Evaluation of the Scrub Practitioners' List of Intraoperative Non-Technical Skills (SPLINTS) system (2012) *International Journal of Nursing Studies, 49,* 201-211

*Lucy Mitchell[1], Rhona Flin[1], Steven Yule[1], Janet Mitchell[2], Kathy Coutts[3], George Youngson[1]

[1] School of Psychology, University of Aberdeen, UK

[2] Aberdeen Royal Infirmary, UK

[3] Royal Aberdeen Children's Hospital, UK

*Corresponding author: Lucy Mitchell, School of Psychology, University of Aberdeen, Aberdeen, AB24 2UB, Scotland

(e-mail: l.mitchell@abdn.ac.uk).

Abstract

Background

The Scrub Practitioners' List of Intraoperative Non-Technical Skills (SPLINTS) system is a new tool for training and assessing scrub practitioner (nurse, technician) behaviours during surgical operations.

Objectives

The aim of the study was to test the psychometric properties including inter-rater reliability of the prototype SPLINTS behavioural rating system.

Methods

Experienced scrub practitioners (n=34) attended a one day session where they received background training in human factors and non-technical skills and were also trained to use the SPLINTS system. They then used SPLINTS to rate the scrub practitioners' non-technical skill performance in seven standardized simulated, surgical scenarios.

Results

Reliability, measured by within-group agreement (r_{wg}) for the three skill categories and six out of nine elements, was acceptable (r_{wg}>0.7). Participants were within one scale point of expert ratings in > 90% of skill categories and elements, and could use SPLINTS to score performance with a reasonable level of accuracy. There was good internal consistency of the system: absolute mean difference was M<0.2 of a scale point for all three categories. Participants were surveyed and they indicated that the system was complete and usable as an assessment tool.

Conclusion

The reliability of the SPLINTS system was deemed to be adequate for assessing scrub practitioners' non-technical skills in simulated, standardized, video scenarios. On the basis of these results, the system can now move on to usability testing in the real operating theatre.

Summary Statement

What is already known about this topic

- Non-technical (cognitive and social) skills are an essential element of safe and efficient task performance for staff working in the operating theatre.
- Previous research has identified taxonomies of non-technical skills for surgeons, anaesthetists and scrub practitioners.
- Behavioural rating systems can provide a structured method for training and rating non-technical skills.

What this paper adds

- The SPLINTS system provides scrub practitioners with a structured method for discussing, training and rating non-technical skills that are required for safe and effective performance, during surgical procedures.
- Even with minimal training, scrub practitioners can use the SPLINTS behavioural rating system to reliably rate the non-technical skills performance of scrub practitioners seen in simulated, standardized video scenarios.
- Empirical evidence gathered from subject matter experts, that the prototype SPLINTS system appears complete and usable.

Keywords

Rating, training, assessment, non-technical skill, operating theatre, nurse, scrub nurse, scrub practitioner

1. Introduction

1.1 Background

Adverse events are unintended injuries or complications caused by the management of the patients' care rather than by the disease itself (Brennan et al., 1991). A systematic review of adverse event studies suggested that approximately 41% of all hospital adverse events occur in the operating theatre (deVries et al., 2008) and retrospective patient record reviews have suggested that around half of all identified adverse events were preventable (Vincent et al., 2001). In the operating theatre, various problems can occur, for example, swabs and instruments are still sometimes retained within patients (Gawande et al., 2003). Reasons for this include breakdown in communication within the nursing team as well as between nurses and surgeons (Riley et al., 2006) or difficulty experienced by nurses in speaking up effectively (Bromiley and Mitchell, 2009), in the hierarchical atmosphere that still pervades some operating theatres.

The operating theatre requires clinicians from different training backgrounds to work together in a coherent manner towards a common goal – the safe surgery of the patient (see Flin and Mitchell, 2009). The scrub practitioner (nurse, operating department practitioner, instrument technician) is a key member of the operating theatre team. The scrub practitioner is scrubbed, obtains and hands instruments and surgical supplies to the surgeon and has many responsibilities. These include ensuring that all equipment used during an operation is accounted for at the end of a surgical procedure.

Behavioural rating systems have already been developed for training and assessing the non-technical skills of anaesthetists (ANTS) (Fletcher et al., 2004) and surgeons (NOTSS) (Yule et al., 2008) but have not yet been produced for scrub practitioners. Non-technical skills are the social and cognitive skills which, combined with good technical expertise, lead to safe and effective performance (see Flin et al., 2008). These behavioural rating systems are hierarchical in structure in that they comprise of a set of skill 'categories', at the highest level, with a second level containing the 'elements', which are the main component skills underpinning each skill category. A

third level provides examples of good and poor behaviours (i.e. behavioural markers), to which the user of the system may refer as a guide when making ratings at the category and element levels (see Flin et al., 2008).

There are also tools for rating operating theatre teamwork, which have adapted the NOTECHS (van Avermaete and Kruijsen, 1998) behavioural rating system for pilots' non-technical skills. For example, the Observational Teamwork Assessment for Surgery (OTAS) (Undre et al., 2007), Oxford NOTECHS (Mishra et al., 2009) and Revised NOTECHS (Sevdalis et al., 2008) methods enable ratings of the three theatre sub-teams, i.e. surgical, anaesthetic and nursing. Example behaviours for the nursing sub-team are provided in these tools although circulating and recovery nurse behaviours are also included in the nursing component, since their purpose is to measure overall theatre team performance.

A review of the literature indicated that a behavioural rating system did not exist specifically for the scrub practitioner (Mitchell and Flin, 2008). To address this, we developed a taxonomy and behavioural rating system of non-technical skills for scrub practitioners (SPLINTS) using methods of task analysis (Kirwan and Ainsworth, 1992) that included a literature review (Mitchell and Flin, 2008), observations and interviews with experienced scrub nurses and consultant surgeons (Mitchell et al., 2011). The emergent, preliminary skill taxonomy was refined using focus group discussions (Whiddett and Hollyforde, 2006) with subject matter experts (n=4 focus groups; total participants n=16) using an iterative process (Gordon, 1994). The resulting skill set contained three categories (situation awareness; communication and teamwork; task management), each with three underlying skill elements. Examples of good and poor performance for each element were also provided to guide users of the system. Figure 1 shows the SPLINTS prototype taxonomy.

Insert Figure 1 about here

The content validity of the SPLINTS system was derived from its systematic development by a multi-disciplinary steering group of subject matter experts: operating theatre clinicians as well as psychologists.

1.2 Evaluation of behavioural rating systems

The need to evaluate behavioural rating tools is recognised in aviation (Flin and Martin, 2001, O'Connor et al., 2008). Both ANTS (Fletcher et al., 2003) and NOTSS (Yule et al., 2008) were evaluated before being used in real operating theatre settings. Evaluation of any rating system is important if it is to be used to assess training effectiveness (O'Connor et al., 2008). The SPLINTS system must be able to measure performance on the skills it is designed to evaluate; moreover it must be usable by different individuals in a consistent manner, to achieve the same ratings for equivalent standards of behaviour (Murdaugh, 2008). The rating framework needs to be complete, within its specified parameters, and the structure, definitions, language and layout of the system must be useable (Gordon, 1994), in this case, by scrub practitioners, with a minimal amount of training. The aim of this study was to investigate the validity, reliability, sensitivity and usability (in a simulated setting) of the SPLINTS system, by developing and delivering Crew Resource Management-style (Kanki et al., 2010) training for participating scrub practitioners.

2. The Study

2.1 Method

2.1.1 Design

Scrub practitioners attended a one day session during which they were introduced to human factors concepts (3hrs) and trained in the use of the SPLINTS system (2hrs). After practice with using the tool to rate behaviours seen in a simulated video scenario (1hr), they used SPLINTS to rate the performance of scrub practitioners seen in seven standardized simulated, surgical video scenarios (1hr). A definitive rating for each of the non-technical skills demonstrated in the scenarios had been obtained from the two subject matter experts on the project team. To achieve this, they provided independent ratings (all of which were within one scale point) before discussing those ratings until a consensus was reached. These agreed ratings are referred to as ‘reference’ ratings and were used as a benchmark in subsequent analyses.

2.1.2 Materials

<u>Simulated video scenarios</u>

Development of the scenarios was guided by the project steering group subject matter experts who considered a range of routine and non-routine surgical events with which the scrub practitioner may be faced. Scenarios were loosely scripted and then filmed in real operating theatres, utilising nursing and medical staff 'acting' in their own roles so as to be as realistic as possible. Scenarios were selected to depict a range of cases and different intraoperative situations; e.g. discovering that he or she is missing a swab during the counting procedure or; during a laparoscopic cholecystectomy, assisting the surgeon who gets into difficulty and has to convert to an open procedure. Each was introduced by a text that described the case type and stage; e.g. "We join as the surgeon is trying to control unexpected blood loss in a groin dissection". There were seven scenarios used to evaluate the SPLINTS system. This number of scenarios enabled depiction of levels of scrub practitioner performance, for the different skill categories and elements, so that raters had the opportunity to discriminate between the four points on the rating scale, for the non-technical skill behaviours in the taxonomy. The seven evaluation scenarios ranged from 1min 58sec to 4min 19sec in length (*M*=3min 2sec). An additional scenario was filmed for enabling participants to practice using the SPLINTS rating form before they rated the evaluation scenarios (12min 13sec) – this was designed in such a way as it could be paused on two occasions for discussion among participants, so that they had an opportunity to compare ratings.

Training package

The training package was designed to include the main aspects of rater training recommended for users of rating systems; e.g. behavioural observation training and rater error training (Baker et al., 2001). There were time constraints associated with the delivery of training since participants had to be released from clinical duties. So they only received five hours of training, and one hour of practice, as opposed to the two days that are recommended by Klamfer et al. (2001), to adequately train assessors in the use of behavioural rating scales. Unfortunately, there was no opportunity for individuals to receive feedback on their own rating performance or to calibrate their rating skills among their training group as has been recommended when training individuals to use behavioural rating systems (Baker et al., 2001). The training incorporated background on human factors, with explanations of the underlying psychological theory (e.g. human error, Reason, 1990). There were also exercises for

participants, e.g. a memory test to illustrate the limited capacity of working memory (Baddeley and Hitch, 1974). Participants were also introduced to each of the non-technical skill categories, elements and behavioural markers within the SPLINTS system.

SPLINTS system handbook

Participants were given a SPLINTS handbook (see **www.abdn.ac.uk/iprc/splints**) which contained detailed information on the background and rationale for the development of the system, as well as comprehensive definitions of the non-technical skill categories and elements contained in the SPLINTS taxonomy. It also suggested examples of good and poor practice for each element (behavioural markers). Participants were able to refer to this handbook throughout the training and evaluation session.

Rating forms

The rating scale that had been chosen for the SPLINTS system was a 4-point rating scale of 1-poor, 2-marginal, 3-acceptable and 4-good, with the additional option of NR for situations when that skill was 'not required' in that particular surgical situation. Each participant was given eight separate rating forms on which they recorded the ratings for scrub practitioner's skills in the training session and then in each of the seven evaluation scenarios.

Background information and evaluation questionnaire

At the conclusion of the evaluation day, participants completed a two-part questionnaire. Part one gathered background information in relation to participants' involvement with training junior members of perioperative staff, assessment activities and knowledge of non-technical skills, as well as any previous involvement in the SPLINTS system development. Basic demographic data (sex, years of experience) were also obtained in this part of the questionnaire.

Part two comprised the SPLINTS evaluation questions. These included; i) 5 questions about completeness of the system and observability of the skills (validity), ii) 5 questions on acceptability and 4 questions on the potential role of SPLINTS in perioperative practice (usability), iii) 4 questions on the rating scale, iv) 5 questions

on the training received during the evaluation day and, v) 3 questions about the film scenarios. The questionnaire included a combination of closed questions (with yes/ no responses or a 5-point scale) and 'free text' response questions.

2.1.3 Procedure

Participant recruitment

Clinical leads at five NHS (National Health Service) hospitals in Scotland agreed to participate in the evaluation study. Recruitment posters were displayed in rest areas of these hospitals to recruit volunteer scrub practitioners with a minimum of two years experience. Participants contacted the researcher who arranged mutually convenient dates at each hospital for participants to attend a one day session. It was hoped to recruit 40-50 nurses as had been recruited in previous studies but it only proved possible to recruit 34, which was adequate for the tests to be performed. The number of participants in each session varied from four to nine.

Training delivery and data collection

A psychologist ran one pilot session, attended by nine participants (including the two experts on the project steering group). Following minor adjustments to the training material, the psychologist conducted seven subsequent evaluation sessions at five teaching hospitals between April and June, 2010. Following training and practice using the SPLINTS system, participants immediately rated the evaluation scenarios (n=7) and, with the SPLINTS handbook for reference, completed a separate SPLINTS rating form for each scenario, without conferring. At the end of the session participants completed the demographic and evaluation questionnaire.

2.1.4 Participants

Participants were 34 scrub practitioners from five Scottish teaching hospitals, of whom 7 were male, with a mean scrub practitioner experience of 17 years (SD = 8.22; range 2-35 years). The majority (91%) indicated they had experience of assessing junior scrub practitioners' performance and 22 of those (70%) had received some form of training for providing assessments although mostly on an informal basis. Less than half (40%) indicated they had no previous knowledge of human factors and 32% had knowledge through involvement in previous stages of the SPLINTS system development process by being interviewed, observed or a focus group member who

refine the SPLINTS skill taxonomy, however, most were unfamiliar with the finalized prototype SPLINTS system.

2.1.5 Data analysis

Data from the rating forms and the questionnaires were analysed using PASW Statistics Version 18 (SPSS Inc., 2009) and Microsoft Excel. Table 1 shows the evaluation study questions, data sources and methods of analyses used to assess the reliability and psychometric properties of the SPLINTS system.

Insert Table 1 about here

Reliability

Within-group agreement (r_{wg}) (James et al., 1984, James et al., 1993) was calculated for the participants' ratings of the SPLINTS elements and categories in each of the seven scenarios. The average across the scenarios was calculated for each category and element and these scores were taken as the overall within-group agreement of the SPLINTS system. The r_{wg} statistic lies between 0=no agreement and 1=complete agreement and represents the degree to which a number of raters agree on the absolute ratings they provided. The generally accepted criterion for acceptable level of agreement is $r_{wg} > 0.7$-0.8 (Nunnally and Bernstein, 1994).

Accuracy/ sensitivity

Basic accuracy of the SPLINTS system was calculated by comparing the participants' ratings with a set of 'reference' ratings. The mean absolute deviation (Goldsmith and Johnson, 2002) from the set of 'reference' ratings, agreed by the subject matter experts, was calculated, e.g. the participant gives a score of 4 (good) and the reference rating is 3 (acceptable) giving an absolute difference of one. The mean of those differences across the seven scenarios was calculated across all participants to give an average 'error' score for each element. Lower numbers indicate a smaller deviation from the expert reference rating, suggesting higher sensitivity, which is desirable. It can be considered to provide a measure of sensitivity because, since the scenarios showed a range of performance levels, if the ratings are accurate then the raters must have been sensitive to the variations in performance.

Internal consistency

As SPLINTS is a hierarchical system, there should be consistency of ratings provided at the element level with the corresponding category under which those elements lie; i.e. if performance at the element level is rated as 1 or 2 then the category rating should also be given a rating of 1 or 2. However, if the category rating was judged as 4, this would infer that those elements do not 'belong' under that category. The mean absolute difference was calculated between ratings at the element level and ratings for the corresponding category. This was achieved by calculating the mean difference among the three elements in each category, across the seven scenarios before calculating the mean absolute difference for each category. Low scores indicate close agreement within the system.

Validity

The SPLINTS system was developed to describe the main (observable) non-technical skills important for good scrub practitioner practice. Having used a systematic empirical method with experienced practitioners to develop the SPLINTS system, a reasonable level of content validity (Litwin, 2003) was expected. The completeness and observability of the skills assessments indicated whether the SPLINTS system actually measures what it is supposed to measure (Holt et al., 2001). The former is whether the SPLINTS system is suitably comprehensive and this was assessed by analysing the responses to the questionnaire using frequency analysis and content review. Observability was assessed from the questionnaire responses and also by calculating percentages of ratings made by participants rather than using the NR (not required) rating or leaving the rating box blank (treated as missing cases).

Usability

Acceptability and usability (Jordan, 1998) of the SPLINTS system, in a simulated setting, were assessed by analysing the questionnaire data. Descriptive analysis and content review of free-text responses were reported.

2.1.6 Ethical considerations

Relevant ethical approval was granted for the study from the University of Aberdeen School of Psychology Ethics Committee and the North of Scotland Research Ethics Committees (refs: pRGF/002/10; 10/S0801/5). All participants were allocated a

participant number which enabled their ratings and questionnaire data to remain anonymous.

3. Results

3.1 Reliability

Within-group agreement

Within-group agreement scores were acceptable for each of the three categories (r_{wg}>0.7) (Nunnally and Bernstein, 1994). Of the nine underlying elements, within-group agreement was acceptable (r_{wg}>0.7) for six of them. The 'providing and maintaining standards' and 'coping with pressure' elements underlying the skill category 'task management' (both; r_{wg}=0.66) and the 'gathering information' element of the category 'situation awareness' (r_{wg}=0.69) did not reach the *a priori* criteria for reliability. Table 2 shows the mean r_{wg} scores across the seven scenarios for the three categories and nine underlying elements.

Insert Table 2 about here

Between scenario differences

Within-group agreement was higher within some scenarios than others. For example, there was perfect agreement (r_{wg}=1.0) in scenarios 4 and 5 for the 'anticipating' element of 'situation awareness' in both those scenarios. Within-group agreement was much lower for the 'anticipating' element in other scenarios (i.e. 3 and 6), where agreement was r_{wg}=0.41 and r_{wg}=0.59, respectively. Within-group agreement in scenario 4 was good (r_{wg}= 1.0, 0.88, 0.91) for the categories of 'situation awareness', 'communication and teamwork' and 'task management', respectively. In scenario 6 however, agreement did not reach *a priori* criteria for those categories (r_{wg} = 0.51, 0.55, 0.60, respectively).

Accuracy/ sensitivity

Column 2 in table 3 displays the mean absolute differences between the participants' ratings and the corresponding reference rating for each of the categories and elements. The average SPLINTS sensitivity was 0.50 and 0.49 of a scale point, at the category and element levels, respectively. To further check accuracy of ratings, percentages of

ratings made to within one scale point of the reference ratings were calculated for categories and elements across all seven scenarios. Mean percentages ranged from 95-97% for categories and 91-98% for elements, shown in column 3 in table 3.

Insert table 3 about here

Internal consistency

The mean absolute difference between raters' element ratings and their ratings for the corresponding categories was calculated. Consistency between the three categories and the corresponding elements was very high with a mean absolute difference (M<0.2 of a scale point, on a four point scale) indicating that there was good internal consistency of the system, and those results can be seen in figure 2.

Insert Figure 2 about here

3.2 Validity

Results for content validity are presented in table 4.

Insert table 4 about here

Completeness

Completeness of the system was assessed by data gleaned from the questionnaire. All participants indicated that the SPLINTS system addressed the key non-technical skills required for the scrub practitioner performance. However, professional conduct and resolving conflict were noted as being skills which were absent from the skill set.

Observability

Non-technical skills of scrub practitioners can be identified (from behaviours seen in realistic scenarios showing scrub practitioner performance in the intraoperative phase of surgery) using the SPLINTS system. The majority of participants reported that it was either, very easy (12%), easy (50%) or average (35%) to associate observed behaviours with elements and very easy (6%), easy (62%), or average (32%) to associate behaviours with categories. Comments indicated that they felt it would get easier to use the SPLINTS system with practice and that when the behaviours were

more extreme, they found it easier to assign ratings to performance. Column 1 in table 3 displays the mean percentages across the seven scenarios of where participants had made a rating as opposed to using the NR option, or leaving the box unmarked. At the category level this was very high; 99% - 100% and at the element level, ranged from 94%-99% observability.

3.3 Usability

Results for usability of the SPLINTS system in a simulated setting are summarised in table 5.

Insert table 5 about here

Acceptability

The SPLINTS system was judged as a useful tool for making observations and for structuring feedback on performance by 94-100% of participants. The only issues related to the sound quality of the scenarios and that more practice using the system would be needed to become more familiar with it. The vast majority indicated that it could be a useful teaching support tool and that it would also provide a record of performance that could subsequently be referred to if required.

Usability

All 34 participants indicated that the descriptions and examples of all the categories and elements were clear and understandable with the exception of one participant who suggested that the behavioural marker; 'arranges for colleague to enter theatre if it appears surgeon would benefit from assistance' in the 'task management' category should have the word 'surgical' inserted before 'colleague' as this is referring to a scrub practitioner discreetly arranging for additional surgical expertise to enter theatre if she or he recognises that the surgeon requires the assistance of a surgical colleague.

Suggested uses for SPLINTS in the free-text section included; to improve performance, professionalism and attitude (n=5); as part of ongoing assessment/ training needs (n=20); for self reflection (n=2).

4. Discussion

The results of this study suggest that the SPLINTS system is an adequately reliable tool to enable progression to testing its usability in the real operating theatre. It can be used by scrub practitioners with a reasonable level of accuracy to rate the non-technical skill performance of a scrub practitioner, seen in simulated surgical video scenarios, and that the system appears to be complete and usable, even with limited training.

Within group agreement was acceptable (r_{wg}>0.7) for the three skill categories and for six of the nine underlying elements. This is encouraging, given the short duration of training that the participants received, particularly since none of the participants had received any formal non-technical skills training or input prior to the training session. There were between scenario differences, which is why the mean agreement score is a better measure of the system's overall inter-rater reliability. When designing and filming the scenarios, the project team tried to ensure that there was at least one scenario that depicted extremely good and one with extremely poor performance and it was accepted that these scenarios were likely to be 'easier' to rate as the behaviours are, by definition, more extreme. This appeared to be the case where the inter-rater agreement for scenario four was extremely high. This scenario had been designed to depict an exemplar performance of a scrub practitioner. Scenario one, on the other hand had been designed to depict extremely poor performance and the results show that the ratings were more variable for that scenario, suggesting that the simulated poor performance may have been more ambiguous for participants to rate.

These results compare favourably with the initial reliability data for other behavioural rating tools, e.g. for anaesthetists (ANTS) and surgeons (NOTSS). In the original ANTS evaluation (Fletcher et al., 2003) where 50 anaesthetists rated performance in eight scenarios, none of the four categories (r_{wg}=0.56-0.65) or 15 elements (r_{wg}= 0.55-0.66) reached the acceptable agreement criteria (r_{wg} = 0.7) (Nunnally and Bernstein, 1994). The evaluation of the NOTSS system for surgeons (Yule et al., 2008) showed acceptable within-group agreement for the social categories of 'leadership' and 'communication and teamwork' but had lower values of r_{wg} for the cognitive skills. These were judged as acceptable for the rating systems to undergo further testing and

both ANTS and NOTSS are now being evaluated in conjunction with performance based assessment tools (Graham et al., 2007, Marriott et al., 2011).

There was low within-group agreement for the 'coping with pressure' element in the SPLINTS system however, this might have been anticipated since advice on developing behavioural marking systems suggests that although it is important to develop skills associated with stress management, they are not normally included in behavioural rating systems as they are difficult to rate unless extreme symptoms are displayed (Flin et al., 2008). However, the NOTSS system for surgeons includes a 'coping with pressure' element that did not meet the acceptable criteria (r_{wg}=0.68) when NOTSS was tested using a similar method to the one used in the present study (Yule et al., 2008). 'Coping with pressure' was viewed as a crucial element of 'task management' during previous stages of the SPLINTS system development which is why it is in the taxonomy and it may be a function of the short video scenarios (*M*=3min 2sec) that this element was not explicitly displayed and that in the real operating theatre, this element will be easier to rate.

The highest average sensitivity score was 0.55, for the 'planning and preparing' element. This means that the largest average deviation was 0.55 of a scale point either higher or lower than the 'accurate' score. This was taken as an adequate measure of the sensitivity of the system, even with minimally trained raters. The percentages of participants who had rated to within one scale point either side of the reference rating were high which provided further evidence. However, the rating system is relatively short; i.e. a 4-point scale which means that there is limited room for flexibility in the scoring as only points 2 and 3 are capable of being rated within one scale point in both directions (higher and lower) so, these particular results of accuracy should be interpreted with caution.

The internal consistency of the SPLINTS system was very good. Since behavioural rating systems are hierarchical in structure (Flin et al., 2008), it is crucial that the elements that underpin the category are in fact, related to that category. In SPLINTS, the mean absolute difference was *M*<0.2 of a scale point for all three categories which indicates that participants were giving similar ratings for the elements and the category to which they relate. However, we do not know whether participants were

providing the category rating by averaging the ratings they had noted for the elements or if they rated the category first.

Participants indicated that the key non-technical skills were all included in the SPLINTS system, although participants were seeing the full skill set for the first time so, although it appears to be complete, further testing of the system would be required to confirm this. Two participants suggested that conflict and professional conduct had been omitted. There is a 'providing and maintaining standards' element underlying the 'task management' category (see figure 1), which could have been utilised for judging professional conduct. Similarly, conflict could be assessed under the 'coordinating with others' element of that category. The participants had limited time using the SPLINTS system and the behavioural markers are not an exhaustive list so it may be that with more practice using the system, users would become better at matching observed behaviours with the categories and elements in the rating system. The comment made by a participant to change the wording of the behavioural marker in the 'providing and maintaining standards' element to explicitly state a that it is a *surgical* colleague that should be called upon when the scrub practitioner recognises that the surgeon at the table would benefit from assistance will be taken into account in future versions of the SPLINTS system.

There were high percentages of observability, indicating that the behaviours in the system are observable. Ratings were provided for the vast majority of categories and elements in the SPLINTS system so participants appeared able to identify the skills demonstrated by the scrub practitioner behaviour in the simulated scenarios. The unmarked boxes were randomly distributed suggesting that they were missing data rather than participants' inability to rate a particular behaviour. However, we do not know whether the missing data were because the participants forgot to make a rating, did not see that particular behaviour on that occasion, or did see the behaviour but could not decide what rating (1-4) to give the scrub practitioner. So, this is worth considering in future testing of the SPLINTS system. There were no major problems with the design and layout since participants indicated that they found the wording and labelling within the SPLINTS system meaningful and clear. They said that it appeared to be a useful tool for various purposes including, structuring feedback, and

providing a record of performance and as a 'back-up' where problems in performance are identified.

4.1 Strengths and limitations

Testing the reliability of the SPLINTS system using standardized scenarios enabled the evaluation method to be consistent when delivered to different groups in various locations. Having designed and filmed the scenarios, we were able to ensure that the fullest range of behaviours and levels of performance could be depicted. By analyzing the results across seven scenarios featuring different scrub practitioners, in different intraoperative situations, the results were based on a range of demonstrated clinical activities.

Participants were highly experienced clinicians and feedback was entirely positive. They felt the system met a professional need because there is no formal training in the United Kingdom curriculum for non-technical skills, which they recognised as very relevant to their practice. These skills are learned in an ad-hoc manner and the participants, 91% of whom indicated that they are currently, or have been involved with training or mentoring junior staff, explained that this would provide a structured means for talking about and training these skills. The results indicate that the SPLINTS system has a consistent internal structure, and can be used with a reasonable level of accuracy to rate performance in simulated scenarios, when compared with subject matter expert ratings. Participants indicated that the SPLINTS system is usable and contains the most important non-technical skills for the scrub practitioner to perform effectively.

The main limitation to the study was that participants received minimal training. It is generally recommended that a 2-day training course be undertaken before one is competent to use this type of training system (Klampfer et al., 2001). However, this study was completed at a time when staffing levels in NHS operating theatres were critical and it was extremely difficult to release the scrub practitioners for the full day required to provide the training and run the evaluation. Neither were the participants given feedback or allowed to discuss or calibrate their ratings (Baker et al., 2001) apart from the practice scenario, as independent judgements were required for the evaluation study. Although we had planned to recruit 40-50 participants, the smaller

number of participants (min n=4; max n=9; total n=34) enabled intimate group sessions which were interactive and generated useful discussions during the training and practice portions of each session. Longer, more comprehensive training and increased opportunities to practise making assessments would improve participants' unfamiliarity with the system, resulting in increased confidence in providing the ratings and feedback. Another limitation related to the 'reference ratings' since these were provided by the two subject matter experts on the project team who, have exceptional nursing experience, but have had no formal or theoretical training in identifying non-technical skills. In the absence of 'gold standards' for non-technical skills, this was the most appropriate method for capturing the level of skill that was actually depicted in the scenarios against which to test participants' ratings.

The use of semi-scripted recorded scenarios rather than live or recorded operating theatre situations meant that some of the scenarios may have seemed more realistic than others. Even though the 'actors' were actual clinicians performing in their own roles, some were better at acting than others. Participants indicated that the scenarios seemed realistic but that, in a real environment there would be more background information available rather than the simply a short text introduction on the simulated scenarios. Also, when observing a real surgical case, there would be periods where there is very little activity, an aspect that was not reflected in the short 'action packed' simulated video scenarios so these are somewhat artificial, even if the behaviours were judged by participants to be authentic. It is accepted that the sound quality in some of the scenarios could have been improved however, in a real operating theatre there are often competing sounds which make hearing critical information difficult.

Despite the limitations to this study, the prototype SPLINTS system offers scrub practitioners a new method for training and assessing this important aspect of their performance. It will require further testing in the operating theatre but we hope that in future, this system will join the range of tools available for assessing the non-technical skills of other individuals, e.g. anaesthetists ((ANTS), Fletcher et al., 2003) and surgeons (NOTSS, Yule et al., 2008), in the operating theatre. These practitioner tools, add to the research tools developed for observing operating theatre teamwork e.g. OTAS (Undre et al., 2006), Oxford NOTECHS (Mishra et al., 2009) and Revised NOTECHS (Sevdalis et al., 2008). Together, these frameworks offer clinicians a

better understanding of the importance of good non-technical skills to their performance, and provide researchers with the means of observing and analysing behaviour (Flin and Mitchell, 2009), in an attempt to establish methods to reduce adverse events in the operating theatre.

5. Conclusion

In spite of the limitations of the study, participants were able to use the SPLINTS system with an acceptable standard of accuracy and reliability and as such, it is now being trialled as a training tool in the operating theatre, in four Scottish hospitals. This will assess the usability of the SPLINTS system in the real operating theatre environment and results from that, combined with this study, may lead to refinements of the SPLINTS system for scrub practitioners' non-technical skills. Providing a common language and a structured method for rating and training non-technical skills could take scrub practitioners one step closer to reducing the still unacceptably high adverse event rate seen in the operating theatre (deVries et al., 2008).

Acknowledgements

Thank you to all the scrub practitioners who took part in the evaluation sessions and to those who facilitated running the sessions in each of the teaching hospitals. Thanks also to the cast and recording crew of the scenarios and to Lauren Ferrier and Stephen Punton for editing the film footage into usable, realistic scenarios. Thanks to Dr Michael Burtscher for his assistance with statistical analysis.

Funding

This work was conducted under the Scottish Research Network with funding from a Scottish Research Development Grant from the Scottish Funding Council. The SPLINTS system was developed under funding from NHS Education for Scotland (NES).

References

Baddeley, A.D., Hitch, G.J., 1974. Working memory. In: Bower, G.H. (Ed.), The Psychology of Learning and Motivation. Academic Press, New York, pp. 47-90.

Baker, D.P., Mulqueen, C., Dismukes, R.K., 2001. Training raters to assess resource management skills. In: Salas, E., Bowers, C., Edens, E. (Eds.), Improving Teamwork in Organisations: Applications of Resource Management Training. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 131-145.

Brennan, T.A., Leape, L.L., Laird, N.M., Herbert, L., Localio, A.R., Lawthers, A.G., Newhouse, J.P., Weiler, P.C., Hiatt, H.H., 1991. Incidence of adverse events and negligence in hospitalised patients. Results of the Harvard Medical practice study I. The New England Journal of Medicine 324, 370-376.

Bromiley, M., Mitchell, L., 2009. Would you speak up if the consultant got it wrong?...and would you listen if someone said you'd got it wrong? . Journal of Perioperative Practice 19, 326-329.

deVries, E.N., Ramrattan, M.A., Smorenburg, S.M., Gouma, D.J., Boermeester, M.A., 2008. The incidence and nature of in-hospital adverse events: a systematic review. Quality and Safety in Health Care 19, 216-223.

Fletcher, G., Flin, R., McGeorge, P., Glavin, R., Maran, N., Patey, R., 2003. Anaesthetists' Non-Technical Skills (ANTS): Evaluation of a behavioural marker system. British Journal of Anaesthesia 90, 580-588.

Fletcher, G., Flin, R., McGeorge, P., Glavin, R., Maran, N., Patey, R., 2004. Rating non-technical skills: Developing a behavioural marker system for use in anaesthesia. Cognition Technology and Work 6, 165-171.

Flin, R., Martin, L., 2001. Behavioural markers for CRM: A survey of current practice. International Journal of Aviation Psychology 11, 95-118.

Flin, R., Mitchell, L., 2009. Safer Surgery: Analysing Behaviour in the Operating Theatre. Ashgate, Farnham.

Flin, R., O'Connor, P., Crichton, M., 2008. Safety at the Sharp End. A Guide to Non-Technical Skills. Ashgate, Aldershot.

Gawande, A.A., Studdert, D.M., Orav, E.J., Brennan, T.A., Zinner, M.J., 2003. Risk factors for retained instruments and sponges after surgery. New England Journal of Medicine 348, 229-235.

Goldsmith, T., Johnson, P., 2002. Assessing and improving evaluation of aircrew performance. International Journal of Aviation Psychology 12, 223-240.

Gordon, S.E., 1994. Systematic Training Program Design: Maximizing Effectiveness and Minimizing Liability. Prentice-Hall, Inc Englewood Cliffs, NJ.

Graham, J., Giles, E., Hocking, G., 2007. Using ANTS for workplace assessment. In: Flin, R., Mitchell, L. (Eds.), Safer Surgery: Analysing Behaviour in the Operating Theatre. Ashgate, Farnham, pp. 189-201.

Holt, R.W., Boehm-Davis, D.A., Beaubien, J.M., 2001. Evaluating resource management training. In: Salas, E., Bowers, C., Edens, E. (Eds.), Improving Teamwork in Organizations. Applications of Resource Management Training. Lawrence Erlbaum Associates, Mahwah, NJ, pp. 165-188.

James, L.R., Demaree, R.G., Wolf, G., 1984. Estimating within-group interrater reliability with and without response bias. Journal of Applied Psychology 69 (1), 85-98.

James, L.R., Demaree, R.G., Wolf, G., 1993. rwg: An assessment of within-group interrater agreement. Journal of Applied Psychology 78, 306-309.

Jordan, P.W., 1998. An Introduction to Usability. Taylor & Francis, London.

Kanki, B., Helmreich, R.L., Anca, J., 2010. Crew Resource Management. Elsevier, San Diego, CA.

Kirwan, B., Ainsworth, L.K., 1992. A Guide to Task Analysis. Taylor & Francis, London.

Klampfer, B., Flin, R., Helmreich, R.L., Hausler, R., Sexton, B., Fletcher, G., Field, P., Staender, S., Lauche, K., Dieckmann, P., Amacher, A., 2001. Group interaction in high risk environments: Enhancing performance in high risk environments, recommendations for the use of behavioural markers. GIHRE, Berlin.

Litwin, M.S., 2003. How to Assess and Interpret Survey Psychometrics. Sage, Thousand Oaks, CA.

Marriott, J., Purdie, H., Crossley, J., Beard, J.D., 2011. Evaluation of procedure-based assessment for assessing trainees' skills in the operating theatre. British Journal of Surgery 98, 450-457.

Mishra, A., Catchpole, K.R., McCulloch, P., 2009. The Oxford NOTECHS System: reliability and validity of a tool for measuring teamwork behaviour in the operating theatre. Quality and Safety in Health Care 18, 104-108.

Mitchell, L., Flin, R., 2008. Non-technical skills of the operating theatre scrub nurse: literature review. Journal of Advanced Nursing 63, 15-24.

Mitchell, L., Flin, R., Yule, S., Mitchell, J., Coutts, K., Youngson, G.G., 2011. Thinking ahead of the surgeon: an interview study to identify scrub nurses' non-technical skills International Journal of Nursing Studies 48, 818-828.

Murdaugh, L.B., 2008. Designing and managing a competence assessment program. In: 4th (Ed.), Competence Assessment Tools for Health-System Pharmacies. American Society for Health System Pharmacists, Inc, Bethesda, MD, pp. 13-21.

Nunnally, J., Bernstein, I., 1994. Psychometric Theory. McGraw Hill, New York.

O'Connor, P., Campbell, J., Newon, J., Melton, J., Salas, E., Wilson, K.A., 2008. Crew Resource Management training effectiveness: A meta-analysis and some critical needs. The International Journal of Aviation Psychology 18, 353-368.

Reason, J., 1990. Human Error. Cambridge University Press, Cambridge.

Riley, R., Manias, E., Polglase, A., 2006. Governing the surgical count through communication interactions: Implications for patient safety. Quality and Safety in Health Care 15, 369-374.

Sevdalis, N., Davis, R., Koutantji, M., Undre, S., Darzi, A., Vincent, C., 2008. Reliability of a revised NOTECHS scale for use in surgical teams. The American Journal of Surgery 196, 184-190.

SPSS Inc., 2009. PASW Statistics, Version 18. . Chicago, IL.

Undre, S., Healey, A.H., Darzi, A., Vincent, C.A., 2006. Observational assessment of surgical teamwork: A feasibility study. World Journal of Surgery 30, 1774-1783.

Undre, S., Sevdalis, N., Healey, A.H., Darzi, A., 2007. Observational Teamwork Assessment for Surgery (OTAS): Refinement and application in urological surgery. World Journal of Surgery 31, 1373-1381.

van Avermaete, J.A.G., Kruijsen, E., 1998. The evaluation of non-technical skills of multi-pilot aircrew in relation to the JAR-FCL Requirements: Final Report NLR-CR-98443. National Aerospace Laboratory (NLR), Amsterdam.

Vincent, C., Neale, G., Woloshynowych, M., 2001. Adverse events in British hospitals: Preliminary retrospective record review. British Medical Journal 322, 517-519.

Whiddett, S., Hollyforde, S., 2006. A Practical Guide to Competencies: How to Enhance Individual and Organisational Performance. Chartered Institute of Personnel and Development, London.

Yule, S., Flin, R., Paterson-Brown, S., Maran, N., Rowley, D.R., Youngson, G.G., 2008. Surgeons' non-technical skills in the operating room: Reliability testing of the NOTSS behaviour rating system. World Journal of Surgery 32, 548-556.

Yule, S., Flin, R., Rowley, D., Mitchell, A., Youngson, G.G., Maran, N., Paterson-Brown, S., 2008. Debriefing surgical trainees on non-technical skills (NOTSS). Cognition Technology and Work 10, 265-274.

Category	*Element*
Situation awareness	• Gathering information • Recognising and understanding information • Anticipating
Communication and teamwork	• Acting assertively • Exchanging information • Coordinating with others
Task management	• Planning and preparation • Providing and maintaining standards • Coping with pressure

Figure 1 SPLINTS system prototype taxonomy

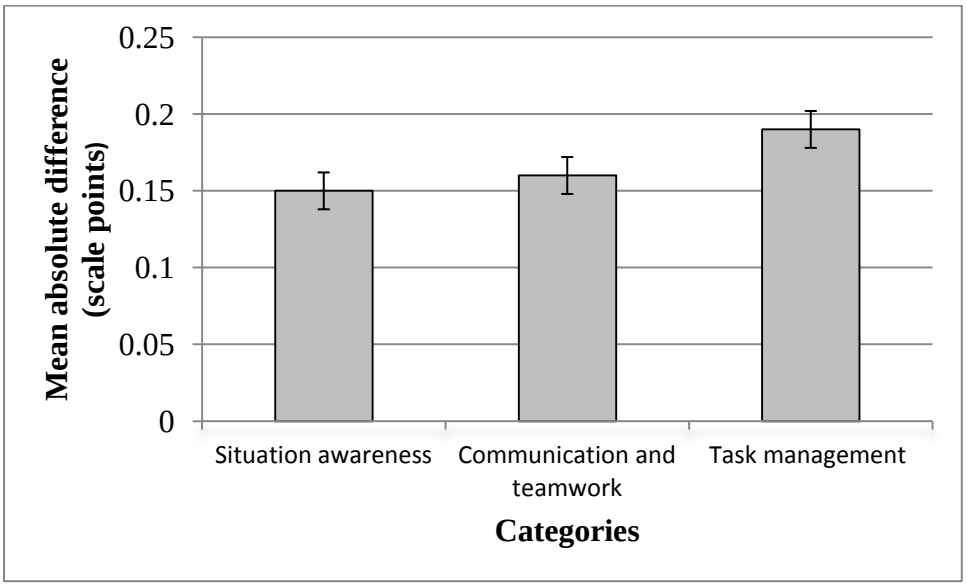

Figure 2 Mean and standard deviation of the absolute difference between the element and category levels

Table 1 Evaluation questions, data sources and analysis techniques

	Test	Evaluation question	Data source and analysis
Reliability	Within-group agreement	Can different raters use SPLINTS to rate performance at the category and element level to an acceptable level of within-group agreement?	Ratings data: within-group agreement statistic[a] to indicate the level of rater consensus (i.e. whether they rate performances the same):
	Accuracy/ sensitivity	Are ratings given at the category and element levels consistent with reference ratings agreed by subject matter experts?	Ratings data: Mean absolute deviation from the reference ratings and basic difference from the reference ratings to establish the level of accuracy or error for ratings.
	Internal consistency	Are ratings provided at the element level consistent with ratings at the category level?	Ratings data: Mean absolute difference[b] between raters' element ratings and their rating for the corresponding category.
Validity	Completeness	Does SPLINTS provide a comprehensive set of categories and elements?	Questionnaire data: basic frequency analysis and content review to identify any unnecessary or missing items.
	Observability	Can scrub practitioners' non-technical skills be identified by observation of behaviour using the SPLINTS system?	Ratings data: basic descriptive statistics. Questionnaire data: frequency analysis and content review.
Usability	Acceptability	Is the SPLINTS system acceptable as a training/ assessment tool?	Questionnaire data: Descriptive statistics and content review of participant opinions/ responses.
	Usability	Is the SPLINTS system usable in a simulated/ training environment?	Questionnaire data: Descriptive statistics and content review. Ratings data: Indication of effective use of system.

[a]James et al., 1984, 1993; [b]O'Connor et al., 2002

Table 3 Results for observability and accuracy/sensitivity averaged across all seven scenarios.

		1	2	3
SPLINTS		Observability (Mean % of observed ratings) [a]	Accuracy (mean absolute difference) [b]	Accuracy (% of ratings accurate ± 1 scale point)
Categories	Situation awareness	100	0.51	96%
	Communication and teamwork	99	0.49	97%
	Task management	99	0.51	95%
Elements	Gathering information	99	0.54	94%
	Recognising and understanding information	98	0.45	96%
	anticipating	98	0.49	97%
	Acting assertively	99	0.46	97%
	Exchanging information	99	0.54	91%
	Co-ordinating with others	98	0.47	97%
	Planning and preparing	98	0.55	92%
	Providing and maintaining standards	94	0.50	91%
	Coping with pressure	94	0.41	98%

[a] high percentages indicate a good level of observability, [b] low numbers indicate a low error rate and good accuracy compared with the reference ratings.

Table 2 Within-group agreement (r_{wg}) across seven experimental scenarios

SPLINTS		Scenario r_{wg} scores							Mean
		Sc 1	Sc 2	Sc 3	Sc 4	Sc 5	Sc 6	Sc 7	r_{wg}
Categories	Situation awareness	0.68	0.82	0.68	1.00	0.88	0.51	0.69	0.75
	Communication and teamwork	0.73	0.75	0.58	0.88	0.86	0.55	0.69	0.72
	Task management	0.70	0.81	0.64	0.91	0.86	0.60	0.67	0.74
Elements	Gathering information	0.70	0.75	0.53	0.88	0.86	0.46	0.67	0.69
	Recognising and understanding information	0.70	0.77	0.51	0.91	0.88	0.48	0.70	0.71
	anticipating	0.69	0.81	0.64	1.00	1.00	0.41	0.59	0.73
	Acting assertively	0.70	0.73	0.61	0.91	0.82	0.50	0.65	0.70
	Exchanging information	0.66	0.75	0.60	0.88	0.84	0.58	0.69	0.71
	Co-ordinating with others	0.72	0.76	0.51	0.91	0.86	0.46	0.65	0.70
	Planning and preparing	0.64	0.79	0.53	0.91	0.91	0.49	0.60	0.70
	Providing and maintaining standards	0.64	0.74	0.55	0.86	0.83	0.51	0.50	0.66
	Coping with pressure	0.64	0.75	0.39	0.82	0.73	0.58	0.69	0.66

Table 4 Summary of results for content validity of the SPLINTS system from the questionnaire

Evaluation criteria	*Result*
Completeness of SPLINTS system (n=34)	(1) Did it address the key non-technical skill behaviours displayed? Yes = 100% Comments; Professional conduct should be addressed
	(2) Do you think any elements and/ or categories are missing? No = 88%; Yes = 9%; Unsure = 3% Comments – Conflict; Professional conduct
	(3) Do you think any elements and/ or categories listed are unnecessary? No = 100%
Observability of NTS using SPLI TS (n=34)	(1) How easy was it to associate observed behaviours with SPLINTS elements? Very easy = 12%; Easy = 50%; Average = 35%; Difficult = 3% Comments; Easy when behaviour in scenario is clearly good or bad; First time using SPLINTS, does it get easier with practice?; Guidelines helped
	(2) How easy was it to associate observed behaviours with the SPLINTS categories? Very easy = 6%; Easy = 62%; Average = 32%; Difficult = 0% Comments; Categories seem more straightforward than elements; there is so some overlap

Table 5 Summary of results from the questionnaire for usability of SPLINTS system in a simulated setting

Evaluation criteria	*Results*
Acceptability of SPLINTS system (n=33-34)	(1) Was the SPLINTS system useful for structuring observation? Yes = 100% Comments; Need to improve sound quality of clips; more practice needed
	(2) Would SPLINTS system be helpful for mentors giving training to junior scrub practitioners? Yes = 97%; No = 3% Comments; Useful for all levels of staff; With practice could be very valuable tool
	(3) Do you think the SPLINTS system would be helpful for assessing junior scrub practitioner performance? Yes = 97%; No = 3% Comment; Would make it easier to give feedback; Provides a record to back me up if causes for concern
	(4) Do you think the SPLINTS system would be helpful for scrub staff in developing the skills needed to be a good perioperative practitioner? Yes = 94%, Not sure = 3%; Missing = 3%
	(5) Do you think the SPLINTS system could be used to support in-theatre teaching? Yes = 97%; no = 3% Comment; Could make staff aware of bad habits; Useful where issues need addressed
Usability of SPLINTS system (n=34)	(1) Was the wording used for the category and element labels meaningful? Yes = 97%, No = 3% Comment; Mostly fine, however task management maintaining standards para 4 needs rewording
	(2) Were the descriptions for each category and element clear? Yes = 100%
	(3) Were the examples of ‘good’ behaviours helpful for identifying the non-technical skill element?