Development and Validation of an Assessment of Regional Anesthesia Ultrasound Interpretation Skills

Glenn E. Woodworth, MD, Patricia A. Carney, PhD,* Joshua M. Cohen, MD,† Sandy L. Kopp, MD,‡ Lindsey E. Vokach-Brodsky, MD,§ Jean-Louis E. Horn, MD,§ Andres Missair, MD,|| Shawn E. Banks, MD,|| Nathan F. Dieckmann, PhD,** and Robert B. Maniker, MD††*

Background: Interpretation of ultrasound images and knowledge of anatomy are essential skills for ultrasound-guided peripheral nerve blocks. Competency-based educational models promoted by the Accreditation Council for Graduate Medical Education require the development of assessment tools for the achievement of different competency milestones to demonstrate the longitudinal development of skills that occur during training.

Methods: A rigorous study guided by psychometric principles was undertaken to identify and validate the domains and items in an assessment of ultrasound interpretation skills for regional anesthesia. A survey of residents, academic faculty, and community anesthesiologists, as well as video recordings of experts teaching ultrasound-guided peripheral nerve blocks, was used to develop short video clips with accompanying multiple choice–style questions. Four rounds of pilot testing produced a 50-question assessment that was subsequently administered online to residents, fellows, and faculty from multiple institutions.

Results: Test results from 90 participants were analyzed with Item Response Theory model fitting indicating that a 47-item subset of the test fits the model well ($P = 0.11$). There was a significant linear relation between expected and predicted item difficulty ($P < 0.001$). Overall test scores increased linearly with higher levels of formal anesthesia training, regional anesthesia training, number of ultrasound-guided blocks performed per year, and a self-rating of regional anesthesia skill (all $P < 0.001$).

Conclusions: This study provides evidence for the reliability, content validity, and construct validity of a 47-item multiple choice–style online test of ultrasound interpretation skills for regional anesthesia, which can be used as an assessment of competency milestone achievement in anesthesiology training.

(*Reg Anesth Pain Med* 2015;40: 306–314)

Ultrasound imaging allows anesthesiologists to monitor needle position and local anesthetic spread relative to visualized nerves and adjacent structures while performing peripheral nerve blocks (PNBs). The addition of ultrasound guidance to the performance of PNBs has been demonstrated to improve efficacy, efficiency, and safety.[1,2] Ultrasound-guided regional anesthesia (UGRA) is growing in popularity and may become the principal technique for performing PNBs.[3–5] Concurrently, medical education is evolving from the traditional time-based model to a competency-based model with defined educational milestones and competency requirements.[6,7] Accreditation and certification bodies are changing requirements for both graduate medical education and postgraduate demonstration of continued competency for maintenance of certification. In the United States, the Accreditation Council for Graduate Medical Education (ACGME) and the American Board of Medical Specialties (ABMS) have developed definitions of competency to stress educational outcomes and include assessment of competency for specific milestones.[8] Residency programs began a phased implementation of the NAS requirements in 2013.

Key elements of competency in UGRA include understanding relevant anatomy, ability to acquire an image of the target nerve or structure with ultrasound, ability to interpret the ultrasound image, ability to safely guide a needle to the target, and the ability to avoid complications by recognizing aberrant anatomy.[9,10] Ultrasound-guided regional anesthesia is a complex task involving multiple skills; however, knowledge of anatomy and the corresponding ultrasound imagery are key components. Because the ability to interpret ultrasound images is critical to UGRA, an assessment of competency in this subskill could be used as a milestone assessment for technical skills in regional anesthesia. Few validated assessments of anesthesia competency have been developed with sound psychometric principles, and none focuses on regional anesthesia ultrasound interpretation skills.[11] We hypothesized that an online assessment using video clip–based multiple-choice questions could distinguish between subjects with varied UGRA skills and could be used to assess learner milestone achievement. The current study also describes the use of psychometric principles to develop and validate the assessment.[11–13]

METHODS

All study activities were approved by the institutional review boards of Oregon Health and Science University; Columbia University; Stanford University; University of California at San Francisco; Mayo Clinic, Rochester; and the University of Miami. These universities are geographically disparate and represent a wide range of faculty and resident clinical experience in the United States (>350 combined anesthesia residents).

The quality of an assessment is based on the psychometric principles of reliability and validity.[11–14] Reliability refers to the reproducibility of a test. External reliability of a test constructed of multiple-choice questions is determined by test-retest reproducibility of results. Internal reliability or consistency is evaluated by

From the *Department of Anesthesiology and Perioperative Medicine, Oregon Health and Science University, Portland, OR; †University of California at San Francisco, San Francisco, CA; ‡Department of Anesthesiology, Mayo Clinic, Rochester, MN; §Department of Anesthesiology, Perioperative and Pain Medicine, Stanford University School of Medicine, Stanford, CA; ||Department of Anesthesiology, University of Miami Miller School of Medicine, Miami, FL; **Statistics Core, School of Nursing, Oregon Health and Science University, Portland, OR; and ††Department of Anesthesiology, Columbia University, New York, NY.

Accepted for publication February 6, 2015.

Address correspondence to: Glenn E. Woodworth, MD, Department of Anesthesiology and Perioperative Medicine, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd, Portland, OR 97239
(e-mail: woodworg@ohsu.edu).

This work was funded in part by a grant from the Foundation for Anesthesia Education and Research, Rochester, MN, and Oregon Health and Science University, Department of Anesthesiology and Perioperative Medicine, Portland, OR.

This work was presented as an abstract/moderated poster at the 39th Annual Regional Anesthesia and Acute Pain Medicine Meeting of the American Society of Regional Anesthesia and Pain Medicine, April 5, 2014, in Chicago, IL.

ISSN: 1098-7339
DOI: 10.1097/AAP.0000000000000236

determining whether examinees of similar competency tend to perform similarly on the same parts of the test.

The validity of a test is determined by evidence indicating that the test is measuring what it intends to measure. Measures of validity and their application are listed in Table 1. Content validity of an assessment is typically determined by steps undertaken to ensure that the appropriate content domains or constructs are covered in the assessment. The extent to which an assessment correlates with a concurrent measure of the skill being assessed is evidence of concurrent validity. This is particularly important if the assessment is correlated with a gold standard criterion.[11,15] When a gold standard criterion does not exist, as in the case with UGRA skill, correlation with multiple criteria that reflect the skill being measured can provide evidence of the assessment's construct validity.[11,15,16] The multistep process for developing the assessment of regional anesthesia ultrasound interpretation skills and the evidence supporting the reliability and validity of the assessment are described below.

Step 1: Domain Generation

The purpose of this step is to determine the sphere of knowledge to be included in the assessment. Domain generation began with 3 academic experts in regional anesthesia reviewing published curricula, residency review committee milestones, published assessments of regional anesthesia knowledge and skills, published keywords, and in-training examination topics.[9,10] This yielded a comprehensive list of regional anesthesia knowledge and skills to be included in an assessment. A survey instrument was developed, and a recruitment e-mail was sent to 20 ACGME-accredited anesthesia residency programs and 3 large community practice groups (Oregon, Washington, California). Academic faculty, PGY 4 anesthesia residents, and community anesthesiologists were eligible to be enrolled. They were asked to identify which items that residents graduating from an ACGME-accredited anesthesia residency should know and which tasks they should be able to perform. A threshold of greater than 75% of participants indicating that graduating residents should be able to perform a particular UGRA block was used to identify domains included in the assessment tool.

Step 2: Item Generation

In the first phase of item generation, 2 experts in regional anesthesia examined anatomy texts and published descriptions of UGRA blocks to create a list of anatomical structures that were within 1 cm of the target nerves identified in domain generation. Additional anatomical structures were included if they were within 1 cm of commonly described ultrasound-guided needle paths to the target nerves.[17–23]

In the second phase of item generation, participants were recruited from 5 ACGME-accredited anesthesiology programs. For residents to be eligible, they were required to have performed fewer than 10 USGR nerve blocks. Two faculty experts and 2 residents were studied at each institution. Each faculty expert was paired with a novice resident and observed during a day of routine clinical care on the regional anesthesia service. Both were blinded to the study's true purpose and were informed that the study involved recording communication styles between the resident and faculty member. During the course of the clinical day, the pair was video recorded performing UGRA blocks in the domains identified in Step 1 (interscalene, supraclavicular, femoral, or sciatic).

In addition to video recording during clinical care, the faculty expert was video recorded giving a short lecture and demonstration using a live model on each of the designated nerve blocks to an audience of novice anesthesia residents. The faculty expert was blinded to the true purpose of this portion of the study and was informed that the study involved recording the different communication styles of faculty members.

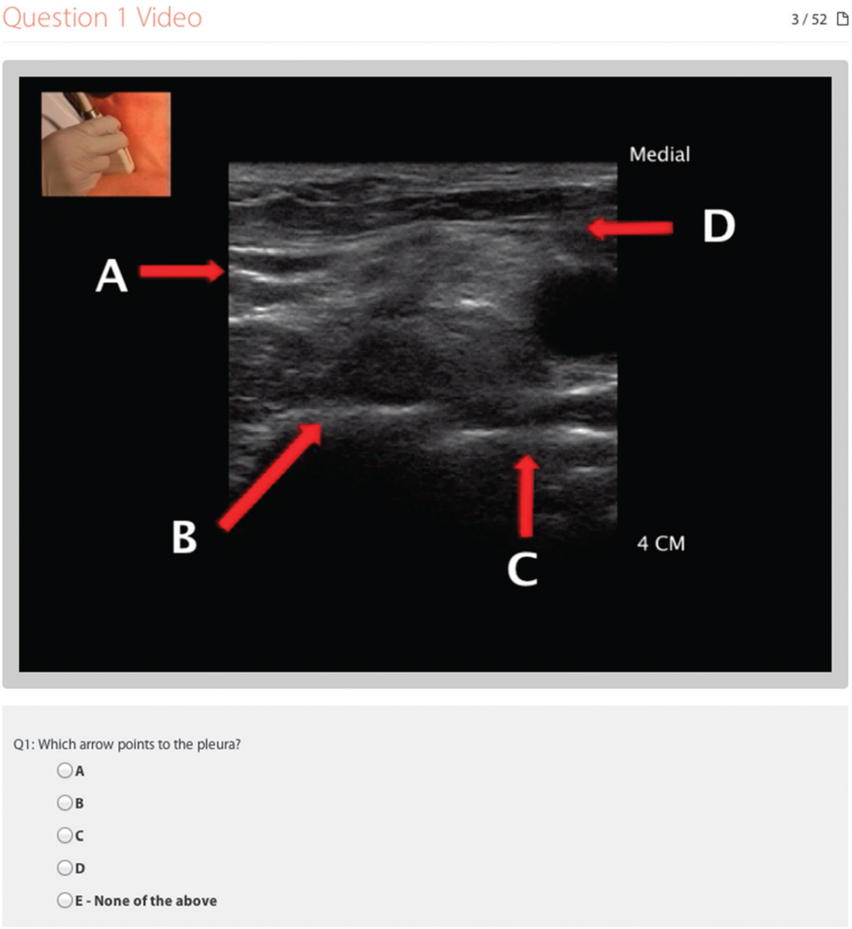
The list of anatomical structures developed by review of anatomy texts and literature was combined with the list of anatomical structures and relationships mentioned during video recording of experts during teaching. The combined list was used to select video clips from ultrasound examinations of the neck, inguinal region, and posterior thigh, which demonstrated relevant structures, features, and relationships. Test questions were created by selecting a 5-second segment from each clip. Each clip was annotated with appropriate depth markers and designations to orient the examinee (eg, medial, lateral, superior). In addition, a picture demonstrating the location and orientation of the ultrasound transducer used to generate the clip was added to the clip. The final frame of the clip was frozen and up to 5 arrows were added pointing to various anatomical structures or distractors. The frozen frame along with any arrows would play for an additional 5 seconds. The questions were presented in a multiple-choice question format with a stem and 5 answer choices (Fig. 1). If the ultrasound transducer was moved during the clip, the stem included a description of the transducer movement (eg, "in this clip, the ultrasound transducer is placed in the groin in an axial plane and gradually translated laterally").

Step 3: Pilot Testing and Revision

For faculty members to be eligible for the pilot testing phase, they were required to self-identify themselves as experts in regional anesthesia. Residents were required to be in an ACGME-accredited residency. The assessment was administered via a Web-based application (MobilePaks, Inc, Portland, Oregon) at a time and place of the subject's own choosing. Participants were unsupervised during testing. Participants were instructed to take

TABLE 1. Measures of Validity

Item	Definition or Application
Face validity	A general impression that the assessment seems appropriate. Evidence for face validity lies in expert and subject agreement that the assessment is appropriate.
Content validity	The assessment tests the content that is being taught.
Concurrent validity	The results of the assessment agree with other established measures of the skill.
Construct validity	If a gold standard for expertise cannot be used for comparison, a surrogate measure for the construct of expertise is used. The assessment should be able to differentiate between various levels of the surrogate measure of expertise. In education evaluations, experience is often used as the surrogate measure.
Predictive validity	The assessment should be able to predict the outcome of some future result (clinical outcome, future assessment results).



Question 1 Video

Q1: Which arrow points to the pleura?

○ A

○ B

○ C

○ D

○ E - None of the above

FIGURE 1. Sample final frame of an ultrasound clip and test question.

no longer than 2 hours to complete the assessment. Initial pilot testing determined that 2 hours was sufficient to complete the test. Participants were advised that all results were confidential, and the investigators were blinded to the identity of the participants. They were further advised that results would only be used to create a pilot test and were not part of their residency or faculty evaluation process. Participants were instructed that they were not expected to know the answers to all the questions; however, they should make their best effort to answer each question without consulting any outside information or individuals. They were also advised that unanswered questions would be scored as incorrect and that consulting outside sources, taking longer than 2 hours, or leaving questions unanswered that they could have answered correctly could affect analysis of the data and substantially affect future expectations of what anesthesiologists should know. In addition to taking the assessment, the faculty experts were asked to rate the difficulty of each question according to what level of trainee should be able to answer the question (Junior, trainee has not completed a rotation in regional anesthesia; Mid, trainee has completed a rotation in regional anesthesia; Senior, PGY 4 who has completed a rotation in regional anesthesia; Advanced, completed residency; Expert, completed a regional fellowship or >5 years of postgraduate experience with UGRA).

After completing the assessment, each faculty expert was interviewed by the principal investigator to discuss each question. The 5 faculty experts were supplied with a summary of the percentage of correct responses for each question from all faculty experts and the resident participants and the average difficulty rating for each question assigned by the experts. The faculty member experts reviewed the results for each question and made suggestions, changing the assigned difficulty rating and revision or elimination of each question. Any question that was answered incorrectly by a faculty expert was eliminated or revised to create the next version of the assessment while assuring the entire test adequately covered the domain. Faculty members could also suggest the inclusion of new questions.

The revised assessment underwent additional rounds of review in the same manner as the first round; however, they did not include residents and used a new expert panel with no prior exposure to the assessment. The process continued with additional rounds until all faculty member experts answered all questions correctly, additional questions were not suggested, and the faculty did not recommend revisions to any questions. The final list of questions was used in the validation phase of the study.

Step 4: Assessment Validation

Faculty, residents, and fellows were recruited via e-mail solicitation from 20 ACGME-accredited anesthesiology residency programs. The assessment was administered via a Web-based application in the same manner and with the same instructions as in Step 3; however, before taking the assessment, all participants filled out a 5-question demographic survey to determine training status and experience with UGRA. Demographic questions identified the level of completed anesthesia training and experience with regional anesthesia.

Statistical Analysis

A sample size in the range of 100 can result in relatively stable Rasch parameter estimates and was used to determine the number of participants enrolled for statistical analysis in the validation phase.[24] The goal of the psychometric analysis was to retain as many items as possible but only include those items that are able to discriminate between test takers of differing skills. Traditional methods of scale construction based on classical test theory (eg, Cronbach α) are ill suited to this task because they do not provide tests of individual item functioning or overall model testing. Methods based on the item-response theory (IRT) allow the examination of overall model fit and the extent to which individual test items discriminate between test takers who vary on the latent trait.[25] One IRT method is the 1-parameter logistic model, also known as the Rasch model.[26] In this model, subject responses are viewed as the result of both the test taker's ability level (ie, standing on the latent trait) and the difficulty of the item. A logistic function is then fit to estimate the probability that an individual test taker will correctly respond to each test item. Through this process, one can characterize the extent to which each item discriminates between test takers of differing ability levels. Items that discriminate between test takers of differing skill can be retained and those who do poorly can be removed from the test. Our goal was to retain a set of items that had high discrimination and, collectively, spanned the range of the latent trait.

Item-response theory analyses were conducted using the ltm and eRm packages available in the R statistical computing environment (R Core Team, Foundation for Statistical Computing 2013, Vienna, Austria).[27,28] The Rasch model was fit to the full 50-item test, and the overall model was evaluated. Model fit tests were done with parametric Bootstrap goodness-of-fit tests using the Pearson χ^2 statistic. In the case of poor model fit ($P < 0.05$), we tested higher parameter IRT models (ie, unconditional Rasch and 2-parameter models). We then examined individual item performance to identify poorly performing items. At the item level, we examined the significance of χ^2 statistics that indicate the contribution of an individual item to poor model fit. We also examined infit and outfit mean-square statistics, where values greater than or equal to 1.5 were indicative of poor item performance (eg, an item that is generally difficult is responded to correctly by several low-ability participants or vice versa). High infit/outfit values indicate an item that is performing in an unpredictable and/or counterintuitive way. Items that had significant P values and MSQ infit or oufit statistics greater than 1.5 were removed from the test. After removal of poor-performing questions, we examine the total test scores with respect to demographic factors as additional evidence of construct validity.

RESULTS

Step 1: Domain Generation

Eighty-six participants were enrolled in the knowledge and skills survey, with at least 12 members from each of 5 stakeholder groups. Greater than 85% of participants indicated that graduating residents should be able to perform femoral, popliteal, interscalene, supraclavicular, and axillary blocks under ultrasound guidance (Table 2). Based on these results, 3 question domains were selected, namely, femoral nerve (inguinal triangle), popliteal sciatic nerve (posterior thigh), and brachial plexus blocks above the clavicle. Although the survey identified axillary blocks as a domain, it was not included to keep the anticipated time to complete the study assessment less than 2 hours. An additional domain for the sciatic nerve in the gluteal and proximal posterior thigh was also included to provide questions that would distinguish expertise from proficiency in regional anesthesia interpretation skills. This process provides evidence of face and content validity for the developed assessment.

Step 2: Item Generation

The anatomical domains identified in Step 1 were used to guide review by experts of anatomy texts and published literature. This analysis yielded 21, 31, and 32 relevant anatomical structures in the inguinal region, posterior thigh, and neck, respectively. The list was subsequently expanded to include anatomical relationships between these structures. For example, the original list included the common femoral artery and vein. The expanded list included the anatomical relationship that the common femoral artery is lateral to the vein. The expanded list included 44, 47, and 43 relevant structures and anatomical relationships in the inguinal region, posterior thigh, and neck, respectively.

In the second phase of item generation, 2 faculty experts and 2 novice residents from 5 different institutions were recruited for the study. One faculty expert dropped out prior to the start of the study due to an unforeseen scheduling conflict. The 9 faculty/resident pairs were recorded performing 36 ultrasound-guided PNBs (9 femoral, 15 popliteal, 1 transgluteal, 11 interscalene). Forty-five, 47, and 64 different anatomical references to items relating to femoral, posterior thigh, and interscalene/supraclavicular blocks, respectively, were identified in the video recordings.

TABLE 2. Knowledge and Skills Survey Results

Group	N	Fem, %	T-Glut, %	I-Glut, %	Pop, %	ISB, %	Supra, %	Infra, %	Ax, %	Tap, %	PVB, %
Community	16	87	37	37	94	94	81	40	80	47	47
Expert community	17	87	49	68	87	87	87	65	88	94	59
Academic	11	91	10	31	70	81	81	30	70	70	30
Expert academic	21	95	47	71	95	95	95	71	95	81	62
Sr. resident	21	91	18	38	76	91	91	33	90	67	33
Total	86	90	34	51	85	90	88	50	86	72	47

Percent of respondents within each stakeholder group indicating that graduating residents should be able to perform these blocks with ultrasound guidance.

Fem indicates femoral; T-Glut, transgluteal sciatic; I-Glut, infragluteal sciatic; Pop, popliteal sciatic; ISB, intersclalene; Supra, supraclavicular; Infra, infraclavicular; Ax, axillary; TAP, transversus abdominis plane; PVB, thoracic paravertebral.

The anatomical list developed by review of anatomy texts and literature, as well as the anatomical structures and relationships mentioned during video recording of experts during teaching, was used to select video clips from ultrasound examinations of the neck, inguinal region, and posterior thigh, which demonstrated relevant structures, features, and relationships. Eighty-one questions were created from these clips.

Step 3: Pilot Testing and Revision

The preliminary 81-question pilot assessment was administered to 7 residents (2 CA3 residents who had completed 8 weeks of regional anesthesia training, 4 CA2 residents who had completed 4 weeks of regional anesthesia training, and 1 CA1 resident who had not participated in any special training regarding regional anesthesia) and 5 regional anesthesia faculty experts. Based on questions answered incorrectly by any expert and debriefing of the experts, the assessment was revised.

Three additional rounds of pilot testing were completed. At the end of the fourth round, all faculty member experts answered all questions correctly without suggesting additional questions or recommending revisions to any questions. The final pilot assessment consisted of 50 questions. The 31 eliminated questions were covered by other questions in the revised assessment. The process of multiple rounds of expert review of the assessment provides further evidence of face and content validity.

Step 4: Assessment Validation

Sample

A convenience sample of 121 participants was enrolled between October 1, 2013, and February 21, 2014. Thirty-one participants were excluded from the primary analysis. Fourteen participants registered but did not complete any of the questions. An additional 11 participants registered but attempted less than 10 of the 50 questions. Six participants did not finish the test and left more than 15 questions blank at the end of the test. The primary analyses were conducted on the remaining 90 participants who finished the test and completed at least 80% of the questions. We conducted sensitivity analyses by including the 6 participants who completed a substantial portion of the items but did not finish the test. The results (not reported) were not substantively altered by the inclusion of these participants. Thus, we report only the primary analysis below. Subject demographic data are reported in Table 3. Sixty-three percent of participants were in residency, 21% in a fellowship, and 16% were postgraduates. Forty-four percent of participants self-identified themselves as novice or proficient with regional anesthesia.

IRT Analyses

A traditional Rasch IRT model was fit to the data to examine item performance and overall model fit. This Rasch fit the test data poorly ($P = 0.04$), although the unconstrained Rasch model and the 2-parameter logistic IRT model provided even poorer fit. Table 4 shows the raw percentage correct, the item difficulty and item fit statistics from the traditional Rasch model, and the predicted difficulty of each item based on judgment by the experts that generated the test items. Item-response theory difficulty parameters are in probability form, indicating the probability that a participant with average ability would get the question correct. Three items (15, 42, and 36) were identified as having significant contributions to a poor model fit and infit or outfit values that were greater than 1.5 and were therefore removed from the test.

The traditional Rasch model on the 47-item test fit the data well ($P = 0.11$). None of the remaining 47 items had significant P values and MSQ infit or outfit statistics greater than 1.5. Figure 2 shows the item characteristic curves (ICCs) and item and test information curves, identifying the existence of items that maximally discriminate at both the high and low levels of ability. This suggests that there are many easy and difficult items to discriminate test takers with different ability levels.

TABLE 3. Demographic Characteristics of Step 4 Validation Participants (N = 90)

Level of formal regional anesthesia training	
Completed clinical base year	20%
Completed CA1	13%
Completed CA2	29%
Completed residency	8%
Current regional fellow	18%
Completed regional fellowship	12%
Regional training during residency	
No specialized training	31%
Completed core	34%
Completed advanced	4%
Current regional fellow	18%
Completed regional fellowship	12%
Current practice status	
In residency	63%
Current fellow	21%
Completed residency	16%
UGRA PNB per year	
0	17%
1–20	29%
20–50	17%
60–100	9%
>100	29%
Regional anesthesia self-rating	
Novice with limited experience	33%
Proficient—comfortable performing "basic" regional anesthesia techniques	21%
Experienced—comfortable with advanced regional anesthesia techniques	32%
Expert—highly competent in advanced blocks, ultrasound, and catheter techniques	13%

We also tested the relation between the expected difficulty of the items based on expert judgment (see the Expected Difficulty column in Table 4) and the observed item difficulty based on percentage correct. There was a significant linear relation ($P < 0.001$) between expected and actual item difficulty (mean percentage correct: Adv/expert, 44.33; Senior, 54.94; Mid/Junior, 68.05), indicating that the observed item difficulty was closely aligned with the expected difficulty assigned by the experts.

Construct Validity

The total score on the 47-item test was calculated for each subject. Table 5 shows the mean and median total test scores on the 47-item test by subject demographic factors. If the test is accurately measuring ultrasound interpretation skill, we would expect participants with more formal training to score higher on the test. Test scores increased linearly at higher levels of formal anesthesia training, regional anesthesia training, ultrasound experience per year, and a self-rating of regional anesthesia skill

TABLE 4. Expected Difficulty and Rasch Item Fit Statistics (N = 90)

Q	Expected Difficulty	% Correct (raw)	IRT Difficulty Parameter	Outfit MSQ	Infit MSQ	*P*
12	Mid	93.3	0.95	0.57	0.90	>0.99
31	Mid	83.3	0.87	0.90	0.94	0.72
15	Junior	82.2	0.86	**1.81**	1.04	**<0.001**
1	Mid	80	0.84	0.95	0.90	0.57
23	Junior	80	0.84	0.89	0.96	0.75
27	Senior	78.9	0.83	0.74	0.92	0.96
21	Mid	78.9	0.83	1.16	1.05	0.13
16	Mid	77.8	0.82	1.06	0.87	0.31
9	Mid	77.8	0.82	0.60	0.82	>0.99
26	Mid	75.6	0.80	1.25	1.22	0.05
8	Mid	75.6	0.80	0.73	0.84	0.97
17	Senior	72.2	0.76	0.82	0.93	0.88
46	Mid	70	0.74	0.83	0.90	0.86
22	Mid	70	0.74	0.85	0.92	0.83
42	Mid	68.9	0.73	**1.69**	1.03	**<0.001**
38	Mid	67.8	0.71	0.93	0.95	0.63
32	Mid	66.7	0.70	0.86	0.92	0.81
24	Senior	66.7	0.70	0.84	0.89	0.84
41	Mid	66.7	0.70	1.09	1.06	0.25
43	Senior	65.6	0.69	0.73	0.83	0.97
29	Senior	62.2	0.65	0.88	0.93	0.77
50	Mid	60	0.63	0.97	1.00	0.53
45	Senior	60	0.63	0.95	0.97	0.59
37	Senior	58.9	0.61	1.06	1.04	0.30
33	Adv	57.8	0.60	1.14	1.15	0.16
7	Mid	57.8	0.60	0.91	0.94	0.70
34	Mid	56.7	0.59	1.01	0.96	0.43
25	Senior	56.7	0.59	0.76	0.83	0.95
49	Senior	55.6	0.57	0.92	0.93	0.68
19	Mid	55.6	0.57	1.21	1.09	0.08
18	Senior	55.6	0.57	1.18	1.20	0.10
44	Adv	54.4	0.56	0.77	0.83	0.94
10	Expert	53.3	0.55	0.85	0.91	0.82
13	Mid	52.2	0.53	0.84	0.88	0.84
6	Adv	50	0.51	1.19	1.17	0.09
20	Senior	48.9	0.49	0.90	0.95	0.71
3	Adv	46.7	0.47	1.23	1.19	0.06
5	Senior	45.6	0.45	1.03	0.96	0.37
28	Mid	44.4	0.44	0.76	0.81	0.95
11	Senior	44.4	0.44	1.15	1.17	0.14
36	Senior	41.1	0.40	**1.73**	**1.55**	**<0.001**
40	Adv	40	0.39	0.96	0.98	0.56
2	Senior	40	0.39	1.17	1.10	0.12
30	Adv	40	0.39	1.21	1.14	0.07
48	Mid	38.9	0.37	1.02	1.04	0.39
4	Adv	37.8	0.36	1.07	1.04	0.29
39	Adv	34.4	0.32	0.82	0.82	0.87
35	Senior	34.4	0.32	0.93	0.95	0.64
14	Senior	33.3	0.31	0.98	1.05	0.50
47	Adv	28.9	0.26	1.27	1.07	0.04

Items sorted by Rasch difficulty parameter (higher numbers indicate easier items). Bold font notes items that reached significance.

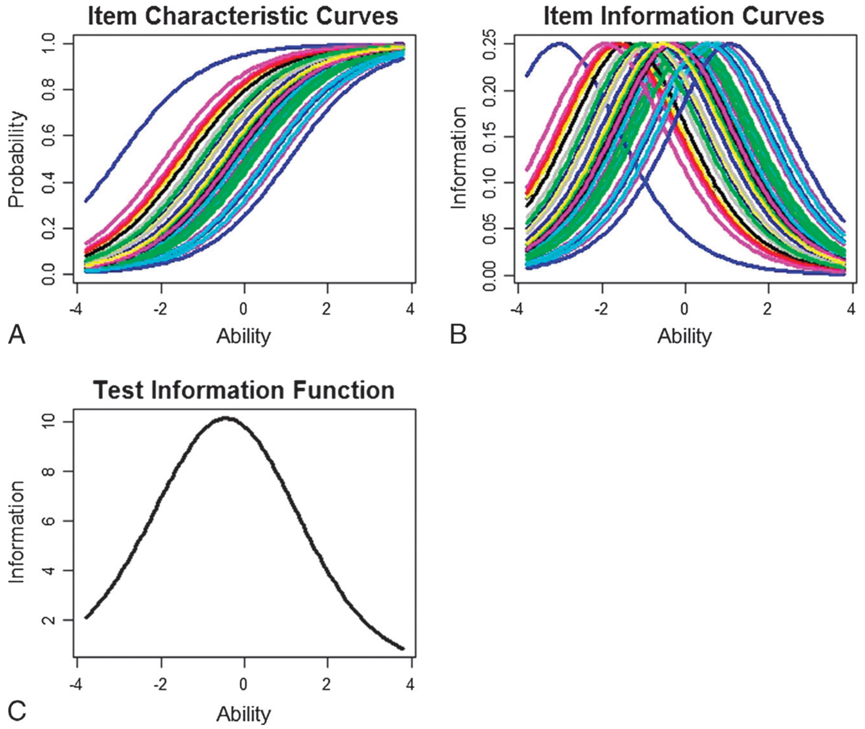


FIGURE 2. Item characteristics and test item information curves (N = 90). Panel A demonstrates the item characteristic curves (ICCs) for each question. The ICC shows the probability of a correct response on the item (*y* axis) as a function of the ability level of an examinee (*x* axis). Ability level is determined by the overall score on the test, with higher scores indicating greater ability. Panel B shows the item information curves. These curves display the amount of information or discriminating ability (*y* axis) that each item contributes. For example, a curve with a peak at the higher end of the ability range (*x* axis) indicates that this item discriminates well among examinees with high ability. Ideally, test questions included in an assessment discriminate along a wide range of ability levels. Panel C is the test information curve, which is the sum of the item information functions in B. The test information function shows the ability of the test as a whole to discriminate between individuals with different levels of the latent trait.

(all $P < 0.001$), demonstrating evidence for construct validity of the assessment.

DISCUSSION

To our knowledge, this is the first study to report the stages of development and validation using psychometric principles of an online assessment of ultrasound interpretation skills for regional anesthesia. The content validity of the assessment was derived in multiple ways. Selection of domains and items was guided by (1) the survey that was administered to graduating anesthesia residents, academic faculty, and community anesthesiologists; (2) the analysis of anatomical structures and relationships; and (3) the analysis of video recordings of attending anesthesiologists teaching UGRA. Thus, the assessment is measuring items of relevant interest to the target audience.

The results of the final validation phase of the study demonstrated that a 47-item subset of the 50-item assessment tool has high internal consistency. Evidence for the reliability of the tool was demonstrated by the significant correlation between the observed and predicted item difficulty. The validation phase also provided strong evidence for the assessment's construct validity as indicated by the correlation of test scores according to 4 different ordinal measures of experience with UGRA. Although a gold standard measure of skill with UGRA does not exist, correlation with the surrogate criterion measures of experience with UGRA indicates that the assessment is measuring the construct it was intended to measure.[11,16] Because the assessment demonstrated the ability to distinguish between learners with different levels of experience with UGRA, the tool can be considered for use as an assessment in the achievement of competency milestones.

Our study differs from those described previously in that it is the first to validate an assessment of ultrasound interpretation skills for regional anesthesia. Despite many requests for the publication of articles describing the development and validation of assessment tools according to sound psychometric principles, few published assessments actually meet these criteria.[11–13]

Although the anesthesiology competency milestones established by the ACGME and the Anesthesiology Residency Review Committee identify a competency milestone for the use of UGRA, they do not identify ultrasound interpretation skills for regional anesthesia as a specific competency milestone.[29] Nonetheless, ultrasound interpretation is considered a key component skill of UGRA and therefore should be subject to assessment in a competency-based educational model.[9,10] By definition, competency-based education identifies knowledge and skills at greater levels of granularity.[30] Thus, more specific (granular) concepts and skills are taught and assessed for mastery in this educational model.

Although we approached the development and validation of the assessment according to psychometric principles, there are several limitations to our study. The extent to which an assessment correlates with a concurrent measure of the skill being assessed is evidence of concurrent validity. This is particularly true if the assessment is correlated with a gold standard criterion.[11,15] Ultrasound-guided regional anesthesia skills and correlation with

TABLE 5. Mean (Median) Total Scores on the 47-Item Test by Demographic Factors (N = 90)

Characteristic	Mean (Median)	Test of Linear Contrast (*P*)
Level of formal anesthesia training		<0.001
Base (n = 18)	18.50 (16.00)	
CA1 (n = 12)	21.00 (18.50)	
CA2 (n = 26)	26.23 (26.00)	
Completed anesthesia residency (n = 7)	31.86 (33.00)	
Current regional fellow (n = 16)	34.94 (36.00)	
Completed regional fellow (n = 11)	38.81 (39.00)	
Residency status*		—
ACGME (n = 2)	29.00 (29.00)	
Non-ACGME US (n = 81)	26.67 (27.00)	
Outside of US (n = 7)	36.86 (39.00)	
Regional training		<0.001
None (n = 33)	23.40 (20.00)	
Completed core (n = 45)	28.53 (29.00)	
Completed advanced (n = 11)	36.81 (38.00)	
Current practice		—
In residency (n = 57)	22.96 (23.00)	
Regional fellow (n = 16)	34.94 (36.00)	
Nonregional fellow (n = 3)	33.33 (36.00)	
Current academic (n = 14)	33.33 (38.00)	
Community (n = 0)	—	
None of the above (n = 0)	—	
Ultrasound procedures per year		<0.001
0 (n = 15)	17.00 (16.00)	
1–20 (n = 26)	22.08 (22.00)	
20–50 (n = 15)	30.80 (30.00)	
60–100 (n = 8)	31.25 (31.00)	
>100 (n = 26)	35.96 (36.50)	
Self-rating		<0.001
Novice (n = 30)	19.00 (18.00)	
Proficient (n = 19)	25.95 (26.00)	
Experienced (n = 29)	33.42 (36.00)	
Expert (n = 12)	37.00 (37.00)	

*There were too few participants in 2 of the categories to make any inferences.

multiple criteria that reflect the skill being measured can provide evidence of the assessment's construct validity.[11,15,16] This was addressed in the current study by providing subjects with clear definitions of skill level for self-assessment and by using both level of training and number of blocks performed as additional criteria for UGRA skill.

An additional limitation is the exclusion of 3 questions from the final assessment, which occurred because of poor model fit for these questions. This occurs when the difficulty of the item (based on the percentage of test takers who answer correctly) does not correlate with test taker proficiency (based on the total number of items they answered correctly).[27,28] For example, question 15 demonstrated an ultrasound clip obtained in the supraclavicular fossa. The subject was asked to identify which arrow pointed to the subclavian artery. This item would seem to be relatively straightforward, and experienced participants would be predicted to answer this question correctly. If enough participants with high overall test scores (highly proficient on the test) did not answer this question correctly, the item would not "fit" the predictive model. This may reflect a poorly written question. In the case of question 15, the video was short and participants may not have been able to appreciate the artery's pulsations. Several round hypoechoic structures in the final image of the clip were used as distractors, including a mirror-image artifact of the artery appearing within the lung. These issues with the question may have caused even experienced participants to be confused by the question, resulting in its poor fit with the predictive model. The statistical Rasch analysis only indicates poor model fit for the item and does not provide insight into the cause of the misfit. To revise the question or create a new one, a further analysis of the actual answers provided by the participants is necessary. Alternatively, debriefing inexperienced participants who answer the question correctly or experienced participants who miss the question could be performed. The stem (description of the ultrasound clip and what the subject should attempt to identify on the clip), the ultrasound clip, or the distractors may all be sources of poor item fit. The 2 other questions that were removed were question 36, asking the subject to identify the lateral circumflex femoral artery in an ultrasound clip of the femoral region, and question 42, asking the subject to identify the femur in an ultrasound clip of the gluteal region. Because this assessment is meant to cover a relevant domain (a certain set of anatomical structures and their sonographic appearance), elimination of questions can compromise the coverage of the domain. These questions may need to be revised and included in a final version of the assessment if it is to be used as a milestone assessment.

Another limitation of the study and of the validated version of the assessment is the lack of complete coverage of the relevant domain. In the survey of graduating residents and anesthesiologists, more than 85% of respondents identified ultrasound-guided axillary block as a skill that graduating residents should possess. The current assessment does not include questions related to the sonographic appearance of structures in the axilla. Questions related to axillary block were not included for concern that the assessment would be too long to maintain feasibility. Questions related to axillary sonoanatomy have been included in subsequent versions of the assessment tool that are able to be completed in less than 1 hour. To comprehensively assess milestone achievement in regional anesthesia, questions in other domains may need to be added using a structured process similar to that used here to develop and validate questions.

Although the validation phase included participants from multiple institutions, an additional limitation of the study is the lack of demographic data to determine if there was broad representation of institutions. The demographic survey did not include a question for the subject to indicate his or her institution. It is possible that the study results were skewed based on an overrepresentation from a single institution. An analysis of the e-mail addresses that the participants used to enroll in the study did not demonstrate significant overweighting of any 1 program; however, not all e-mail addresses contained information that would link a subject to an institution.

This study describes the use of psychometric principles to develop and validate an assessment of ultrasound interpretation skills for UGRA. The validation process described here may prove useful in the future development of other milestone assessment tools designed to evaluate competency achievement of residency trainees or lifelong learners. We have provided evidence for the internal reliability, content validity, and construct validity of the developed assessment. Such an assessment is a candidate for use in anesthesiology training for formative or summative assessment of milestone achievement. Trainee assessment results could be coupled with other direct observation assessments or evaluation of needle guidance skills with ultrasound before permitting trainees to perform UGRA procedures on patients. In addition, assessment results could be used to provide focused remediation during training or as a summative assessment before graduation from residency. Furthermore, an assessment of demonstration of specific competencies in UGRA may become part of future credentialing or maintenance of certification activities. Further study will be necessary to improve the assessment, to identify the time course for trainee acquisition of UGRA competency during residency, and to map assessment scores to specific levels of competency. In addition, future research will be needed to correlate UGRA component skills assessment results with patient-centered outcomes, for example, block success or complication rates.

REFERENCES

1. Abrahams MS, Aziz MF, Fu RF, Horn JL. Ultrasound guidance compared with electrical neurostimulation for peripheral nerve block: a systematic review and meta-analysis of randomized controlled trials. *Br J Anaesth*. 2009;102:408–417.
2. Barrington MJ, Kluger R. Ultrasound guidance reduces the risk of local anesthetic systemic toxicity following peripheral nerve blockade. *Reg Anesth Pain Med*. 2013;38:1289–297.
3. Clendenen SR, Robards CB. The role of ultrasound and regional anesthesia. *Int Anesthesiol Clin*. 2010;48:13–20.
4. Gray AT. Ultrasound-guided regional anesthesia: current state of the art. *Anesthesiology.* 2006;104:368–373.
5. Marhofer P, Willschke H, Kettner S. Current concepts and future trends in ultrasound-guided regional anesthesia. *Curr Opin Anaesthesiol*. 2010;23:632–636.
6. Aggarwal R, Darzi A. Technical-skills training in the 21st century. *N Engl J Med*. 2006;355:2695–2696.
7. Reznick RK, MacRae H. Teaching surgical skills—changes in the wind. *N Engl J Med*. 2006;355:2664–2669.
8. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—rationale and benefits. *N Engl J Med*. 2012;366:1051–1056.
9. Sites BD, Chan VW, Neal JM, et al.The American Society of Regional Anesthesia and Pain Medicine and the European Society Of Regional Anaesthesia and Pain Therapy Joint Committee recommendations for education and training in ultrasound-guided regional anesthesia. *Reg Anesth Pain Med*. 2009;34:40–46.
10. Smith HM, Kopp SL, Jacob AK, Torsher LC, Hebl JR. Designing and implementing a comprehensive learner-centered regional anesthesia curriculum. *Reg Anesth Pain Med*. 2009;34:88–94.
11. Boulet JR, Murray D. Review article: assessment in anesthesiology education. *Can J Anaesth*. 2012;59:182–192.
12. Bould MD, Crabtree NA, Naik VN. Assessment of procedural skills in anaesthesia. *Br J Anaesth*. 2009;103:472–483.
13. Tetzlaff JE. Assessment of competency in anesthesiology. *Anesthesiology.* 2007;106:812–825.
14. Turnbull J, Gray J, MacFadyen J. Improving in-training evaluation programs. *J Gen Intern Med*. 1998;13:317–323.
15. Cronbach LJ, Meehl PE. Construct validity in psychological tests. *Psychol Bull*. 1955;52:281–302.
16. Clauser BEMM, Swanson DB. Issues of validity and reliability for assessments in medical education. In: Holmboe ESHR, ed. *Practical Guide to the Evaluation of Clinical Competence*. Philadelphia, PA: Mosby/Elsevier, 2008:10–23.
17. Chan VW. Applying ultrasound imaging to interscalene brachial plexus block. *Reg Anesth Pain Med*. 2003;28:340–343.
18. Neal JM, Gerancher JC, Hebl JR, et al.Upper extremity regional anesthesia: essentials of our current understanding, 2008. *Reg Anesth Pain Med*. 2009;34:134–170.
19. Soares LG, Brull R, Lai J, Chan VW. Eight ball, corner pocket: the optimal needle position for ultrasound-guided supraclavicular block. *Reg Anesth Pain Med*. 2007;32:94–95.
20. Barrington MJ, Lai SL, Briggs CA, Ivanusic JJ, Gledhill SR. Ultrasound-guided midthigh sciatic nerve block—a clinical and anatomical study. *Reg Anesth Pain Med*. 2008;33:369–376.
21. Sinha A, Chan VW. Ultrasound imaging for popliteal sciatic nerve block. *Reg Anesth Pain Med*. 2004;29:130–134.
22. Hadzic A, Vloka JD. *New York School of Regional Anesthesia. Peripheral Nerve Blocks*. 1st ed. New York, NY: McGraw-Hill Health Professions Division, 2004.
23. Snell RS. *Clinical Anatomy by Regions*. 9th ed. Baltimore, MD: Lippincott Williams & Wilkins, 2012.
24. Wright BD, Mok M. Rasch models overview. *J Appl Meas*. 2000;1:83–106.
25. De Ayala RJ. *The Theory and Practice of Item Response Theory.* New York, NY: Guilford Press, 2009.
26. Rasch G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Expanded ed. Chicago, IL: University of Chicago Press, 1980.
27. Rizopouls D. Itm. An R package for latent variable modeling and item response analysis. *J Stat Softw.* 2006;17:25.
28. Mair P, Hatzinger R. Extended Rasch modeling: the eRm package for the application of IRT models in R. *J Stat Softw.* 2007;20:1–20.
29. Schartel SA, Kuhn C, Culley DJ, Wood M, Cohen N. Development of the anesthesiology educational milestones. *J Grad Med Educ*. 2014;6(Suppl 1):12–14.
30. Frank JR, Mungroo R, Ahmad Y, Wang M, De Rossi S, Horsley T. Toward a definition of competency-based education in medicine: a systematic review of published definitions. *Med Teach*. 2010;32:631–637.