Clinical skills cover a broad spectrum of competences beyond medical expertise.\textsuperscript{1,3} Nontechnical skills such as situation awareness, decision making, task management, and teamwork are as important for safe and effective task management as is medical expertise.\textsuperscript{4} Together, technical and nontechnical skills can reduce the likelihood of error and consequently of accidents.\textsuperscript{5,6} Studies have identified poor nontechnical skills as important components of incidents.\textsuperscript{7,9} Several studies have shown that teamwork and communication affect the quality and safety of healthcare.\textsuperscript{5,6} Intrasurgical communication failures are frequent and occur in almost every procedure.\textsuperscript{5} Communication failures are associated with low teamwork quality, procedural problems, and complications.\textsuperscript{9} Studies show that the interplay between technical and nontechnical skills can be improved by training and practice.\textsuperscript{8}

In Denmark, traditional training placed great emphasis on technical performance, knowledge, and practical skills. During the last 10 years, however, the specialty-training curriculum for nurse anesthetists has also included training of nontechnical skills, often in simulation-based courses. The focus in assessment was mainly on technical skills. In Denmark, competence cards are used widely for workplace-based assessment of healthcare professionals.\textsuperscript{10,11} They also partly include some nontechnical skills, such as communication. The assessments are formative throughout the education, which means that feedback and guidance from supervisors are provided to facilitate the trainees’ learning. At the end of the education, the trainees need to have passed all assessments (summative part).

Nurse Anaesthetists’ Non-Technical Skills (N-ANTS) is a tool to assess and train nurse anaesthetists’ nontechnical skills in the operating room.\textsuperscript{11} This tool was adapted from the Anaesthetists’ Non-Technical Skills (ANTS) assessment tool\textsuperscript{4} to develop a system for nurse anaesthetists. The N-ANTS system comprises a 3-level hierarchy of 4 categories, 15 elements, and numerous behavioral markers. The categories are situation awareness, decision making, task management, and teamwork. Each category comprises 3 to 5 elements to describe the category in more detail. Each element has numerous behavioral markers that illustrate positive and negative behaviors relating to safety in the operating room. A 5-point rating scale ranging from 1 (poor performance) to 5 (excellent performance) is used to assess performance at both the category and the element level (Figure). A global rating
score with a 7-point rating scale is included to measure if the overall impression is aligned with category and element ratings. A handbook is available to guide the use of N-ANTS. Currently, N-ANTS is mainly intended to stimulate formative feedback discussions that should help nurse anesthetist trainees develop their interplay between technical and nontechnical skills. Thus, N-ANTS can supplement the existing competence cards in Denmark, by including a broader range of nontechnical skills in a more systematic fashion. The tool allows supervisors to provide feedback on the strengths and weaknesses of the participants’ nontechnical skills and in this way enable them to reflect on their own nontechnical skills in relation to their technical skills.

Formative assessment is dependent on good observational skills because the observations form the basis for feedback. Research is needed to explore whether the use of N-ANTS provides reliable (robust in interobserver agreement) and valid (reflects what it intended to measure) assessments in the operating room. Furthermore, raters should be adequately trained in the use of assessment tools.

The aims of the study were to (1) evaluate the psychometric properties (reliability and validity) of the use of N-ANTS in an experimental setting and (2) investigate the effect of training nurse anesthetist supervisors in the use of N-ANTS in an experimental setting.

**Materials and Methods**

- **Workshop Participants.** We designed a 1-day workshop for 22 nurse anesthetist supervisors working clinically and as supervisors performing work-based assessment of nurse anesthetist trainees. The nurse anesthetist supervisors were recruited by email from all hospitals in the eastern educational region of Denmark, where most Danish nurse anesthetists are educated.

  The sampling was purposeful; all nurse anesthetist supervisors from the eastern educational region of Denmark were targeted. Knowledge of educational needs was regarded beneficial for the evaluation of the psychometric properties of the use of N-ANTS.

- **Development of Videos.** Thirteen videos were produced demonstrating a range of elective and acute situations in a simulated general anesthesia setting; 9 to collect assessment data and 4 for rater training. The loosely scripted videos allowed the role players to perform according to their own experience. The video scripts were developed by a nurse anesthetist and a trainee anesthesiologist to show different levels of anesthetic performance in the 4 N-ANTS categories. For example, one video was scripted to illustrate the anesthetic team being good at teamwork and decision making, but poor at situation awareness and task management.

  The videos were recorded in a full-scale simulation environment. A life-sized, computer-controlled manikin was placed in a replica of a clinical setting. All people involved wore clinical uniforms and strived to behave clinically realistically. A consultant anesthesiologist with extensive human factors experience and a work and organizational psychologist were present during the recordings to ensure that the videos displayed a sufficient amount of nontechnical skills to allow rater training. The videos were cut to last 3 to 5 minutes.

- **Reference Ratings of Videos.** An experienced anesthesiologist and an experienced nurse anesthetist, who were both familiar with nontechnical skills, evaluated the video recordings for realism and authenticity. The videos were found to be appropriate for the intended use.

  To investigate research question 2, a set of reference ratings was produced to explore whether workshop participants rated the video similarly to “expert” raters. A set of reference ratings for the role-playing nurse anesthetist’s nontechnical skills was developed by 4 experts in nontechnical skills: a nurse anesthetist and an anesthesiologist from the research group and 2 experienced nontechnical skills supervisors with an anesthetic background. All 4 raters independently assessed the acting nurse anesthetist in all 9 videos and discussed ratings until consensus was reached. They established for each video a global score and a rating for each category and element.

- **Workshop Procedures.** Before the workshop day the participants received a program and a written description of the N-ANTS categories and elements, including the behavioral markers describing each element. The workshop took place at the Danish Institute for Medical Simulation in Copenhagen (now called Copenhagen Academy for Medical Education and Simulation), Denmark, in June 2012. At the beginning of the workshop day background information, including familiarity with assessments and nontechnical skills, was collected in a questionnaire. Then there was a pretraining session, followed by training and a posttraining session. At the end of the workshop the participants evaluated N-ANTS in writing.

- **Pretraining Session.** This session consisted of a short introduction to the rating scale, the assessment form, and the structure of N-ANTS (Figure). Nine videos were shown in random order. Five minutes between each video were given to take notes and complete global, category, and element ratings on the N-ANTS assessment form. The session lasted 90 minutes. Participants were not allowed to discuss the videos but could use the written description of N-ANTS they received before the test to assist their ratings.

- **Rater Training.** The educational session lasted 2 hours and covered several aspects in accordance with accepted guidelines on how to train assessors in using observational tools. The topics covered were:

  1. The importance of nontechnical skills and how these skills can be trained.
  2. The impact of nontechnical skills on patient
outcome and safety. This was demonstrated by showing a video on the use of human factors principles in the operating room to enhance performance.

3. Detailed explanations of the categories and elements were provided in a discussion.

4. An overview of the adaptation of N-ANTS, including several behavioral markers to illustrate how they help to indicate the presence or absence of the skill elements in a given situation.

5. Explanation of how to observe nontechnical skills by paying attention to both verbal and nonverbal communication. It was noted to concentrate on observable behavior and avoid interpretation and assumptions regarding motives.

6. Rater bias such as leniency-stringency bias, end-aversion bias (ie, the reluctance to use the extreme ends of the scale), and cognitive and social biases such as fixation and confirmation bias.

The remaining training focused on providing participants with experience in observing and rating the 4 training videos. The participants rated one video at a time and then presented their ratings on category level using an audio response system (Turning Point, Turning Technologies). All ratings on the category level were displayed in a summary, and the considerations and explanations behind the ratings were explored and discussed in the plenum after each rating.

• Posttraining Session. The same 9 videos as in the pretraining session were shown in a different order. Again, participants could use the written description of N-ANTS to assist their ratings.

• Ethical Considerations. Danish law exempts this kind of research from ethical board approval because no patients were involved. An exemption letter from the Regional Ethical Committee of the Capital Region, Denmark, was obtained (H-2014-FSP-013). The role players in the videos gave written informed consent for the videos to be used in the study and for educational purposes. The workshop participants gave written informed consent for rating data to be published and used for educational purposes. Participation was voluntary and could be terminated at any time. Rating data were anonymized using participant numbers on the rating forms and the background information sheet. Data were handled in accordance with the regulations of the Danish Data Protection Agency.

• Statistical Analysis. To assess the correlation across all 22 raters (intrarater reliability), we calculated intraclass correlation coefficients (ICCs) for average measures (Cronbach $\alpha$) and single measures for pre- and posttraining ratings for the global rating score, the categories, and the elements. This represented the consistency in ratings. Raters with many missing items (more than 10% of all possible scores) were not included in the analysis.

Participants’ ability to use N-ANTS to assess nontechnical skills was explored by comparing the global rating and the ratings on the category and element level with the reference ratings. Points were assigned to reflect how close participants’ ratings were to the reference ratings. The ratings were scored according to the difference between the reference rating and the participant’s rating, with 5 points for no difference, 4 points for a 1-point difference, and so on. A missing response resulted in a score of zero points. A total point score was calculated for pre- and posttraining ratings to compare the level of concurrence with the reference ratings. Maximum scores possible for global rating scores were 63 (7 points × 9 videos × 1 global score), for categories 180 (5 points × 9 videos × 4 categories), and for elements 675 (5 points × 9 videos × 15 elements). Scores are reported as percentages of maximum score that any individual rater achieved compared with the rating of the reference group.

Paired $t$ tests were used to compare pre- and posttraining ratings. Differences were considered statistically significant if $P < .05$. Statistical analyses were performed using a statistical software package (SPSS version 19.0, IBM SPSS).

To get an overview of the evaluation comments that...
participants offered at the end of the workshop, we summarized the topics from the papers collected.

Results
We collected 9 pre- and 9 postraining rating forms from 22 workshop participants.

Participants were 18 female and 4 male nurse anesthetists; 7 of the participants were responsible for the nurse anesthetist-training program in their department. All participants were nurse anesthetist supervisors involved in clinical assessment of competence. The mean experience in nursing was 21 years (range, 4-35 years). Nineteen participants were actively involved in clinically supervising trainees, and 3 were simulation instructors at a simulation center. Four had received prior training in using an observation tool for providing feedback to trainees but had no training in rating performance. All participants except 4 were familiar with the ANTS concepts at a general level. Six participants had been involved in the N-ANTS project at earlier stages.

The qualitative statements from the participants clearly indicated that N-ANTS was considered a useful tool for structuring daily practice, especially if rater training would be provided. The tool can be used in courses during specialist training by supervisors but was also considered useful for educated nurse anesthetists to direct attention to nontechnical skills. The common terminology in particular was considered helpful.

The combined reliability (Cronbach α) of all 22 raters was high in both pre- and postraining ratings on all 3 levels (global rating score, category, and element level). A single rater using the tool (ICC single measures) resulted in a measurable increase in reliability after training at all 3 levels. The reliability improved considerably after training at the element level. The Table shows the ICC values.

Nurse anesthetist supervisors’ ability to use the tool (rater accuracy) was good before and after training. For the global rating score they reached 84% of the maximum score before the training and 80% after the training. For the categories they achieved 82% of the maximum score before and 73% after the training. At the elements level they reached 78% of the maximum score before and 74% after the training. Paired t-tests showed a statistically significant reduction in accuracy after the training for the global rating score (P = .005) and the category level (P < .001). At the element level there was no statistically significant change between the measurement points (P = .079).

Discussion
In this study we found that nurse anesthetists with prior knowledge of nontechnical skills concepts and terms, but no prior specific training in using N-ANTS, could use N-ANTS for making reliable ratings and valid assessments of videos showing simulations of clinical performance. The participants’ ratings were highly comparable between raters and to the reference ratings. Interrater reliability at the N-ANTS element level did improve after training. For the comparison of the participants’ ratings there was no statistically significant effect on training. We found a significant reduction in rating accuracy after the training on the global rating and category level, but not the elements. Reasons for the study findings will be explored in the following paragraphs.

• Reliability of Nurse Anaesthetists’ Non-Technical Skills Before and After Training. The interrater reliability was high before training. One explanation might be that the videos were easy to rate, as some of the videos showed extreme—very good or poor—performance. About one third of the participants were already familiar with the ANTS concepts at a general level, because the ANTS system has been used for several years in clinical competence assessments in Denmark. Nontechnical skills training has also been an integral part of simulation-based training, which is mandatory in the specialty-training program for nurse anesthetists in Denmark. There are regular multidisciplinary in-hospital training programs such as “trauma team training” and “cardiac arrest and resuscitation team training”. All these programs address the use of nontechnical skills for nurse anesthetists.

The high pretraining reliability likely prevented the detection of an overall improvement in the participant’s assessment skills during the course. Their ceiling effect did not leave much room for improvement. On the other hand there is the possibility that the training session was too short for participants to significantly improve their rating skills.

At the element level, the interrater reliability increased after training, especially for the single measures. This measurement assesses the psychometric features of the tool, when just one rater uses it. On this level the figures were lower in the pretraining and improved with the postraining. We suggest that rater training was helpful for the raters to discuss the nontechnical skills and striving for consensus on standards on the more detailed level of behaviors. The more detailed the descriptions are, the higher should be the likelihood of actually seeing different views by different persons. Although it seems reasonable to assume that different raters might assess an action on a higher level comparably, differences in their views might show up when they would relate their assessment to very concrete actions (and this is what the elements are). In this sense the elements might be the most sensitive part of the tool.

The patterns of changes in the psychometric features between the pre- and postraining emphasize the need to continually develop the supervisors’ understanding of N-ANTS at the element level. Rater training seems necessary despite the fact that the levels of agreements were high compared with other studies of nontechnical ratings, such as ANTS evaluated in similar simulated
Using more than 1 rater would also increase the reliability of the assessment.16,23

• Validity of Nurse Anaesthetists’ Non-Technical Skills Before and After Training. The workshop participants had acceptable ratings for rater accuracy compared with the reference ratings before and after the training,3 despite the fact that they were not systematically trained in the use of N-ANTS before rating the pretraining videos. Our findings seem to indicate that the videos were easy to rate for the raters, who might have been somewhat familiar with the ANTS and nontechnical skills terminology.

The findings in this study reveal that N-ANTS produces reliable and valid measurements. This means that the tool sets the focus on a description of nurse anesthetists’ nontechnical skills. The fact that the workshop participants could use the tool even without additional training suggests that N-ANTS can be used in the clinic to provide formative feedback. The psychometric features of the tool point to a potential to use N-ANTS also in summative assessment situations. However, there are more requirements to establish N-ANTS in such a way (eg, the logistical side of creating the time and space for the assessment).

To further enhance the validity of formative competence assessments, the Danish Health and Medicine Authorities recommend the use of a variety of methods and that these methods should be uniform in all national health regions. The results of this study suggest that N-ANTS could be a useful addition to the established competence cards for the provision of feedback.

This study has limitations. There were more missing data in the posttraining session than in the pretraining session (Table). Measuring agreement with the reference ratings penalized “missing data” (an item not answered resulted in 0 points). This might indicate rater fatigue and/or a decrease in motivation to use the N-ANTS tool as participants were asked to perform many ratings at short intervals. This might explain the lower postraining agreement between the workshop participants and the reference ratings. Because N-ANTS is intended for formative feedback, it is unlikely that assessors will be requested to fill out 22 assessment forms a day, and as such, rater fatigue will be of less concern in the clinical context.

We produced the videos to train our assessors in a simulated environment. The choice to use videos rather than clinical settings was due to the wish to minimize factors such as shortage of time, the probability of interruptions, variety in patient categories, and environmental influence such as the work of collaborators, all of which might potentially challenge the supervisors’ abilities to learn how to assess and provide feedback on observed behavior.14 We acknowledge that the short, partly scripted videos were likely easier to rate than the “muddy waters” of clinical practice would have been, and thus the psychometric features of N-ANTS and its usability in the clinic setting remain to be tested. However, the current study suggests that the tool is designed in a way that it is ready for such tests.

The rater training was shorter than usually recommended for this type of training4,22 because we had to fit the workshop into clinical schedules. Danish clinical supervisors attended this workshop, and they might not be fully representative of an international forum. As stakeholders’ acceptance of assessment procedures is crucial to ensure validity and reliability of an assessment tool,13 we suggest that the N-ANTS system could be used as an initial tool for formative assessment in different settings.

In conclusion, Danish nurse anesthetist supervisors were able to make comparable assessments of video-recorded simulation scenarios and to rate them similarly.

<table>
<thead>
<tr>
<th>Level</th>
<th>Measure</th>
<th>Before training</th>
<th>After training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global rating scores</td>
<td>Number of raters, making ratings in minimum 90% of the global rating scores (rating a minimum of 9 of 9 global rating scores)</td>
<td>19/22</td>
<td>18/22</td>
</tr>
<tr>
<td></td>
<td>Intraclass correlation coefficient, average measures</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>Intraclass correlation coefficient, single measures</td>
<td>0.69</td>
<td>0.79</td>
</tr>
<tr>
<td>Categories</td>
<td>Number of raters, making ratings in minimum 90% of the categories (rating a minimum of 33 of 36 category ratings)</td>
<td>19/22</td>
<td>16/22</td>
</tr>
<tr>
<td></td>
<td>Intraclass correlation coefficient, average measures</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Intraclass correlation coefficient, single measures</td>
<td>0.66</td>
<td>0.70</td>
</tr>
<tr>
<td>Elements</td>
<td>Number of raters, making ratings in minimum 90% of the elements (rating a minimum of 122 of 135 element ratings)</td>
<td>17/22</td>
<td>19/22</td>
</tr>
<tr>
<td></td>
<td>Intraclass correlation coefficient, average measures</td>
<td>0.89</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Intraclass correlation coefficient, single measures</td>
<td>0.33</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table. Interrater Reliability, Shown as Intraclass Correlation Coefficients, Average Measures (Cronbach α) and Single Measures

Values are shown for all 3 levels (global ratings, category levels, and elements levels) before and after rater training.
to how a reference group rated the videos. They were able to do this without additional training besides their work experience. Rater accuracy did not improve after training, which perhaps could be explained by rater fatigue or an insufficient training period.

REFERENCES


AUTHORS

Helle Teglgård Lyk-Jensen, MSc, CRNA, RN, is employed by Copenhagen Academy for Medical Education and Simulation in Copenhagen, Capital Region of Denmark, Herlev University Hospital, Herlev, Denmark. Email: helle.teggaard@gmail.com.

Peter Dieckmann, Dipl Psych, PhD, is employed by Copenhagen Academy for Medical Education and Simulation, Capital Region of Denmark, Herlev University Hospital, Herlev, Denmark.

Lars Konge, MD, PhD, is employed by Copenhagen Academy for Medical Education and Simulation, Capital Region of Denmark and Copenhagen University, Copenhagen, Denmark.

Rikke Malene H.G. Jepsen, MD, PhD, is employed by Copenhagen Academy for Medical Education and Simulation, Capital Region of Denmark, Herlev University Hospital, Herlev, Denmark.

Lene Spanager, MD, PhD, is employed by Copenhagen Academy for Medical Education and Simulation, Capital Region of Denmark, Herlev University Hospital, Herlev, Denmark.

Doris Østergaard, MD, DMSc, is employed by Copenhagen Academy for Medical Education and Simulation, Capital Region of Denmark and Copenhagen University, Denmark.

DISCLOSURES

The authors have declared they have no financial relationships with any commercial interest related to the content of this activity. The authors did not discuss off-label use within the article.