Neonatal Intubation Competency Assessment Tool: Development and Validation

Lindsay Johnston, MD, MEd; Taylor Sawyer, DO, MEd; Akira Nishisaki, MD, MSCE; Travis Whitfill, MPH; Anne Ades, MD, MSEd; Heather French, MD, MSEd; Kristen Glass, MD; Rita Dadiz, DO; Christie Bruno, DO; Orly Levit, MD; Sandeep Gangadharan, MD; Daniel Scherzer, MD; Ahmed Moussa, MD, MEd; Marc Auerbach, MD, MSCI; on behalf of the INSPIRE Research Network[1]

From the Department of Pediatrics (L Johnston, T Whitfill, C Bruno, O Levit, and M Auerbach), Yale University School of Medicine, New Haven, Conn; Department of Pediatrics (T Sawyer), University of Washington School of Medicine, Seattle; Departments of Anesthesiology and Critical Care Medicine (A Nishisaki); Departments of Pediatrics (A Nishisaki, A Ades, and H French), University of Pennsylvania Perelman School of Medicine, Philadelphia; Department of Pediatrics (K Glass), Penn State College of Medicine, Hershey, Pa; Department of Pediatrics (R Dadiz), University of Rochester School of Medicine and Dentistry, Rochester, NY; Department of Pediatrics (S Gangadharan), Cohen Children's Medical Center-Northwell Health/Hofstra University, New Hyde Park, NY; Division of Emergency Medicine (D Scherzer), Nationwide Children's Hospital, Columbus, Ohio; and Department of Pediatrics (A Moussa), CHU Sainte-Justine, University of Montreal, Canada
[1]A complete list of nonauthor contributors appears in Supplementary Appendix B.
The authors have no conflicts of interest to disclose.
Address correspondence to Lindsay Johnston, MD, MEd, Department of Pediatrics, Yale University School of Medicine, 333 Cedar St, PO Box 208064, New Haven, CT 06520-8064 (e-mail: Lindsay.johnston@yale.edu).
Received for publication April 13, 2018; accepted July 23, 2018.

Abstract

Background: Neonatal tracheal intubation (NTI) is an important clinical skill. Suboptimal performance is associated with patient harm. Simulation training can improve NTI performance. Improving performance requires an objective assessment of competency. Competency assessment tools need strong evidence of validity. We hypothesized that an NTI competency assessment tool with multisource validity evidence could be developed and be used for formative and summative assessment during simulation-based training.

Methods: An NTI assessment tool was developed based on a literature review. The tool was refined through 2 rounds of a modified Delphi process involving 12 subject-matter experts. The final tool included a 22-item checklist, a global skills assessment, and an entrustable professional activity (EPA) level. The validity of the checklist was assessed by having 4 blinded reviewers score 23 videos of health care providers intubating a neonatal simulator.

Results: The checklist items had good internal consistency (overall $\alpha = 0.79$). Checklist scores were greater for providers at greater training levels and with more NTI experience. Checklist scores correlated with global skills assessment ($\rho = 0.85$; $P < .05$), EPA levels ($\rho = 0.87$; $P < .05$), percent glottic exposure ($r = 0.59$; $P < .05$), and Cormack-Lehane scores ($\rho = 0.95$; $P < .05$). Checklist scores reliably predicted EPA levels.

Conclusions: We developed an NTI competency assessment tool with multisource validity evidence. The tool was able to discriminate NTI performance based on experience. The tool can be used during simulation-based NTI training to provide formative and summative assessment and can aid with entrustment decisions.

Keywords: global skills assessment; entrustable professional activities assessment; neonatal intubation; procedural skills checklist; validity

Academic Pediatrics 2019;19:157−164

What's New

No available assessment tools for neonatal intubation have strong validity evidence. This novel neonatal intubation assessment tool demonstrated multisource validity and can be used for simulation-based assessment and entrustment decisions.

Neonatal tracheal intubation (NTI) performance by pediatric trainees and providers is suboptimal.[1] Reported success rates for pediatric residents range from 20% to 60%.[1–5] Limited months in neonatal intensive care unit (NICU) rotations,[6] increased utilization of nurse practitioners (NPs) and physician assistants (PAs),[7] changes in Neonatal Resuscitation Program guidelines,[8] and increased use of noninvasive ventilation[9] may decrease procedural opportunities and thus contribute to decreased NTI competency. Competition for procedures among trainees has limited other options, such as anesthesia rotations, to supplement clinical NTI experience. Suboptimal NTI performance can have serious consequences. Neonates undergoing multiple NTI attempts are at risk for adverse events, intracranial hemorrhage, and impaired neurodevelopmental outcomes.[10–12] Given limitations in NTI experience, it is unlikely that the majority of pediatric residents develop competency to safely perform NTI during residency.[13] However, this skill remains important for pediatricians, particularly those practicing in rural settings or entering fellowship training in a critical care subspecialty.[14–16] This has led pediatric educators to ask how NTI competency can be improved in the current training environment.[13]

Previous studies have shown improved competency in clinical procedures, such as central-line placement, performance of advanced cardiac life support, and infant lumbar puncture, after simulation-based training using deliberate practice (DP) and mastery learning (ML).[17–19] With DP and ML, learners receive immediate formative feedback directed at improving performance to a predefined standard or minimum passing score on an assessment tool.[20] Learners continue training until they reach a level of "mastery" on the assessment tool.[20] An evidence-based pedagogical framework for procedural skills training—Learn-See-Practice-Prove-Do-Maintain—includes DP and ML in simulation to ensure a basic level of competency before trainees are entrusted to perform procedures clinically.[21] Literature reviews and meta-analyses have shown that simulation-based medical education with DP and ML yields improved patient care/outcomes and cost savings compared with traditional clinical education.[22] Given the evidence supporting DP with ML, it makes sense to apply this strategy to NTI training. A key step in conducting such training is developing a competency assessment tool with multisource validity evidence.

Validity is a complex topic. Many authors have contributed to the theories and frameworks that form the contemporary approach to validity.[23–26] Messick's framework includes multiple sources of evidence to support construct validity, where "intangible attributes," or constructs, are linked to "observable attributes."[26] Five sources of validity described by Messick are: 1) content, 2) response process, 3) internal structure, 4) relations to other variables, and 5) consequences. Ideally, an NTI competency assessment tool should have evidence of validity in each area.

In this report, we describe the development and testing of an NTI competency assessment tool with multisource validity evidence. We first describe how the checklist was created, then provide multiple sources of validity evidence for its use. We hypothesize that the competency assessment tool could be used for formative and summative assessment during simulation training in NTI and could aid in entrustment decisions.

METHODS

STUDY DESIGN

The NTI competency assessment tool was developed through the International Network for Simulation-based Pediatric Innovation, Research, and Education Research Network following established guidelines,[27] which have been used in a previous study.[28] These guidelines are based on evidence and theory for simulation-based education and assessment[21] and provide an overview of necessary steps for assessment tool development. Sources of validity evidence for the assessment tool were evaluated using the framework described by Messick.[26] The study was approved by the Human Subjects Committee at Yale University.

CHECKLIST DEVELOPMENT

INITIAL CHECKLIST DRAFT

The initial assessment tool draft was developed based on a literature review. Three investigators (L.J., T.S., A.N.) searched for "intubation," "pediatric intubation," and "neonatal intubation" and reviewed textbooks on pediatric/neonatal procedures and resuscitation to identify published NTI competency assessment tools with validity evidence. One 13-item dichotomous procedural checklist with a nonanchored global skills assessment (GSA) for neonatal nasotracheal intubation was identified and used as a reference.[4] The newly drafted NTI competency assessment tool used a previously published template[27] and included 21 dichotomously rated (eg, "done correctly" vs "done incorrectly, or not done") performance elements, an anchored 5-point GSA (describing performance over a spectrum of novice to expert), and a 5-point entrustable professional activity (EPA) level. The 5 GSA ratings were novice, advanced beginner, competent, proficient, and expert.[27] The 5 EPA levels were as follows: 1) "ready to observe the activity only, not ready to perform procedure on patient even with direct supervision"; 2) "ready to perform procedure with direct supervision present in the room"; 3) "ready to perform the procedure with supervision available within minutes"; 4) "ready to perform procedure without direct supervision (ie, under clinical oversight)"; and 5) "ready to provide supervision to juniors learning the procedure."[27,29]

MODIFIED DELPHI PROCESS

A modified Delphi process was used to optimize the content of the draft.[30,31] Twelve subject-matter experts (SMEs) in NTI with training in neonatal–perinatal medicine, critical care medicine, emergency medicine, otolaryngology, and anesthesiology participated in the Delphi process. These specialties were chosen because they include the intubation of neonates. The SME group was international, with members from United States, Canada, and Japan. All SMEs were board-certified, fully licensed, and practicing independently for a minimum of 5 years.

Delphi rounds were conducted via e-mail using a standardized template. SMEs were instructed to review each checklist item on the assessment tool and rate it on a 7-point Likert scale, with scores of 1, 2, and 3 indicating "not important"; 4, 5, and 6 indicating "somewhat important"; and 7 indicating "mandatory."[27] SMEs also were asked to provide comments regarding checklist item alteration or need for additional items. Results were aggregated and analyzed. Suggestions for item alteration/addition were collated by 1 investigator (L.J.). The initial checklist was revised to reflect the comments and suggestions provided, and the revised draft checklist was then

Table 1. Demographics of Study Participants (n = 23)

	Student (n = 1)	Residents (n = 6)	Fellows (n = 4)	PA/NP (n = 7)	Attending Neonatologists (n = 5)	*P* Value
Years of experience, mean (SD)	1.5 (0)	1.8 (0.8)	2.5 (0.6)	11.5 (8.5)	14 (7)	.05
Previous neonatal intubations, median (IQR)	0 (0)	0 (0)	32.5 (27.5–40)	52.5 (8–90)	150 (97.5–225)	.002

PA indicates physician assistant; NP, nurse practitioner; SD standard deviation; and IQR, interquartile range.

redistributed to the SMEs for the second round of Delphi review. After 2 rounds, consensus was reached and no additional comments were provided. All 12 SMEs participated in both Delphi rounds.

CHECKLIST VALIDATION

VALIDITY AND RELIABILITY TESTING

The final version of the competency assessment tool was evaluated for validity and reliability. For this assessment, 23 subjects of variant training levels and with different NTI experience were recruited from Yale University. Demographic data on the subjects are presented in Table 1. Each subject performed a single NTI attempt on a neonatal simulator (SimNewB; Laerdal Medical, Wappinger Falls, NY). The SimNewB manikin was selected based on its realistic airway. Subjects used a video laryngoscope (VL) to perform the NTI (Storz C-MAC; Karl Storz Inc, Germany) but intubated using a direct laryngoscopic view (ie, not using the video screen). For the simulated NTI, the SimNewB manikin was positioned on a radiant warmer (with its lower extremities facing the head of the warmer), and appropriate NTI equipment was available. Subjects were instructed to perform an NTI as part of a brief scenario involving a 3-kg infant undergoing nonemergent NTI for retinopathy of prematurity surgery. One additional provider played the role of a bedside nurse. The “nurse” could assist the laryngoscopist if requested but could not offer suggestions or perform the NTI. Each subject performed a single NTI attempt and was encouraged to verbalize actions as they were performed. Subjects were provided a copy of the assessment tool that would be used to rate performance and reviewed it before starting the procedure.

The NTI performance of each subject was video-recorded using both a wide-angle video camera and VL. The video camera (Vixia HSM31; Canon USA, Melville, NY) was set up to record across the warmer from the manikin’s right side at a distance of 3 feet. The VL recorded a view of the airway during the NTI attempt. Representative video images of the simulated NTIs are provided in the Figure.

Four blinded reviewers used the neonatal NTI competency assessment tool to score each of the 23 video-recorded NTIs by assigning a checklist score, a GSA, and an EPA level. Reviewers also scored each subject’s airway visualization from VL images using the Cormack-Lehane (C-L) scale and percentage of glottic opening (POGO).[32,33] The C-L scale was graded on a 4-point scale, with Grade 1 indicating that most of the glottis was visible, Grade 2 indicating that only the posterior extremity of the glottis was visible, Grade 3 indicating that only the epiglottis was seen and no part of the glottis was seen, and Grade 4 indicating that not even the epiglottis was seen.[32] The POGO was scored on continuous range from 0% to 100% of glottic exposure.[33] The raters were board-certified neonatologists who practice in level IV academic centers. Raters were not familiar with study subjects and were blinded to subjects’ level of training and experience. NTI videos were sent to the reviewers via a secure server.

RATER TRAINING AND CALIBRATION

The 4 blinded reviewers participated in a 2-hour training session that introduced them to the study’s purpose, provided information on the development of the competency assessment tool, and instructed them on how score performance using the tool. To ensure consistency, raters were provided with an assessment tool that included operational definitions for each checklist item. After the initial training, each rater scored 3 NTI calibration videos that had previously been independently rated by 3 investigators (L.J., T.S., A.N.). Each reviewer subsequently

Figure. *(A) Video image of intubation used by reviewers to complete checklist scores, GSA, and EPA. (B) Video image of airway used by reviewers to assign POGO and C-L scores. GSA indicates global skill assessment; EPA, entrustable professional activities; POGO, percent of glottic opening; and C-L, Cormack–Lehane.*

discussed their scores with a study investigator (L.J.). Rater training was stopped when the interrater reliability (IRR) among 4 reviewers reached an a priori−defined threshold (intraclass correlation coefficient [ICC]: ICC >0.8).[25]

STATISTICAL ANALYSIS

The Cronbach alpha (α) was used during each round of the Delphi process to determine the internal consistency of checklist items by the 12 SMEs. An a priori α of 0.7 to 0.9 was used to define SME consensus.[30,31] Differences in experience between study subjects were calculated using 1-way analysis of variance (ANOVA), or Kruskal-Wallis 1-way ANOVA on ranks with all pairwise multiple comparison procedures (Dunn method). A 2-way, mixed-effects, consistency, average-measures ICC among the 4 blinded reviewers and 3 investigators was used during rater training to determine IRR. Competency assessment tool scores from the blinded reviewers were evaluated in several ways. The IRR of the reviewers was calculated based on the ICC of checklist scores. Correlation between checklist scores, GSA, EPA level, POGO, and C-L score were calculated using the Pearson correlation coefficient (r) for continuous variables or Spearman rho (ρ) coefficient for noncontinuous (ordinal) variables. Statistical significance of correlations was calculated using 1-way ANOVA for normally distributed variables or a Mann-Whitney U test for nonparametric variables. An acceptable ICC was considered >0.7.[34] Statistical significance was set at $P < .05$. Statistical analyses were completed with SPSS 22.0 software (IBM Corp, Armonk, NY).

RESULTS

CHECKLIST DEVELOPMENT

The initial checklist contained 21 items. In the first round of Delphi, no item scored <4 on average; therefore, all items were maintained for the second round. The Cronbach α of the first checklist draft was 0.73 (95% confidence interval, 0.44−0.91). One item, "employs appropriate technique to open mouth prior to inserting blade," was added, and several other items were modified. During the second round, no item scored <4 on average, and no additional items were added. The Cronbach α for the second (final) checklist draft increased to 0.79 (95% confidence interval, 0.48−0.96). The final checklist had 22 items (including 9 general procedural items and 13 discrete NTI performance elements) (Supplementary Appendix A).

CHECKLIST VALIDATION

A summary of the sources of validity used in the study is provided in Table 2. The ICC during rater training ranged from 0.8 to 0.92. The overall ICC of the 4 blinded reviewers on the 23 neonatal NTI videos was 0.64, indicating moderate IRR.

Checklist scores, GSA ratings, and EPA levels correlated with subject training level (Table 3). NICU-specialized providers (ie, fellows, attendings, NP/PA) outperformed general providers (ie, students and residents) ($P < .001$). Nine subjects (39%) were not able to intubate the manikin successfully. Attending neonatologists had the greatest GSA and EPA ratings, but their checklist scores were lower than those of neonatal fellows and NP/PAs. Checklist scores, GSA ratings, and EPA levels correlated with subject NTI experience (Table 4). Discriminatory ability, however, appeared to be lost with NTI experience of >30.

Correlations between checklist scores, GSA, EPA level, POGO, and C-L scales are provided in Table 5. The greatest levels of correlation were seen with GSA rating ($\rho = 0.85$), EPA level ($\rho = 0.87$), and C-L scores ($\rho = 0.95$). All correlation coefficients were statistically significant with $P < .05$. Checklist scores and EPA level were positively correlated at a level of statistical significance ($P < .001$, $\rho = 0.770$). A checklist score of 6 equaled an EPA level of 1, a score of 10 equaled an EPA level of 2, a score of 12 equaled an EPA level of 3, a score of 13 equaled an EPA level of 4, and a score of 14 equaled an EPA level of 5.

DISCUSSION

We developed an NTI competency assessment tool that included a procedural checklist, a GSA, and an EPA level. An international group of multidisciplinary experts in NTI participated in the Delphi process to develop the tool. The assessment tool demonstrated acceptable IRR and multiple sources of validity evidence, including content, response process, internal structure, relation to other variables, and consequences. We believe this tool has sufficient evidence of validity for use in both formative and summative assessment during simulation-based NTI training.

The mean scores of the itemized checklist successfully discriminated subjects based on their NTI skill, as defined by a greater level of training and a greater number of successful clinical NTIs. Interestingly, the mean scores of the attending neonatologists on the itemized checklist were lower than the scores of the fellows and the NP/PAs. This discrepancy may be due to the fact that more experienced providers often score lower on itemized checklists as the result of their lesser focus on individualized procedural steps.[35] An expert's high-level performance may be the result of the ability to assess a situation and to skip nonessential steps based on pre-existing knowledge and experience. This also highlights the value of global assessment scales, in which experts are more likely to score greater even when they omit several steps on a checklist. Our study finding was consistent with this phenomenon; that is, the attending neonatologists had the greatest scores on the GSA and EPA assessments. Another contributing factor might be that fellows and NP/PAs intubate more frequently in

Table 2. Sources of Validity Evidence for the Neonatal Tracheal Intubation Checklist

Sources of Evidence[31]	Examples	Sources Used in This Study	Results
Content	Representativeness of assessment items to domain • Literature review • Expert opinion Quality of items on tool Qualifications of authors of tool	Literature review performed International, multiprofessional subject matter experts participated in modified Delphi	Initial draft was 21 items No item scored <4 on average on either of 2 Delphi rounds One item added and another item broken up into subitems Final checklist had 22 items, with GSA and EPA ratings
Response process	Rater training and calibration Data storage and security Quality control of scoring/reporting	Rater training and calibration Standardized, simulated setting and videos of neonatal intubation Data stored on secure computer	IRR during rater training was 0.8 to 0.92 ICC among the 4 blinded reviewers ranged from 0.79 to 0.91
Internal structure	Item analysis data Score scale reliability Generalizability	Internal consistency IRR	Cronbach α for the first checklist draft was 0.73 (95% CI, 0.44–0.91) Cronbach α for the second (final) checklist draft was 0.79 (95% CI, 0.48–0.96) Cronbach α of each procedural checklist item ranged from 0.85 to 0.87 Overall ICC between 4 raters on 23 videos was 0.64
Relations to other variables	Correlation with other relevant variables • Measures of experience (role, years in practice) • Scores on other exams/assessments	Correlation with other relevant variables • Experience • Checklist scores • GSA • EPA • POGO • C-L scores	Checklist scores, GSA ratings, and EPA levels correlated with experience levels Checklist scores correlated with GSA, EPA, POGO, and C-L scores
Consequences	Standard setting procedures Impact of scores on participants/society	Acceptable standard set for ICC EPA levels	An ICC >0.7 was considered acceptable Checklist score <10 suggests not entrusting individual to intubate a neonate, even with direct supervision

GSA indicates global skill assessment; EPA, entrustable professional activities; IRR, interrater reliability; ICC, interclass correlation coefficient; CI, confidence interval; POGO, percentage of glottic opening; and C-L, Cormack–Lehane score.

the academic NICUs than attending neonatologists.[1,5] This may result in fewer recent NTI attempts in the attending neonatologist group, resulting in NTI skill decay. The discrepant results of the different assessment methods highlight the need for multimodal competency assessment techniques, as different methods have strengths and weaknesses.[36]

We noted positive and statistically significant correlations between checklist scores, GSA ratings, EPA levels, POGO, and C-L scores. The greatest correlations were seen between checklists scores, GSA ratings, EPA levels, and CL scores. GSA ratings were also highly correlated with EPA levels. The POGO, however, had only moderately positive correlation with the other metrics. This lower degree of correlation may reflect the different skills evaluated by the different assessment metrics. POGO reflects a specific psychomotor skill involved in direct laryngoscopy, for example, the ability to expose the glottis. The checklist scores, GSA rating, and EPA levels evaluate the entire NTI process.

Entrustment to perform a procedure without direct supervision is a key outcome of graduate medical education.[29] Our data on the relationship between checklist scores and EPA levels may allow the checklist scores to guide entrustment decisions regarding NTI. There was a significant positive association between checklist scores and EPA levels, providing strong consequential validity evidence. When the results of the current study are used, to entrust a trainee to attempt a clinical NTI under direct

Table 3. Subject Training Level and Checklist Scores, GSA Ratings, and EPA Level*

	Student (n = 1)	Residents (n = 6)	Fellows (n = 4)	PA/NP (n = 7)	Attending Neonatologists (n = 5)	*P* Value (ANOVA)	*P* Value (Students/ Residents vs. Fellows/PA/NP/ Attending)
Checklist scores, mean (SD)	2.0 (0)	6.7 (1.6)	12.8 (2.9)	12.5 (1.8)	11.4 (3.0)	<.001	<.001
GSA rating, mean (SD)	1.0 (0)	1.6 (0.5)	3.8 (1.5)	4.0 (0.8)	4.4 (1.3)	<.001	<.001
EPA level, median (IQR)	1.0 (1.0, 1.0)	1.0 (1.0, 1.0)	4.0 (3.0, 4.8)	3.0 (2.0, 4.0)	4.0 (2.3, 5.0)	<.001	<.001

GSA indicates global skill assessment; EPA, entrustable professional activities; PA, physician assistant; NP, nurse practitioner; ANOVA, analysis of variance; SD, standard deviation; and IQR, interquartile range.

*Ratings were provided by 4 independent raters, and the median or mean is reported.

Table 4. Subject Previous Intubation Experience and Checklist, GSA, and EPA Ratings*

Previous Intubation Experience	0 Intubations	1–29 Intubations	30–100 Intubations	≥100 Intubations	*P* Value
Checklist scores, mean (SD)	6.3 (2.3)	10.0 (3.2)	13.8 (1.6)	10.5 (2.9)	<.001
GSA rating, median (IQR)	2.0 (1.0, 2.0)	3.0 (1.9, 4.0)	5.0 (4.0, 5.0)	5.0 (2.0, 5.0)	0.002
EPA level, median (IQR)	1.0 (1.0, 1.3)	3.0 (1.8, 4.0)	5.0 (3.3, 5.0)	4.0 (2.0, 5.0)	0.003

GSA indicates global skill assessment; EPA, entrustable professional activities; SD, standard deviation; and IQR, interquartile ratio.

*Ratings were provided by 4 independent raters, and the median or mean is reported.

Table 5. Correlations Between Checklist Scores, GSA, EPA Level, POGO, and C-L Scores

	Checklist Score	GSA	EPA	POGO	C-L Score
Checklist score	1.00	0.85	0.87	0.59	0.95
GSA	–	1.00	0.94	0.55	0.53
EPA	–	–	1.00	0.46	0.65
POGO	–	–	–	1.00	0.67
C-L score	–	–	–	–	1.00

GSA indicates global skill assessment; EPA, entrustable professional activities; POGO, percentage of glottic opening; and C-L, Cormack–Lehane.

All correlation coefficients were statistically significant, with $P < .05$.

supervision, the subject would need to score at least 10 points on the checklist (equal to an EPA level of 2). Scores <10 would indicate a need for continued training before entrusting the trainee to perform NTI, even with direct supervision. A score of >13 suggests the ability to perform NTI without direct supervision. However, making ultimate decisions regarding entrustment, or trustworthiness in application of knowledge, skills, and attitudes during performance of an EPA, requires the discernment of numerous additional attributes, including conscientiousness, awareness of limitations, and truthfulness, in addition to consideration of clinical intubation psychomotor technique and success rates.[37]

This study has notable limitations. First, the subjects enrolled in this study were from a convenience sample at a single academic medical center. There was a wide range of experience with NTIs among the subjects. Several subjects had no previous experience, several subjects had performed >100 NTIs, but fewer subjects were between these 2 extremes. This may limit the generalizability of the scores across a wider spectrum of provider experience. Second, raters in this study were trained extensively on the use of the assessment tool. Although this degree of training is not feasible for most educators, the tool was developed to focus on objective behavior and should be relatively simple to use after a brief orientation. Next, there is risk of halo effect, where an opinion of an individual's attributes (appearance, mannerisms, experience) may inadvertently influence a rater's assessment.[38] Finally, the data reported here were derived from a simulation-based setting. Providers may perform differently in the simulated environment as compared with a clinical environment. A future clinical study to assess the validity evidence of this NTI competency assessment tool in the neonatal intensive care unit and/or delivery room is warranted.

CONCLUSIONS

We developed an NTI competency assessment tool that includes a checklist, GSA, and EPA level. The tool displayed multiple sources of validity evidence. The NTI competency assessment tool can serve as a valuable tool to evaluate NTI competency during simulation-based training and may be considered when making entrustment decisions. A separate manuscript is planned to further evaluate the relationship between NTI checklist, GSA, and EPA scores. Additional investigations on this topic, including test-retest reliability, generalizability analysis, and correlations of scores on the assessment tool to clinical NTI success, would be useful to further elaborate on the validity evidence presented. Future studies are needed to collect additional validity evidence in the clinical environment.

ACKNOWLEDGMENTS

The authors acknowledge the participants in the Delphi process, Ellen Deutsch, Hiroshi Kurosawa, Anne Brayer, Laurie Steiner, Dawn Sweeney, Maria Esperanza, Lee Smith, Antoine Payot, Geraldine Pettersen, Chantal Hickey, Jeranil Nunez, and Arlene Garingo.

Financial disclosure: The authors are members of the International Network for Simulation-based Pediatric Innovation, Research, and Education Research Network, which receives infrastructure support from the Rbaby Foundation, and funding for biannual meetings from the Society for Simulation in Healthcare and the International Pediatric Simulation Society.

Authorship Statement: L.J. conceptualized and designed the study, coordinated the development of the checklist and the Delphi process, designed the data collection instruments, coordinated and collected data, drafted the initial manuscript and critically reviewed the manuscript for important intellectual content. T.S. conceptualized and designed the study, participated in the development of the checklist, designed the data collection instruments for the Delphi process, carried out the initial analysis, and critically reviewed the manuscript for important intellectual content. M.A. and A.N. conceptualized and designed the study, and reviewed and revised the manuscript. T.W. conceptualized and designed the study, analyzed the data, and reviewed and revised the manuscript. A.A., H.F., K.G., R.D., C.B., and O.L. participated in the conceptualization and design of the study, underwent rater training, collected data, and reviewed and revised the manuscript. S.G, D.S., and A.M. conceptualized and designed the study, participated in the development of the checklist and conduct of the Delphi process, and reviewed and revised the manuscript. All authors approved the final manuscript as submitted and agree to be accountable for all aspects of the work.

SUPPLEMENTARY DATA

Supplementary data related to this article can be found online at doi:10.1016/j.acap.2018.07.008.

REFERENCES

1. Haubner LY, Barry JS, Johnston LC, et al. Neonatal intubation performance: room for improvement in tertiary neonatal intensive care units. *Resuscitation*. 2013;84:1359–1364.
2. Leone TA, Rich W, Finer NN. Neonatal intubation: success of pediatric trainees. *J Pediatr*. 2005;146:638–641.
3. O'Donnell CPF, Kamlin COF, Davis PG, et al. Endotracheal intubation attempts during neonatal resuscitation: success rates, duration, and adverse effects. *Pediatrics*. 2006;117:e16–e21.
4. Bismilla Z, Finan E, McNamara PJ, et al. Failure of pediatric and neonatal trainees to meet Canadian neonatal resuscitation program standards for neonatal intubation. *J Perinatol*. 2010;30:182–187.
5. Foglia EE, Ades A, Napolitano N, et al. Factors associated with adverse events during tracheal intubation in the NICU. *Neonatology*. 2015;108:23–29.
6. Accreditation Council for Graduate Medical Education. Program requirements for residency education in pediatrics. Available at: www.acgme.org/acWebsite/downloads/RRC_progReq/320pediatrics07012007.pdf. Accessed January 1, 2012.
7. Gozzo YF, Cummings CL, Chapman RL, et al. Who is performing medical procedures in the neonatal intensive care unit? *J Perinatol*. 2011;31:206−211.
8. Weiner G. Textbook of Neonatal Resuscitation. 7th ed. Elk Grove, IL: American Academy of Pediatrics; 2016.
9. DeMauro SB, Millar D, Kirpalani H. Noninvasive respiratory support for neonates. *Curr Opin Pediatr*. 2014;26:157–162.
10. Hatch LD, Grubb PH, Lea AS, et al. Endotracheal intubation in neonates: a prospective study of adverse safety events in 162 infants. *J Pediatr*. 2016;168:62–66.
11. Wallenstein MB, Birnie KL, Arain YH, et al. Failed endotracheal intubation and adverse outcomes among extremely low birth weight infants. *J Perinatol*. 2016;36:112–115.
12. Sauer CW, Kong JY, Vaucher YE, et al. Intubation attempts increase the risk for severe intraventricular hemorrhage in preterm infants: a retrospective cohort study. *J Pediatr*. 2016;177:108–113.
13. Gilhooly J, Redden HR, Leonard DT. Competency in neonatal endotracheal intubation: mission impossible? *Pediatrics*. 2015;135: e1290–e1291.
14. Gaies MG, Landrigan CP, Hafler JP, et al. Assessing procedural skills training in pediatric residency programs. *Pediatrics*. 2007;120:715–722.
15. Wood AM, Jones J, Douglas M, et al. Neonatal resuscitation skills among pediatricians and family physicians: is residency training preparing for postresidency practice? *J Grad Med Educ*. 2011;3:475–480.
16. Jones M, McGuinness GA. Residency Review and Redesign in Pediatrics (R3P) Committee. The future of pediatric residency education: prescription for more flexibility. *J Pediatr*. 2009;154:157–158.e3.
17. Barsuk J. Simulation-based mastery learning reduces complications during central venous catheter insertion in a medical intensive care unit. *Crit Care Med*. 2009;37:2697–2701.
18. Wayne DB, Butter J, Siddall VJ, et al. Mastery learning of advanced cardiac life support skills by internal medicine residents using simulation technology and deliberate practice. *J Gen Intern Med*. 2006;21:251–256.
19. Kessler DO, Auerbach M, Pusic M, et al. A randomized trial of simulation-based deliberate practice for infant lumbar puncture skills. *Simul Healthc*. 2011;6:197–203.
20. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med*. 2004;79:S70–S81.
21. Sawyer T, White M, Zaveri P, et al. Learn, see, practice, prove, do, maintain: an evidence-based pedagogical framework for procedural skill training in medicine. *Acad Med*. 2015;90:1025–1033.
22. McGaghie WC, Issenberg SB, Cohen ER, et al. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Acad Med*. 2011;86:706–711.
23. Cook DA, Brydges R, Ginsburg S, et al. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ*. 2015;49:560–575.
24. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull*. 1955;52:281–302.
25. Downing SM, Yudkowsky R. Assessment in Health Professions Education. New York, NY: Routledge; 2009.
26. Messick S. Validity. In: Linn RL, ed. Educational Measurement, 3rd ed, New York, NY: American Council on Education and Macmillan; 1989:13–103.
27. Sawyer TL, White ML. INSPIRE procedural checklists development and validation instructions. Available at: http://www.inspiresim.com/?ddownload=178. Accessed March 3, 2017.
28. Gerard JM, Kessler DO, Braun C, et al. Validation of global rating scale and checklist instruments for the infant lumbar puncture procedure. *Simul Healthc*. 2013;8:148–154.
29. Ten Cate O, Hart D, Ankel F, et al. Entrustment decision making in clinical training. *Acad Med*. 2016;91:191–198.
30. Ogden SR, Culp Jr WC, Villamaria FJ, et al. Developing a checklist: consensus via a modified Delphi technique. *J Cardiothorac Vasc Anesth*. 2016;30:855–858.

31. Morgan PJ, Lam-McCulloch J, Herold-McIlroy J, et al. Simulation performance checklist generation using the Delphi technique. *Can J Anaesth*. 2007;54:992–997.
32. Cormack RS, Lehane J. Difficult tracheal intubation in obstetrics. *Anaesthesia*. 1984;39:1105–1111.
33. Levitan RM, Ochroch EA, Kush S, et al. Assessment of airway visualization: validation of the percentage of glottic opening (POGO) scale. *Acad Emerg Med*. 1998;5:919–923.
34. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;12:155–163.
35. Hodges B, Rehehr G, McNaughton N, et al. OSCE checklists do not capture increasing levels of expertise. *Acad Med*. 1999;74:1129–1134.
36. Ilgen JS1, Ma IW, Hatala R, et al. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ*. 2015;49:161–173.
37. Kennedy TJT, Regehr G, Baker GR, et al. Point-of-care assessment of medical trainee competence for independent clinical work. *Acad Med*. 2008;83:S89–S92.
38. Nisbett RE, Wilson TD. The halo effect: evidence for unconscious alteration of judgments. *J Pers Soc Psychol*. 1977;35:250–256.