Blinded Evaluation of Interrater Reliability of an Operative Competency Assessment Tool for Direct Laryngoscopy and Rigid Bronchoscopy

Stacey L. Ishman, MD, MPH; James R. Benke, BS; Kaalan Erik Johnson, MD; Karen B. Zur, MD; Ian N. Jacobs, MD; Marc C. Thorne, MD, MPH; David J. Brown, MD; Sandra Y. Lin, MD; Nasir Bhatti, MD, MPH; Ellen S. Deutsch, MD

Objectives: To confirm interrater reliability using blinded evaluation of a skills-assessment instrument to assess the surgical performance of resident and fellow trainees performing pediatric direct laryngoscopy and rigid bronchoscopy in simulated models.

Design: Prospective, paired, blinded observational validation study.

Subjects: Paired observers from multiple institutions simultaneously evaluated residents and fellows who were performing surgery in an animal laboratory or using high-fidelity manikins. The evaluators had no previous affiliation with the residents and fellows and did not know their year of training.

Interventions: One- and 2-page versions of an objective structured assessment of technical skills (OSATS) assessment instrument composed of global and a task-specific surgical items were used to evaluate surgical performance.

Results: Fifty-two evaluations were completed by 17 attending evaluators. The instrument agreement for the 2-page assessment was 71.4% when measured as a binary variable (ie, competent vs not competent) (κ=0.38; P=.08). Evaluation as a continuous variable revealed a 42.9% percentage agreement (κ=0.18; P=.14). The intraclass correlation was 0.53, considered substantial/good interrater reliability (69% reliable). For the 1-page instrument, agreement was 77.4% when measured as a binary variable (κ=0.53, P=.0015). Agreement when evaluated as a continuous measure was 71.0% (κ=0.54, P<.001). The intraclass correlation was 0.73, considered high interrater reliability (85% reliable).

Conclusions: The OSATS assessment instrument is an effective tool for evaluating surgical performance among trainees with acceptable interrater reliability in a simulator setting. Reliability was good for both the 1- and 2-page OSATS checklists, and both serve as excellent tools to provide immediate formative feedback on operational competency.


Currently, otolaryngology resident surgical performance is typically evaluated on a subjective scale (“poor, fair, good, and outstanding”) determined by the evaluating faculty members. Typically, in training programs for otolaryngology residents, faculty members assess the surgical performance of the residents at the end of their resident rotation, which is often weeks or even months after performance of the surgical tasks. In the current system, evaluators must rely strictly on memory and recall, which introduces risk of improper and mistaken assessment of critical details regarding surgical performance. The lapse of time between surgery completion and assessment also inhibits timely and detailed feedback in reference to the procedure.

The Accreditation Council for Graduate Medicine Education (ACGME) recognizes the flaws in the current system for surgical evaluation and recommends procedures that promote objective, detailed, and immediate feedback to residents to improve surgical learning over the course of the residency rotation. In an attempt to combat the shortcomings of the current evaluation system, our institution has developed a number of objective structured assessment of technical skills (OSATS) tools, including one evaluating pediatric direct laryngoscopy and rigid bronchoscopy. The modified delphi technique involving a panel of pediatric otolaryngologists was used to...
create this OSATS evaluation tool. The assessment tool includes global and task-specific checklists based on collectively identified steps integral to the effective completion of direct laryngoscopy and rigid bronchoscopy and emphasizes maintenance of feasibility, validity, and interrater agreement.

Because previous literature has suggested that global rating tools may be superior to task-specific checklists, our OSATS tools have included both types of checklists. It has been suggested that the primary difference between the 2 tools is that the task-specific checklist documents the occurrence of discrete steps during the procedure, while the global checklist is intended to document how effectively these tasks were executed.

A pilot study of these tools was completed in 2010 and found them to be quite valid and reliable. Seven faculty members completed 44 assessments to evaluate 19 residents performing direct laryngoscopy and rigid bronchoscopy procedures performed by residents over a 3-year period. The assessment instruments demonstrated a quick, effective feedback mechanism for evaluators. The tools required only 3 to 5 minutes to complete. Evaluators commented on the tools’ ease of use, comprehensiveness, and practicality. However, in the pilot study, the faculty evaluators were familiar with the residents’ abilities and training, allowing for the presence of bias and introducing doubt as to whether the study was an impartial test of the instruments’ objectivity. To further test the effectiveness of the tools as objective instruments, the evaluators should have no previous affiliation with the performing residents. Other OSATS measures have been developed for pediatric airway endoscopy and tested in a simulation environment,

The purpose of the present investigation was to analyze the effectiveness of the OSATS tools in a setting where the surgical evaluators were not familiar with those completing the surgical tasks. This was intended to remove confirmation bias and ensure that evaluators were able to assess performance, and the tools themselves, more objectively. In addition, a 1-page version of the assessment tool was piloted, integrating elements of both the global and task-specific checklists, in an attempt to create a form that is easier and quicker to complete while incorporating the most salient components of the previous version.

**METHODS**

**OSATS TOOLS**

Task-specific and global assessment tools, as originally described in the validation study, were used to evaluate direct laryngoscopy and rigid bronchoscopy procedures performed by residents or fellows. During the administration of the OSATS tool in 2010, 2 instruments were used to evaluate surgical competency. The first page, the task-specific checklist (Figure 1), evaluated residents during integral steps of the surgical task using a 5-point Likert scale accompanied by 3 anchors. The first anchor, “Unable to perform,” corresponded to a rating of 1. A rating of 3 corresponded to “Performs w/minimal prompting,” while a rating of 5 corresponded with the anchor “Performs easily w/good flow.” This assessment tool was used to provide immediate and constructive feedback to residents and is referred to herein as the 2-page assessment tool.

The second page of this OSATS evaluation instrument (Figure 2) assessed the overall surgical performance. The form consisted of 10 global items that linked specific, concrete descriptors to a 5-point Likert scale. A score of 3 was used to de-
note procedure competence and represents the minimally acceptable surgical performance score.

Based on faculty feedback from 2010 and in collaboration with attending otolaryngologists representing multiple training programs using rating scales for surgical performance, an updated 1-page OSATS tool was developed titled “Pediatric Rigid Airway Endoscopy Performance Scale” (Figure 3). These collaborators include the authors of all previously published rating scales for rigid airway endoscopy performance.1,3,5 The new instrument combined questions from both checklists and integrated them into a single form. This revised task-specific tool was again based on a 5-point Likert scale but also included a description for each of the 5 anchors in addition to the “N/A” column (not applicable). A rating of 3 was again designated the minimally acceptable surgical performance score denoting competence. The first anchor, “Verbal Instruction and Demonstration,” corresponded to a rating of 1, while a rating of 2 accompanied the anchor “Verbal Prompts with Errors.” “Independent with Errors” corresponded to a rating of 3. A rating of 4 corresponded with “Independent w/o Errors,” while a rating of 5 corresponded with “Independent and Efficient.”

EVALUATION

Both the original 2-page and revised 1-page OSATS instruments were used to evaluate surgical performance of residents and fellows with whom raters had no previous affiliation at a regional simulation and skill course for pediatric airway management. Attending-level faculty members observed trainees while evaluating surgical competency during simulated pediatric direct laryngoscopy and rigid bronchoscopy with both animal and full-body high-technology simulators. A 10-minute oral instruction session and written instructions (Figure 4) were used to train the faculty on proper use of the assessment instruments during an informational session prior to the course. Only 2 of the faculty members were familiar with the tool prior to the course. All of the evaluations were performed simultaneously by 2 separate assessors who were paired to test the interrater reliability of the tools. At the conclusion of the evaluations, faculty members judged the instruments’ feasibility and ease of use.

STATISTICAL ANALYSIS

The distribution of values was examined through the use of descriptive characteristics. Because most of the trainees were second- and third-year residents, we did not look at construct validity because there were few unexposed or expert participants in this course. The interrater agreement measured the agreement between observers who evaluated the same resident during the same surgical performance. Binary and continuous variables were calculated comparing the scores of
individual questions between evaluators to determine interrater agreement. The significance of interrater agreement as a binary variable was determined by the $\kappa$ statistic. The intraclass coefficient was used to evaluate the reliability of measures between evaluators assessing trainees at the same time when evaluated as a continuous variable. The Cronbach $\alpha$ was used to evaluate the new 1-page tool and served as a measure of internal consistency and reliability, ie, comparing the likelihood that different survey items reliably assessed the same characteristic.

This study was found to have exempt status by the institutional review boards at the Johns Hopkins School of Medicine and the Children's Hospital of Philadelphia. Significance was set at $P < .05$. Stata Statistical Software, Release 12.1 (StataCorp LP) was used to analyze data.

### RESULTS

Seventeen attending-level pediatric otolaryngology faculty members representing 14 different training programs performed 52 paired assessments of residents and fellows during a pediatric endoscopy course. Complete data were available for 45. Feasibility was judged based on the ability of the faculty members to complete the forms (100%) for those asked to perform the ratings. In addition, the length of time to fill out the forms was recorded: 3 to 5 minutes for the 2-page form, and 2 to 3 minutes for the 1-page version. For the 1-page form with the new anchors, the faculty evaluators found the form to be easy to use. However, there was some discussion of how teaching style influenced ratings between 2 and 3 because faculty members who use verbal prompts as a teaching method may not have allowed trainees the opportunity to demonstrate competence before giving verbal prompts.

Measurement of interrater reliability was carried out for each survey version. For the 2-page version, agreement on competence was 71.4% when measured as a binary variable (ie, competent vs not competent) ($\kappa = 0.38, P = .08$). Evaluation as a continuous variable revealed a 42.9% agreement ($\kappa = 0.18; P < .001$). The intraclass correlation was 0.53, considered moderate agreement with a reliability of 69% (Figure 4).

Evaluation of the 1-page version found that agreement on competence was 77.4% when measured as a binary variable (ie, competent vs not competent) ($\kappa = 0.54; P < .002$). Evaluation as a continuous variable revealed a 71.0% percentage agreement ($\kappa = 0.53; P < .001$). The intraclass correlation was 0.73, considered strong agreement with a reliability of 85%. Overall, there was 75.6% agreement between evaluators when analyzing the ques-
tion of competence, regardless of assessment tool format ($\kappa=0.47$, $P<.001$) (Figure 4).

The 1-page checklist was also evaluated for internal consistency (Table). There were few responses for question 1 regarding history and physical examination findings because this was completed in a simulation setting; therefore, it was excluded from this analysis. For evaluation of questions 2 through 11, the Cronbach $\alpha$ was 0.71, considered acceptable; however, there were limited responses to question 10, likely significantly affecting this measure of internal consistency. When the analysis excluded questions 1 and 10, the Cronbach $\alpha$ was excellent at 0.92.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Items included</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>Reliability coefficient</td>
<td>0.71</td>
<td>0.92</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Interpretation</td>
<td>Acceptable</td>
<td>Excellent</td>
<td>Excellent</td>
<td>Excellent</td>
</tr>
</tbody>
</table>

Abbreviation: OSATS, objective structured assessment of technical skills.

$^a$From Ishman et al.$^1$
high as was desired, debriefing of the faculty revealed that some of this variability may be owing to education of the evaluators themselves as many of the faculty members were naïve to the tools. Further training of faculty evaluators and earlier introduction to the tools would likely benefit reliability and should be studied in the future.

These competency assessment tools can be used in animal laboratory and simulation settings as well as the operating room setting, as was demonstrated in our first study. Regular and specific assessment of a resident’s surgical preparedness is paramount in an environment where mistakes can lead to tragic consequences. Objective, structured laboratory assessments can be used to identify residents who are prepared to graduate from simulation and animal laboratory settings to live surgery procedures. For those who have not yet achieved competence in a specific procedure, these evaluation tools can act as learning tools and a basis for correcting specific skills to increase competence using those skills outlined in the OSATS questionnaires. The instrument also serves as an impetus to generate discussion and feedback between the trainee and the attending physician. The use of the OSATS tool in the training curriculum also satisfies the requirements of accreditation agencies to monitor the level of proficiency provided during the residency program to ensure that competency is achieved.

Valid, feasible, and reliable OSATS tools can be systematically integrated into training programs to ensure effective feedback for residents. The 1- and 2-page OSATS tools outline and clearly define what is expected of the residents during the surgical procedures. Deficiencies in specific tasks can then be highlighted during the procedure, allowing for the resident to implement specific improvements and to work on directed subtasks to achieve surgical competency. With 100% compliance and integration into surgical training, a program can increase the effective use of time and resources used to train surgical residents. We hypothesize that the inclusion of OSATS tools would decrease the time required for a resident to achieve competency, since the tools would allow a resident to target weak areas and are consistent with theories of adult learning. We also suggest that future study should be carried out to test this hypothesis.

One limitation of this study is the limited sample size. A larger number of residents evaluated by an even larger number of faculty members would provide a greater understanding of the OSATS instruments’ impact on surgical competency. Continuing the study over a longer period would illustrate the effect of its use on the length of time for a resident to reach surgical competency or advance from simulation and animal laboratory to human surgery. Also, while this validation study was carried out in a simulation setting, the 1-page OSATS tool still requires validation in real patients. Finally, the use of evaluators naïve to the instrument may have decreased its reliability and validity, although levels of agreement were quite good. This is less likely to be an issue when the tool used in individual programs because training and familiarity would be expected to be much higher with regular use. Additional exposure and training will be integrated in future studies.

In conclusion, in the context of work-hour limitations decreasing operative experience and clinical surgical exposure, and society demanding greater accountability for technical surgical competence, there is increasing need for timely and specific surgical competency evaluation and feedback for trainees. While the traditional system of evaluation provides residents with a subjective surgical assessment, it is anticipated that regular use of OSATS tools will provide objective, qualitative evaluation and allow for more efficient improvement of specific skills. A blinded evaluation of tools designed to objectively assess the surgical skills and competence of those performing direct laryngoscopy and bronchoscopy suggests that the tools can be used to provide effective and informative evaluations to residents with good validity and feasibility. Reliability was reasonable but not optimal and may be improved with greater evaluator training. In addition, both the 1- and 2-page versions of this OSATS tool had similarly good reliability. Future studies will focus on improving interrater reliability and assessing how these tools may affect time to surgical competency.

Submitted for Publication: April 24, 2012; final revision April 24, 2012; accepted May 4, 2012.

Published Online: September 17, 2012. doi:10.1001/2013.jamaoto.115

Correspondence: Stacey L. Ishman, MD, MPH, Department of Otolaryngology—Head and Neck Surgery, Johns Hopkins School of Medicine, 601 N Caroline St, Sixth Floor, Baltimore, MD 21287 (Sishman1@jhmi.edu).

Author Affiliations: Department of Otolaryngology—Head and Neck Surgery, The Johns Hopkins School of Medicine, Baltimore, Maryland (Drs Ishman, Lin, and Bhatti and Mr Benke); Division of Otolaryngology—Head and Neck Surgery, Cincinnati Children’s Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, Ohio (Dr Johnson); Department of Otolaryngology—Head and Neck Surgery (Drs Zur and Jacobs) and the Center for Simulation, Advanced Education, and Innovation, Department of Anesthesiology and Critical Care (Dr Deutsch), The Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania; and Department of Otolaryngology—Head and Neck Surgery, University of Michigan, Ann Arbor (Drs Thorne and Brown).

Financial Disclosure: None reported.

Previous Presentation: This article was presented at the American Society of Pediatric Otolaryngology 2012 Annual Meeting; April 20, 2012; San Diego, California.

REFERENCES


