Surgical Education

Reliable assessment of general surgeons’ non-technical skills based on video-recordings of patient simulated scenarios

Lene Spanager, M.D. a,*, Randi Beier-Holgersen, M.D. b, Peter Dieckmann, Ph.D. a, Lars Konge, Ph.D. c, Jacob Rosenberg, D.M.Sc. d, Doris Oestergaard, D.M.Sc. a

aDanish Institute for Medical Simulation, Capital Region of Denmark and University of Copenhagen, Copenhagen, Denmark; bDepartment of Surgery, Hillerød Hospital, Hillerød, Denmark; cCenter for Clinical Education, Capital Region of Denmark and University of Copenhagen, Copenhagen, Denmark; dDepartment of Surgery, Herlev Hospital and University of Copenhagen, Copenhagen, Denmark

KEYWORDS:
Surgery; Surgeon; Nontechnical skills; Operating room; Training; Assessment

Abstract

BACKGROUND: Nontechnical skills are essential for safe and efficient surgery. The aim of this study was to evaluate the reliability of an assessment tool for surgeons’ nontechnical skills, Non-Technical Skills for Surgeons dk (NOTSSdk), and the effect of rater training.

METHODS: A 1-day course was conducted for 15 general surgeons in which they rated surgeons’ nontechnical skills in 9 video recordings of scenarios simulating real intraoperative situations. Data were gathered from 2 sessions separated by a 4-hour training session.

RESULTS: Interrater reliability was high for both pretraining ratings (Cronbach’s $\alpha = .97$) and post-training ratings (Cronbach’s $\alpha = .98$). There was no statistically significant development in assessment skills. The D study showed that 2 untrained raters or 1 trained rater was needed to obtain generalizability coefficients >.80.

CONCLUSIONS: The high pretraining interrater reliability indicates that videos were easy to rate and Non-Technical Skills for Surgeons dk easy to use. This implies that Non-Technical Skills for Surgeons dk (NOTSSdk) could be an important tool in surgical training, potentially improving safety and quality for surgical patients.

© 2013 Elsevier Inc. All rights reserved.

Errors and mishaps occur in surgery, despite efforts to minimize them. Studies have identified poor nontechnical skills, such as communication, teamwork, and decision making, as important components of incidents.1,2 Observational studies have shown evidence that complications occur with a higher rate for operating room (OR) teams with poor teamwork than for OR teams with good teamwork.3 The causal relationship between nontechnical skills, errors and patient outcome remains to be fully explored. A review of the impact of nontechnical skills on technical performance concluded that surgeons’ situational awareness is important for the outcomes of surgery and that good leadership can shorten the duration of an operation.4 However, the identified studies were heterogeneous and from simulated environments rather than real ORs.
Although high-profile safety initiatives such as the World Health Organization’s Surgical Safety Checklist indirectly address some of these issues, they do not capture the entire intraoperative phase of surgery. Tools to assess and train nontechnical skills in the OR are also required and have been developed for surgeons, anesthetists, scrub nurses, and the full surgical team. Assessment can serve several purposes: individual purposes, such as giving structured feedback to trainees; institutional purposes, such as measuring effects of training; and societal purposes, such as certification and recertification. Regardless of the aim, assessment instruments require psychometric robustness (ie, they need to allow reliable and valid measurements of nontechnical skills).

In this article, we present an evaluation study of Non-Technical Skills for Surgeons (NOTSSdk), which is an assessment tool for surgeons’ nontechnical skills during surgery. The specific objectives were: (1) to explore whether surgeons, with no prior training, could assess nontechnical skills in standardized video-recorded simulation scenarios; (2) to test whether surgeons’ assessment skills would improve after 4 hours of training in assessment of nontechnical skills; and (3) to estimate how many raters would be needed to provide reliable assessment of nontechnical skills in video-recorded simulation scenarios.

Methods

We designed a 1-day assessment course for surgeons to explore the psychometric properties of NOTSSdk and the effect of training using video simulated scenarios.

NOTSSdk is a behavioral marker system designed for use as a competence assessment tool. It comprises 4 categories (situational awareness, decision making, communication and teamwork, and leadership), 13 elements, and numerous behavioral examples, such as “acts according to the operative findings,” illustrating good situational awareness, and “does not ensure team’s focus is centered on task,” illustrating poor communication and teamwork. A user guide to NOTSSdk can be found at http://www.regionh.dk/dims/menu/NOTSSdk. The tool uses a 5-point rating scale ranging from 1 (“very poor performance”) to 5 (“very good performance”) (Fig. 1). A rating scale was chosen opposed to a checklist because of the complexity of the skills, and 5 points were chosen to allow an average score, balancing ratings of good and poor behavior. We included a global score to explore content validity of the tool indicated by the correlation between global scores and category scores. The global score will not be in the “clinical version” of the tool.

Videos

We used 13 video-recorded scenarios: 9 to collect assessment data and 4 for rater training. The scenarios showed a broad range of both elective and acute situations in general surgery and were designed to show different levels of surgical performance in the 4 categories of NOTSSdk: situation awareness, decision making, communication and teamwork, and leadership (Fig. 2). Videos were filmed at the Danish Institute for Medical Simulation using real clinicians acting in their own roles. The clinicians were 2 general surgeons, 2 scrub nurses, 3 trainee surgeons, 2 nurse anesthetists, and 2 anesthesiologists dressed clinically and treating a simulated patient (an intubated Laerdal SimMan displaying vital signs on its monitor). The loosely scripted scenarios were developed by the research group, comprising a trainee surgeon (L.S.), a consultant general surgeon (R.B.-H.), a consultant anesthesiologist with extensive human factors experience (D.O.), and a work and organizational psychologist (P.D.). The videos each lasted between 3 and 5 minutes.

The realism of the videos was surveyed by a surgeon not involved in the project, who considered the videos to be sufficiently realistic.

Reference ratings

A set of reference ratings of the “acting” surgeon’s nontechnical skills was developed. This was to evaluate whether surgeons’ assessment skills improved during the course by comparing their ratings against the references. The reference ratings were obtained in the research group. The researchers independently rated each video using NOTSSdk and then discussed the videos until consensus ratings on the category level were agreed upon for all 9 evaluation videos.

Sample and participants

General surgeons were recruited for the course from all 9 hospitals in the eastern region of Denmark. An e-mail was sent to the head of the department at each hospital with information about the project, asking for the participation of 1 surgeon at the specialist level and 1 trainee surgeon in the last years of specialist training. Fifteen surgeons participated from 8 hospitals.

Procedure

The course was run at the Danish Institute for Medical Simulation and consisted of 2 rating sessions separated by an educational session. First, participants were given a short introduction to the aim of the study, the rating form, and how the videos were developed. Background information was obtained by exploring participants’ familiarity with assessments and nontechnical skills.

The videos were then shown in random order in the first rating session. Participants were asked to use NOTSSdk to rate the performance of the lead surgeon on all 4 categories, 13 elements, and 1 global rating. After watching each video, the participants had 4 minutes to provide ratings.
before seeing the next. Participants did not discuss their ratings during the rating sessions.

The educational session lasted 4 hours and was focused on the use of NOTSSdk, not on providing participants with an in-depth understanding of the underlying concepts. Training covered aspects of rater training according to accepted guidelines. This included a 90-minute explanation of the rationale for the development of NOTSSdk, and the project methods and results. Detailed explanations of the categories and elements using behavioral examples from NOTSSdk were provided. A lesson on how to observe nontechnical behaviors by paying close attention to verbal and nonverbal communication and by avoiding interpretations and assumptions of motives was included. Furthermore, potential rater biases such as leniency-stringency bias and end-aversion bias (ie, the tendency to ignore the
extreme response options because they are viewed as too strong) were addressed.\textsuperscript{13,14} The remaining 150-minute training session focused on providing participants with experience in observing and rating using the training videos, allowing them to discuss their ratings afterward. The participants saw 1 training video at a time and provided individual ratings that they presented using an audio response system (Turning Point; Turning Technologies, Youngstown, Ohio). Figure 2 shows the structure used in the design of the scenarios. Each column represents 1 NOTSSd category, and the y axis shows the level of performance of the surgeon. CT = communication and teamwork; DM = decision making; L = leadership; SA = situational awareness.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Laparoscopic appendectomy</td>
</tr>
<tr>
<td>3</td>
<td>Laparoscopic cholecystectomy with unclear anatomy in which concerns of trainee are considered by the surgeon</td>
</tr>
<tr>
<td>2</td>
<td>Prolonged laparoscopic appendectomy</td>
</tr>
<tr>
<td>4</td>
<td>Laparoscopic cholecystectomy with unclear anatomy in which concerns are being neglected</td>
</tr>
<tr>
<td>5</td>
<td>Gastroscopy for acute upper gastrointestinal bleeding with a surgeon showing “tunnel vision”, i.e ignores/suppresses important signs that do not comply with own view</td>
</tr>
<tr>
<td>6</td>
<td>Surgical Safety Check with a hasty, unwilling surgeon</td>
</tr>
<tr>
<td>7</td>
<td>Left sided hemicolectomy with unanticipated bleeding from the spleen</td>
</tr>
<tr>
<td>8</td>
<td>Right sided hemicolectomy with unanticipated venous bleeding</td>
</tr>
<tr>
<td>9</td>
<td>Low anterior resection with difficulty creating the anastomosis</td>
</tr>
</tbody>
</table>
Ethical considerations

The clinicians involved as actors in the videos signed informed consent forms. Participation in the evaluation study was voluntary and could be terminated at any time, and all participants gave oral informed consent. Data were anonymized using participant numbers on the rating forms and the background information sheet. The regional ethics committee was contacted and stated that the study was exempt from ethical approval (journal no. H-2-2012-FSP). The study was registered at http://www.clinicaltrials.gov (NCT 01562054).

Statistical analyses

To assess the average correlation across all 15 raters (intrarater reliability), we calculated Cronbach’s α values for pretraining and posttraining assessments on both the element and category levels and for the global scores. The correlation between global and category scores was then calculated using Pearson’s correlation coefficient as a measure of the test’s content validity.

Surgeons’ ability to assess nontechnical skills was explored by comparing the category scores for all raters, with the reference rating category scores. First, points were assigned to reflect how close participants’ ratings were to the reference rating (e.g., a participant rating of 4 got 4 points if the reference rating was 4 but only 3 points if there was a 1-point deviation, 2 points if there was a 2-point deviation, 1 point if there was a 3-point deviation, and 0 points in case of a 4-point deviation). A total point score was then calculated for pretraining and posttraining ratings (the maximum score possible was 144: 4 points × 9 videos × 4 categories), and these were compared using paired t tests. Differences were considered statistically significant at P < .05.

We used generalizability analyses to explore how many raters were needed for reliable assessment. First, we generated estimated variance components for a G study (2-facet, balanced design with videos crossed with raters crossed with category items). Then the estimated variance components were used in a D study to estimate generalizability coefficients corresponding to the number of raters. Both the G study and the D study were conducted for pretraining and posttraining ratings.

Statistical analyses were performed using SPSS version 19.0 (SPSS Inc, Chicago, IL).

Results

Participants comprised 7 female surgeons and 8 male surgeons: 8 trainees and 7 specialists, for a mean surgical experience of 12 years (range, 3 to 20 years); 14 had experience working as supervisors for younger colleagues, 11 had occasional experience with competence assessment, only 1 performed assessment on a regular basis, and 3 had no experience in the field. Eight participants said that assessment tools were lacking in their daily practice as surgeons; 11 reported that they had not received any formal education or training in assessment and that they had no prior knowledge of “human factors” or “nontechnical skills.” Those who did report familiarity with the terms had been introduced to NOTSSdk on a 1-day obligatory course during their training (n = 2), had been an interview participant during NOTSSdk development (n = 1), or had come across the terms during literature studies (n = 1).

 Interrater reliability was high for the pretraining ratings on both the element and category levels (with Cronbach’s α values of .96 and .97, respectively) and also high in the posttraining ratings (Cronbach’s α values of .97 and .98, respectively). Interrater reliability of global scores was equally high before and after training (Cronbach’s α = .98 and .99, respectively) and was correlated with the category score (Pearson’s correlation coefficient = .95). There was no statistically significant development in the surgeons’ assessment skills as measured by accuracy in ratings compared with reference ratings (120 points of 144 possible for pretraining vs 121 points of 144 possible for posttraining, P = .20). Fig. 3 shows the D study (i.e., the effect of the number of raters on the generalizability coefficient). It demonstrates that 2 untrained or 1 trained rater was necessary to obtain a generalizability coefficient >.80.

The qualitative statements clearly indicated that NOTSSdk was considered a useful tool for daily clinical practice. For comments, see Table 1.

Comments

In this study, we explored the reliability of the NOTSSdk tool using video recordings of simulated scenarios and found that surgeons who were novice raters could assess other surgeons’ nontechnical skills with high intrarater reliability. We could not detect any improvement in their assessment skills during the course. The tool showed signs of content validity, and generalizability coefficients suggest that it can be used reliably using 2 untrained raters or 1 trained rater.

The levels of agreement between raters were high compared with other studies of nontechnical skills ratings,
such as NOTSS evaluated in similar simulated videos,\textsuperscript{16} NOTSS evaluated in clinical settings,\textsuperscript{17} and Observational Teamwork Assessment for Surgery examined in clinical settings.\textsuperscript{18}

We found no improvement in assessment skills after training. This reflects that pretraining interrater reliability was very high, which is likely explained either by the fact that NOTSSdk is easy to use or that the nontechnical skills in the videos were easy to observe and rate even without prior training. We did not formally obtain the participants’ evaluations of the realism of the videos, but anecdotal statements from the participants indicated that they were viewed as highly realistic. This was supported by the realism evaluation by the independent surgeon and suggests that although easy to rate, the videos did not display caricatures of performance. Potential selection biases could also have contributed to the high pretraining interrater reliability because it is possible that the participating surgeons volunteered because they were particularly interested in nontechnical skills, education, and assessment, but this was not explored.

Two participants stated familiarity with NOTSSdk, as they were briefly introduced to it during a mandatory course in their specialist training. This could have raised the pretraining reliability, because some of the participants could not be considered true novice raters. However, their previous introduction to NOTSSdk involved only a plenum introduction and 1 simulation. Participants were given no opportunity to use NOTSSdk for assessment of another surgeon. Therefore, we do not consider this potential bias to be influential.

The D study showed that 2 untrained raters or 1 trained rater would be needed to obtain ratings with a generalizability coefficient $> .8$, which is the generally accepted threshold for reliable ratings for formative assessment.\textsuperscript{18} Moreover, the D study also indicated that using 4 untrained raters or 3 trained raters would provide ratings $> .9$, which is considered sufficient for high-stakes assessments (ie, for certification purposes or selection into surgery). However, caution is advised when trying to generalize from this study using simulated videos to real clinical settings. We recommend using NOTSSdk for formative assessment, as a means of giving trainees structured feedback on the basis of direct observations.\textsuperscript{6} We consider the tool an important potential adjunct to the current surgical curriculum because the assessment of surgical trainees greatly relies on supervisors’ statements and questionnaire-based 360° multisource feedback, both susceptible to recall bias and rarely based on direct observation.\textsuperscript{11}

![Figure 3](image-url) Effect of the number of raters on the generalizability coefficient. The dotted line marks the generally accepted threshold for reliable ratings when used for formative assessment.

---

### Table 1

<table>
<thead>
<tr>
<th>Theme</th>
<th>Examples</th>
<th>Written comments</th>
</tr>
</thead>
</table>
| Usefulness of NOTSSdk      | • To acquire skills sooner  
|                            | • In courses  
|                            | • In discussions with supervisor  
|                            | • To direct attention to nontechnical skills  
|                            | • For evaluation and learning  
|                            | • Not as a checklist                                                     | • “Important attributes that surgeons should appreciate. Could be used in courses during specialist training, locally in short teaching sessions or as simulation training. Cannot be used as a checklist.” |
|                            |                                                                         | • ‘Good instrument to quantify and assess the ‘soft skills.’ With experience comes the surplus to make different decisions. Perhaps this instrument helps generate these competences sooner.” |
|                            |                                                                         | • “Assessment by more persons and not just the supervisor. Also as a tool for discussions with the supervisor.” |
|                            |                                                                         | • “Can be useful—but important to carefully consider who assess and follows up.” |
| Use of NOTSSdk             | • By trained persons  
|                            | • By more persons than the supervisor  
|                            | • In conjunction with assessment of technical skills and the CanMEDS framework |                                                                         |

---

L. Spanager et al. Assessment of surgeons’ nontechnical skills 815
Content validity was sought and ensured through the development of NOTSSdk, during which the prototype instrument was discussed with a panel of surgeons, who commented on and revised the tool. The strong correlation between category and global scores provides further evidence of content validity, because a weaker correlation could have indicated that participants felt that not all aspects of surgeons’ nontechnical skills were covered in the 4 categories. The global score will not be included in the “clinical version” of the tool.

In this study, we used a broad sample of participants in general surgery with varying surgical experience and different levels of assessment experience. We consider it a strength that participants rated people they did not know in a simulated setting, thereby reducing potential rater bias. A study exploring the assessment of technical skills found that blinded assessments reduced the bias by approximately 10%. It is acknowledged that stakeholders’ acceptance of assessment procedures is crucial to ensure validity and reliability of the instrument. Offering this course to both trainers and trainees at different hospitals meant exposure of a broad range of surgeons to NOTSSdk and was therefore the beginning of implementation of NOTSSdk as a formative assessment tool. It provided knowledge of the evidence behind nontechnical skills in surgery and allowed participants to discuss safety and performance issues across hospitals. The qualitative evaluations from the participants suggested that the course helped raise awareness of nontechnical skills and that the added value of NOTSSdk was appreciated.

We used 9 different video scenarios, allowing the depiction of a range of different clinical situations and different levels of nontechnical performance. This showed the fullest range of behaviors. However, the use of extreme ends means that caution should be taken before generalizing the results to clinical use, because studies from the United Kingdom have reported somewhat homogeneous and above-average trainee performance during observations in real-life settings. This study had limitations. The reference ratings we obtained were not true gold standards. They were obtained in the research group because no Danish-speaking external experts existed. Neither was it possible to obtain evidence of concurrent validity of the system, because no other assessment tools on nontechnical performance existed in Danish. The training session was shorter than usually recommended for this type of training because we had to fit the course into clinical schedules. This possibly add to the explanation of the lack of improvement, as the training might have been too short to properly learn how to both observe and sort the skills according to categories and elements and how to rate them. Furthermore, although the limited available time allowed plenum discussion, it did not allow individual feedback on ratings or calibration of raters. This study did not allow testing of trainee variance or variance among cases, which could have affected interrater reliability. Studies done in real-life settings in the OR during daily clinical work have suggested that observing a trainee during >1 operation improves the reliability of the assessment, even though the nontechnical skills seemed to be somewhat procedure independent.

Conclusions

This evaluation study showed that NOTSSdk is an assessment tool with high interrater reliability in the simulated setting. It also showed signs of content validity. We found no statistically significant improvement of assessment skills after a 4-hour training session of novice surgical assessors. Two untrained raters or 1 trained rater appeared sufficient for the reliable assessment of trainee surgeons. This implies that NOTSSdk could be an important tool in surgical training, potentially improving safety and the quality of treatment of surgical patients.

References


