The Journal of Education in Perioperative Medicine

Original Research

Development of an Endotracheal Intubation Formative Assessment Tool

Adam Ryason, PhD
Emil R. Petrusa, PhD
Uwe Kruger, EngD
Zhaohui Xia, PhD
Vanessa T. Wong, BS
Daniel B. Jones, MD, MS
Suvranu De, PhD
Stephanie B. Jones, MD

Introduction

Endotracheal intubation (ETI) is a critical procedural skill in many areas of medicine. It is most commonly learned on static mannequins or in the clinical setting with patients. Existing tools to assess ETI performance predominantly consist of binary item checklists intended for summative, rather than individualized and formative, assessment of competency.[1-4]

For example, RESCAPE (Resuscitation and Emergency Simulation Checklist for Assessment in Pediatrics)[1] contains 16 binary items in its oral ETI checklist. Silverman et al[2] described a nonsurgical airway management curriculum for surgical trainees. Performance was evaluated based on the procedure followed on an airway mannequin, irrespective of how well individual tasks were performed or the relative importance of a given task.

Kuszajewski et al[3] used a modification of the National Registry Emergency Medical Technicians Advanced Level Psychomotor Examination for Ventilatory Management-Adult, a 21-point checklist, as both a training and testing tool. Practice was guided by the checklist, with a member of the training team giving feedback. Subsequent training phases used a low-fidelity manikin and high-fidelity simulator. The final assessment took place while intubating a patient in the operating room.

The studies in this article describe the development and measurement characteristics of observable assessment items for performing an ETI with feedback to the training clinician. The goal of this study was to develop a technical metric that measures the performance of direct laryngoscopy during airway management training. It is hypothesized that assessment items can be agreed on when viewed by different experts and will correlate with the reference standards of rating and rank-ordering performance.

Materials and Methods

The study protocol was approved by the Beth Israel Deaconess Medical Center Institutional Review Board (IRB). Written informed consent was waived by the IRB.

Metric Development

To derive the metric, we used a combination of existing literature[2,4–6] and expert opinion from 5 subject matter experts (clinically active anesthesiologists). Subject matter experts were asked to supply the main objectives for a typical oral intubation with direct laryngoscopy, each with associated items. The 5 main objectives were as follows: positioning of the patient, insertion of direct laryngoscopy blade, achieving the optimal laryngeal view, inserting the endotracheal tube, and avoiding injury to the patient. All items had corresponding feedback to the trainee on how they can improve that specific objective. The aim of this metric is to allow an informed observer to provide reliable and valid data regarding the quality of ETI performance with minimum subjectivity and give useful feedback during training. The next step (see Figure 1) for metric development consisted of 6 expert raters (Pilot Group 1, $N_p = 6$), new to the study, who were asked to complete the metric regarding the quality of video recordings of actual ETI performances (Video Collection A) and complete the pilot metric (Figure 1).

Experts in this research consisted of board-certified anesthesiologists who are also faculty in anesthesia residency training programs with a median of 12.5 years post board certification. Modifications of some metric items were made based on low inter-rater agreement and feedback about items from the experts. A final metric (Appendix A) was established and used in subsequent steps of this research.

Video Environment

Two sets of videotapes of ETI performed in the operating room were captured to use as standard performances against which to assess the measurement quality of the new metric. The first set (Video Collection A) was a convenience sample of 14 ETIs by 14 clinicians using a Mac 3 C-MAC blade with actual patients (3 of the original 14 videos were excluded from the collection because of incomplete visualization). Following a metric analysis conducted using the original collection, a second set (Video Collection B) of 16 ETIs performed by 16 different clinicians was obtained to ensure adequate power to analyze our hypothesis. Those performing the ETIs were interns (1), anesthesiology residents (17 CA1, 6

continued on next page

continued from previous page

CA3), certified registered nurse anesthetists (CRNAs) (1), and anesthesiology faculty (2). Given the range of expertise, we expected that the variability would be enough to allow good correlations among 3 measures of quality. These measures are rank ordering, global scoring (ratings from 1 to 100, 100 being the best), and average points from the metric (22 items).

Before recording the videos, verbal approval was obtained from the patients and all operating room personnel. Adult patients (≥ 18 years old) undergoing elective procedures in which general endotracheal anesthesia was planned were eligible for inclusion. Patients undergoing cardiac, neurosurgical, or intrathoracic procedures were excluded, as were patients with anticipated difficult intubation or need for rapid sequence induction. The identity of the clinician performing the ETI was hidden from view, but voices were audible.

Each video contains audio and 3 video views, one from a camera mounted on the head of the clinician (GoPro HERO 3; GoPro, Inc. San Mateo, CA, Figure 2A), the C-MAC screen view (Figure 2B and 2C), and a lateral view captured with a digital camcorder. The C-MAC screen was not visible to the clinician during the intubation; it was available to the reviewers to assess performance. Each video begins before the physician adjusts the patient's head and ends once the breathing circuit is attached. While performing the procedure, it was requested that the operator "think aloud" to express actions or intentions that may not be obvious from the video recording. This narration was subtitled in the videos presented to the evaluators.

Clips from the 3 different views were synced and shown in series, such that only one view is being seen at a time by the viewer. The expert raters were permitted to pause, rewind, and rewatch the videos as desired. The identity of the patients was kept confidential by blurring faces and any identifying features such as tattoos.

Metric Analysis

For evidence of the metric's reliability and validity, we performed 2 between-subject studies (Figure 3): one to assess the measurement characteristics of the metric (Metric Study), and the other to assess the validity of the metric against a reference standard global assessment (Reference Study). These studies were conducted using the 2 video collections, with 2 different groups analyzing each collection (4 total expert groups). Literature shows that panels of at least 3 experts have been used in the past to develop grading systems when an established standard does not exist.[7,8] Experts in Metric Group and Reference Group 1 were recruited by emailing all attending physician and CRNA members of the Department of Anesthesia, Critical Care, and Pain Medicine at Beth Israel Deaconess Medical Center. The experts in the Metric Group and Reference Group 2 were recruited by emailing the Society for Education in Anesthesia committee members and chairs.

Measurement Characteristics of Metric—Study I

The goal for the ETI metric groups was to test for agreement between the evaluators for each video and for individual items in the metric. Two different groups of experts, referred to as Metric Group 1 and 2 ($N_{M1} = 6$ and $N_{M2} = 3$), completed the metric for each of the ETI procedures in Video Collection A and B, respectively.

Reference Standard—Study II

The goal of the reference standard groups was to provide a score on a 100-point scale of operator performance for each of the videos and a rank ordering of the videos based on operator performance. We had 2 groups of experts, referred to as Reference Group 1 and 2 ($N_{R1} = 7$ and $N_{R2} = 3$), use their personal standard of quality, as they would while teaching in the operating room, to rank order and rate the ETI procedures in Video Collection A and B, respectively. For each video in a collection, experts assigned a global score from 0 to 100 using a digital slider scale, where 100 is optimal performance of laryngoscopy and intubation. Next, experts rank-ordered the same videos in descending order from optimal/best-ranked number 1 through the worst for that collection. Finally, metric-based data from Study I was correlated with rating and rank data from Study II to test the validity of the metric data.

Statistical Analysis

All statistics and models were generated using IBM SPSS (IBM Corp, Armonk, NY) and MATLAB (The MathWorks Inc., Natick, MA). Agreement across those ranking and rating videos was assessed with an intraclass correlation coefficient (ICC) using an absolute agreement model for ranks and a consistency model for ratings. For the observational metric, the highest proportion of agreement from the 6 (Metric Group 1) and 3 (Metric Group 2) observers in each group was calculated for each item. Finally, the proportion of agreement for each item was averaged to produce an estimate of overall observer agreement. An average rank, an average rating, and an average metric score was calculated for each video. A nonparametric correlation (Spearman ρ) was calculated between pairs of these measures. A flowchart of how the results from each study were used in the calculation of agreement and quality of the metric can be seen in Figure 4.

There are at least 2 approaches for setting a score that separates competent and not competent performance. One involves identifying only actions that optimally separate competent from not competent performance. The other uses all clinically important items, as determined by expert clinical educators, where competent performance requires documenting that each action is done correctly. Feedback on each action is key for training to competent performance. When all required actions are completed correctly, and no harmful actions made, the performance is declared competent. As both assessment approaches might be needed when training to competency (summative versus formative), we first determined which items optimally discriminate competent from not competent performance. To assess whether different item weighting resulted in higher correlation with global ratings and ranks, we used 3 different models: binary scoring, expert-based weighting, and partial least square (PLS) regression. The first model awarded the participant 1 point for doing an item correctly and no points if the participant partially or incorrectly completed an item. The expert-based weights were determined during the metric development phase. PLS[9] is a statistical method that de-

continued on next page

continued from previous page

termines a linear regression model by identifying the relationship between 2 matrices. It is known to perform well if the number of observations relative to the number of variables is small,[10-12] which follows from the fact that it obtains latent variables using a covariance-based objective function that balances model accuracy with capturing variance information from the predictor variable set. For our regression problem, we have 22 predictor variables (the recorded metric items) and 27 observations (videos). For the PLS model, we sought to determine weightings of the individual items in a regression analysis with the global scores. To develop the model, we averaged each metric item for an ETI participant across all raters from the Metric Study and regressed it to the average global scores found in the Reference Study. Our algorithm optimized the combination of metric items based on the lowest attainable error of the predicted score while achieving a power greater than 0.8 ($P < .05$) for our given size of observations (n = 27).[12] To evaluate the regression-based models, we used an independent cross-validation model approach, where a model was created by leaving out the scores from one video and then used to predict the global score. Error was then calculated by taking the difference from the average from the Reference Study. This was done for all videos, and the error for the set of items was calculated by averaging the difference over all videos.

Results

Three of the original 14 videos were not usable for rating because a mechanical error by one camera left only 2 views and some items on the metric required the third view. Thus, with the second collection of 16 videos, a total of 27 videos were available for these studies.

The percentage of observer agreement for each item is presented in Table 1. The average of individual item agreement yielded an overall agreement of 80% for the metric. ICCs may be calculated 2 ways: one for consistency and the other for absolute agreement. Using the absolute agreement approach, ICCs (2-way mixed, average measures, absolute agreement) from the first rater group pair (Video Collection A, Metric Group 1 and Reference Group 1) were 0.85 and 0.73, for ranks and ratings, respectively. The second rater group pair (Video Collection B, Metric Group 2 and Reference Group 2) had ICCs of 0.88 and 0.87 for ranks and ratings, respectively. Correlations between ranks and ratings for each group are shown in Table 2. The correlation coefficient between global ranks and ratings for the 2 collections were −0.95 ($P < .05$) and −0.96 ($P < .05$). Figure 5 shows a plot of the average rank versus average score for each of the videos in Video Collection A and B. Because of the high correlation coefficients, a minimum sample size of 9 (r = 0.81) is met for all the comparisons to achieve a power greater than 0.8 using a 2-tailed hypothesis test.[13]

Results from different weighting models showed that the binary weighting and expert weighting strongly correlated with the global score (0.87 and 0.86, respectively). Both models had a strong correlation with ranks in Video Collection A (−0.86 and −0.87, respectively) and in Video Collection B (−0.82 and −0.87, respectively). When predicting global score, the expert-based weighting had an R^2 of 0.7095 and the binary weighting had an R^2 of 0.7120. Two PLS models were calculated, one using all metric items and the other using a subset of items that produced the lowest error when predicting the score. When using all the original items, the sum of R^2 is 0.7242, which is considered a fair model (1 is optimal). The calculated weights for the full model can be seen in Figure 6.

When using the reduced set of items to create the model (3 items: 11, 15, and 20), the R^2 is 0.8218 ($P < .05$). The absolute weighting coefficients from the reduced set model consisting of items 11, 15, and 20 are 0.3917, 0.4987, and 0.3818, respectively. This is a significantly better model than using all items because its lowest weighting coefficient of 0.3818 achieves a power greater than 0.8 for a sample size of 27. An independent cross-validatory assessment is accomplished in which scores for the *i*th video is tested against the *i*th model using all items and using a reduced set. The scores are based on the average score given for each video across the raters in Reference Group 1 and 2. Figure 7 shows each item's error between the actual and the predicted models. Smaller errors indicate a better fit. In the full set model (left graph), videos 2, 3, 6, 8, and 24 show high variability. When using the reduced set model, all videos show smaller errors (ie, a better fit). The reduced set model achieves lower errors when cross validating. Results from the 2 PLS models showed that both models correlated strongly with ranks and ratings (see Table 2).

The weighting for each item option and how a global score would be calculated for the full and reduced metric can be seen in Appendix A and B, respectively. An individual's score starts at a base score of 82 (reduced metric) or 66 (full metric) and depending on their actions for each metric item, points are earned or deducted. Considering that larger absolute weightings had a more profound impact in predicting the score value, this implies that the accuracy of carrying out the associated tasks for items 11, 15, and 20 is predictive of overall proficiency of the procedure.

Discussion

We developed and analyzed a 22-item observational metric for use in assessing the quality of ETI performance with sufficient detail to provide formative feedback. High observer agreement and high correlations between metric and rank data support the validity of using these items to assess ETI quality. Analysis of 2 weighting models, a binary model of a single point awarded for each correct action and weighting based on expert opinion, yielded calculated scores that correlated strongly with the ratings and ranks from global assessment (Reference study). We showed that calculated weights from the PLS model were able to better predict a score compared with the binary and expert weighted model (R^2 of 0.7242, r =0.93) and further improved when using a subset of the metric items (R^2 of 0.8218, r =0.91). It was also seen that the reduced set model correlated more strongly with rating and ranking from the Reference study than the full set model.

Following the weighted data analysis, we discussed the results with expert anesthesiologists to examine the correlation with what is expected in practice. Results from the PLS model demonstrated that one of the best predictors of a final performance score were "multiple blade insertion attempts to achieve proper view" (item 15). This re-

continued on next page

continued from previous page

sult parallels a study by Martin et al,[14] who showed that higher skill levels correlated with fewer attempts, and repeated intubation attempts may lead to tissue trauma, edema, and bleeding. Although this is useful for differentiating high and low performance, it does not necessarily inform trainees on how to improve, as multiple attempts may be due to several factors, such as unsuccessfully achieving an appropriate view (item 13), having the blade in the incorrect location when lifting (item 10), or the angle of lift on the first attempt (item 14). Another predictor that parallels well with existing studies[15–17] was "excessive force while interacting with the vocal cords," as a major goal of ETI training is to reduce the likelihood and incidence of laryngeal injuries.

The items in the metric were designed to be measurable using existing technology, observable by an expert and quantifiable using an algorithm. To use the metric in practice, an observer, whether it be human or computer-based, must detect and indicate when an action is performed. From a learning perspective, the items were intended to provide feedback for performance improvement and to differentiate performance between users. When using the metric for formative clinical training, the full-weighting model should be used, as a learner must demonstrate all important actions, even if some are not efficient predictors of the final score. When using the metric to measure competency, the reduced PLS model (3 items) is most accurate in predicting a global score while maintaining a power of at least 0.8 ($P < .05$).

Next steps include testing the metric for effectiveness in learning and incorporating into ETI training. In addition, we are planning to continue collecting videos to improve the accuracy of the models. Further work is needed to demonstrate that the scoring reliably separates competent from not competent performance, and to add force and angular measurements to the assessment model.

There are some limitations to this work. One is our choice to videotape actual ETI performance in the operating room by providers at various stages of training and expertise. Evaluators were viewing video recordings of the ETI rather than directly viewing the procedure being performed. Despite presenting multiple views, evaluators were limited to those perspectives that may have been different from direct observation in the operating room. Another limitation was that we used only a Mac 3 C-MAC blade for the trials in this study, and although the same metric could be used to provide feedback, the weighting may differ for another blade. Recruitment of experts with sufficient experience in education was also a limitation. We needed at least 3 experts to measure for agreement in each group, and we did not accept partial expert reviews, hence the difference in expert group numbers.

The intention was to use actual operating room performance to evaluate our metric for use with actual training. With clinicians at various levels of training and expertise, we obtained a considerable range of ETI performance quality that allowed us to find high correlations with different scoring approaches. Other samples of actual performance are needed to confirm the applicability of our metric. The groups of expert observers were sufficient for observer agreement,[7] and although another group of experts may yield different results, we hypothesize that selecting another sample of seasoned anesthesiologists would produce the same result. A formal standard-setting process will need to be undertaken to determine the score that separates competent from not competent performance. The set of metric items only references ETI quality based on what can be observed and does not include sensors such as force tracking. Research conducted by Bishop et al,[18] Hastings et al,[19,20] and Garcia et al[21] provide force and torque parameters applied to the laryngoscope for tracheal intubation. Finally, checklist metric items were established from a review of literature and recommendations from experienced anesthesiology educators at one institution. Other groups of anesthesiologists may determine different items and/or weightings when judging the quality of ETI performance. A limitation that we experienced with the generation of the regression-based model was that as the number of metric items increased, the weighting of each item would decrease. A low regression weighting for a given item would require a very high number of observations to achieve a power greater than 0.8.[12] For instance, a weighting coefficient of $\beta = 0.1$ would require at least 599 observations for a power of 0.8. Due to the low weighting coefficients of the 22-item PLS model, the power is well below the threshold of 0.8 for a significance level of .05 and sample size of n = 27. Although this weighting was not the primary goal of the paper, more observations can be collected to see if the error and number of metrics plateau at a certain sample size.

In conclusion, the results of these studies provide evidence that the checklist metric items and descriptions developed by our experts have strong measurement characteristics. Use of the metric may take place in training within simulators and the operating room to give feedback to individuals. Further testing is needed to demonstrate utility and effectiveness as a feedback tool. The metric potentially provides a way to give more detailed technical feedback than currently available ETI checklists.

References

1. Faudeux C, Tran A, Dupont A, et al. Development of reliable and validated tools to evaluate technical resuscitation skills in a pediatric simulation setting: resuscitation and emergency simulation checklist for assessment in pediatrics. *J Pediatr* 2017;188:252-7.e6.
2. Silverman E, Dunkin BJ, Todd SR, et al. Nonsurgical airway management training for surgeons. *J Surg Educ* 2008;65(2):101-8.
3. Kuszajewski ML, O'Donnell JM, Phrampus PE, et al. Airway management: a structured curriculum for critical care transport providers. *Air Med J* 2016;35(3):138-42.
4. Apfelbaum JL, Hagberg CA, Caplan RA. Practice guidelines for management of the difficult airway. *Anesthesiology* 2013;(5):1269-77.
5. Miller RD. *Miller's Anesthesia,* 7th ed, Churchill Livingstone, Philadelphia, PA, 2010.
6. Horton WA, Fahy L, Charters P. Defining a standard intubating position using "Angle Finder." *Br J Anaesth* 1989;62(1):6-12.
7. van Houten CB, Naaktgeboren CA, Ashkenazi-Hoffnung L, et al. Expert panel diagnosis demonstrated high reproducibility as reference standard in infectious diseases. *J Clin Epidemiol* 2019;112:20-7.
8. Tampin B, Broe RE, Seow LL, et al. Field testing of the revised neuropathic pain grading system in a cohort of patients with neck and upper limb pain. *Scand J Pain* 2019;19(3):523-32.
9. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intell Lab Syst* 2001;58(2):109-30.
10. Höskuldsson A. PLS regression methods. *J Chemom*. 1988;2(3):211-28.

continued on next page

continued from previous page

11. Hair JF, Ringle CM, Sarstedt M. PLS-SEM: indeed a silver bullet. *J Mark Theory Pract* 2011;19(2):139-52.
12. Kock N, Hadaya P. Minimum sample size estimation in PLS-SEM: The inverse square root and gamma-exponential methods. *Inf Syst J.* 2016;28:227-61.
13. Hulley SB, Cummings SR, Browner WS, et al. *Designing Clinical Research: An Epidemiologic Approach*, 4th ed. Lippincott Williams & Wilkins, Philadelphia, PA, 2013.
14. Martin LD, Mhyre JM, Shanks AM, et al. 3,423 Emergency tracheal intubations at a university hospital. *Anesthesiology* 2011;114(1):42-8.
15. Pacheco-Lopez PC, Berkow LC, Hillel AT, Akst LM. Complications of airway management. *Respir Care* 2014;59(6):1006-19; discussion 1019-21.
16. Benjamin B, Holinger LD. Laryngeal complications of endotracheal intubation. *Ann Otol Rhinol Laryngol* 2008, 117(9):2-20.
17. Santos PM, Afrassiabi A, Weymuller EA. Risk factors associated with prolonged intubation and laryngeal injury. *Otolaryngol Neck Surg* 1994;111(4):453-9.
18. Bishop MJ, Harrington RM, Tencer AF. Force applied during tracheal intubation. *Anesth Analg* 1992;74:411-4.
19. Hastings RH, Hon ED, Nghiem C, Wahrenbrock EA. Force, torque, and stress relaxation with direct laryngoscopy. *Anesth Analg* 1996;82(3):456-61.
20. Hastings RH, Hon ED, Nghiem C, Wahrenbrock EA. Force and torque vary between laryngoscopists and laryngoscope blades. *Anesth Analg* 1996;82(3):462-8.
21. Garcia J, Coste A, Tavares W, et al. Assessment of competency during orotracheal intubation in medical simulation. *Br J Anaesth* 2015;115(2):302-7.

The following authors are in the Center for Modeling, Simulation and Imaging in Medicine (CeMSIM) at Rensselaer Polytechnic Institute, Troy, NY: ***Adam Ryason*** *is graduate student;* ***Uwe Kruger*** *is a professor of practice.* ***Zhaohui Xia*** *is an assistant professor at Huazhong University of Science and Technology, National Enterprise Information Software Engineering Research Center, Wuhan, Hubei, China.* ***Emil R. Petrusa*** *is a professor of surgery at Harvard Medical School, Department of Surgery, Learning Lab, Massachusetts General Hospital, Boston, MA.* ***Vanessa T. Wong*** *is a project coordinator, Department of Anesthesia, Critical Care and Pain Medicine, Beth Israel Deaconess Medical Center, Boston, MA.* ***Daniel B. Jones*** *is the Vice Chair of Surgery, Harvard Medical School, Department of Surgery, Beth Israel Deaconess Medical Center, Boston, MA.* ***Stephanie B. Jones*** *is the Chair, Department of Anesthesiology. Albany Medical Center, Albany, NY.*

Corresponding author: *Adam Ryason, 110 8th Street, CII 9th floor, Center for Modeling Simulation and Imaging in Medicine. Troy, NY, 12180. Telephone: 845-803-1738*

Email address: *Adam Ryason: ryasoa@rpi.edu*

Funding: *National Institutes of Health 5R01HL119248-04*

Abstract

Background: Valid methods for providing detailed formative feedback on direct laryngoscopy and endotracheal intubation (ETI) performance do not exist. We are developing an observation-based assessment tool for measuring performance and providing feedback during ETI.

Methods: Based on the literature and interviews of experts, we proposed an initial ETI metric with 22 items. Six anesthesiology experts used it to assess the quality of ETI performance in videotaped intubations. Following metric revisions, 2 expert groups assessed 2 collections of videos (27 total) using the revised metric. Two reference standards for comparison with metric scores were created with a third and fourth group of experts; (1) an average global rating (1-100) of each ETI performance and (2) average rank-ordered performance from best to worst. Rater agreement and correlations between the 2 methods were calculated. Regression analysis determined items that optimally discriminated quality. When calculating a score based on all clinically important terms, multiple weightings were evaluated.

Results: Metric items had high average rater agreement (80%) with intraclass correlation coefficients averaging 0.83. Correlations of the reference rank and score were high for both video collections (−0.96, $P < .05$, and −0.95, $P < .05$). Regression coefficients for different item weighting methods indicated strong relationships with global ratings (averaging $r = 0.89$, $P < .05$) and rankings averaging −0.85, $P < .05$). Prediction of global ratings using regression achieved high accuracy ($R^2 = 0.8218$).

Conclusions: High observer agreement and strong correlations between metric and rank data support the validity of using this metric to assess ETI performance. Different weighting models yielded scores that correlated strongly with the ratings and ranks from global assessment. When using the metric to predict competency, a 3-item regression model is most accurate in predicting a global score.

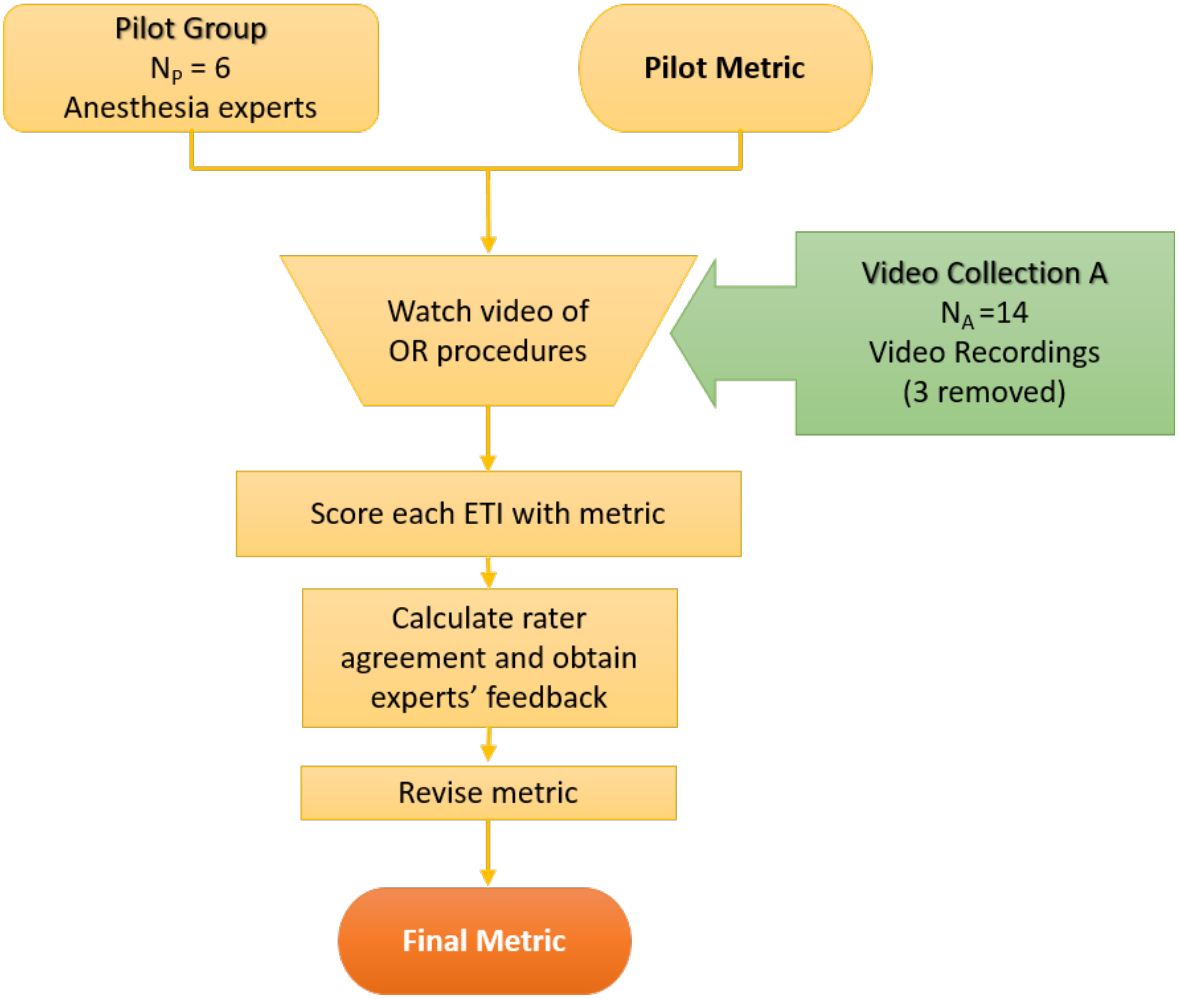
Keywords: Regression analysis, endotracheal intubation, training rubric, video analysis

continued on next page

continued from previous page

Figures

Figure 1. *Study design for deriving the Endotracheal Intubation Performance Metric.*

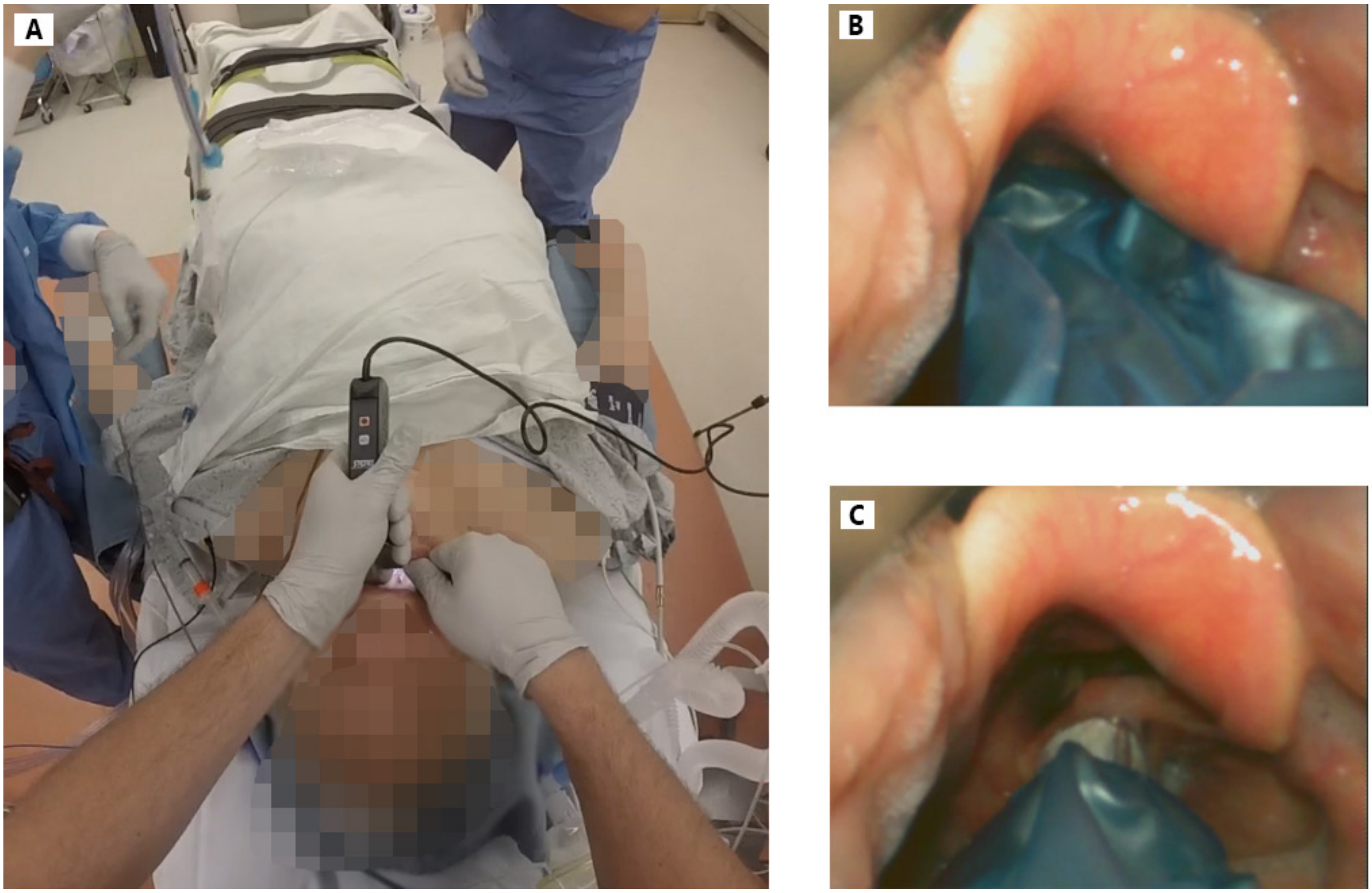


continued on next page

continued from previous page

Figures *continued*

Figure 2. *(A) Screenshot of the view from the clinician. (B) Laryngoscope view showing the clinician properly inserting the endotracheal tube. (C) Laryngoscope view showing the clinician incorrectly inserting the endotracheal tube.*



continued on next page

continued from previous page

Figures *continued*

Figure 3. *Study design for developing a grading standard based on multiple panels of experts. Left (orange) is the study testing the measurement characteristics of the metric. Right (blue) is the study developing the reference standard. Metric and Reference Group 1 and 2 analyzed Video Collection A and B, respectively.*

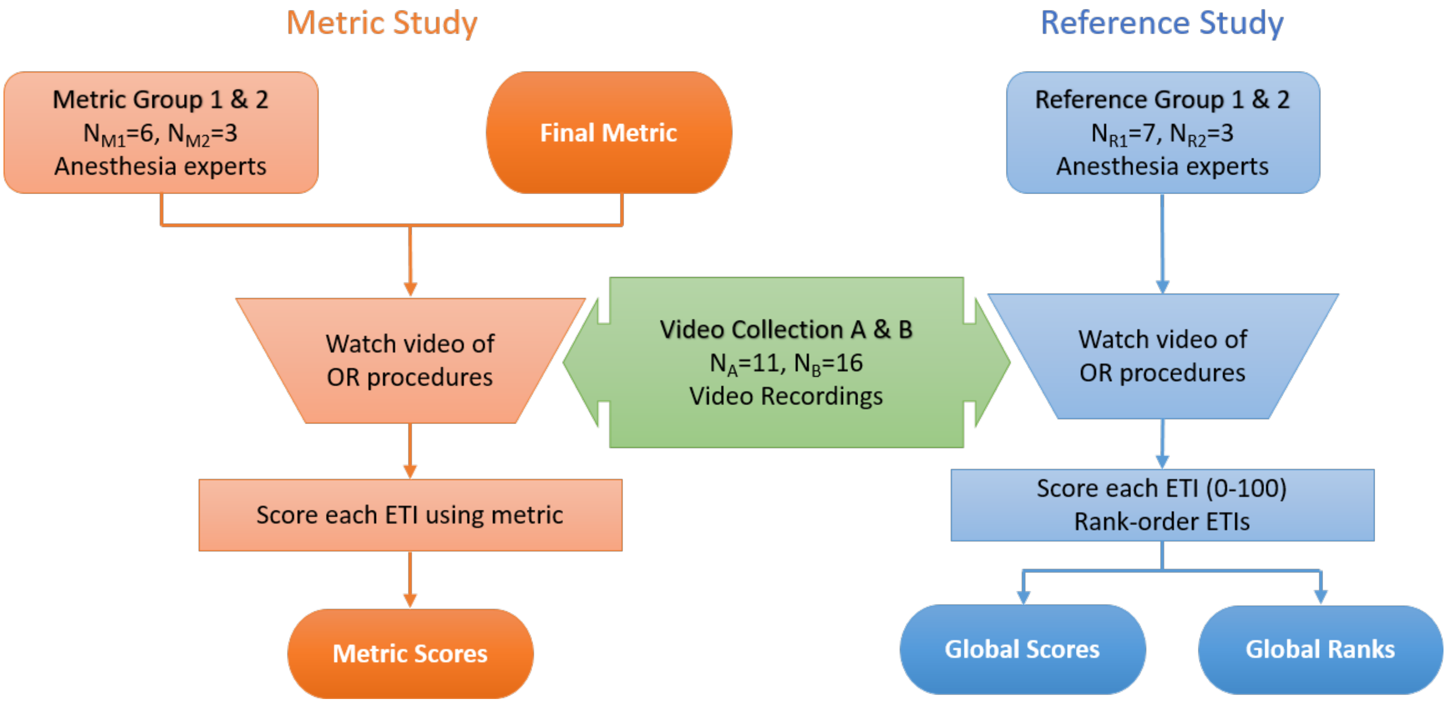
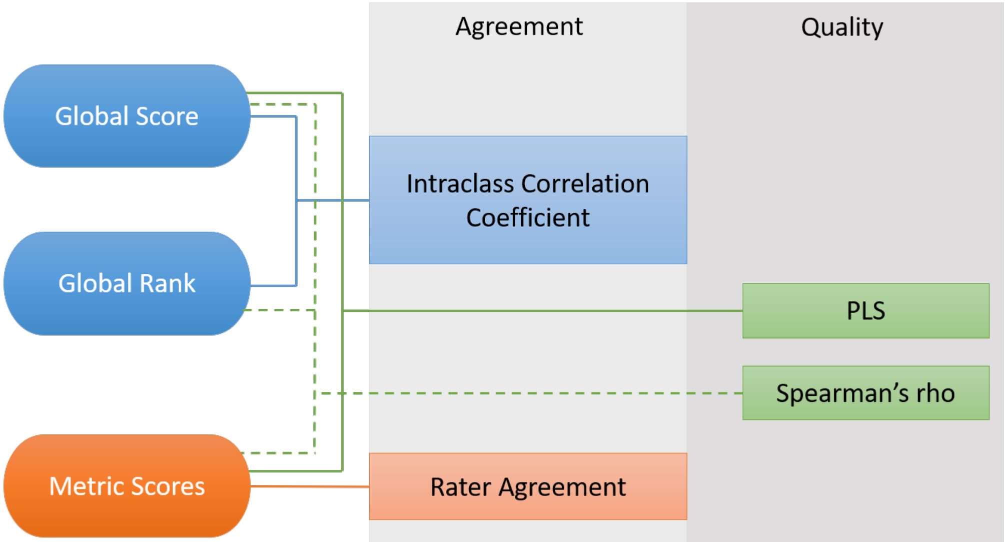


Figure 4. *Analysis performed to check for agreement and quality of the metric.*

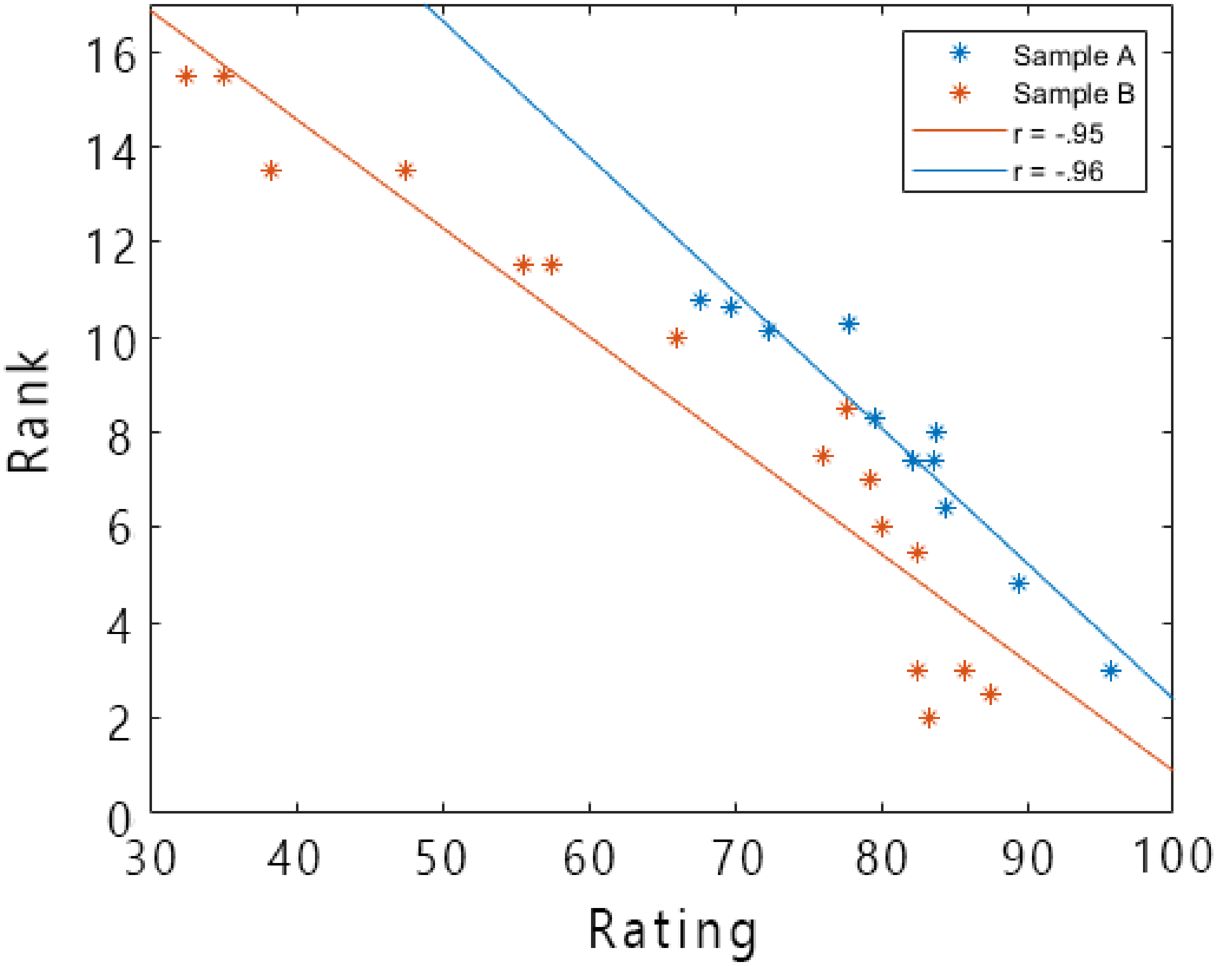


continued on next page

continued from previous page

Figures *continued*

Figure 5. *Plot of the average global score versus average global ranking for the 2 sample groups.*



continued on next page

continued from previous page

Figures *continued*

Figure 6. *Bar plot of the normalized weights for the 22-item partial least squares model when predicting global score.*

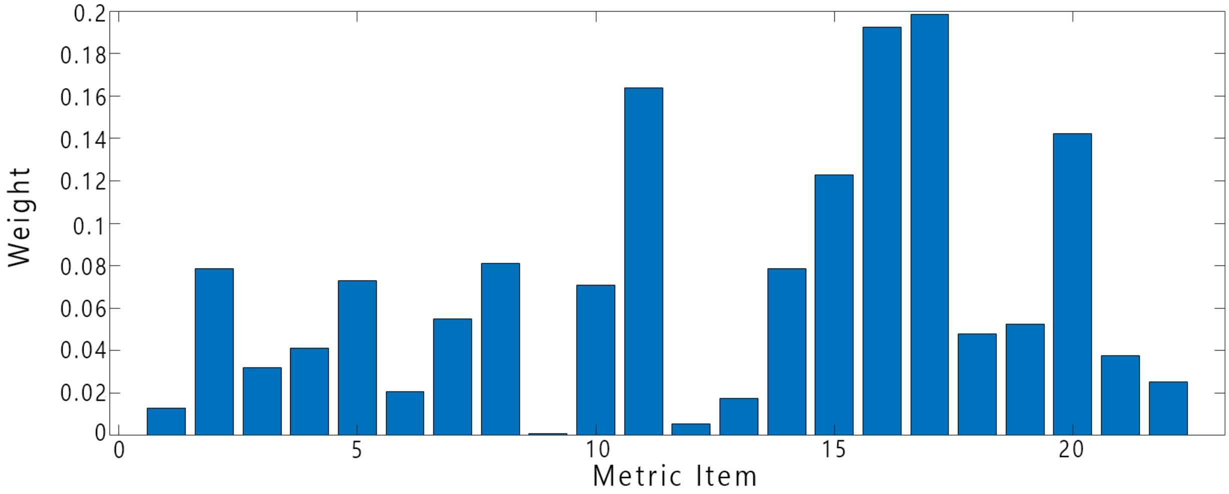
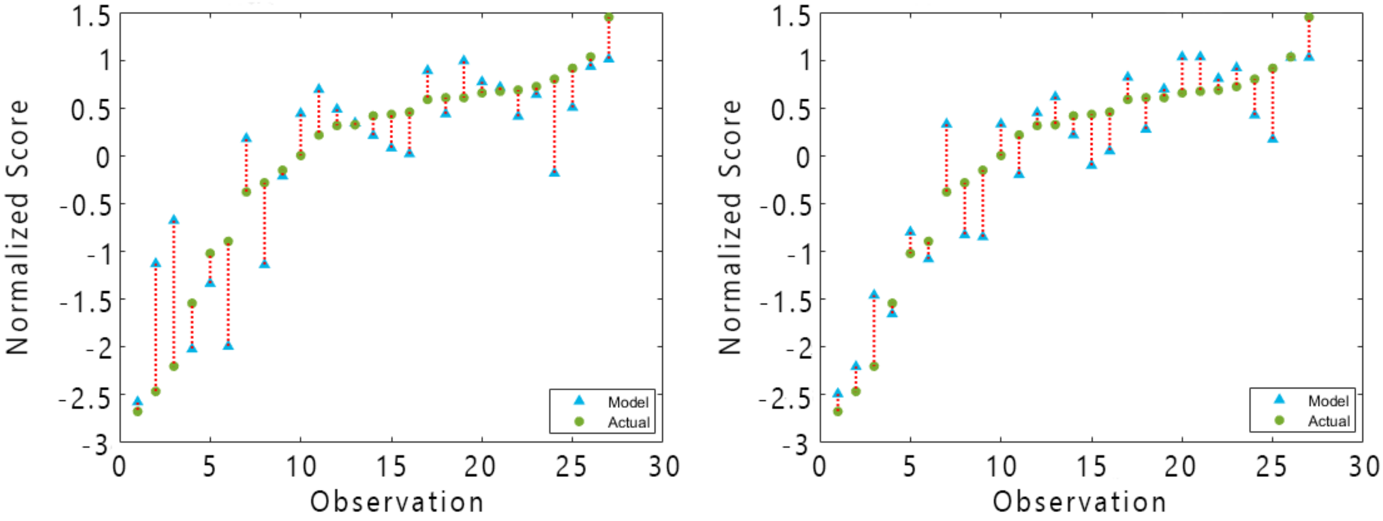


Figure 7. *Scores as predicted by the model compared with the actual scores for the 27 videos (rank ordered) when using independent cross-validatory assessment with all metric items (left) and as predicted with a reduced set of 3 items (right).*



continued on next page

continued from previous page

Tables

Table 1. *Rater Agreement for Each Item in the Metric*

Item	1	2	3	4	5	6	7	8	9	10	11
Agreement	0.75	0.80	0.90	0.74	0.71	0.89	0.78	0.83	0.94	0.67	0.77
Item	12	13	14	15	16	17	18	19	20	21	22
Agreement	0.83	0.79	0.65	0.77	0.83	0.84	0.86	0.79	0.81	0.79	0.79

Table 2. *Correlations Between Pairs of Measures of ETI Quality*

Pairs of Data	Spearman ρ[a]		
	Global Score	Rank (Collection A)	Rank (Collection B)
Binary weighting	0.87	−0.86	−0.82
Expert-based weighting	0.86	−0.87	−0.87
PLS model (22 items)	0.93	−0.87	−0.88
PLS model (3 items)	0.91	−0.76	−0.89

Abbreviations: ETI, endotracheal intubation; PLS, partial least square.

[a] $P < .001$.

continued on next page

continued from previous page

Appendix A

ETI Performance Metric

Operator ID: Date: Trial Number: Reviewer ID:	Base Score 66 +	Metric Scores =	Final Score /100

Positioning of the Patient		
1. Positioning of Patient Head	Expert Based Points	Score
The operator properly tilted the head into the sniffing position	**1**	
The operator did not tilt the head into the sniffing position *Feedback: The angle of neck flexion should be placed at approximately 35 degrees[1].*	**0**	
2. Elevation of Patient's Head		
The operator properly elevated the patient's head or did not need to elevate the patient's head	**1**	
The operator should have elevated the patient's head but did not *Feedback: The angle of face extension should be approximately 15 degrees[1].*	**-3**	

Insertion of Direct Laryngoscopy Blade		
3. Grip of Laryngoscope	Expert Based Points	Score
The operator had a proper grip on the laryngoscope	**2**	
The operator had an improper grip on the laryngoscope *Feedback: The laryngoscope needs to be in the left hand and high enough such that it's not obstructing the entry of the blade in to the mouth*	**0**	
4. Method to Open Mouth		
The operator adequately opens the mouth by scissoring their finger and thumb	**2**	
The operator adequately opens the mouth some other way (e.g., using the blade)	**2**	
The operator does not adequately open the mouth *Feedback: Apply opposing pressure onto lower and upper teeth using thumb and middle finger, respectively.*	**-1**	
5. Location of the Blade While Inserting Into the Mouth		
Right side of the mouth and sweep the blade and tongue left until midline is reached	**1**	
Middle of the mouth but still able to sweep tongue *Feedback: Start at the right side of the mouth to sweep tongue to the left.*	**-1**	
Any other insertion location (please describe in comment) *Feedback: Start at the right side of the mouth to sweep tongue to the left.*	**-3**	

[1] Horton WA, Fahy L, Charters P. Defining a Standard Intubating Position Using "Angle Finder." *Br J Anaesth*. 1989;62(1):6-12.

continued on next page

continued from previous page

Appendix A *continued*

6. Blade Insertion with Respect to the Vallecula	Expert Based Points	Score
Insert blade into the vallecula	**0**	
Insert blade under the vallecula *Feedback: Pull blade back into the vallecula*	**-1**	
7. Force Used while Interacting with Vallecula		
The force appeared excessive *Feedback: Correct direction, use less force*	**-3**	
The force used appeared appropriate	**1**	
The force used appeared insufficient *Feedback: Verify correct direction, use more force*	**-1**	
8. Contact with Teeth During Lifting the Blade		
The blade was lifted without contacting the teeth	**1**	
The blade was lifted while hitting teeth with no damage *Feedback: Reduce leveraging the blade toward the operator. Blade should be lifted at a 45-degree angle away from the operator.*	**-2**	
The blade was lifted while hitting teeth with damage to the teeth *Feedback: Reduce leveraging the blade toward the operator. Blade should be lifted at a 45-degree angle away from the operator.*	**-4**	
9. Order of Events for the Insertion of the Laryngoscope		
The order of events was correct	**0**	
The order of events was incorrect *Feedback:* *1. Position and raise the head* *2. Properly grip the laryngoscope* *3. Scissor open the mouth* *4. Insert the blade on the right side of the mouth and sweep the tongue left* *5. Enter the vallecula*	**-1**	

Achieving the Optimal Laryngeal View		
10. Final Blade Position in the Vallecula When Lifting for Optimal View	Expert Based Points	Score
Blade is in the correct position in the vallecula	**1**	
Blade is too shallow in the vallecula *Feedback: Place the blade deeper in the vallecula*	**0**	
Blade is too deep in the vallecula *Feedback: Do not insert the blade as deep in the vallecula*	**-2**	
Blade is not in the vallecula *Feedback: Reposition the blade to be in the vallecula*	**-3**	

continued on next page

continued from previous page

Appendix A *continued*

11. Blade Position with Respect to the Oropharynx		
Blade is in the midline of the patient's oropharynx	**2**	
Blade is not in the midline of the patient's oropharynx, but the operator started again from the right (correct) side *Feedback: Optimize attempts to position the blade in the midline of the oropharynx.*	**-2**	
Blade is not in the midline of the patient's oropharynx and is not adjusted *Feedback: The blade should be in the midline.*	**-7**	
12. Lift on Laryngoscope for Proper View		
The blade lifted up on tongue/vallecula enough for sufficient view	**1**	
The blade did not lift up on the tongue/vallecula enough for a sufficient view *Feedback: Increase the lift at a 45-degree angle.*	**-1**	
13. Quality of the Vocal Cords View		
The vocal cords were in view before intubating	**1**	
The vocal cords were not in view before intubating *Feedback: The vocal cords should be in view before endotracheal intubation. Check that the appropriate lift angle was used.*	**-1**	
14. Angle of Lift on First Attempt		
The laryngoscope had a backward angle onto the teeth *Feedback: Reduce the angle of the blade by keeping the angle of the handle around 45 degrees*	**-2**	
The laryngoscope was angle appropriately (approx. 45°)	**4**	
The laryngoscope was angle too shallow (0-45°) *Feedback: Increase the angle of the blade by keeping the angle of the handle around 45 degrees*	**0**	
15. Multiple Blade Insertion Attempts to Achieve Proper View		
The proper view was achieved the first time the blade was inserted (no removal & reentry of blade)	**2**	
The blade was removed from the patient and reentered 2-3 times before achieving proper view *Feedback: Optimize early attempts and guarantee that previous steps are completed properly before continuing.*	**0**	
The blade was removed from the patient and reentered 4+ times before achieving proper view *Feedback: Optimize early attempts and guarantee that previous steps are completed properly before continuing.*	**-3**	
A proper view was not obtained. *Feedback: Improve other metric items based on their feedback.*	**-5**	

continued on next page

continued from previous page

Appendix A *continued*

Inserting the Endotracheal Tube		
16. Number of Contacts of Tube During Insertion	Expert Based Points	Score
The tube was inserted with no or negligible number of contacts to surrounding anatomy	**3**	
The tube was inserted with an excessive number of contacts to surrounding anatomy *Feedback: Avoid excessive contact with surrounding anatomy. Observe end of tube during insertion to avoid excessive contact with surrounding anatomy.*	**-2**	
The tube was not inserted. *Feedback: Improve other metric items based on their feedback.*	**-8**	
17. Multiple Intubation Attempts		
The clinician successfully intubated the patient on the first attempt	**3**	
Had to perform one additional intubation attempt *Feedback: Optimize early attempts and guarantee that there is a clear view of the vocal cords before intubating*	**0**	
Had to perform at least 2 additional intubation attempts *Feedback: Optimize early attempts and guarantee that there is a clear view of the vocal cords before intubating*	**-4**	
The intubation was not successful *Feedback: Improve other metric items based on their feedback.*	**-7**	

Avoiding Injury to the Patient		
18. Was Excessive Force Used to Insert the Laryngoscope into the Oropharynx?	Expert Based Points	Score
No, the force appeared appropriate	**1**	
Yes, at one of more times the force appeared excessive *Feedback: Reduce the rate of approach and observe surrounding tissues during insertion.*	**-2**	
19. Was Excessive Force Used to Insert the ETT into the Oropharynx?		
No, the force appeared appropriate	**4**	
Yes, at one of more times the force appeared excessive *Feedback: Reduce the rate of approach and observe surrounding tissues during insertion. Do not force the ETT into the oropharynx.*	**1**	
The tube was not inserted. *Feedback: Improve other metric items based on their feedback.*	**-2**	
20. Was Excessive Force Used While Interacting with the Vocal Cords?		
No, the force appeared appropriate.	**2**	
Yes, at one or more times the force appeared excessive. *Feedback: Reduce the force applied when interacting with the vocal cords. Do not force the ETT through the vocal cords; the ETT should pass smoothly through the vocal chords. If not, may need to alter angle of approach or consider using smaller ETT.*	**-8**	

continued on next page

continued from previous page

Appendix A *continued*

21. Laryngoscope Manipulation Around Lip(s)		
There was no pinching of the lips	**0**	
There was pinching of the lips *Feedback: Ensure to clear lips from around the laryngoscope blade*	**-1**	
22. Laryngoscope and ETT Contact with Tissue and Structures		
The contact with tissue and structures was appropriate	**1**	
The contact with tissues and structures was excessive *Feedback: Minimize contact with surrounding tissue and structures. Observe laryngoscope and ETT during insertion to avoid excessive contact with surrounding anatomy.*	**0**	

Appendix B

ETI Performance Reduced Metric

Operator ID: Date: Trial Number: Reviewer ID:	Base Score 82 +	Metric Scores =	Final Score /100

Achieving the Optimal Laryngeal View		
11. Blade Position with Respect to the Oropharynx	Points	Score
Blade is in the midline of the patient's oropharynx	**6**	
Blade is not in the midline of the patient's oropharynx but the operator started again from the right (correct) side *Feedback: Optimize attempts to position the blade in the midline of the oropharynx.*	**-5**	
Blade is not in the midline of the patient's oropharynx and is not adjusted *Feedback: The blade should be in the midline.*	**-16**	
15. Multiple Blade Insertion Attempts to Achieve Proper View		
The proper view was achieved the first time the blade was inserted (no removal & reentry of blade)	**7**	
The blade was removed from the patient and reentered 2-3 times before achieving proper view *Feedback: Optimize early attempts and guarantee that previous steps are completed properly before continuing.*	**-3**	
The blade was removed from the patient and reentered 4+ times before achieving proper view *Feedback: Optimize early attempts and guarantee that previous steps are completed properly before continuing.*	**-12**	
The blade did not lift up on the tongue/vallecula enough for a sufficient view *Feedback: Attempt to re-sweep tongue until the blade reaches the midline and increase the lift at a 45-degree angle.*	**-22**	

continued on next page

continued from previous page

Appendix B *continued*

Avoiding Injury to the Patient		
20. Was Excessive Force Used While Interacting with the Vocal Cords?	Points	Score
No, the force appeared appropriate.	**5**	
Yes, at one or more times the force appeared excessive. *Feedback: Reduce the force applied when interacting with the vocal cords. Do not force the ETT through the vocal cords; the ETT should pass smoothly through the vocal chords. If not, may need to alter angle of approach or consider using smaller ETT.*	**-21**	