Assessment of Residents Readiness to Perform Lumbar Puncture: A Validation Study

Mikael Johannes Vuokko Henriksen, MD[1,2], Troels Wienecke, MD, PhD[2,3], Helle Thagesen, MD[3], Rikke Vita Borre Jacobsen, MD, PhD[2,4], Yousif Subhi, MD[2,5], Charlotte Ringsted, MD, PhD[6], and Lars Konge, MD, PhD[1,2]

[1]Copenhagen Academy for Medical Education and Simulation, The Capital Region of Denmark, Rigshospitalet section 5404, Copenhagen, Denmark; [2]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark; [3]Department of Neurology, Zealand University Hospital, Roskilde, Denmark; [4]Center for Head and Orthopedic/UFU 4231 Anesthesiology, Rigshospitalet, Copenhagen, Denmark; [5]Department of Ophthalmology, Zealand University Hospital, Roskilde, Denmark; [6]Centre for Health Science Education, Faculty of Health, Aarhus University, Aarhus, Denmark.

BACKGROUND: Lumbar puncture is a common procedure in many specialties. The procedure serves to diagnose life-threatening conditions, often requiring rapid performance. However, junior doctors possess uncertainties regarding performing the procedure and frequently perform below expectations. Hence, proper training and assessment of performance is crucial before entering clinical practice.

OBJECTIVE: To develop and collect validity evidence for an assessment tool for lumbar puncture performance, including a standard to determine when trainees are ready for clinical practice.

DESIGN: Development of a new tool, based on clinician interviews and a literature review, was followed by an explorative study to gather validity evidence.

PARTICIPANTS AND MAIN MEASURES: We interviewed 12 clinicians from different specialties. The assessment tool was used to assess 11 doctors at the advanced beginners' level and 18 novices performing the procedure in a simulated, ward-like setting with a standardized patient. Procedural performance was assessed by three content experts. We used generalizability theory to explore reliability. The discriminative ability of the tool was explored by comparing performance scores between the two groups. The contrasting groups method was used to set a pass/fail standard and the consequences of this was explored.

KEY RESULTS: The interviews identified that in addition to the technical aspects of the procedure, non-technical elements involving planning and conducting the procedure are important. Cronbach's alpha = 0.92, Generalizability-coefficient was 0.88 and a Decision-study found one rater was sufficient for low-stakes assessments (G-coefficient 0.71). The discriminative ability was confirmed by a significant difference between the mean scores of novices, 40.9 (SD 6.1) and of advanced beginners, 47.8 (SD 4.0), p = 0.004. A standard of 44.0 was established which was consistent with the raters' global judgments of pass/fail.

CONCLUSION: We developed and demonstrated strong validity evidence for the lumbar puncture assessment tool. The tool can be used to assess readiness for practice.

Received August 16, 2016
Revised December 13, 2016
Accepted December 30, 2016
Published online February 6, 2017

KEY WORDS: medical education; medical education-assessment/evaluation; medical education – clinical skills training.
J Gen Intern Med 32(6):610–8
DOI: 10.1007/s11606-016-3981-y
© Society of General Internal Medicine 2017

INTRODUCTION

Lumbar puncture is a skill used to diagnose a variety of diseases, including life-threatening conditions such as central nervous system infections[1] and subarachnoid hemorrhage.[2] Residents are expected to learn the procedure,[3] however, despite clinical training new graduates and residents remain uncertain about how to perform lumbar puncture and their skills do not comply with stakeholders' expectations[4,5]. Residents' disproportionate concerns about the risk of iatrogenic harms may lead to reluctance to perform the procedure.[6]Hence, patient safety and quality of care may benefit from more structured training of the procedure and standardized assessment of competence.[7] Advanced technology in simulation may be valuable in preparing medical students and junior doctors for clinical practice. Sound assessment is a cornerstone in simulation based mastery learning[8] (SBML), as it provides clear learning objectives and can be used for systematic and structured feedback.[9], Current assessment tools tend to focus entirely on the technical aspects of the procedure[4,5,10]. However, recent studies have identified that the most complex aspects of lumbar puncture are patient-related and environmental factors.[11] Therefore, the aims of this study were: To develop and collect validity evidence for an assessment tool for lumbar puncture performance, incorporating both technical and non-technical aspects of performance and to set a credible pass/fail standard to determine when trainees are ready for clinical practice.

METHODS DESIGN

This was an explorative study using Messick's contemporary framework[12] for gathering validity evidence from five sources: content, response process, internal structure, relation to

other variables, and consequences.[12] The study included two steps. First, content for the new assessment tool was obtained from interviews of clinicians and a review of the literature on existing assessment tools. Second, an experimental study was set-up in a simulated setting, in order to gather evidence related to response process, international structure, relation to other variables, and consequences from using the new tool to assess performance.

The study was conducted in 2014-15.

PARTICIPANTS

This study had no biomedical or patient participation and hence the local ethics committee of Capital Region of Denmark waived the need for approval (Journal number: H-15018242). Participants took part voluntarily and provided verbal and written informed consent.

Participants in the interviews were 12 clinicians representing various levels of clinical experience and medical specialties, see Table 1. Participants in the experimental assessment study were 18 novices and 11 advanced beginners performing lumbar puncture in the simulated setting, see Table 2. Finally, the study included three clinical experts serving as raters of performance.

APPROACH

Content

Collecting content validity evidence for the assessment tool was initiated by semi-structured interviews of experts and novices regarding performing lumbar puncture. The aim of the interviews was to identify differences in approach to procedure performance related to experience level. This approach aligns with the established concept of a Cognitive Task Analysis[13]. The physicians recruited for the interviews were purposefully sampled to represent various experience levels and a variety of specialties: internal medicine, neurology and anesthesia. A semi-structured interview-guide was designed with open-ended questions to investigate how the physicians approach the lumbar puncture procedure. In addition to the questions designed for all participants, the novices were asked specific questions about barriers to performance of the procedure and they were asked to reflect upon changes in relation to increasing experience. Experts were asked to reflect upon common obstacles to skill development and typical mistakes made by novices. The interviews lasted approximately 30 minutes each, were audio-recorded, and transcribed verbatim.

A thematic content analysis[14] was applied to the data to identify themes and establish categories which incorporated similar meanings[15] and to achieve a condensed and broad description of the categories[16]. Data analysis followed the steps for inductive content analysis including open coding, creating categories and abstraction[16]. First, MH, YS, and CR independently read through all interview material to familiarize themselves with the entire dataset. Subsequently, MH, YS and CR independently completed open coding of one interview from each group to identify the codes emerging from the material. Following discussions and agreement on codes, MH then coded all interviews. A final coding process was achieved through discussion with YS and CR and key issues relevant to

Table 1 Results of the Interviews with 12 Clinicians Representing Different Experience Levels

Categories	Example quotes
Goal setting and strategic planning	[Novice] "*My first thought would be where to puncture*." [Expert] " *It's about setting a scene and inducing patient confidence, because one thing that makes the procedure difficult is having a tense and extremely nervous patient. And it* [*nervousness*] *transmits, so therefore I establish some calmness on the setting.*" [Expert] "*I usually say I have made some LEAN* [Optimizing procedure flow] *on the procedure, and thought it through – how can I optimize the procedure so things flow*?" [Expert] "*I always use local anaesthesia. I feel sorry for the patient if midway I have to change strategy and use a larger needle. And you can't always count on being successful at the first attempt - you may need to make another needle insertion. I always disinfect a large area and use a large draping with a hole. Then I can still palpate while keeping sterile*."
Initial considerations and self-efficacy regarding performance	[Novice] "*I think that in the beginning I thought that if I missed in the first attempt it was because I lacked the skills and was a novice. Now I know that sometimes it's just difficult and I just need to make another attempt*." [Expert] "*It's about taking your time … you know the most important thing is actually not the puncture*." [Novice] "*Another problem was that if the patient started to feel pain and get uneasy … I panicked. I have tried to work on that …in order to avoid getting panic. … And then I ensure having a nurse present that can help with the samples and the proper equipment.* "
Integration of anatomical and clinical knowledge	[Novice] "*The lumbar puncture… I think I fear it more than other procedures, because it is so connected to the brain … it is not just something you do. The fact that you are working nearby the meninges induces a fear or … maybe a respect.* " [Novice] "*Other things you have better feeling with, but the back is so massive and the spine so away from the surface so you lose the feeling.*" [Expert] "*Because it is my impression that if you are not experienced, then you have a great fear of sticking* [*the needle in*] *too high and thereby causing damage to the medulla*." [Expert] "*It's that feeling, you must have columna in your mind, and how the bones look in relation to each other, and then it's about having the feeling of whether you hit a bone? Or are you in a cavity? Or are you touching a spinosi overlying? This can be really difficult to pass on.*"

Table 2 The Content and Design of the LumPAT. Lumbar Puncture Assessment Tool (LumPAT)

		Poor 1	2	Accept-able 3	4	Perfect 5
Planning and preparation						
1	Secures patient identification and checks for allergies (Lidocaine and disinfection). Considers contraindications (especially intracranial pressure and INR/thrombocytes) Informs about potential side effects					
2	Positions patient appropriately					
3	Palpation of anatomical landmarks including: iliac crests and spinous processes L3, L4, L5. Marks site with pen or indent					
4	Prepares material using sterile technique Requestsappropriate needle size					
5	Prepares and optimizes procedure including e.g: - Ensures bedrest atappropriate height - Use of assistant - Patient co-operation					
Performance of procedure						
6	Preps the skins with disinfection and injects lidocaine at site of puncture					
7	Insertion and handling of the needle: Site in midline, immediately above lower spinous process • Aims needle 15°-20° cephalad • Needle bevel parallel to midline Feels for a "pop" sensation as it passes ligamentum flavum Removes stylet and notes any fluid from needle					
8	If no fluid, replaces stylet and inserts 3-5 mm deeper • Redirects needle using a different angle • Repeats 2 to 3 times if necessary					
9	Able to obtain fluid and to fill tubes in the appropriate sequence and with appropriate volume(eg. 20 droplets in each).					
Finalization						
10	Applies pressure after needle withdrawal to avoid spilling fluid . Applies bandage to site. Cleans up correctly using sharps container for needle					
Communication						
11	Informs the patient about the procedure initially and throughout performance, inducing patient to feel confident					

Global rating of the entire procedure

1	2	3	4	5	6	7

Poor Excellent

Global judgment of the entire performance

Pass	Borderline	Fail

the design of the assessment tool were identified.

In addition to the interviews, a literature search of published assessment tools for lumbar puncture was performed, in order to further inform the content of the assessment tool. We designed the lumbar puncture assessment tool (LumPAT) based on an integration of the results of our interviews and the literature search.

Response Process, Internal Structure, Relation to Other Variables, and Consequences

In order too collect evidence of the assessment tool related to response process, internal structure, relations to other variables, and consequences, we recruited two groups of participants. A novice group represented PGY-1 doctors with no previous experience of lumbar puncture. Recruitment was conducted with an e-mail invitation to newly graduated doctors from the University of Copenhagen. An advanced beginners group represented medical students or residents having performed ≥ 20 lumbar punctures. Performers at this level of experience were expected to demonstrate performance significantly above baseline.[17] The advanced beginners were recruited from the Departments of Neurology and Hematology at Rigshospitalet.

Study participants carried out the lumbar puncture procedure individually in a simulated setting and performances were video-recorded for assessment. The setup was a ward-like setting including a standardized patient (SP), a standardized assistant (SA), and a lumbar puncture phantom, KyotoKagaku Lumbar Puncture Simulator II (Kyoto, Japan). The phantom simulated the lumbar anatomy (including landmarks for palpation), provided life-like skin and tissue resistance (standard insertion was used for all cases), and allowed for the drainage of a liquid sample. The only preparation participants received was 15 minutes of written instructions which included illustrations. This occurred immediately prior to entering the procedural room.They were instructed to interact as in a real clinical setting and were provided with authentic medical records including laboratory results and brain Computer Tomography (CT) results. The SP pretended to be uninformed about the procedure, with moderate fear, and initially assumed an inadequate position. Just before needle insertion, the phantom was introduced. In order to allow for ongoing communication, the SP remained on the bed with the phantom strapped to her back during the rest of the procedure. The SA assisted only upon request and did not participate in decision-making or procedural performance. The participants were instructed to inform the SP, position her, and mark the puncture sites after palpation before the phantom was introduced for needle insertion. Participants who failed to obtain liquid had to inform the SP and terminate the session. We used the same SP in all scenarios and the SA was portrayed by one of a group of three. The items in the LumPAT were unknown to the participants, the SP, and the SA.

Evidence of the response process was collected from a team of three raters representing content experts who were associate professors with significant teaching experience (two neurologists and one anesthesiologist). Initially the content experts rated and commented on five pilot cases of novice performances in order to identify ambiguity and missing items. Raters participated in a two-hour training session which included instructions on using the full scope of each item's rating scale to minimize the risk of end-aversion bias. Further, we instructed raters to make individual assessments for each item and to avoid being influenced by the preceding items, in order to minimize the halo effect. Finally, three additional pilot-ratings were conducted to confirm acceptable inter-rater reliability. Pilot-videos were not included in the study.

To ensure blinding of participants' experience levels, we recruited participants in both study groups, who were of similar age, ensured that none of the participants' were known by the content experts, and utilized a web-based solution[18] which provided the raters with a masked sample of the videos in a random order. The raters independently reviewed all video-recordings, scored each individual item of the LumPAT, made a global assessment on the 7-point Likert scale, and concluded with an overall judgment on pass, borderline, or fail.

Statistical Analysis

Internal consistency was explored using Cronbach's alpha.[19] We correlated the mean item score with the corresponding global assessment using Pearson's r to determine if the items in the LumPAT were representative of the complete procedure. We used descriptive statistics to assess the raters' use of the full range of all response options within the scale. We used generalizability theory to give a combined estimate of the reliability of the assessment tool and to explore the impact of the different sources of variance[20]. We further conducted a Decision-study to explore the number of raters needed to ensure sufficient reliability.[20]

Relation to other variables was explored by comparing LumPAT mean items scores between the two groups. We used an independent samples t-test to assess the ability to discriminate between the performances of the two study groups. Having more than 10 participants per group made it reasonable to assume that our data could be considered as normally distributed.[20] Cohen's *d* was calculated to estimate the effect sizes of the differences in the mean between the groups.

Consequences of the assessment were explored by establishment of a pass/fail score using the contrasting groups standard setting method.[21] This was defined by the intersection of a distribution plot of the mean scores of the two groups. To explore the consistency between the consequences of the standard to the raters global judgment we inserted the allocations from the standard setting and rater judgments into a 2x3 contingency table. Pearson's chi-square test was chosen as no cells in the table had an expected count less than five. P-values

below 0.05 were considered statistically significant. For statistical analysis we used SPSS vers. 22. (SPSS Inc, Chicago, IL). For calculating the generalizability-coefficient we used G_string IV vers 6.1.1 software package (Ontario, Canada).

RESULTS

Twelve physicians agreed to participate in the interviews. The five experts were senior physicians having performed more than 300 procedures, representing the specialties of: anaesthesia, neurology and internal medicine. Novices were represented by seven PGY-1 and PGY-2 doctors from Neurology, Internal Medicine and Emergency Medicine, with experience in the range 0-40 prior procedures. Below, we provide a summary of the validity evidence from each of the five sources in Messick's framework. Eighteen novices and eleven advanced beginners completed the procedure, but data from one participant in the advanced beginners group was lost for technical reasons. The participants in the two groups were of equal age: Novices mean 30 years, range 27-34 years; advanced beginners mean 29 years, range 27-31 years. Novices had no previous experience, but had observed the procedure being performed in a mean of 2.5 observations, range 0-6 observations. The advanced beginners had experience corresponding to a mean of 52.4 procedures, range 20-100.

The 28 successful video-recordings were independently assessed by the three expert raters resulting in 84 LumPAT forms for analysis.

Content

Three categories where novices and experts contrasted were identified: 1) Goal setting and strategic planning; 2) Initial considerations and self-efficacy regarding performance; and 3) Integration of anatomical and clinical knowledge.

The primary goal for novices was to get the sample and they focused on which needle to use and where to insert it. Thus, novices had an outcome focus and their main considerations were associated with the technical skills of the procedure. In contrast, experts' task analysis and strategic planning included more attention to the clinical environment than to the technical aspects of the procedure. Preparation included having an assistant available and having the right tools at hand. Moreover, they emphasized the importance of good communication and cooperation with the patient to ensure a calm environment. Finally, the correct positioning of the patient and sufficient management of pain were mentioned as important factors. All of these considerations were initiated the moment the expert clinicians entered the room and made an environmental scan of the situation and the patient. Experts were aware that first attempt was not always successful. Therefore, they took precautions to optimize the procedure, including provision of local anesthesia, use of the thinnest needle possible, and keeping the patient informed throughout the procedure. The findings and the quotations are summarized in Table 1.

The literature search identified three relevant studies describing checklists. However, they all primarily dealt with technical aspects of procedure performance. The selected studies were based on a Delphi approach,[10] a task analysis approach based on expert faculty opinion[4] and consensus[5].

By combining already published checklists for assessment of lumbar puncture with the results from the interviews, optimal content for the assessment tool was ensured in the design phase. The LumPAT included four major categories: Planning and Preparation, Performance of Procedure, Finalization, and Communication. These categories covered eleven individual items to be assessed on a five-point Global Rating Scale (GRS). The content of the assessment tool is presented in Table 2.

Response Process

During pilot testing two situations were identified as particularly challenging to assess: 1) A participant who performed well, with proper needle insertion, relevant attempts to correct failure, but who did not obtain liquid, and 2) A participant who was lucky and obtained liquid despite inappropriate palpation and needle insertion technique. We agreed that situation one should be rated perfect in items 7 and 8 but poor in item 9; situation two should be rated poor in item 7 but good in items 8 and 9. Raters' agreements on items were written down and distributed to the group.

After participating in rater training, the three raters demonstrated acceptable agreement when evaluating pilot videos.

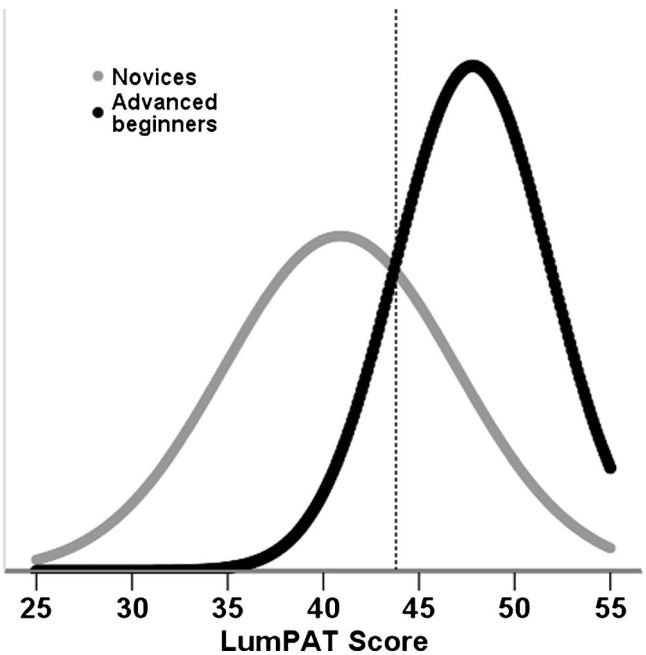

Fig. 1 Standard-setting by the contrasting groups method. The X-axis demonstrates the distribution of the total scores for the two study groups which differ based on experience level (novice and advanced beginner). The intersection representing the cut-off standard is marked by the vertical line.

Table 3 Distribution of Raters' Global Judgment (Pass, Borderline or Fail) compared to the LumPAT Score Based on the Established Standard

	Rater Fail	Rater Borderline	Rater Pass	Total
LumPAT Pass	1 (3%) AB:1 Novice:0	3 (8%) AB:2 Novice: 1	32 (89%) AB: 21 Novice:11	36 100%
LumPAT Fail	12 (25%) AB:2 Novice:10	16 (33%) AB:3 Novice:13	20 (42%) AB:1 Novice:19	48 (100%)
Total	13	19	52	84

The distribution is based on the three content experts' ratings of a total of 84 assessments. AB = Advanced beginners

The recruited participants were similar in age and therefore facilitated an effective blinding process.

Internal Structure

The internal consistency of the LumPAT was high: Cronbach's alpha = 0.92. Internal structure of the assessment tool was supported by a good correlation between raters' means items score and global assessment score: Pearson's correlation = 0.83, ($p < 0.001$). All response options within the 5-point scale were used for all items except item 2 and 5, which had scores from 2-5 points. Generalizability coefficient using three raters was 0.88 and the magnitude of the different sources of variance showed that 71% of the variance originated from difference among the participants. Rater variance accounted for 11% and the unexplained variance was 18%. A D-study demonstrated that using one or two raters would result in generalizability coefficients of 0.71 and 0.83, respectively.

Relations to Other Variables

The assessment results demonstrated sufficient discrimination between the two groups of different experience levels, with a mean for the novice group of 40.9 (SD 6.1) and a mean for the advanced beginners group of 47.8 (SD 4.0), $p = 0.004$. The impact of the identified differences was Cohen's d effect size 1.34.

Consequences

The contrasting groups' method established a pass/fail standard of 44.0 points (Fig. 1). Pearson's chi-square test found that the observed allocations did not arise by chance ($p < 0.001$), demonstrating coherence between the standard setting and raters' global judgment on pass/borderline/fail (Table 3).

DISCUSSION

In this study we developed a new assessment tool, the LumPAT, for evaluating the performance of lumbar puncture and gathered validity evidence in accordance with the contemporary framework for validation.[12] The main contribution of this study to existing assessment formats is the incorporation of contextual aspects of performing the procedure. Recent studies have identified that such contextualization of the technical and non-technical aspects in the assessment and training of clinical procedures benefits performance[22]. The experimental study setup resembled the intended clinical setting and participants in the study represented the target group of performers, emphasizing the transition from novice to advanced beginner. The results using the LumPAT provided a standard for performance that will be helpful in judging readiness for clinical practice. Although the American Board of Internal Medicine (ABIM) does not require all internists to master the lumbar puncture procedure, ABIM does expects thorough evaluation and credentialing of physician skills prior to independent clinical practice.Our study demonstrated that the LumPAT can be used by a single rater for low-stakes assessments.

Our interviews revealed that the experts emphasized the importance of contextualizing the procedure and having a strategic plan, including setting specific goals for procedural performance. The inclusion of different specialties and experience levels increase the external validity of these results.

The LumPAT differs from current checklists for lumbar puncture in two ways. First, the LumPAT includes items regarding patient communication, positioning, and procedure optimization, (including the use of an assistant). These aspects are reported to have the highest impact on the complexity of the procedure for the novice performer: excessive motion in the patient and lack of cooperation, as well as environmental aspects, such as inexperienced assistants, time constraints, and poor preparation.[11] By dedicating specific items in the tool to patient-related and environmental elements, we highlighted their importance in the process of judging clinical readiness. Second, we used the GRS design rather than a checklist design. Reviews have found that GRS is superior in capturing more nuanced elements of expertise[23,24] and it is also superior to checklists in identifying weak or even dangerous performances.[25] For the infant lumbar puncture the GRS design is found to outperfom the checklist design in its discriminant ability and interrater agreement[24].

The contemporary framework for validity implies that "all sources of error associated with test administration are controlled and eliminated to the maximum extent possible".[26] An important part of validation of assessment tools relates to the statistical and psychometric characteristics of the tool investigated.[27] Internal consistency of the LumPAT was high[28] and the generalizability study[20] found that most of the variance was associated with the performers (71%) rather than disagreement between the raters (11%). The generalizability-coefficient for our setup with three raters was 0.88, which is above the level of 0.80 for high-stakes assessments.[28] The D-study demonstrated that two raters would be sufficient for obtaining the 0.8 G-coefficient and additional raters result only in minor increases. A G-coefficient of above 0.7 is considered acceptable for low-stakes assessment[28]. The D-

study demonstrates that one rater would result in a G-coefficient of 0.71 which makes the LumPAT ideally suited to provide valuable, structured feedback to trainees after supervised performance of the procedure. None of the existing assessment tools have investigated generalizability-coefficient or provided data on how the number of raters impact on the reliability of the tool[4,5].

The correlation between the rating scores of the items with the overall global judgment of the performances, 0.83 ($p < 0.001$), reflects how well the items included in the assessment tool reflect the clinicians' overall impression of the trainees' performance. This supports our aim to assess novices' ability to contextualize the procedure as opposed to assessing only technical tasks. No other studies has included global judgments[4,5] precluding comparisons between assessment tools' ability to assess the intended clinical context.

Earlier validation studies have been criticized for including only expert-novice comparisons which should automatically result in a large difference in scores.[29] We minimized this spectrum bias by including participants who resembled the intended target population. A prior study has demonstrated that having performed more than 45 spinal punctures leads to a 90% success rate,[17] and this was the level of our advanced beginners group. The LumPAT tool could discriminate between novices and advanced beginners with differences of 6.9 points ($p = 0.004$) between mean scores, corresponding to a large effect-size of 1.34. The contrasting groups method, which is a well-established method for standard setting,[21] revealed that a participant should obtain >44.0 points to pass the test (Fig. 1). This standard was confirmed by the raters' overall judgment as the Pearson chi square test demonstrated significant consistency with the passing standard. (Table 3). In this sample, there was only a single case in which a judgment of fail obtained a passing LumPAT score. By contrast, 20 cases that were given judgments of pass by the raters obtained a score below cut-off for the LumPAT. These results indicate that raters may bee too lenient on participants with suboptimal performance. However, further studies are needed to investigate whether or not this finding is confirmed in the clinical setting.

Limitations

The sample size in this study was limited but, nonetheless, the study was sufficiently powered to establish significant results. The potential risk of systematic rater bias favouring the experienced group[30] was minimized by blinding experience levels. The intended context for our assessment tool is the clinical setting with patient interaction. However, since our study included participants with no procedure experience and only limited training, it was not ethically and practically feasible to conduct procedures on real patients. However, all efforts were made to have a simulation scenario that was similar to a clinical setting and we ensured realistic ongoing communication by having a SP present throughout. Previous studies utilizing an integrated design using a SP and mannequin have demonstrated that such an approach is valuable[31]. The primary discrepancy between our simulated setting and clinical practice is structural fidelity, characterized by patients' anatomical variability and the consequences of procedural failure for the patient. However, Hamstra et al. recommend a shift from emphasis on physical resemblance of the simulator to functional alignment with the entire simulation and the intended applied context.[32] The standardized set-up minimized random effects, making it suitable for gathering validity evidence. On the other hand, the standardized set-up might compromise the external validity regarding application to the clinical context. A recent systematic review and meta-analysis found that simulation-based assessments correlated well with patient outcomes.[33] However, the review did not include any studies on lumbar puncture performance and, hence, future studies should investigate the correlation of the LumPAT score obtained in a simulated environment with performance in a clinical setting.

Implications

In 2003 a hypothetical formula for improving patient outcomes was developed that is comprised of three interactive factors: Medical Science x Educational Efficiency x Local Implementation.[34] For lumbar puncture the first factor has been optimized by introducing atraumatic needles[35] and sonography guided punctures.[36] Education has been guided by the expectation that experience will lead to mastery.[3] This assumption has been disproved by simulation-based and observational studies[5,37] which call for more educational research, including assessments of the impact on patient outcomes. A contrast to the maxim of "see one, do one, teach one" is mastery learning (ML)[38]. According to ML principles, we should create and implement a set of educational conditions, course curricula, and assessment plans that yield mastery level achievements among learners.[38] ML implies that learners should practice and re-test until they reach a designated mastery level, making the final level the same for all, although the time taken to reach that level may vary.[39] Assessment tools are a cornerstone in ML as they guide the individual learner in optimizing their skills to meet the expected standard.[8] Combining the ML principles with simulation-based training of central venous catheterization reduce patient related complications[7].

To-date simulation-based training of the lumbar puncture procedure has focused nearly exclusively on its technical aspects[5,40]. However, this present study incorporates the non-technical aspects of the procedure and, thereby addresses previous concerns that simulation based training is too distant from the clinical context.[3] The LumPAT and our study design with the SP reduce the gap between simulation context and clinical context by integrating aspects of patient communication and contextualization of the procedure. The LumPAT demonstrates sufficient validity evidence suitable for integration in a SBML program that will prepare residents for clinical practice. However, studies to assess the retention of skills

achieved through SBML are required in order to optimize the curriculum. Finally, there is a need for translational studies to evaluate the impact that integrating the LumPAT into SBML will have on patient outcomes.

CONCLUSION

Based on Messicks framework we demonstrated strong validity evidence for a lumbar puncture assessment tool, the LumPAT. The tool can be used to assess readiness for clinical practice.

Acknowledgements: *The study was funded by TrygFonden Grant no 105112; being a non-medical non-governmental organization. The funding sources had no role in design and conduct of the study; collection, management, analysis or interpretation of the data; preparation, review or approval of the manuscript; the decision to submit the manuscript for publication.*
Authorization has been obtained for disclosure of persons on the supplementary video.
The study was presented as a poster at the 2016 IMSH conference, San Diego at the 17th of January 2016.

Corresponding Author: *Mikael Johannes Vuokko Henriksen, MD; Copenhagen Academy of Medical Education and Simulation, The Capital Region of Denmark, Rigshospitalet section 5404, Blegdamsvej 9 2100, Copenhagen East, Denmark (e-mail: mikael.johannes.vuokko.henriksen@regionh.dk).*
Compliance with Ethical Standards:

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

REFERENCES

1. **Fitch MT, van de Beek D.** Emergency diagnosis and treatment of adult meningitis. Lancet Infect Dis. 2007;7(3):191–200.
2. **Martin SCG, Teo MKCH, Young AMH, et al.** Defending a traditional practice in the modern era: the use of lumbar puncture in the investigation of subarachnoid haemorrhage. Br J Neurosurg. 2015;29(6):799–803.
3. **Nathan BR, Kincaid O.** Does experience doing lumbar punctures result in expertise? A medical maxim bites the dust. Neurology. 2012;79(2):115–116.
4. **Lammers RL, Temple KJ, Wagner MJ, Ray D.** Competence of new emergency medicine residents in the performance of lumbar punctures. Acad Emerg Med. 2005;12(7):622–628.
5. **Barsuk JH, Cohen ER, Caprio T, McGaghie WC, Simuni T, Wayne DB.** Simulation-based education with mastery learning improves residents' lumbar puncture skills. Neurology. 2012;79(2):132–137.
6. **Kneen R.** The role of lumbar puncture in suspected CNS infection–a disappearing skill? Arch Dis Child. 2002;87(3):181–183.
7. **Barsuk JH, McGaghie WC, Cohen ER, Balachandran JS, Wayne DB.** Use of simulation-based mastery learning to improve the quality of central venous catheter placement in a medical intensive care unit. J Hosp Med. 2009;4(7):397–403.
8. **Lineberry M, Soo Park Y, Cook DA, Yudkowsky R.** Making the case for mastery learning assessments: key issues in validation and justification. Acad Med. 2015;90(11):1445–1450.
9. **McGaghie WC, Siddall VJ, Mazmanian PE, Myers J.** Lessons for continuing medical education from simulation research in undergraduate and graduate medical education: effectiveness of continuing medical education: american college of chest physicians evidence-based educational guidelines. Chest. 2009;135(3 Suppl):62S–68S.
10. **Berg K, Riesenberg LA, Berg D, et al.** The development of a validated checklist for adult lumbar puncture: preliminary results. Am J Med Qual. 2013;28(4):330–334.
11. **Haji FA, Khan R, Regehr G, Ng G, de Ribaupierre S, Dubrowski A.** Operationalising elaboration theory for simulation instruction design: a Delphi study. Med Educ. 2015;49(6):576–588.
12. **Cook DA, Beckman TJ.** Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med. 2006;119(2):166. **e7-16**.
13. **Sullivan ME, Yates KA, Inaba K, Lam L, Clark RE.** The use of cognitive task analysis to reveal the instructional limitations of experts in the teaching of procedural skills. Acad Med. 2014;89(5):811–816.
14. **Krippendorff K.** Content analysis. 3rd ed. Thousand Oaks, California: SAGE Publications, Inc; 2013.
15. **Hsieh H-F, Shannon SE.** Three approaches to qualitative content analysis. Qual Health Res. 2005;15(9):1277–1288.
16. **Elo S, Kyngäs H.** The qualitative content analysis process. J Adv Nurs. 2008;62(1):107–115.
17. **Kopacz DJ, Neal JM, Pollock JE.** The regional anesthesia "learing curve": What is the minimum number of epidural and spinal blocks to reach consistancy? Reg Anesth Pain Med. 1996;21(3):182–190.
18. **Subhi Y, Todsen T, Konge L.** An integrable, web-based solution for easy assessment of video-recorded performances. Adv Med Educ Pract. 2014;5:103–105.
19. **Cronbach LJ.** Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16(3):297–334.
20. **Bloch R, Norman G.** Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach. 2012;34(11):960–992.
21. **Yudkowsky R, Downing SM, Tekian A.** Standard setting. In: Downing S, Yudkowsky R, eds. Assesment in health professions education. Secondth ed. New York: Roudledge Taylor and francis Group; 2009:119–148.
22. **Brunckhorst O, Shahid S, Aydin A, et al.** The relationship between technical and nontechnical skills within a simulation-based ureteroscopy training environment. J Surg Educ. 2015;72(5):1039–1044.
23. **Ilgen JS, Ma IWY, Hatala R, Cook DA.** A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. Med Educ. 2015;49(2):161–173.
24. **Gerard JM, Kessler DO, Braun C, Mehta R, Scalzo AJ, Auerbach M.** Validation of global rating scale and checklist instruments for the infant lumbar puncture procedure. Simul Healthc. 2013;8(3):148–154.
25. **Ma IWY, Zalunardo N, Pachev G, et al.** Comparing the use of global rating scale with checklists for the assessment of central venous catheterization skills using simulation. Adv Heal Sci Educ. 2012;17(4):457–470.
26. **Downing SM.** Validity: on the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830–837.
27. **Downing SM, Haladyna TM.** Validity and its threats. In: Downing S, Yudkowsky R, eds. Assesment In health professions education. firstth ed. New York: Roudledge Taylor and francis Group; 2009:21–55.
28. **Downing SM.** Reliability: on the reproducibility of assessment data. Med Educ. 2004;38(9):1006–1012.
29. **Cook DA.** Much ado about differences: why expert-novice comparisons add little to the validity argument. September: Adv Health Sci Educ Theory Pract; 2014.
30. **Konge L, Vilmann P, Clementsen P, Annema JT, Ringsted C.** Reliable and valid assessment of competence in endoscopic ultrasonography and fine-needle aspiration for mediastinal staging of non-small cell lung cancer. Endoscopy. 2012;44(10):928–933.
31. **Kneebone R, Kidd J, Nestel D, Asvall S, Paraskeva P, Darzi A.** An innovative model for teaching and learning clinical procedures. Med Educ. 2002;36(7):628–634.
32. **Hamstra SJ, Brydges R, Hatala R, Zendejas B, Cook DA.** Reconsidering fidelity in simulation-based training. Acad Med. 2014;89(3):387–392.
33. **Brydges R, Hatala R, Zendejas B, Erwin PJ, Cook DA.** Linking simulation-based educational assessments and patient-related outcomes: a systematic review and meta-analysis. Acad Med. 2015;90(2):246–256.

34. **Søreide E, Morrison L, Hillman K, et al.** The formula for survival in resuscitation. Resuscitation. 2013;84(11):1487–1493.
35. **Tung CE.** Education research: changing practice: Residents' adoption of the atraumatic lumbar puncture needle. Neurology. 2013;80(17):e180–2.
36. **Shaikh F, Brzezinski J, Alexander S, et al.** Ultrasound imaging for lumbar punctures and epidural catheterisations: systematic review and meta-analysis. BMJ. 2013;346:f1720.
37. **Edwards C, Leira EC, Gonzalez-Alegre P.** Residency training: a failed lumbar puncture is more about obesity than lack of ability. Neurology. 2015;84(10):e69–e72.
38. **McGaghie WC.** Mastery learning: It is time for medical education to join the 21st century. Acad Med. 2015;90(11):1438–1441.
39. **Yudkowsky R, Park YS, Lineberry M, Knox A, Ritter EM.** Setting mastery learning standards. Acad Med. 2015;90(11):1495–1500.
40. **Conroy SM, Bond WF, Pheasant KS, Ceccacci N.** Competence and retention in performance of the lumbar puncture procedure in a task trainer model. Simul Healthc. 2010;5(3):133–138.