA), with specialist training (ST) level, with years of UK and non-UK training, and with age. Pearson's method was used for these comparisons,

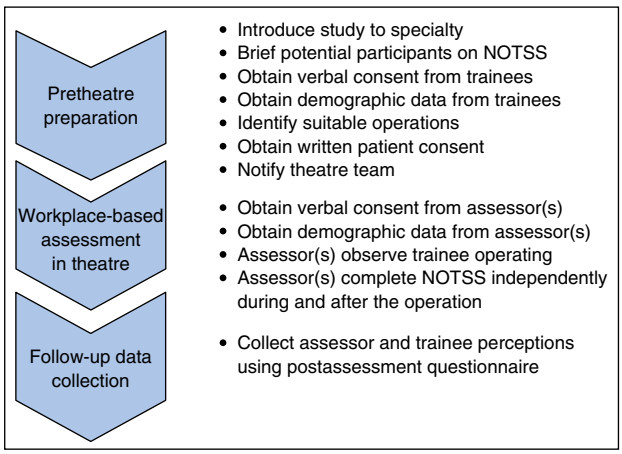

Fig. 2 Study implementation. NOTSS, Non-Technical Skills for Surgeons

and correlations were presented to two decimal places. Where several variables were explored, the significance threshold was adjusted using the Bonferroni method.

The reliability evaluation applied generalizability theory[21]. A generalizability analysis first uses regression modelling (variance component analysis) to estimate how much each and every relevant factor in the assessment, and their interactions, has influenced the observed scores (G-study). The data in this study were naturalistic (incomplete and unbalanced), so the G-study used the minimum norm quadratic unbiased estimation (MINQUE) procedure to provide the best estimates[22], and the regression model was a partially nested random-effects model. The regression model estimated the independent contribution to GS variance of: trainee ability ($V_{trainee}$), trainee case-to-case variation (V_{case}), index procedure difficulty ($V_{procedure}$), assessor stringency/leniency ($V_{assessor}$), the stringency/leniency of the different assessor groups ($V_{designation}$), trainee aptitude for a particular procedure ($V_{trainee \times procedure}$), and assessor subjectivity over an individual case or a particular type of procedure ($V_{assessor \times case}$, and $V_{assessor \times procedure}$). The D-study (reliability modelling) then used Cronbach's original equations to combine these sources of variance and to estimate reliability with various assessment designs and varying numbers of cases or assessors[23]. A G coefficient of 0·8 or more demonstrated an acceptable level of reliability.

Results

A total of 85 surgical trainees gave their verbal consent to participate. Study information was sent to 832 patients listed for surgery; 260 (31·3 per cent) were not approached for consent predominantly because of schedule changes affecting the list or the availability of the trainee. Fourteen (1·7 per cent) refused to participate. The remaining 558 (67·1 per cent) gave their consent to participate. Of these, 404 (72·4 per cent) were included. Consented patients were not included for the same reasons as above.

The 404 operations were assessed by 56 anaesthetists, 39 scrub nurses, two SCPs and three IAs, and included all 15 index procedures across all six specialties. In total, assessors provided 715 assessments.

Demographics

All but one trainee provided complete demographic data. Of those who responded, the majority of trainees were male (55 of 84, 65 per cent), half had graduated in the UK (43 of 84, 51 per cent), and all ST levels were represented.

Forty-eight of the 56 anaesthetists provided complete demographic data. They had a median age of 41 years, and a median consultant experience of 8 years; 35 of 48 were men, and 42 were UK-trained. Thirty-three of the 39 scrub nurses provided complete demographic data. They had a median age of 39 years, and a median scrub nurse experience of 10 years; seven of 39 were men, and 32 were UK-trained. The two SCPs who provided NOTSS assessments were both UK-trained men; they were aged 33 and 51 years, and had 4 and 8 years of SCP experience respectively. The three IAs were: a SCP with 4 years' experience who was a 37-year-old female UK graduate; a consultant vascular surgeon with 17 years' experience who was a 51-year-old male UK graduate; and an ST4 trainee in obstetrics and gynaecology who was a 29-year-old female UK graduate.

NOTSS assessments and scores

Of the 715 assessments performed in total, 424 (59·3 per cent) were performed by an IA, 192 (26·9 per cent) by an anaesthetist, 96 (13·4 per cent) by a scrub nurse and three (0·4 per cent) by a SCP (*Table 1*).

There were 16 NOTSS scores per assessment (4 category and 12 element scores), giving a total of 11 440 responses. Across the trainee cohort, non-technical performance was scored as being good (4) in 2146 (18·8 per cent) of responses, acceptable (3) in 5618 (49·1 per cent), marginal (2) in 2500 (21·9 per cent), poor (1) in 105 (0·9 per cent), and 'not applicable' in 1071 (9·4 per cent).

Table 2 presents the proportion of scores that fell into each category across the four assessor groups (anaesthetists, IAs, nurses and SCPs). The experts (IAs) more frequently gave lower scores across almost all elements in comparison with the novice groups. However, the regression model applied in the G-study showed that any scoring differences between rater groups were not significant in terms of reliability, once all variables had been accounted for. IAs were no more frequently able to score the cognitive domains (situation awareness and decision-making) than the novice groups (*Table 2*). The behavioural domain of leadership was most frequently regarded as being not applicable in the judgement of all the assessor groups.

Feasibility and acceptability

It proved possible to recruit 404 of the target 450 surgical cases for this study; the vast majority of patients, trainees and assessors agreed to participate in a voluntary assessment process, and the vast majority of assessment items were scored as applicable. However, implementation was challenging.

Table 1 Distribution of cases assessed

	No. of cases	Anaesthetist	Independent assessor	Scrub nurse	Surgical care practitioner
Cardiothoracic					
Aortic valve replacement	4	3	4	0	1
Coronary artery bypass graft	31	26	31	5	2
Colorectal					
Anterior resection	13	6	13	6	0
Right hemicolectomy	11	9	11	8	0
Gastrointestinal					
Hernia	16	10	19	1	0
Laparoscopic cholecystectomy	35	20	35	3	0
Orthopaedic					
Hip replacement	18	3	20	0	0
Knee replacement	16	0	16	1	0
Obstetrics and gynaecology					
Diagnostic laparotomy	72	25	77	32	0
Elective caesarean section	60	26	67	20	0
Urgent caesarean section	5	5	5	4	0
Evacuation of uterus	45	11	46	11	0
Vascular					
Aortic aneurysm	14	10	14	4	0
Carotid endarterectomy	21	14	23	1	0
Varicose veins	43	24	43	0	0
Total	404	192	424	96	3

Table 2 Item responses by assessor designation

	Anaesthetist ($n = 192$)					Independent assessor ($n = 424$)					Scrub nurse ($n = 96$)					Surgical care practitioner ($n = 3$)				
	1	2	3	4	NA	1	2	3	4	NA	1	2	3	4	NA	1	2	3	4	NA
Situation awareness	1·6	12·5	50·0	29·2	6·8	2·6	30·7	55·7	9·7	1·4	1	9	60	24	5	0	0	33	67	0
Gathering information	2·1	14·6	50·0	28·6	4·7	1·2	40·3	44·3	12·5	1·7	2	14	54	28	2	0	0	67	33	0
Understanding information	1·6	7·3	47·4	31·3	12·5	0·7	17·7	59·4	16·3	5·9	0	9	59	25	6	0	0	33	33	33
Projecting and anticipating future state	2·1	12·5	46·4	27·6	11·5	1·9	32·5	43·4	9·9	12·3	2	11	56	26	4	0	0	33	33	33
Decision-making	0·5	14·1	52·1	25·0	8·3	0·2	26·2	54·5	12·3	6·8	0	18	54	25	3	0	0	33	67	0
Considering options	0·5	10·9	50·5	27·1	10·9	0·2	19·8	55·9	14·2	9·9	0	14	59	24	3	0	0	33	67	0
Selecting and communicating option	0·5	15·6	48·4	25·0	10·4	0·2	27·4	48·3	16·3	7·8	0	19	54	20	7	0	0	67	33	0
Implementing and reviewing decisions	0·5	11·5	46·4	25·0	16·7	0	22·9	47·4	15·1	14·6	1	18	51	18	13	0	0	33	33	33
Communication/teamwork	1·6	19·3	43·2	31·8	4·2	0·7	38·4	51·7	8·3	0·9	0	20	48	30	2	0	33	0	67	0
Exchanging information	1·6	17·2	46·4	32·8	2·1	0·7	37·0	46·7	14·9	0·7	0	24	44	31	1	0	33	33	33	0
Establishing a shared understanding	1·0	20·3	44·3	28·6	5·7	0·2	42·5	42·9	11·3	3·1	1	20	56	19	4	0	33	0	67	0
Coordinating team activities	2·6	15·6	44·8	33·9	3·1	0·7	31·6	55·4	11·3	0·9	0	17	47	34	2	0	33	0	67	0
Leadership	0·5	7·3	51·6	27·6	13·0	2·4	24·8	59·4	11·3	2·1	0	11	61	24	3	0	33	0	67	0
Setting and maintaining standards	0	9·4	47·9	29·7	13·0	2·8	25·5	55·9	14·6	1·2	2	9	56	30	2	0	33	0	67	0
Supporting others	0·5	8·9	39·6	25·5	25·5	0	13·9	42·0	12·0	32·1	0	14	50	24	13	0	33	0	67	0
Coping with pressure	0	3·6	35·4	30·2	30·7	0	12·7	29·0	11·6	46·7	1	9	42	19	29	0	0	33	67	0

Values are cumulative percentages for each score. NA, not applicable.

Study records

Study field notes recorded the following feasibility and acceptability challenges: it was difficult to access clinicians' time to introduce the process and to provide training in advance of assessment; where a consultant prompted trainees or took over leadership of the case, assessment became difficult; planned assessments were often cancelled because of changes to lists or changes in the availability

Table 3 Views regarding NOTSS (Non-Technical Skills for Surgeons) for 56 assessors

	No response	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1 NOTSS provides a common language to discuss non-technical skills	0 (0)	0 (0)	1 (2)	12 (21)	39 (70)	4 (7)
2 It was easy to rate cognitive skills (situation awareness, decision-making)	0 (0)	0 (0)	14 (25)	12 (21)	29 (52)	1 (2)
3 It was easy to rate interpersonal skills (communication and teamwork, leadership)	0 (0)	0 (0)	5 (9)	9 (16)	40 (71)	2 (4)
4 Using NOTSS added too much time to my list	0 (0)	9 (16)	25 (45)	17 (30)	4 (7)	1 (2)
5 NOTSS is a useful tool to support reflective practice or to provide insight	0 (0)	0 (0)	0 (0)	9 (16)	38 (68)	9 (16)
6 NOTSS is a valuable adjunct to tools that assess surgical skills (e.g. PBA/OSATS)	2 (4)	0 (0)	0 (0)	15 (27)	36 (64)	3 (5)
7 Routine use of the NOTSS system will enhance safety in the operating theatre	1 (2)	1 (2)	9 (16)	20 (36)	20 (36)	5 (9)
8 NOTSS provides useful feedback for the trainee	17 (30)	0 (0)	0 (0)	0 (0)	31 (55)	8 (14)

Values in parentheses are percentages. PBA, procedure-based assessment; OSATS, Objective Structured Assessment of Technical Skills.

of trainees; both anaesthetists and scrub nurses had to discontinue NOTSS assessments on some occasions because clinical priorities took over.

Respondents' views

Thirty (54 per cent) of the 56 anaesthetists and 26 (67 per cent) of the 39 scrub nurses completed a final questionnaire about the validity, feasibility and acceptability of NOTSS (*Table 3*). Questions 4–8 relate to feasibility and acceptability. Only five (9 per cent) agreed or strongly agreed that NOTSS added too much time to the operating list, whereas the majority perceived NOTSS to be useful for supporting insight (47; 84 per cent) and for providing feedback (39; 70 per cent). Most regarded NOTSS as an important adjunct to surgical skills assessment methods (39; 70 per cent). Twenty-five (45 per cent) felt that the routine use of NOTSS would enhance patient safety in the operating theatre.

Validity

Respondents' views

Questions 1–3 of the final questionnaire asked assessors about the validity of NOTSS. At least 75 per cent agreed that NOTSS provided a common language for assessing non-technical skills and found it easy to assess the interpersonal domains. However, only 54 per cent found it easy to rate the cognitive domains (situation awareness and decision-making) (*Table 3*).

Internal structure

Table 4 displays the internal structure of the NOTSS category and element scores as a rotated factor component matrix. The internal structure of the instrument matched the four-domain structure without exception; only element score 14 (setting and maintaining standards) loaded on to a second factor as strongly as its own domain. The loadings suggest that this element is an aspect of situation awareness as well as leadership.

Table 4 NOTSS (Non-Technical Skills for Surgeons) rotated factor component matrix

		Factor			
Item	Category or element	1	2	3	4
1	Situation awareness			0·76	
2	Gathering information			0·83	
3	Understanding information			0·59	
4	Projecting and anticipating future state	0·50		0·60	
5	Decision-making	0·82			
6	Considering options	0·72			
7	Selecting and communicating option	0·76			
8	Implementing and reviewing decisions	0·76			
9	Communication/teamwork		0·83		
10	Exchanging information		0·73		
11	Establishing a shared understanding		0·69		
12	Coordinating team activities		0·67		
13	Leadership				0·76
14	Setting and maintaining standards			0·54	0·54
15	Supporting others				0·73
16	Coping with pressure				0·68

For clarity, factor loadings less than 0·4 are not displayed.

Relationships to external variables

Parallel assessments using NOTSS and PBA were performed for 317 of the 404 cases. All four domain scores were significantly positively correlated with the PBA global summary score. Pearson's coefficient was 0·48 ($P < 0·001$), 0·55 ($P < 0·001$), 0·43 ($P < 0·001$) and 0·49 ($P < 0·001$) for SDS, DDS, CDS and LDS respectively.

Parallel assessments using NOTSS and OSATS were performed for 90 of the 404 cases. All four domain scores

Table 5 Domain score correlations with measures of experience

	Age	Specialist training level	Years of UK surgical training	Years of non-UK surgical training
Situation awareness				
r	0·29	0·57	0·49	−0·15
*P**	0·010	< 0·001	< 0·001	0·192
Decision-making				
r	0·31	0·57	0·47	−0·04
*P**	0·007	< 0·001	< 0·001	0·726
Communication/teamwork				
r	0·22	0·40	0·36	−0·14
*P**	0·060	< 0·001	0·001	0·209
Leadership				
r	0·18	0·46	0·40	−0·14
*P**	0·114	< 0·001	< 0·001	0·224

*Two-tailed test.

were significantly positively correlated with the generic part of the OSATS score. The corresponding Pearson's coefficients were 0·58 ($P < 0·001$), 0·57 ($P < 0·001$), 0·40 ($P < 0·001$) and 0·50 ($P < 0·001$). Thus, the decision-making domain was most strongly correlated with technical performance.

Table 5 displays the relationships between trainee NOTSS scores and demographic measures of their experience. Because there were four measures of experience (age, ST level, years of UK training and years of non-UK training), the significance threshold was adjusted to 0·0125. Across all four NOTSS domains, performance was significantly positively correlated with ST level and years of UK training. Older age, however, was not associated with improved scores in the behavioural domains (communication and teamwork, and leadership). More years of non-UK training were not associated with better performance in any of the domains.

Reliability

G-study

Variance component analysis shows which factors have the greatest influence on any given score. The ability of the trainee being assessed (consistent across cases and assessors) had the greatest impact on a score, explaining 30·9 per cent of score variance (*Table 6*). However, the stringency or leniency of the assessor (hawk or dove) and the subjectivity of assessors (partiality) contributed significantly (27·0 and 20·1 per cent of score variance respectively). In addition, a given trainee's performance varied from case to case (9·6 per cent). Most other effects, including the procedure being performed and the designation of the assessor (anaesthetist, IA, nurse or SCP), exerted only either a very small, or no effect on the score given. This shows that the relative stringency and leniency of IAs and SCPs was insignificant to the overall reliability of NOTSS in relation to other factors affecting scores, once case mix had been accounted for.

Table 6 G-study: variance component analysis

Component	Estimate	%	Explanation
$V_{trainee}$	0·111	30·9	Trainee ability
V_{case}	0·035	9·6	Trainee case-to-case variation
$V_{procedure}$	0·012	3·4	Procedure difficulty
$V_{assessor}$	0·097	27·0	Assessor stringency
$V_{designation}$	0	0	Assessor designation stringency
$V_{trainee \times procedure}$	0·019	5·4	Trainee procedure aptitude
$V_{assessor \times case}$	0·072	20·1	Assessor subjectivity over case
$V_{assessor \times procedure}$	0·013	3·6	Assessor subjectivity over procedure
V_{error}	0	0	Residual variation

D-study

The D-study results show how the reliability of trainees' scores increased when they were based on several cases or several assessors' scores (*Table 7*). Because each trainee's case-to-case variation was fairly small, increasing the number of cases without increasing assessors will have relatively little impact on reliability. However, because assessors are variable, combining the scores of several assessors rapidly improves reliability so that the combined scores of six assessors (each seeing a single case) provided a highly reproducible reflection of the combined scores of any six assessors and so achieved a reliability coefficient of 0·82 even in a 'nested' design where each trainee saw six different assessors. If trainees were assessed by the same group of assessors (crossed design), the reliability would be much higher. When trainees were compared across different procedures, the reliability fell slightly because

Table 7 D-study: reliability modelling

	Assessors per case		
No. of cases	1	2	3
1	0·35	0·48	0·55
2	0·57	0·69	0·74
3	0·68	0·78	0·82
4	0·75	0·83	0·86
5	0·79	0·86	0·89
6	0·82	0·88	0·90
7	0·85	0·90	0·92
8	0·86	0·91	0·93

Different assessor(s) were assumed for each case.

trainee ability showed slight procedure-specificity or aptitude (see $V_{trainee \times procedure}$ in *Table 6*). Comparing across procedures, eight assessors, each assessing a single case, would be required to achieve a G coefficient of 0·8 or more.

Discussion

This study set out to evaluate an instrument (NOTSS) developed for the purpose of assessing surgeons' non-technical skills, in a real-world setting, with minimally trained assessors. There were several difficulties similar to those in earlier evaluations. The fact that most scores were either acceptable (3) or good (4) may reflect a reluctance to provide negative ratings; nearly half of assessors who responded did not find it easy to rate the cognitive domains, as in previous work[15,16]. In addition, the study records revealed that consultant surgeons sometimes prompted trainees or took over leadership, and that clinical commitments sometimes diverted anaesthetists and scrub nurses from NOTSS assessments.

Unlike previous NOTSS evaluations[15], the novice assessors gave more 'acceptable' and 'good' scores than the experts. These observations are more consistent with the wider WBA literature which suggests that more senior or expert assessors are increasingly willing to provide low scores[24]. Reassuringly, however, the G-study showed that any stringency differences were insignificant in their contribution to the overall reliability of NOTSS. Unlike previous NOTSS evaluations, the novices rated cognitive items just as often as the experts. Rather, it was the behavioural domain of 'leadership' that both novice and expert assessors most often regarded as not applicable to a cohort of trainees in a WBA context. It is likely that this reflected the involvement of consultant surgeons that has already been highlighted. This has not been evident in simulated evaluations or evaluations of fully trained surgeons, and it has implications for the value of NOTSS for assessing trainees.

The present study adds to the controlled evaluations by highlighting some of the difficulties of implementing large-scale assessment in the workplace[25]. Difficulties included accessing clinicians' time to prepare them for delivering or receiving assessment, relying on staff as assessors, who also have clinical tasks, and schedule changes that undermined planned assessments. The present study quantified these difficulties in the real world, to find out whether they impacted on the validity and reliability of the resulting scores. Almost all responding anaesthetists and scrub nurses felt that NOTSS assessment could be fitted into their operating list; three-quarters perceived that it provided a common language for assessing non-technical skills, and the majority considered it useful for supporting insight and providing feedback. Most recognized its importance as an adjunct to technical skills assessment, and nearly half felt that, used routinely, it would enhance patient safety in the operating theatre. The half that did not feel it would enhance patient safety may perceive a problem with the assessment, a cynicism about change, or a lack of enthusiasm about the patient safety agenda. The questionnaire did not explore reasons for the responses.

Crucially, almost a quarter of NOTSS scores identified a trainee as performing an element or a category at a 'marginal' or 'poor' standard. In this respect, NOTSS was much more sensitive to performance weaknesses than the majority of WBA methods.

Despite the fact that nearly half of respondents did not find it easy to rate the cognitive domains, almost all assessors in every assessor group were able to provide category scores for situation awareness and decision-making. The factor analysis shows that those scores reflected separate constructs from the behavioural domains and from each other; in other words, they were not simple extrapolations of easier domains. It shows, for example, that a trainee who scored relatively well on one of the 'situation awareness' items scored relatively well across all of them, but not necessarily across other domains.

The moderate correlation between NOTSS scores and the technical skill scores on PBA and OSATS provided some evidence of validity and reflected the fact that non-technical skill is a separate attribute, but that the two are related. The strongest correlations between NOTSS and PBA/OSATS were in the decision-making domain. This is to be expected as several items in both PBA and OSATS relate to decision-making.

The positive association between scores and ST level or years of UK training also provided encouraging evidence that NOTSS is measuring a training-related skill.

Finally, and critically, the reliability analysis showed that a trainee who performed well on one case, in the judgement of one assessor, performed relatively well on another in the judgement of another assessor. Consequently, whilst assessors' stringencies and biases, and trainees' case-to-case variation, made a contribution to any given score, a relatively small sample of assessors' judgements provided a reliable indicator of how a trainee performed relative to other trainees.

The reliability of NOTSS compared favourably with many other WBA instruments, many of which require samples of ten to 20 assessments before achieving similar levels of reliability[26]. Unlike the technical instruments, NOTSS scores are relatively procedure-independent.

Non-technical skill transfers across procedures in a way that technical skill does not[6].

The study had a number of limitations. The scope was limited to the question of assessing non-technical performance. There were no data in this study to show whether trainees who displayed better non-technical skills performed safer or more efficient operations.

Although trainees and assessors from three hospitals were included, all three were located in one city and were associated with one deanery. There is a chance that these trainees or assessors were sufficiently unrepresentative of the national population as to affect the generalizability of the results; the present assessors might have been more consistent or more positive about NOTSS than the national population. The trainees might have been more heterogeneous (and therefore easier to separate by assessment), although this is unlikely.

Evidence about feasibility and acceptability was based on participants' perceptions alone, but there is no other valid way to evaluate perceived outcomes. In addition, the questionnaire response rates (54 per cent of anaesthetists and 67 per cent of nurses) raise the possibility of non-response bias. Anaesthetists and nurses had to use some conjecture in responding to some of the questionnaire items (especially items 6 and 8). The results must be interpreted in light of these caveats. Furthermore, because trainee feedback was based on PBA and OSATS scores rather than NOTSS scores, and because trainees were asked to evaluate only their technical skills assessments, no trainee data on the feasibility and acceptability of the process can be provided. The factor structure of the scores may be a result of the organization of the NOTSS items within four categories rather than independent meaningful interpretation of each item. If so, this reduces the strength of the evidence for validity.

The overall conclusion of this study was that NOTSS can be implemented in the real world of the operating theatre and that novice assessors can provide scores with acceptable measurement characteristics for assessment. Non-technical skill is relatively procedure-independent and can thus be assessed on any sample of appropriate procedures using NOTSS. However, on a procedure-specific basis, three to four different assessors observing one case each should achieve a reliability of 0·7 for each important procedure. No WBA process is undemanding, but this should be achievable within most training programmes.

This study has not added to the evidence about the potential importance of non-technical skills, but it has added to the evidence that they are assessable in the workplace. The broad implication is that the ISCP and other surgical training programmes worldwide should consider including non-technical skills in their curriculum and assessment framework, and could have confidence in NOTSS as an assessment instrument.

Acknowledgements

This project was funded by the National Institute for Health Research Health Technology Assessment Programme and has been published in full in the Health Technology Assessment series[17]. The views and opinions expressed are those of the authors and do not necessarily reflect those of the Department of Health. The authors declare no conflict of interest.

References

1 Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ* 2002; **36**: 800–804.
2 Elstein AS, Shulman LS. *Medical Problem-solving: an Analysis of Clinical Reasoning*. Harvard University Press: Cambridge, 1978.
3 Rethans J, Sturmans F, Drop R, van der Vleuten C, Hobus P. Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. *BMJ* 1991; **303**: 1377–1380.
4 Kopelow M, Schnabl G, Hassard TH, Tamblyn R, Klass DJ, Beazley G *et al*. Assessment of performance in the office setting with standardized patients. *Acad Med* 1992; **67**(Suppl 10): S19–S21.
5 Academy of Medical Royal Colleges. *Improving Assessment*. AoMRC Press: London, 2009.
6 Marriott J, Purdie H, Crossley J, Beard JD. Evaluation of procedure-based assessment for assessing trainees' skills in the operating theatre. *Br J Surg* 2011; **98**: 450–457.
7 Carthey J, de Leval M, Wright D, Farewell V, Reason J. Behavioural markers of surgical excellence. *Safety Science* 2003; **41**: 409–425.
8 Gawande A, Zinner M, Studdert D, Brennan T. Analysis of errors reported by surgeons at three teaching hospitals. *Surgery* 2003; **133**: 614–621.
9 Lingard L, Regehr G, Orser B, Reznick R, Baker G, Doran D *et al*. Evaluation of a preoperative checklist and team briefing to reduce failures in communication. *Arch Surg* 2008; **143**: 12–17.
10 Mishra A, Catchpole K, Dale T, McCulloch P. The influence of non-technical performance on technical outcome in laparoscopic cholecystectomy. *Surg Endosc* 2008; **22**: 68–73.
11 Mishra A, Catchpole K, 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**: 109–115.
12 Healey A, Undre S, Vincent C. Developing observational measures of performance in surgical teams. *Qual Saf Health Care* 2004; **13**(Suppl 1): i33–i40.

13 Yule S, Flin R, Paterson-Brown S, Maran N, Rowley D. Development of a rating system for surgeons' non-technical skills. *Med Educ* 2006; **40**: 1098–1104.

14 Yule S, Flin R, Maran N, Rowley D, Youngson G, Paterson-Brown S. Surgeons' non-technical skills in the operating room: reliability testing of the NOTSS behavior rating system. *World J Surg* 2008; **32**: 548–556.

15 Yule S, Rowley D, Flin R, Maran N, Youngson G, Duncan J *et al*. Experience matters: comparing novice and expert ratings of non-technical skills using the NOTSS system. *ANZ J Surg* 2009; **79**: 154–160.

16 Yule S, Flin R, Maran N, Youngson G, Mitchell A, Rowley D *et al*. Debriefing surgeons on non-technical skills. *Cogn Technol Work* 2008; **10**: 265–274.

17 Beard JD, Marriott J, Purdie H, Crossley J. Assessing the surgical skills of trainees in the operating theatre: a prospective observational study of the methodology. *Health Techn Assess* 2011; **15**: i–xxi, 1–162.

18 Yule S, Flin R, Maran N, Rowley D, Youngson G, Duncan J *et al*. Development and evaluation of the NOTSS behaviour rating system for intraoperative surgery. In *Safer Surgery: Analysing Behaviour in the Operating Theatre*, Flin R, Mitchell L (eds). Ashgate: Farnham, 2009; 7–25.

19 Industrial Psychology Research Centre, University of Aberdeen. *NOTSS: Non-Technical Skills for Surgeons*, 2010; http://www.abdn.ac.uk/iprc/notss [accessed 6 September 2010].

20 Royal College of Surgeons of England. *Training the Trainers*. http://www.rcseng.ac.uk/education/courses/training-the-trainers [accessed 10 June 2010).

21 Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002; **36**: 972–978.

22 Crossley J, Roberts C, Jolly B, Humphries G, Ricketts C, Norcini J *et al*. 'I'm pickin' up good regressions': the governance of generalisability analyses. *Med Educ* 2007; **41**: 926–934.

23 Cronbach LJ, Gleser GC, Nanda H, Rajaratnam N. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley: New York, 1972.

24 Bullock AD, Hassell A, Markham WA, Wall DW, Whitehouse AB. How ratings vary by staff group in multi-source feedback assessment of junior doctors. *Med Educ* 2009; **43**: 516–520.

25 Marriott J, Purdie H, Crossley J, Beard J. Implementing the assessment of surgical skills and nontechnical behaviours in the operating room. In *Safer Surgery: Analysing Behaviour in the Operating Theatre*, Flin R, Mitchell L (eds). Ashgate: Farnham, 2009; 47–66.

26 Kogan J, Holmboe E, Hauer K. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA* 2009; **302**: 1316–1326.