Academic Emergency Medicine
Official Journal of the Society for Academic Emergency Medicine

BRIEF REPORT

Assessing the Validity Evidence of an Objective Structured Assessment Tool of Technical Skills for Neonatal Lumbar Punctures

Maya S. Iyer, MD, Sally A. Santen, MD, PhD, Michele Nypaver, MD, Kavita Warrier, MD, Stuart Bradin, DO, Rachel Chapman, MD, Jennifer McAllister, MD, Jennifer Vredeveld, MD, and Joseph B. House, MD

Abstract

Background: The lumbar puncture (LP) is a procedural competency deemed necessary by the Accreditation Council for Graduate Medical Education and the Emergency Medicine and Pediatric Residency Review Committees. The emergency department (ED) is a primary site for residents to be evaluated performing neonatal LPs. Current evaluation methods lack validity evidence as assessment tools.

Objectives: This was a pilot study to develop an objective structured assessment of technical skills for neonatal LP (OSATS-LP) and to document validity evidence for the instrument in regard to five sources of test validity: content, response process, relation to other variables, inter-rater reliability, and consequences of testing.

Methods: Pediatric residents were videotaped in the fall of 2011 for comparison of faculty evaluation of resident performance during a neonatal LP using a video-delayed format. Residents completed a demographic experience survey evaluating relations to other variables. Content and response process validity was obtained through expert panel meetings and resulted in the following seven domains of performance for the OSATS-LP: preparation, positioning, analgesia, needle insertion, cerebrospinal fluid (CSF) collection, management of laboratory studies, and sterility. t-tests assessed significance between level of training, previous intensive care unit experience, and residents' self-assessed confidence in comparison with their total performance score. The inter-rater agreement of the OSATS-LP was obtained using the Fleiss' kappa for each domain.

Results: Sixteen pediatric residents completed the simulation with six raters evaluating each resident (96 ratings). The domains of sterility and CSF collection had moderate statistical reliability ($\kappa = 0.41$ and 0.51, respectively). The domains of preparation, analgesia, and management of laboratories had substantial reliability ($\kappa = 0.60$, 0.62, and 0.62, respectively). The domains of positioning and needle insertion were less reliable ($\kappa = 0.16$ and 0.16, respectively). Individuals who had completed one or more rotations in the neonatal intensive care unit (NICU) had a higher total score (12.5 vs. 16.9; $p < 0.01$). The residents' own perception of ability to perform an LP unsupervised did not result in a higher total score.

Conclusions: The OSATS-LP has reasonable evidence in four of the five sources for test validity. This study serves as a launching point for using this tool in clinical environments such as the ED and, therefore, has the potential to provide real-time formative and summative feedback to improve resident skills and ultimately lead to improvements in patient care.

ACADEMIC EMERGENCY MEDICINE 2013; 20:321–324

From the Department of Pediatrics, Division of Emergency Medicine, Children's Hospital of Pittsburgh (MSI), Pittsburgh, PA; and the Department of Medical Education (SS), the Department of Emergency Medicine (SS, MN, SB, JBH), the Department of Pediatrics and Communicable Diseases (MSI, KW, JV), the Division of Neonatal-Perinatal Medicine (RC, JM), and the Department of Internal Medicine (JV), University of Michigan, Ann Arbor, MI.

Received June 27, 2012; revision received September 9, 2012; accepted October 2, 2012.

The authors have no relevant financial information or potential conflicts of interest to disclose.

Supervising Editor: Daniel L. Theodoro, MD, MSCI.

Address for correspondence and reprints: Maya Subbarao Iyer; e-mail: maya.iyer@chp.edu.

The lumbar puncture (LP) is an important competency noted as part of the 2009 Model of the Clinical Practice of Emergency Medicine and deemed necessary by the Emergency Medicine and Pediatric Residency Review Committees.[1,2] Residents are reported to "fail" at the pediatric LP up to 26% of the time.[3] An assessment tool for LPs with validity evidence would allow an educator to provide formative and summative assessment to the graduate medical learner. While not widely used in pediatrics, the Objective Structured Assessment of Technical Skills (OSATS) is an evaluation instrument commonly used in the surgical field that outlines the key portions of the procedure and allows for global assessment of the skill.[4] The

doi: 10.1111/acem.12093
ISSN 1069-6563
PII ISSN 1069-6563583

OSATS has been adapted for other emergency medicine (EM) procedural skills including pediatric rapid sequence intubation and has documented validity evidence.[5] The goal of this pilot study was to develop and document validity evidence for an OSATS instrument for neonatal LP (OSATS-LP).

METHODS

Study Design

We developed and prospectively tested the OSATS-LP rating instrument for assessing the performance of the LP by pediatric residents in a video-delayed format.[5] In the development and testing of the instrument, we collected four of the five domains of validity evidence: content, response process, relation to other variables, and internal structure.[6] This educational research study was reviewed and exempted by the institutional review board.

Study Setting and Population

This study was conducted at the University of Michigan Health System (UMHS). The pediatric residency program is a 3-year program at UMHS and was comprised of a total of 62 categorical residents in the time period of 2011 and 2012. Five residents from the PGY1 class, five residents from the PGY2 class, and six residents from the PGY3 class were asked via e-mail to volunteer in this study. These participants were randomly selected by using the e-mail list for each class and selecting every fourth name until the approximately five subjects per class were recruited. All those who were selected agreed to participate.

Study Protocol

Tool Development. To establish content validity, an expert panel (one neonatologist, one neonatal intensive care unit [NICU] hospitalist, two pediatric hospitalists, and two pediatric EM physicians), using EM texts and checklists from previous studies, convened to determine the critical domains for performing a neonatal LP.[7] Using a modified Delphi method, the panel determined seven critical domains: preparation, positioning, analgesia, needle insertion, cerebrospinal fluid (CSF) collection, diagnostic management of laboratory studies, and creating and maintaining a sterile field. A two-point anchor for each domain was thought to insufficiently differentiate degrees of performance. Thus, a three-point Likert-scale with specific anchors was provided for each score.

The response process validity was achieved by careful design and revision of the instrument alongside rater training. The expert panel also served as raters. These individuals were trained by first rating and discussing two videos. For domains in which there was not 100% agreement, the raters provided reasoning behind their scoring. Modifications were made to the OSATS-LP tool, including clarification of instructions, which modified raters' future ratings. This process was repeated with four additional videos. This led to the final OSATS-LP as shown in Figure 1. The final instrument included seven domains, a global overall rating, and a total score (sum of all domains).

Data Collection. Sixteen pediatric resident volunteers were prospectively enrolled. Each resident completed a demographic and experience survey, including months in the NICU and pediatric ICU (PICU), and range of previous LPs, factors hypothesized for improved performance. The range of LPs was reduced to having performed zero to two LPs or more than two LPs. This provided the basis for the third source of test validity, relations to other variables.

Each resident was videotaped performing an LP on a simulation model (Laerdal Baby Stap, Laerdal Medical, Wappingers Falls, NY). Residents were instructed that this pilot study was to validate an assessment tool for neonatal LPs; however, they were not given the OSATS-LP and did not receive any formalized training in this simulation. The simulation began with the reading of a clinical scenario describing a febrile infant requiring an LP to evaluate for possible meningitis. Residents were instructed to verbalize what equipment was needed, their actions during the procedure, and laboratory studies desired. The residents were permitted to ask questions for clarification prior to the start of the simulation only. The six raters reviewed all of the videos (96 ratings). The raters were able to re-watch the videos as necessary for accurate scoring.

Data Analysis

The inter-rater reliability of the OSATS-LP was determined by calculating the Fleiss' kappa.[8] Interpretation of kappa values was as follows: <0, poor; 0 to 0.2, slight; 0.21 to 0.4, fair; 0.41 to 0.6, moderate; 0.61 to 0.80, substantial; and 0.81 to 1, almost perfect. These measurements provided the fourth source of validity, internal structure validity evidence, or the inter-rater reliability. t-tests were used to assess if there was significance between level of training, previous NICU and PICU experience, previous performance on LPs, and resident self-assessed confidence in comparison with their total performance scores. Statistical significance was set at alpha = 0.05.

RESULTS

The ratings using the OSATS-LP by year of training are found in Table 1. There was an increase in the percentage of observations deemed to be competent (score of 2 or 3) in the domains of preparation, positioning, needle insertion, CSF collection, and sterile field, with increasing PGY year.

The domains of preparation, sterile field, and CSF collection had kappa values of 0.60, 0.41, and 0.51, respectively. The domains of analgesia and management of laboratories had kappa values of 0.62 and 0.62, respectively. Finally, the domains of positioning and needle insertion had kappa values of 0.16 and 0.16, respectively.

There was no difference in mean scores between residents who had performed more than two LPs and those who had performed two or fewer (17.0 vs. 14.5, $p = 0.14$). Individuals who had completed NICU rotations had higher mean total score in the OSATS-LP (12.5 vs. 16.9; $p < 0.01$). However, individuals who had completed rotations in the PICU, or those who had previously performed simulated LPs, were not found to

Please circle the number corresponding to the candidate's performance in each category.

Preparation LP Kit Sterile Gloves Support Staff BONUS: Extra Needles	**1** Does not know necessary supplies	**2** Knows all of the supplies needed and identifies a "holder"	**3** Knows all of the supplies needed, identifies support staff, and has extra supplies
Positioning of the Patient	**1** Cannot position patient or find landmarks	**2** Positions patient, identifies anatomical landmarks with readjustments	**3** Demonstrates positioning/landmarks, asks for repositioning when needed, anticipates/monitors status of pt
Analgesia	**1** Does not use any analgesia	**2** Has a plan for analgesia	**3** Has alternative plan for administering analgesia if original method is not effective
Needle Insertion Technique	**1** No attention to bevel orientation or needle angle/aim	**2** Bevel angled parallel to dural fibers, aimed toward umbilicus but with significant subcutaneous movement or more than one insertion attempt	**3** Needle bevel parallel to dural fibers, aimed toward umbilicus with minor readjustments if needed, but all in one attempt
CSF Fluid Return/Collection	**1** Does not access CSF fluid or attempts to readjust without withdrawing needle first	**2** Accesses CSF fluid but does not replace stylet when removing and reinserting needle.	**3** Accesses CSF fluid, replaces stylet when removing and reinserting needle
Diagnostic Purpose/Lab Management of CSF	**1** Does not obtain a CSF culture	**2** Obtains at least a CSF culture plus additional labs(i.e., cell count/differential, gram stain, protein, glucose)	**3** Obtains all pertinent labs (culture, cell count/differential, gram stain, protein, glucose)
Creates and Maintains Sterile Field	**1** Does not create or maintain sterile field, does not recognize when contamination has occurred.	**2** Creates but cannot maintain sterility, knows when contamination has occurred, takes steps to correct	**3** Creates sterile field and maintains sterility throughout procedure

TOTAL SCORE OBTAINED:

Would you allow this student to perform an LP unsupervised and independently on your next patient?

Yes No

OVERALL PERFORMANCE	**1** Not Competent	**2** Competent Technically	**3** Clearly Technically Superior

Figure 1. Objective Structured Assessment of Technical Skills (OSATS) Global Rating Scale of Lumbar Puncture in the neonate. CSF = cerebral spinal fluid; LP = lumbar puncture; pt = patient.

have higher scores (15.8 vs. 17.3, p = 0.16) and (16.3 vs. 16.5, p = 0.82), respectively. A resident's own perception of ability to perform an LP unsupervised was not associated with a higher total score.

DISCUSSION

As discussed by Cook and Beckman,[9] "validity is not a property of the instrument, but of the instrument's scores and their interpretations." Interpreting the kappa values above, the domains of preparation, analgesia, and management of laboratory studies showed substantial reliability. The domains of CSF collection and sterility showed moderate reliability. Finally, the domains of positioning and needle insertion were slightly reliable. For these domains, the raters commented that it was difficult on the video to view the exact positioning and needle insertion. If the resident verbalized positioning and monitoring of the patient, he or she was given credit. However, if the resident did not verbalize this, the raters were limited to the visual performance on the videotape. Nevertheless, five of the seven domains showed greater than moderate reliability. Previous studies have considered that the inter-rater reliability for the tool itself is "good" based on such kappa values.[4]

LIMITATIONS

This study evaluated only pediatric residents and, therefore, evaluation in EM residents is warranted. This study was also limited by its small sample size. Moreover, this tool was evaluated in a simulated setting and therefore, applicability to clinical environments needs further investigation.

This pilot study also documented validity evidence for the OSATS-LP in four of the five sources of test validity. However, there were limitations in each of these sources. Content validity often describes the meticulous process outlining the steps needed to ensure appropriate representation of the construct. In this study, there was construct underrepresentation given that that the OSATS-LP did not assess communication competency, such as obtaining informed consent and other features that are important in a clinical environment.[10] Furthermore, there was a potential for construct irrelevant variance if the raters inadvertently used personal constructs

Table 1
Prior Experience in Procedure and Resident Assessment Using the OSATS-LP

Variable	PGY1	PGY2	PGY3
Range of LPs done	1–2	6–10	>10
NICU	3/5 (60)	5/5 (100)	6/6 (100)
PICU	0/5 (0)	0/5 (0)	6/6 (100)
Procedural confidence	0/5 (0)	4/5 (80)	6/6 (100)
Preparation*	23/30 (77)	24/30 (80)	32/36 (89)
Positioning*	22/30 (73)	23/30 (77)	28/36 (78)
Analgesia*	30/30 (100)	12/30 (40)	17/36 (47)
Needle insertion*	28/30 (93)	28/30 (93)	33/36 (92)
CSF collection*	24/30 (80)	26/30 (87)	34/36 (94)
Diagnostic management*	30/30 (100)	30/30 (100)	30/30 (100)
Sterile field*	15/30 (50)	19/30 (63)	35/36 (97)
Total score, mean (±SD)	15.5 (±3.1)	16.2 (±1.4)	17.3 (±1.1)

*The numerator is the number of observations scoring either a 2 or 3, and the denominator is the total number of observations for each postgraduate year (PGY) class.
CSF = cerebrospinal fluid; LP = lumbar puncture; NICU = neonatal intensive care unit; OSATS = Objective Structured Assessment of Technical Skills; PICU = pediatric intensive care unit.

instead of the provided OSAT-LP to evaluate residents.[10]

In regard to the response process, there was a potential for "halo bias" if the raters recognized the residents. The voices of the residents were not altered, but their faces were not shown.

When evaluating external variables affecting performance, residents who had completed rotations in the NICU, as expected, overall had higher scores. This was not seen for residents who had also completed PICU rotations, likely due to the small sample size. Furthermore, this study did not evaluate how the OSATS-LP compares with other assessment tools. However, currently, there are not other validated assessment tools for neonatal LPs.

Finally, this pilot study did not evaluate the fifth source of validity: "the impact on examinees from the assessment scores, decisions, and outcomes and the impact of assessments on teaching and learning," as described by Downing and Haldyna.[10] This tool has the potential to provide formative feedback. There was no statistical difference in the total scores between residents who perceived that they could perform an LP unsupervised compared to those who were not confident. The discrepancy in perceived and actual ability is the crux of why assessment needs to be tailored to each learner.

CONCLUSIONS

This pilot study demonstrates that the Objective Structured Assessment of Technical Skills for neonatal lumbar punctures demonstrates reasonable validity in a simulated setting in regard to content, response process, inter-rater reliability, and relationship to external variables. Although two domains of performance exhibited less reliability, and despite not evaluating consequences of testing, this study and the Objective Structured Assessment of Technical Skills for neonatal lumbar puncture serve as launching points for the development of a neonatal lumbar puncture assessment tool.

The authors thank Larry Gruppen, PhD, for providing assistance calculating kappa statistic. The authors would also like to thank the University of Michigan Clinical Simulation Center for providing use of the simulator for this study.

References

1. Core Content Task Force. The Model of the Clinical Practice of Emergency Medicine. Acad Emerg Med. 2001;8:660–81.
2. Accreditation Council for Graduate Medical Education. ACGME Program Requirements for Graduate Medical Education in Pediatrics. Available at: http://www.acgme.org/acgmeweb/Portals/0/PFAssets/ProgramRequirements/320_pediatrics_07012007.pdf. Accessed Dec 15, 2012.
3. Baxter AL, Fisher RG, Burke BL, Goldblatt SS, Isaacman DJ, Lawson ML. Local anesthetic and stylet styles: factors associated with resident lumbar puncture success. Pediatrics. 2006;117:876–81.
4. Ishman SL, Benke JR, Johnson KE, et al. Blinded evaluation of interrater teliability of an operative competency assessment tool for direct laryngoscopy and rigid bronchoscopy. Arch Otolaryngol Head Neck Surg. 2012;17:1–7.
5. House JB, Dooley-Hash S, Kowalenko T, et al. Prospective comparison of live evaluation and video review in the evaluation of operator performance in a pediatric emergency airway simulation. J Grad Med Educ. 2012;4:312–6.
6. Andreatta PB, Gruppen LD. Conceptualizing and classifying validity evidence for simulation. Med Educ. 2009;43:1028–35.
7. Carlson DW, Digulio, GA, Gewitz, MH, et al. Pediatric lumbar puncture. Procedures. In: Fleischer G, Ludwig S, Henretig FM (Eds). Textbook of Pediatric Emergency Medicine, 4th ed. Philadelphia, PA: Lippincott Williams & Wilkins, 1999, pp 1812–3.
8. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.
9. Cook DA, Beckman TJ. Current Concepts in validity and reliability in psychometric instruments: theory and application. Am J Med. 2006;119:116.
10. Downing SM, Haladyna TM. Validity and its threats. In: Downing SM, Yudkowsky R (Eds). Assessment in Health Professions Education. New York, NY: Routledge, 2009.