Observational Study

Medicine®

OPEN

Validity evidence of a task trainer for normal and difficult lumbar puncture

A cross-sectional study

Yu Akaishi, MD, PhD[a], Yuito Okada, RN, MS, MPH[b], Jannet Lee-Jayaram, MD[c], Jun Seok Seo, MD[d], Toru Yamada, MD, PhD[a], Benjamin Worth Berg, MD[c,*]

Abstract

Validation of the anatomically complex configurations of the Lumbar Puncture Simulator II (KYOTO KAGAKU CO., LTD., 15 Kitanekoya-cho Fushimi-ku Kyoto, Japan 612-8388) have not been reported. Previous validation of the normal anatomic configuration has been reported. This study aims to evaluate evidence for construct and content validity of 4 interchangeable lumbar puncture (LP) complex anatomic configurations of this simulator.

We performed a cross-sectional study between April 2018 and May 2019. Novice volunteer medical students and expert physicians who had performed over 30 LP procedures performed sequential LP procedures on each of 4 simulated interchangeable anatomic LP puncture blocks (normal, obesity, geriatric, combined geriatric/obesity). Primary outcome measures compared between groups for each LP procedure were return of cerebrospinal fluid within 5 minutes and a calculated performance score. Subjective face validity and content validity 5-point Likert questionnaires were completed by participants.

35 novice (n = 19) and expert (n = 16) subjects completed 140 procedures. Significant differences were found between novice and expert groups for both cerebrospinal fluid success rates and performance scores for normal ($P = .001$/$P = .001$) geriatric ($P = .005$/$P = .002$) and obesity ($P = .003$/$P = < .001$) configurations. There were no differences for the geriatric/obesity configuration. Expert median score of simulator realism (face validity) was 4 (range 3–4); median score of utility as a training tool (content validity) was 4 (range 4–5).

We provide evidence for construct validity for each of the complex LP configurations, except combined geriatric/obesity. Expert physicians found the simulator sufficiently realistic to effectively teach LP skills.

Abbreviations: CSF = cerebrospinal fluid, LP = lumbar puncture.

Keywords: construct validity, lumbar puncture, medical education, task trainer, validation

1. Introduction

Simulation-based education is widely used for training and assessment in healthcare.[1,2] From the perspective of patient safety, simulation-based education provides an opportunity for learners to develop and demonstrate core competency in skills and knowledge, before application of skills in the clinical setting.[3–5] Lumbar puncture (LP) is an invasive procedure skill taught to medical students and residents. Traditional training includes didactic content (including lectures, reading and video) and mentored/supervised procedural training with patients. More recently, improvement in LP skills has been demonstrated using task trainer simulation-based training.[6–9]

Mastery learning incorporating simulation-based skill performance assessment before performing real-life invasive procedures is an established method of improving patient safety.[10–12] There is no data regarding safety outcomes or clinical procedure performance by learners who have completed LP simulator-based training using mastery learning design with established minimum passing score competency determination. To accurately assess simulation-based competency, simulators incorporated in assessment protocols must be studied and validated.[13,14] Simulator validity is ultimately the characteristic of the simulator which confirms that simulated skill performance translates to performance in real life.[15,16] A classical validity framework includes Content, Criterion, Construct, and Face validity. Rigorous validation of contemporary simulation-based training assessment requires use of an identified validation framework, such as Messick or Kane[15]; both of which have largely supplanted the classical validity framework, because they more rigorously define the elements of validity and the descriptions of the relative strengths of each element. Face validity concepts are incorporated in the content validity construct of contemporary validation

Editor: Gunjan Arora.

Yu Akaishi belonged to SimTiki Simulation Center, John A. Burns School of Medicine, University of Hawaii at Manoa.

The authors have no conflicts of interest to disclose.

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

[a] Department of Family Medicine, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan, [b] Cancer Epidemiology Program, University of Hawaii Cancer Center, [c] SimTiki Simulation Center, John A. Burns School of Medicine, University of Hawaii at Manoa, Honolulu, HI, [d] Department of Emergency Medicine, Dongguk University Ilsan Hospital, Dongguk University College of Medicine, Seoul, Republic of Korea.

[*] Correspondence: Benjamin Worth Berg, SimTiki Simulation Center, John A. Burns School of Medicine, University of Hawaii at Manoa, 651 Ilalo Street Honolulu, HI 96813 (e-mail: bwberg@hawaii.edu).

How to cite this article: Akaishi Y, Okada Y, Lee-Jayaram J, Seo JS, Yamada T, Berg BW. Validity evidence of a task trainer for normal and difficult lumbar puncture: a cross-sectional study. Medicine 2020;99:41(e22622).

Received: 14 May 2020 / Received in final form: 28 July 2020 / Accepted: 8 September 2020

http://dx.doi.org/10.1097/MD.0000000000022622

frameworks. Assessment validation studies require determination of both the quality of evidence supporting validation of the overall educational construct and of multiple discrete validation elements which may include simulators, assessment tools, raters and rater training, curricular/scenario/instructional design, learner factors, and more. Validity frameworks classically describe a variety of sources of evidence. Messick 5 sources of evidence for articulating the strength of a validation study are Content, Internal Structure, Relationships with other variables, Response Process, and Consequences. This framework has been adopted in validation standards of the American Psychological Association.[17] Kane more recent validity framework incorporates 4 validation inferences: Scoring, Generalization, Extrapolation, and Implications/Decisions.[16,18] Overlapping types of evidence are used to support both Messick and Kane frameworks. For example, Expert-Novice comparisons are considered a source of validity evidence in all 3 frameworks; Classical, Messick, and Kane. Reliability is a construct related to validity and is used to measure quality of assessment tools in conjunction with validity. Interrater reliability is an element of validation evidence in both Kane and Messick frameworks. High reliability is a necessary but not sufficient singular element of validity determination of assessment paradigms. Multi-point evidence is required to establish validity and credibility of an overall assessment construct of interest. Validity studies construct an evidence-based argument regarding how well an overall assessment paradigm or element of that design, such as an assessment tool, measures the intended outcome.[19] The aggregate strength of evidence for each aspect of validity ultimately informs the strength of validity of the overall assessment paradigm. The strength of evidence in published simulation-based assessment validation studies remains low.[20]

Valid simulation-based assessment paradigms may incorporate expert-novice comparisons as one element of overall validity evidence in frameworks of Construct validity (classical), Relationships with other variables (Messick), and Extrapolation (Kane). Validation of a paired simulation device and assessment tools for assessment of procedural skills frequently report expert-novice comparisons.[21,22] One source of evidence for validation of a training assessment device (eg, simulator) is to determine if novices and experts demonstrate detectable performance differences upon performance of a standardized simulated procedure.[13,15] If expert scores are higher than novice scores, this validation framework defined evidence element is confirmed, and designated "high quality evidence," because the paired device and rating instrument can distinguish expert from novice skill. There are few published reports evaluating construct validity with a LP simulator.[23,24]

The Society of Simulation in Healthcare research agenda identifies validation of simulation in assessment as a priority; "... we anticipate aspects of validity, reliability, and standard setting to be implicit regarding any future studies on measurement instruments"[25]

The Lumbar Puncture Simulator II (Kyoto Kagaku CO., LTD.) used in this study has 4 interchangeable LP "blocks," representing clinical conditions with varying degrees of LP performance difficulty: I. Normal, II. Geriatric, III. Obesity, and IV. Combined Geriatric/Obesity (Fig. 1). Validation study of the normal anatomy configuration was verified,[24] and expert physician impressions of this simulator were evaluated in a prior study.[26] No validation studies of the anatomically complex components of this simulator (Blocks II, III, IV) have been published.

This study sought to evaluate evidence for construct validity by comparing LP skill performance metrics of expert physician and novice medical students on each of 4 anatomic variations of varying clinical complexity and procedural difficulty of the Lumbar Puncture Simulator II (Kyoto Kagaku, LTD.). Content validity evidence was assessed by user surveys. Internal structure and generalization were incorporated by interrater reliability of performance assessments.

2. Methods

2.1. Study design and setting

We performed a Institutional Review Board (IRB) approved cross-sectional study at the John A. Burns School of Medicine SimTiki Simulation Center, University of Hawaii at Manoa, USA between April 2018 and May 2019. STROBE reporting guidelines for health care simulation research informed study design and are incorporated in this report.[27]

2.2. Participants

Two distinct participant groups were recruited to evaluate construct validity of an LP task training simulator; 3rd and 4th year volunteer medical students who had observed or performed less than 10 LP procedures (novices) and practicing physicians who had performed at least 30 LP procedures in clinical practice (experts).

2.3. Task trainer

The Lumbar Puncture Simulator II (Kyoto Kagaku CO., LTD.) used in this study has 4 interchangeable LP "blocks," representing anatomic conditions which represent varying degrees of LP performance difficulty: Normal, Geriatric, Obesity, and Combined Geriatric/Obesity. Manufacturer terminology for the geriatric model is "Senior." Palpable anatomic landmarks represented by this task trainer include midline spinous processes and posterior superior iliac crests located at the level of the L3–4 intervertebral space. Simulated cerebrospinal fluid (CSF) under adjustable pressure can be accessed by insertion of a spinal needle at levels L2–3, L3–4, and L4–5. Successful CSF access is recognized by the appearance of free-flowing clear fluid at the needle hub. Simulator functionality includes simulation of CSF pressure measurement; this functionality was not utilized in this study. The model as used in this study did not simulate traumatic/bloody LP procedures. Each anatomic block permits access to simulated CSF when a 20G spinal needle is passed from skin to the simulated spinal canal. The obesity configured blocks represent lumbar spinal elements at greater depth from the skin than other blocks. The geriatric configured blocks represent increased tissue resistance and a spinal osteoarthritic bone configuration.

2.4. Study procedures

Participants completed a demographic questionnaire and were provided a standardized 5-minute scripted verbal orientation to the simulator, equipment, and the protocol. Participants performed 4 sequential LP procedures using the LP task trainer in the simulated upright/sitting position. A fixed LP configuration sequence procedure was utilized: normal, geriatric, obesity, and combined geriatric/obesity. The sequence of anatomic blocks was not known to subjects. Sterile technique, local anesthetic techniques, and draping were not performed by participants in this study.

Figure 1. The Lumbar Puncture Simulator II (Kyoto Kagaku CO., LTD.). This simulator has 4 interchangeable LP "blocks," representing anatomic conditions which represent varying degrees of LP performance difficulty: Normal, Geriatric, Obesity, and Combined Geriatric/Obesity. LP = lumbar puncture.

LP procedure time was limited to 5 minutes for each of the 4 anatomic configurations, with a break of 2 to 5 minutes between successive LP attempts. No feedback was provided regarding performance and no access to training resources was available to participants between successive LP attempts. CSF access success or failure was confirmed in real-time under direct visualization by a single rater (YA). A skills performance rating sheet was completed by the rater in real-time immediately after completion of each LP attempt. Each LP attempt was terminated upon visualization of CSF flow from the hub of a 20G 3.5" LP needle or if no CSF was visualized within 5 minutes. No verbal or non-verbal coaching or cuing regarding LP technique was provided by the rater at any time. All procedures were recorded by a digital video recorder positioned behind the participant, outside of the participant's visual field. Trials in which the needle was placed outside of the boundaries of the anatomic block were excluded from analysis. Immediately following completion of 4 LP procedures, participants completed a 5-point Likert scale regarding simulator fidelity/realism and perceived utility for teaching and learning.

2.5. LP skill rating tool

A task-specific LP skill performance 9-item rating tool (range: 0–13 total points) incorporating modified items from previously reported and validated LP skill assessment tools (Fig. 2)[6,9,24] was developed by the authors. Rated items scores were weighted based on the authors consensus regarding prioritization of individual critical items. The tool was designed to detect differences in novice versus expert LP skill performance under direct observation.

2.6. Outcomes and measurements

Participants completed a pre-procedure demographic questionnaire, including details of gender, age, expert or novice status, adult and pediatric LP procedure experience, including LP training, and/or LP simulation experience Content validity of perceptions of simulator realism and utility as a training tool, were assessed using a post procedure 5-point Likert scale. Three raters completed observation and scoring of each LP procedure. A single rater (YA) with expert level LP skill and LP teaching

RATER EVALUATION SHEET

Subject #	**Date/Time**			
PROCEDURE SCORING				
Block Type	**Normal**	**Geriatric**	**Obesity**	**Geriatric / Obesity**
Trial #	**I**	**II**	**III**	**IV**
Did the subject touch the pelvic landmarks †	**Yes / No**	**Yes / No**	**Yes / No**	**Yes / No**
Insertion site (first insertion)*	◯ **L2/3** ◯ **L3/4** ◯ **L4/5** ◯ **Unknown**	◯ **L2/3** ◯ **L3/4** ◯ **L4/5** ◯ **Unknown**	◯ **L2/3** ◯ **L3/4** ◯ **L4/5** ◯ **Unknown**	◯ **L2/3** ◯ **L3/4** ◯ **L4/5** ◯ **Unknown**
Needle insertion in midline †	**Yes / No**	**Yes / No**	**Yes / No**	**Yes / No**
Manual stabilization technique utilized †	**Yes / No**	**Yes / No**	**Yes / No**	**Yes / No**
The number of insertions (# of punctures) ‡				
# of times direction of needle insertion was changed §				
# of times needle insertion site was changed ‖				
Procedure time (maximum 300s) ¶	_____**Sec**	_____**Sec**	_____**Sec**	_____**Sec**
CSF sample obtained	**Yes / No**	**Yes / No**	**Yes / No**	**Yes / No**
TOTAL SCORE (Maximum 13)::				

RATER: _____

SCORING GUIDE
* L3/4 or L4/5; 2 points, L2/3; 0 point, above L2/3 or below L4/5; -2points
† Yes; 1 point, No; 0 point
‡ Under 3 times; 1 point, More than 3 times; 0 point
§ Under 4 times ; 1 point, More than 4 times; 0 point
‖ Under 2 times; 1 point, More than 2 times; 0 point
¶ <90 seconds; 5 points, 91-180 seconds; 4 points, 181-300 seconds; 3 points, 0 Points >300 seconds

Figure 2. Rater evaluation sheet. A task-specific LP skill performance 9-item rating tool (range: 0–13 total points) incorporating modified items from previously reported and validated LP skill assessment tools. LP = lumbar puncture.

experience, scored participant performance using the skill performance rating tool by direct observation, and completed independent video review for reconciliation of ambiguous or unclear rating points as needed. Two experienced physician-educators, who were enrolled in a 1-year post-graduate simulation methods focused medical education fellowship independently viewed and scored video recordings of all LP attempts. Video-only raters received rater training and independently reviewed and scored participant videos in a predetermined random order. Video-only raters blinded to participant novice or expert status. Primary outcome measures were success rate of CSF access within 5 minutes and the LP skill performance rating total score.

2.7. Rater training

Video-only raters completed rater training using a frame-of-reference training approach[28] conducted by the primary author in a single group session for all raters. Frame-of-reference training rater training included a standardized lecture, item by item review of the rating tool, and interactive discussion regarding scoring of specific rating tool items and participant actions during observation. Following the lecture, a single familiarization LP video was reviewed by raters in training who then independently scored standardized videos including completion of written skill performance total scores for each video. Raters in-training discussed rationale for scoring and interrater discrepancies after viewing each rater training video.

2.8. Sample size calculation

Sample size was estimated based on pilot study data from 4 experts and 3 novices. Sample size was estimated at 32 subjects (Novice 16, Expert 16) for significance $\alpha = 0.05$, power $1\text{-}\beta = 0.80$.

2.9. Statistical analysis

Chi-square was used to compare between group (novice vs expert) LP CSF success rate and Mann–Whitney *U* test to compare median LP skill performance rating score. Cronbach α and intraclass correlation coefficient were computed to assess interrater reliability within and between rater total scoring for video-only vs. direct observation raters. All analyses were completed using R (R core Team, 2018). A *P*-value of less than .05 was considered statistically significant.

3. Results

3.1. Study population

Thirty-five subjects (novice 19, expert 16) completed a total of 140 LP procedures (Table 1). The 16 experts were physicians with self-reported experience of performing more than 30 LP procedures. Expert physician specialties included emergency medicine (6; 37.5%), critical care (7; 46.7%), anesthesiology (2; 13.3%), and radiology (1; 6.7%). Novices were US medical students in 3rd (16) or 4th (3) year of medical school. Five novice LP attempts by 2 participants were excluded due to spinal needle placement outside of the boundaries of the manufacturer recommended simulated LP model block. One each using the normal, geriatric, and obesity blocks, and 2 for the geriatric/obesity block (Fig. 3). All participant LP attempts were via a midline approach. A total of 135 LP procedures were included in the analysis.

3.2. Construct validity

Expert and novice normal block CSF success rates were 16/16 (100%) and 9/18 (50%) respectively ($P = .001$); median total scores reported by all observers were 13 (range 11–13) and 7

Table 1

Demographics.

Characteristics	Novice (N = 19)	Expert (N = 16)
Mean age ± SD - yr	26.3 ± 2.4	44.1 ± 11.2
Male - no. (%)	7 (36.8)	12 (75)
3rd year medical student - no. (%)	16 (84.2)	
4th yr medical student - no. (%)	3 (15.8)	
Years in practice ± SD		15.5 ± 8.1
Specialty - no. (%)		
Critical care		7 (44)
Emergency		6 (38)
Anesthesiology		2 (13)
Radiology		1 (6.3)
Adult LP's performed - no. (%)		
0	18 (94.7)	0
Less than 10	1 (5.3)	0
Between 10 and 49	0	1 (6.3)
Between 50 and 99	0	8 (50)
More than 100	0	7 (43.8)
Adult LP's observed - no. (%)		
0	8 (42.1)	0
Less than 10	10 (52.6)	3 (18.8)
Between 10 and 49	1 (5.3)	4 (25)
Between 50 and 99	0	6 (37.5)
More than 100	0	3 (18.8)
LP training experience - no. (%)*		
Yes	4 (21.1)	16 (100)
No	15 (78.9)	0
LP simulator experience- no. (%)*		
Yes	3 (15.8)	5 (31.3)
No	16 (84.2)	11 (68.8)

LP = lumbar puncture, SD = standard deviation.
* Formal and informal training.

(range 5–11) respectively ($P = .001$). Expert and novice Geriatric block CSF success rate were 12/16 (75%) and 4/18 (22%), respectively ($P = .005$); median total scores were 12 (range 6–12) and 5 (range 4–6) ($P = .002$). Expert and novice Obesity block CSF success rates were 16/16 (100%) and 10/18 (56%), respectively ($P = .003$); median total scores were 13 (range 12–13) and 9 (range 4–12) ($P < .001$). Combined Geriatric/Obesity block expert and novice success rates were 6/16 (37.5%) and 2/17 (11.8%), respectively ($P = .12$); median total scores were 6 (range 5–11) and 6 (range 5–6) ($P = .44$) (Table 2).

3.3. Content and face validity

Expert median Likert scores were 4 (range 3–4) for simulator realism and 4 (range 4–5) for utility as a training tool. Novice Likert score was 4 for simulator realism. Only 1 novice rated realism of simulator, other novices had no real patient LP experience. Novice median score of utility as a training tool was 4 (range 4–5).

3.4. Assessment tool reliability

Cronbach α for all rater scoring was 0.96. Intraclass correlation [intraclass correlation coefficient (2, K)] was 0.98, indicating high intra- and inter-rater reliability.

4. Discussion

This cross-sectional study investigated construct and content validity of an LP simulator. This validation study confirms and extends results from prior studies to now include validation evidence regarding complex components of the LP simulator. Expert skill performance measured by LP CSF retrieval success rates and total skill performance scores for each block were higher than those of novice medical students. Performance outcomes differences for the combined Geriatric/Obesity block were not statistically significant. These results constitute strong evidence for construct validity of the LP simulator normal, obesity, and geriatric configurations. The trend in findings for the combined geriatric/obesity configuration suggests construct validity. Expert physicians found the LP simulator to be similar to real patients and effective for teaching LP skills, contributing evidence of content validity.

Our findings confirm prior validation of the normal anatomic configuration for this simulator,[24,26] and extend evidence of construct validity to the complex anatomy blocks. Validation of simulators is infrequently reported. Commercially available simulators were found to be lacking evidence of validation in 93.5% of cases in a 2014 report,[20] which concludes that lack of validity risks learning improper and incorrect skills among other educational hazards. A small proportion of task-trainer simulators were reported as validated in that report and in more recent LP simulator validation studies.[23,24,26]

Validity comprises a critical element of instructional design using mastery learning assessment and high-stakes simulation assessment.[29,30] Our validation evidence includes strong interrater reliability evidence, expert-novice comparison evidence, and content validation evidence through use of previously validated assessment items. The study represents a prototype simple approach for establishing evidence to support validity of a simulator and assessment tool as combined elements in the overall validity argument for effective employment in education activities. Our study examined only these accessible elements of the full range of available sources of evidence. As validity arguments are accumulated the strength of a simulator-based assessment system grows incrementally. Missing from our validity evidence are elements such as correlation with other measures of similar outcomes, clinical outcomes, and consequences such as impact on learner educational or clinical performance. We believe that our results justify further validation of the simulator/assessment tool for use in high stakes or formative evaluation studies including establishment of performance thresholds for incorporation in mastery learning protocols.

In obesity and geriatric configurations expert skill performance total scores and LP CSF retrieval success rates were significantly higher than those of novice medical students. However, there was no significant difference between experts and novices for the most complex and technically difficult obesity/geriatric anatomic configuration. This result indicates that this training model is not valid for skill assessment of LP with a high degree of technical difficulty. The CSF success rate and skill performance total score trend for this model did favor higher performance by experts and a larger study may reveal statistical support for construct validity for the obesity/geriatric high complexity model. Inter-rater reliability for skill performance total score confirmed reliability at a level constituting strong internal structure and evidence of validity. Extending the strength of interrater reliability with a larger cohort of trained raters would comprise a stronger generalizability validation argument. The total score results can be further validated and refined through consideration of psychometric analysis of each item, which would require a larger cohort of subjects. Total score validation could be

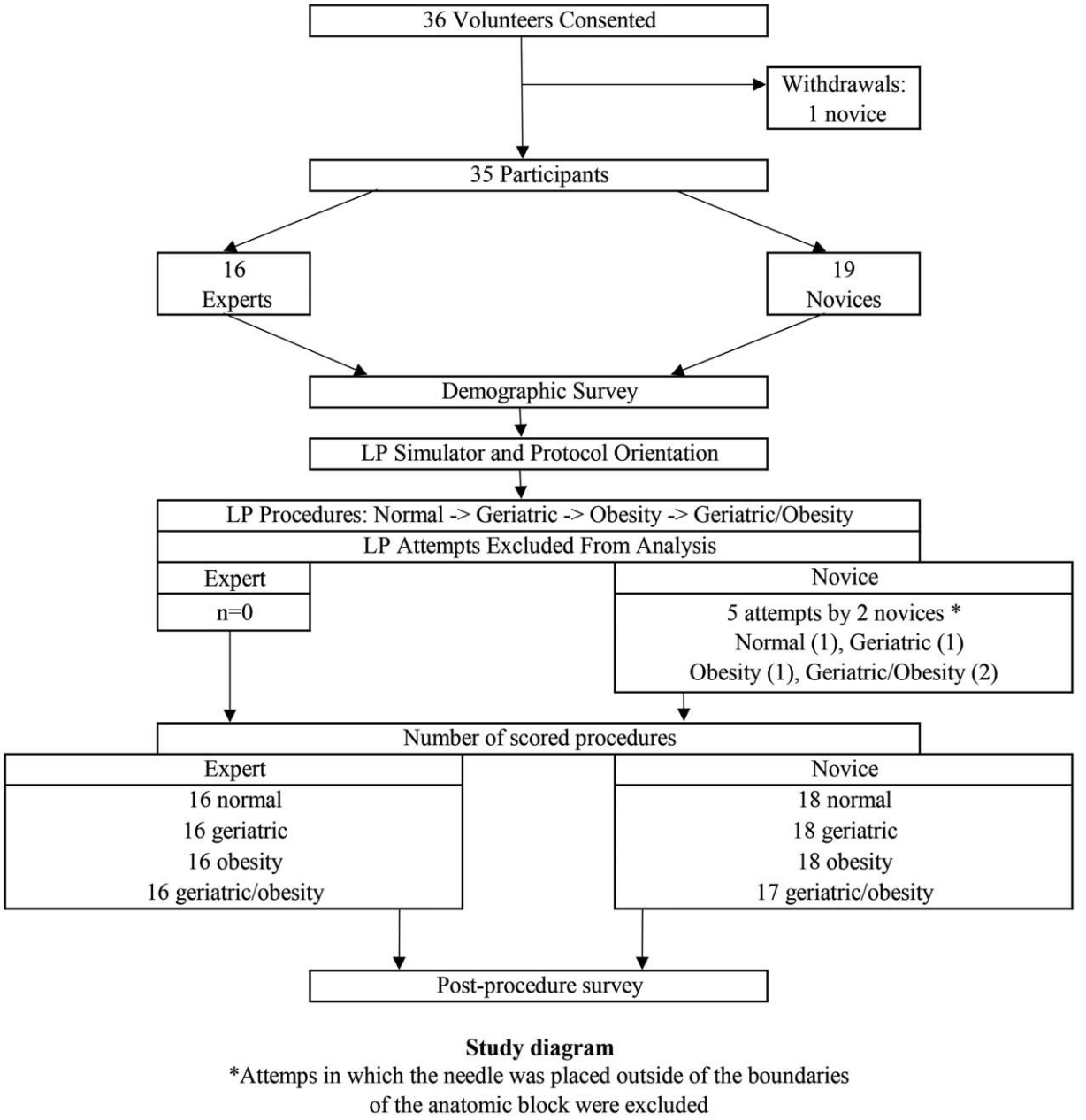

Figure 3. Study diagram. Thirty-five subjects (novice 19, expert 16) completed a total of 140 LP procedures. Five novice LP attempts by 2 participants were excluded due to spinal needle placement outside of the boundaries of the manufacturer recommended simulated LP model block. One each using the normal, geriatric, and obesity blocks, and 2 for the geriatric/obesity block. LP = lumbar puncture.

accomplished by completion of a similar study using the same LP skills performance checklist and one or more different LP task training devices.

Experts reported on the post procedure survey that the LP simulator is sufficiently realistic for use as a training and assessment device. Freeform written and verbal response to inquiries regarding perceptions included comments that realism was lacking in domains regarding motion and pain reactions. This represents Face Validity, a construct recognized as a relatively weak form of evidence in validation studies, yet which serves a useful purpose in the early stage development of simulators and for assessment protocols utilizing simulators. Incorporation of rigorous Face Validity evidence in future studies may be strengthened by application of quantitative analytic methods.

4.1. Limitations

Novice-expert cohort comparisons such as this study support validation through generation of widely divergent performance scores and other assessment parameters. This strategy leaves a void in understanding the ability of the simulator and other elements of assessment protocols to distinguish less pronounced performance differences.[31] We report the first validation study which incorporates this simulator's complex anatomy. We also gathered validity evidence to inform the utility and advisability of

Table 2

CSF success rate and total skill performance score.

Construct validity	Novice (N = 19)	Expert (N = 16)	*P*-value
Success rate - no. (%)			
Normal	9/18 (50)	16/16 (100)	.001
Geriatric	4/18 (22.2)	12/16 (75)	.005
Obesity	10/18 (55.6)	16/16 (100)	.003
Geriatric obesity	2/17 (11.8)	6/16 (37.5)	.118
Median total score (IQR)			
Normal	7 (5–11)	13 (11–13)	.001
Geriatric	5 (4–6)	12 (6–12)	.002
Obesity	9 (4–12)	13 (12–13)	<.001
Geriatric obesity	6 (5–6)	6 (5–11)	.44

Five novice LP attempts were excluded (n).
Normal (1), Geriatric (1), Obesity (1), Geriatric/Obesity (2).
CSF = cerebrospinal fluid, IQR = interquartile range.

more detailed study with a larger and more representative subject learner cohort, and for more comprehensive performance, for example including sterile technique, and anesthetic administration, and CSF pressure measurement skills. Our results suggest the assessment tool/performance checklist can be used to assess the construct validity of other LP simulators but should not be considered validated for individual learner comprehensive LP performance assessment, without additional evidence incorporating other key elements of LP performance.

The lack of validation evidence for the combined obesity/geriatric configuration is likely a reflection of our small sample size, based on an estimate from a limited cohort pilot study and published data regarding the normal anatomic configuration.

Identification and scoring of specific lumbar spine vertebral interspace site selection for LP needle insertion was subjective. This LP simulator has no insertion site sensor to definitively identify interspace needle penetration. Investigators confirmed that the center of the simulated skin attachment overlay the simulated L3/4 interspace, but no identifying surface marking was apparent to participants or raters. Video-only raters were oriented to this location as the primary reference point for scoring of needle insertion site selection by participants. However, this specific rater skill was not verified during rater training, introducing the possibility of rater error as a limitation. Future considerations for use of this combined rating tool and simulator for high stakes or mastery learning assessments should include rater calibration including accuracy of rater identified intervertebral space needle puncture sites.

A learning effect was likely present since participants completed 4 sequential LP exercises, potentially impacting differential success rates and time to success with the various anatomic models; our study was not powered to detect these potential differences. This limitation was recognized a-priori and addressed in the study protocol design, in which an identical sequence of LP skill performance anatomic variants was presented to both novices and experts.

Finally, we cannot exclude bias based upon the inability to blind the direct observer to a novice or an expert participant status, this was mitigated by the blinding of video-only observers.

5. Conclusion

For the simple task of obtaining CSF via standard LP technique for normal and complex anatomy the combined assessment protocol elements of the LP simulator, rating tool, and rater training protocol utilized in this study have strong evidence for construct validity for each of the anatomic models, except for the combined obesity/geriatric complex anatomic representation. Face and content validity are likewise supported by the results of this study. Our findings suggest that the 3 components of the simulator-based skill performance assessment can be used for assessment. This simulator-based assessment protocol validation supports further investigation of validity in studies designed to establish high stakes testing parameters for clinical skills validation and to establish valid minimum passing scores for mastery learning protocols including simple and complex LP anatomy.

Author contributions

Conceptualization: Yu Akaishi, Benjamin Worth Berg.
Data curation: Yu Akaishi.
Formal analysis: Yu Akaishi, Yuito Okada.
Investigation: Yu Akaishi, Jun Seok Seo, Benjamin Worth Berg.
Methodology: Yu Akaishi, Yuito Okada, Jannet Lee-Jayaram, Jun Seok Seo, Toru Yamada, Benjamin Worth Berg.
Project administration: Jannet Lee-Jayaram, Benjamin Worth Berg.
Resources: Yu Akaishi, Benjamin Worth Berg.
Software: Yu Akaishi, Yuito Okada.
Supervision: Benjamin Worth Berg.
Validation: Yu Akaishi, Benjamin Worth Berg.
Visualization: Yu Akaishi, Benjamin Worth Berg.
Writing – original draft: Yu Akaishi.
Writing – review & editing: Yu Akaishi, Yuito Okada, Jannet Lee-Jayaram, Jun Seok Seo, Toru Yamada, Benjamin Worth Berg.

References

[1] Wayne DB, Didwania A, Feinglass J, et al. Simulation-based education improves quality of care during cardiac arrest team responses at an academic teaching hospital: a case-control study. Chest 2008;133:56–61.
[2] 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;86:706–11.
[3] Ziv Stephen D Small Paul Root Wolpe A. Patient safety and simulation-based medical education. Med Teach 2000;22:489–95.
[4] Berg KT, Mealey KJ, Weber DE, et al. Are medical students being taught invasive skills using simulation? Simul Healthc 2013;8:72–7.
[5] Kneebone RL. Practice, rehearsal, and performance: an approach for simulation-based surgical and procedure training. JAMA 2009;302:1336–8.
[6] Barsuk JH, Cohen ER, Caprio T, et al. Simulation-based education with mastery learning improves residents' lumbar puncture skills. Neurology 2012;79:132–7.
[7] Sun C, Qi X. Evaluation of problem- and simulator-based learning in lumbar puncture in adult neurology residency training. World Neurosurg 2018;109:e807–11.
[8] McMillan HJ, Writer H, Moreau KA, et al. Lumbar puncture simulation in pediatric residency training: improving procedural competence and decreasing anxiety. BMC Med Educ 2016;16:198.
[9] Conroy SM, Bond WF, Pheasant KS, et al. Competence and retention in performance of the lumbar puncture procedure in a task trainer model. Simul Healthc 2010;5:133–8.
[10] Barsuk JH, McGaghie WC, Cohen ER, et al. Simulation-based mastery learning reduces complications during central venous catheter insertion in a medical intensive care unit. Crit Care Med 2009;37:2697–701.
[11] Barsuk JH, McGaghie WC, Cohen ER, et al. 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;4:397–403.

[12] Evans LV, Dodge KL, Shah TD, et al. Simulation training in central venous catheter insertion: improved performance in clinical practice. Acad Med 2010;85:1462–9.
[13] Van Nortwick SS, Lendvay TS, Jensen AR, et al. Methodologies for establishing validity in surgical simulation studies. Surgery 2010;147: 622–30.
[14] Ramos P, Montez J, Tripp A, et al. Face, content, construct and concurrent validity of dry laboratory exercises for robotic training using a global assessment tool. BJU Int 2014;113:836–42.
[15] Cook DA, Hatala R. Validation of educational assessments: a primer for simulation and beyond. Adv Simul (Lond) 2016;1:31.
[16] Cook DA, Brydges R, Ginsburg S, et al. A contemporary approach to validity arguments: a practical guide to Kane's framework. Med Educ 2015;49:560–75.
[17] American Educational Research Association APAStandards for Educational and Psychological Testing. Washington, DC: National Council on Measurement in Education; 2014.
[18] Kane MT. Validating the interpretations and uses of test scores. J Educ Meas 2013;50:1–73.
[19] Sullivan GM. A primer on the validity of assessment instruments. J Grad Med Educ 2011;3:119–20.
[20] Stunt J, Wulms P, Kerkhoffs G, et al. How valid are commercially available medical simulators? Adv Med Educ Pract 2014;5:385–95.
[21] McDougall EM, Corica FA, Boker JR, et al. Construct validity testing of a laparoscopic surgical simulator. J Am Coll Surg 2006;202:779–87.
[22] Duffy AJ, Hogle NJ, McCarthy H, et al. Construct validity for the LAPSIM laparoscopic surgical simulator. Surg Endosc 2005;19:401–5.
[23] Corvetto MA, Fuentes C, Araneda A, et al. Validation of the imperial college surgical assessment device for spinal anesthesia. BMC Anesthesiol 2017;17:131.
[24] Henriksen MJV, Wienecke T, Thagesen H, et al. Assessment of residents readiness to perform lumbar puncture: a validation study. J Gen Intern Med 2017;32:610–8.
[25] Issenberg SB, Ringsted C, Ostergaard D, et al. Setting a research agenda for simulation-based healthcare education: a synthesis of the outcome from an Utstein style meeting. Simul Healthc 2011;6:155–67.
[26] Uppal V, Kearns RJ, McGrady EM. Evaluation of M43B Lumbar puncture simulator-II as a training tool for identification of the epidural space and lumbar puncture. Anaesthesia 2011;66:493–6.
[27] Cheng A, Kessler D, Mackinnon R, et al. Reporting guidelines for health care simulation research: extensions to the CONSORT and STROBE statements. Simul Healthc 2016;11:238–48.
[28] Feldman M, Lazzara EH, Vanderbilt AA, et al. Rater training to support high-stakes simulation-based assessments. J Contin Educ Health Prof 2012;32:279–86.
[29] Lineberry M, Soo Park Y, Cook DA, et al. Making the case for mastery learning assessments: key issues in validation and justification. Acad Med 2015;90:1445–50.
[30] Boulet JR. Summative assessment in medicine: the promise of simulation for high-stakes evaluation. Acad Emerg Med 2008;15:1017–24.
[31] Cook DA. Much ado about differences: why expert-novice comparisons add little to the validity argument. Adv Health Sci Educ Theory Pract 2015;20:829–34.