The Ultrasound Competency Assessment Tool (UCAT): Development and Evaluation of a Novel Competency-based Assessment Tool for Point-of-care Ultrasound

Colin Bell, MD¹ , Andrew K. Hall, MD, MMED¹ , Natalie Wagner, PhD^{2,3} , Louise Rang, MD¹, Joseph Newbigging, MD¹, and Conor McKaigney, MD⁴

ABSTRACT

Objectives: Point-of-care ultrasound (POCUS) has become an integral diagnostic and interventional tool. Barriers to POCUS training persist, and it continues to remain heterogeneous across training programs. Structured POCUS assessment tools exist, but remain limited in their feasibility, acceptability, reliability, and validity; none of these tools are entrustment-based. The objective of this study was to derive a simple, entrustment-based POCUS competency assessment tool and pilot it in an assessment setting.

Methods: This study was composed of two phases. First, a three-step modified Delphi design surveyed 60 members of the Canadian Association of Emergency Physicians Emergency Ultrasound Committee (EUC) to derive the anchors for the tool. Subsequently, the derived ultrasound competency assessment tool (UCAT) was used to assess trainee ($N = 37$) performance on a simulated FAST examination. The intraclass correlation (ICC) for inter-rater reliability and Cronbach's alpha for internal consistency were calculated. A statistical analysis was performed to compare the UCAT to other competency surrogates.

Results: The three-round Delphi had 22, 26, and 26 responses from the EUC members. Consensus was reached, and anchors for the domains of preparation, image acquisition, image optimization, and clinical integration achieved approval rates between 92 and 96%. The UCAT pilot revealed excellent inter-rater reliability (with ICC values of 0.69-0.89; $p < 0.01$) and high internal consistency ($\alpha = 0.91$). While UCAT scores were not impacted by level of training, they were significantly impacted by the number of previous POCUS studies completed.

Conclusions: We developed and successfully piloted the UCAT, an entrustment-based bedside POCUS competency assessment tool suitable for rapid deployment. The findings from this study indicate early validity evidence for the use of the UCAT as an assessment of trainee POCUS competence on FAST. The UCAT should be trialed in different populations performing several POCUS study types.

From the ¹Department of Emergency Medicine; the ²Department of Biomedical & Molecular Sciences, Queen's University, Kingston, Ontario; the ³Office of Professional Development & Educational Scholarship, Queen's University, Kingston, Ontario; and the ⁴Department of Emergency Medicine, University of Calgary, Calgary, Alberta, Canada.

Received April 13, 2020; revision received July 15, 2020; accepted August 9, 2020.

Presented at CAEP 2019 National Conference, Halifax, Nova Scotia, Canada, May 2019.

This project received financial support from the SEAMO AFP Innovation Research Fund No. 6023777 and from the University of Calgary Office of Medical and Health Education Scholarship (OHMES) Research and Innovation Funding Grant, ID 2016-03.

The authors have no potential conflicts to disclose.

Author contributions: concept and design—CB, AH, LR, and CM; data acquisition—CB, LR, and CM; analysis and Interpretation—NW and AH; drafting of manuscript—CB, AS, and MW; critical revision—CB, AH, NW, LR, JN, and CM; statistical expertise—NW; and acquisition of funding—CB, AH, LR, JN, and CM.

Supervising Editor: Sally Santen, MD, PhD.

Address for correspondence and reprints: Colin Bell; e-mail: colinbell85@gmail.com.

AEM EDUCATION AND TRAINING 2021;5:1–12

Point-of-care ultrasound (POCUS) has become an integral diagnostic and interventional tool across multiple medical specialties.¹ As POCUS has become more routinely and broadly employed, there has been a growing appetite for POCUS training among medical trainees, specifically within emergency medicine (EM).² Traditionally, credentialing of POCUS in North America has been based on performing a fixed number of scans, direct observation, and written examinations.^{3,4} There is some evidence to suggest that the completion of the predetermined number of POCUS studies can be associated with an adequate image acquisition and interpretation;⁵ however, other studies suggest otherwise.⁶ POCUS education, assessment, and supervision are variable among EM residency training programs.^{7,8} In addition to this variability, there are a number of feasibility challenges for POCUS education within the current system.

Numerous trainees are interested in POCUS training; however, there are a limited number of experts available to supervise and credential POCUS operators.² Of the several POCUS assessment tools that already exist,^{9–14} none are in widespread use,⁸ and they have limitations in feasibility, acceptability, reliability and validity.¹⁵ Thus, even if learners completed a requisite number of scans, their POCUS education, assessment, and supervision is so variable between training programs,⁸ that trainees progression toward competence is unlikely to be uniform at graduation.

Competency-based medical education (CBME) calls for direct observation and objective assessments of competency to ensure a minimum standard of competence for all trainees entering into independent practice.¹⁶ Given the above-mentioned limitations in current POCUS competency assessment processes, and the global shift toward CBME, now is the time for the development of more robust and valid tools to use for POCUS assessment. Other modern competency-based assessment tools and processes have harnessed entrustment scales to capture assessor judgments of a trainee's proficiency to care for patients under decreasing levels of supervision.^{17,18} Entrustment scales harness the tacit knowledge of experts for focused competency assessments and have been shown to improve reliability compared to traditional reductive checklist assessment methods or more nonspecific global rating scales.^{17–19} The use of entrustment scales for workplace-based assessment are typically received well by frontline faculty as they are already employing entrustment-based principles in their daily work,

continuously making judgments about the degree of supervision their trainees required.²⁰

The objective of this study was to derive a straightforward tool to assess POCUS competence, using an entrustment scale that could be applicable to a wide variety of POCUS examinations suitable for rapid bedside deployment. Further, we aimed to evaluate the tool in a pilot assessment setting and acquire some early validity evidence, using Messick's validity framework,²¹ for the tool in the assessment of trainee POCUS competency.

METHODS

Study Design

This study was composed of two phases. Phase one focused on the initial tool derivation process using a three-step modified Delphi design. In phase two, the developed tool was used to assess trainee performance on ultrasounds completed in simulation, and data were collected to inform an early validity argument for the tool. Ethics approval for the derivation was provided by the University of Calgary (No. 18-0373). The Queen's University Health Science Research Ethics Board (No. 6023366) approved both the derivation and the pilot.

Phase 1: Ultrasound Competency Assessment Tool Derivation

The tool derivation process incorporated a three-step modified Delphi design that took place between September and December 2018. The Delphi is an iterative process that uses repetitive rounds of survey and is an effective and reliable process for determining expert group consensus.²² The table of recommendations to ensure methodologic rigor when using consensus group methods from Humphrey-Murto et al.²² is included as Data Supplement S1, Table S1 (available as supporting information in the online version of this paper, which is available at <http://onlinelibrary.wiley.com/doi/10.1002/aet2.10520/full>).

The initial framework for the proposed Ultrasound Competency Assessment Tool (UCAT) was derived by authors CB, CM, JN, and LR, who all have advanced training in EM ultrasound, and AH, a clinician educator. Authors started with a tool previously developed by Dashi et al.¹² (Data Supplement S1, Figure S1) which was based on the Queen's Simulation Assessment Tool (QSAT). This tool sought targeted assessment across four main domains: preparation, image

acquisition, image optimization, and clinical integration and a final overall entrustment score.²⁰ The domains chosen correspond to the different stages of an ultrasound study and reflect both the psychomotor and the cognitive skills needed for POCUS competence. Under each domain were a set of anchors previously derived to facilitate assessment. The anchors were intended to be key behaviors related to each domain, the presence or absence of which would inform judgment of competence. These domains and corresponding anchors from Dashi et al. and suggestions from authors CB, CM, JN, and LR were chosen as the starting point for round 1 of the Delphi process.

The aim of the Delphi process was purposely inclusive of as many relevant national stakeholders as possible to establish consensus on a tool with maximum acceptability and to facilitate its national uptake.¹⁵ Sixty members of the Canadian Association of Emergency Physicians (CAEP) emergency ultrasound committee (EUC) were invited to participate in the Delphi to maximize inclusion, and facilitate uptake and implementation across Canada. In round 1, participants were asked to select the behavioral anchors they felt would most importantly contribute to their assessment of trainee competence in each domain and to provide any additional suggestions for anchors they believed were missing. All round 1 anchors and new anchor suggestions were then carried forward into round 2. In the second round, results of the first round were reflected back to participants (anchor approval rate) and participants were asked again to indicate which anchors they felt were critical for assessment in each domain. No anchor suggestions were put forward by the survey respondents in round 2. Anchors that were approved by 65% or more of participants were carried forward into round 3. In round 3, participants were provided with a final list of anchors for each domain and asked if they believe that the anchors generated via rounds 1 and 2 for each domain represented an acceptable set of anchors to guide assessment. If they indicated “no,” then they were offered a free text box to provide rationale. The survey was anonymous and all rounds of the Delphi process were completed electronically using the Qualtrics survey platform from Qualtrics (Provo, UT).

The final UCAT was comprised of four domains (preparation, image acquisition, image optimization, and clinical integration), with the Delphi-derived anchors and an overall entrustment score (Figure 1).

Each of the four domains was assessed on a three-point scale (competent performance of SOME, MOST, or ALL components). The scale was used to simplify rater assessment of performance within domains and to facilitate formative feedback related to the domain anchors. As the focus of the assessment tool, a final entrustment score, the Ottawa Surgical Competency Operating Room Evaluation (OSCORE) was chosen.²³ The OSCORE scale has a strong and growing argument for validity in procedural contexts.²⁴ In addition, it has rater comfort and familiarity, since it is the most commonly utilized entrustment score in Canadian EM CBME assessment systems.²⁵

Phase 2: Pilot Testing and Evaluation

Design. In February 2019, the Department of Emergency Medicine at Queen’s University conducted a POCUS assessment station for all EM trainees as an add-on to an existing simulation-based Objective Structured Clinical Examination (OSCE).²⁶ Participation in the POCUS assessment component was voluntary for all residents and lasted approximately 15 minutes. In the POCUS station, participants had 1 minute to read a standardized stem, and 10 minutes to complete the station followed by feedback. The stem requested a FAST of a simulated patient involved in a high-speed rollover motor vehicle collision. The FAST was selected because novices are typically able to generate interpretable images and it is highly transferrable to the clinical environment. Two fellowship-trained sonographers watched each performance live (CB and CM) and independently assessed the trainees’ POCUS competence using the derived UCAT. Both CB and CM independently scanned the live human model and verified the absence of free fluid. One assessor (CM) was from a different training site and had not previously interacted with the residents, making him blind to participants’ year of training and previous POCUS experience.

Prior to the POCUS station, trainees also completed a brief questionnaire quantifying their past ultrasound experience, including number of scans completed, perceived confidence, and previous exposure to ultrasound training. A second POCUS simulated assessment was held in July 2019, for incoming residents and for those unable to participate in the February 2019 session. Identical methods were followed, although in July 2019, both reviewers were blinded to the participants’ ultrasound experience.

Domain	Performance Rating				
	Competent performance of SOME criteria	Competent performance of MOST criteria	Competent performance of ALL criteria		
Preparation -Patient & Sonographer -Positioning -Probe selection -Appropriate clinical indication -Initial settings (depth, pre-sets) -Ensures clean transducer	1	2	3		
Image Acquisition -Hand & probe position -Identify appropriate landmarks -Thorough visualization of target -Efficiency of probe motion -Appropriate measurements -Troubleshoots technical limitations	1	2	3		
Image Optimization -Centers area of interest -Overall image quality for interpretation -Troubleshoots patient obstacles -Optimizes machine settings (gain, focal zone, depth, frequency)	1	2	3		
Clinical Integration -Appropriate interpretation -Understands limitations of US scan -Utilizes information from PoCUS to appropriately inform management decisions -Performs multiple PoCUS exams when appropriate -Communicates findings to care team and documents appropriately -Considers false positives/negatives as cause of findings	1	2	3		
Entrustment	I had to do	I had to talk them through	I had to prompt from time to time	I needed to be there just in case	I did not need to be there
	1	2	3	4	5

Figure 1. Ultrasound Competency Assessment Tool (UCAT).

Data Analysis. Assessment scores from the February and July 2019 OSCEs were combined into one dataset. A two-way random, intraclass correlation coefficient (ICC) was calculated for each performance domain on the UCAT to assess inter-rater reliability. Cronbach's alpha was calculated to measure internal

consistency or agreement across performance domains. Finally, the raters' scores were averaged, and scores for each domain were combined (and weighted equally) with the entrustment score to create a composite score for each participant. All subsequent analyses were run twice, once using the composite score as the UCAT

Table 1
Participant Demographics (Tool Derivation and Initial Implementation)

Tool Derivation							
Phase		Time		Total Number of Participants		Overall Proportion Represented	
Delphi process: round 1		September 2018		22		22/60 (37%)	
Delphi process: round 2		November 2018		26		26/60(43%)	
Delphi process: round 3		December 2018		26		26/63 (43%)	
Initial Implementation							
OSCE 1		February 2019		25		37/38 (97%)	
OSCE 2		July 2019		12			
Level of Training*	N	Previous FAST Scans		Previous TOTAL Scans		Self-rated POCUS Confidence (1 to 5), Mean (±SD)	Self-rated POCUS Entrustment (1 to 5), Mean (±SD)
		Mean (±SD)	Range	Mean (±SD)	Range		
PGY1	9	24.22 (±23)	0-55	91 (±84)	0-228	2.56 (±1.23)	3.33 (±0.87)
PGY2	4	171 (±109)	96-332	540 (±327)	289-571	4.25 (±0.50)	4.75 (±0.5)
PGY3	4	150 (±71)	82-249	486 (±193)	268-459	4.50 (±0.58)	4.75 (±0.5)
PGY4	2	348 (±214)	197-500	1088 (±707)	588-1588	5.0 (±0.0)	5.0 (±0.0)
PGY5	3	367 (±128)	245-500	1283 (±434)	830-1696	5.00 (±0.0)	5.0 (±0.0)
CCFP-EM	15	40.73 (±39)	4-103	135 (±128)	16-367	3.07 (±1.16)	3.6 (±0.98)

CCFP-EM = Canadian College of Family Physicians Emergency Medicine; OSCE = Objective Structured Clinical Examination; POCUS = point-of-care ultrasound.

*There are two training pathways for EM physicians in Canada. A 5-year residency accredited by the Royal College of Physicians and Surgeons or a 1-year certificate of added competence following completion of a family medicine residency accredited by the CCFP.

performance outcome and once using the average entrustment score alone.

Initially a one-way analysis of variance (ANOVA) with a Bonferroni correction was conducted to determine whether differences between postgraduate year (PGY) exist for the UCAT entrustment and UCAT composite scores. As the results suggested no significant differences across PGY, level of training was condensed into junior (PGY1-2), senior (PGY3-5) and Canadian College of Family Physicians Emergency Medicine (CCFP-EM) trainees. A two-way ANOVA was used to compare UCAT performance by level of training and number of previous scans reported. Previous number of scans was divided into two groups: those who reported completing less than 50 previous FAST scans and those who reported 50 or more. The analyses were also performed by grouping previous number of scans by <200 total POCUS scans or >200 total POCUS scans. We explored both cutoff points in our analyses to represent the variable credentialing requirements.^{3,4} Participant's self-rated POCUS confidence was also compared to previous number of scans and UCAT score using a Spearman's rank correlation. All statistical analyses were conducted with SPSS Version 12 (IBM SPSS Statistics, Armonk, NY). Statistical significance was considered at $p \leq 0.05$.

RESULTS

Phase 1: UCAT Derivation

The initial distribution of the survey had 22 respondents from the 60-member CAEP EUC committee. In rounds 2 and 3, each received 26 committee responses (Table 1). Responses were received from IP addresses corresponding to 27 different Canadian postal codes and one zip code. Mean ratings for each anchor in every domain were calculated and published in subsequent distributions (Table 2). The final set of suggested anchors for each of the domains of preparation, image acquisition, image optimization, and clinical integration received approval rates of 96, 92, 96, and 92%, respectively, from the EUC respondents (Figure 1). For the few respondents who disapproved of the final set of anchors, most did not provide rationale. Only two participants offered a suggestion with each requesting removal of an anchor in either preparation (ensures clean transducer) and image optimization (overall image quality for interpretation). However, the anchors were not removed as consensus had already been achieved.

Phase 2: Pilot Testing

Descriptive Statistics. Across the two OSCEs, a total of 37 to 38 (97%) of the Queen's EM trainees participated; there were no repeat participants between

Table 2
Delphi Process – Approval Rates per Anchor

Anchor Suggestion	Round 1 Approval N (%)	Round 2 Approval N (%)	Final Approval of domain anchors N (%)
Preparation			25 (96)*
Ergonomics	17 (73) Consolidated for round 2	15 (58)	
Patient position	19 (86) Consolidated for round 2	22 (85)	
Patient and sonographer positioning*	Consolidated Suggestion added for round 2	22 (85)	
Probe selection*	19 (86)	22 (85)	
Gel application	11 (50)	4 (15)	
Draping	14 (64)	15 (58)	
Patient engagement	14 (64)	15 (58)	
Initial machine settings*	19 (86)	23 (88)	
Appropriate clinical indication*	19 (86)	23 (88)	
Ensures clean transducer*	17 (77)	19 (86)	
Patient ID entry	Suggestion added for round 2	10 (30)	
Image Acquisition			23 (92)*
Starting Location	14 (65)	14 (54)	
Hand/Probe position*	19 (86)	23 (88)	
Identifies appropriate landmarks*	17 (77)	19 (73)	
Efficiency of probe motion*	15 (68)	17 (65)	
Appropriate measurement(s) if indicated*	17 (77)	17 (65)	
Troubleshoots technical limitations*	19 (86)	21 (81)	
Explores target with appropriate thoroughness	19 (86) Reworded for round 2		
Visualizes and explores target of exam with appropriate thoroughness*	Reworded anchor from round 1	17 (65)	
Probe accuracy	11 (50)	16 (62)	
Image Optimization			23 (96)*
Centres area of interest on screen*	19 (86)	22 (85)	
Optimizes gain	19 (86)	22 (85)	
Optimizes focal zone	8 (36) Consolidated for round 2	3 (12)	
Frequency adjustment	9 (41) Consolidated for round 2	4 (15)	
Troubleshooting patient obstacles (Obesity, bowel gas, scars, etc)*	19 (86)	22 (85)	
Optimizes depth	19 (86) Combined for round 2	22 (85)	
Optimizes machine settings (gain, focal zone, depth, frequency)*	Consolidated anchor suggestion added for round 2	22 (85)	
Overall Image quality for interpretation*	19 (86)	22 (85)	
Activation/Deactivation of artifact filters	Suggestion added for round 2	5 (19)	
Clinical Integration			22 (92)*
Appropriate interpretation*	19 (86)	22 (85)	
Understands limitations of POCUS	19 (86) Combined		
Interprets results in the context of the patient	19 (86) Combined		
Understands limitations of POCUS and interprets results appropriately (positive/negative/indeterminate)*	Suggestion added for round 2	22 (85)	
Utilizes information from POCUS to appropriately inform management decisions*	19 (86)	22 (85)	
Performs multiple POCUS studies when appropriate*	18 (82)	19 (73)	
Communicates findings to care team and documents appropriately*	17 (77)	17 (65)	
Identifies need for serial exams	17 (77)	15 (54)	
Considers false positives and negatives*	19 (86)	22 (85)	

(Continued)

Table 2 (continued)

Anchor Suggestion	Round 1 Approval N (%)	Round 2 Approval N (%)	Final Approval of domain anchors N (%)
Identifies when alternate work up indicated	Suggestion added for round 2	15 (58)	
Understands when POCUS may delay care	Suggestion added for round 2	12 (46)	
Number of items in each Delphi round			
Total number of items in each round	33	35	21

*Bolded anchors with asterisk were included in the final tool.

Items without bolding or Asterik were either combined in subsequent rounds or did not meet the Delphi consensus for inclusion in the final tool.

February and July 2019 (Table 1). Participants represented all PGYs, variable previous POCUS experience, and variable self-rated POCUS confidence (Table 1). Of the 37 participating trainees, 21 of 37 (57%) had completed 50 or more FAST studies prior to the OSCE session. All participants stated that no free fluid was identified in the OSCE patient.

Inter-rater Reliability and Internal Consistency. Results of the ICC analyses suggest that the UCAT had excellent inter-rater reliability,²⁷ with ICC values of 0.69 to 0.89 and p values of <0.01 (Table 3). The results of the Cronbach's alpha suggest high internal consistency among performance categories in the UCAT ($\alpha = 0.91$).

Level of Training, Previous POCUS Experience, and UCAT Score. The results of the two-way ANOVA suggest a main effect of previous FAST

scans completed on UCAT composite score [$F(1,32) = 10.93$; $p = 0.002$]. We did not find a significant main effect of level of training [$F(2,32) = 1.18$; $p = 0.32$] or interaction between level of training and number of previous FAST scans completed [$F(1,32) = 0.01$; $p = 0.92$; Figure 2A)]. Similar analyses were performed using the entrustment scores alone as the dependent variable and grouping previous number of scans by <200 total POCUS scans or >200 total POCUS scans. All analyses yielded nearly identical results (Figure 2B).

Self-rated Confidence. Self-rated confidence was strongly correlated with number of FAST examinations previously performed ($r = 0.91$, $p < 0.001$). However, the number of previously performed FAST examinations and self-rated confidence was only moderately correlated with entrustment and composite score on the UCAT (r 's = 0.60 and 0.56–0.57, respectively; Table 3).

Table 3

Mean Scores Across Raters, ICCs Coefficients per Domain, and Spearman Rank Correlations

Mean Scores Across Raters, ICC Coefficients per Domain				
UCAT Component	Rater 1 (CB), Mean Rating ($\pm$ SD)	Rater 2 (CM/LR), Mean Rating ($\pm$ SD)	ICC (Absolute Agreement)	p-value
Preparation	2.32 ($\pm$ 0.78)	2.5 ($\pm$ 0.65)	0.89	<0.01*
Image acquisition	2.21 ($\pm$ 0.82)	2.4 ($\pm$ 0.60)	0.72	<0.01*
Image optimization	2.15 ($\pm$ 0.82)	2.16 ($\pm$ 0.65)	0.81	<0.01*
Clinical integration	2.57 ($\pm$ 0.60)	2.32 ($\pm$ 0.58)	0.69	<0.01*
Entrustment	3.60 ($\pm$ 1.30)	3.73 ($\pm$ 1.04)	0.85	<0.01*
Spearman rank correlations				
Correlation	Spearman's Rho (r)		Correlation Strength	Significance
Number of FAST examinations and entrustment	0.60		Moderate	<0.001†
Number of FAST examinations and composite UCAT	0.60		Moderate	<0.001†
Number of FAST examinations and self-rated confidence	0.91		Strong	<0.001†
Self-rated confidence and UCAT entrustment score	0.56		Moderate	<0.001†
Self-rated confidence and UCAT composite score	0.57		Moderate	<0.001†

ICC = intraclass coefficient; UCAT = Ultrasound Competency Assessment Tool.

*Significance considered at $p \leq 0.05$.

†Significance considered at $p < 0.005$ due to multiple comparisons.

Figure 2. (A) UCAT performance by level of training and number of FAST scans. (B) UCAT entrustment by level of training and number of FAST scans. (C) UCAT Performance by level of training and number of total scans. CCFP-EM = Canadian College of Family Physicians Emergency Medicine; UCAT = Ultrasound Competency Assessment Tool.

DISCUSSION

To become consistent with modern education principles of assessment in CBME,²⁸ POCUS assessment

must become individualized and deemphasize fixed quantities.^{29,30} The assessment of trainees, and subsequent delivery of effective feedback, requires high-

quality assessment tools. While structured assessment methods for POCUS assessment exist, few have undergone robust evaluation, and none are in widespread use.^{9–14} In conjunction with the content experts on the CAEP EUC, we used modified Delphi methodology to create a universally applicable, simple, entrustment-based, POCUS competency assessment tool and subsequently piloted it in an OSCE scenario. Through the development and pilot process, we aim to articulate an early argument for validity of the tool trainee POCUS competency, focusing on components of Messick's validity framework.²¹ Specifically, we present the rigorous development and design process as evidence of content validity, employment of a broadly studied and utilized modern entrustment-based scale as support for response process validity, measures of inter-rater reliability and internal consistency as evidence validity pertaining to internal structure, and comparison of performance scores with several other relevant variables as a final marker of validity.

UCAT Design

The UCAT was intentionally developed to be simple and easy for assessors to use, even by those unfamiliar with it. We chose to collaborate with the CAEP EUC to ensure broad participation in the Delphi process and thus bolster the tool's content validity. Moreover, by engaging with national ultrasound leaders, we hoped to facilitate its acceptance and implementation of the tool at POCUS education centers across Canada.³¹

The focus of assessment is an overall entrustment score, the OSCORE,²³ that is currently employed daily in most EM training programs for workplace-based assessment.²⁵ Entrustment scales like this aim to capture the tacit capacity of faculty to make decisions regarding the need for supervision and degree of independence provided; a process that physicians engage in constantly in the workplace.¹⁸ There are robust data demonstrating both the reliability and the validity of global assessment scores like this when compared to checklist-based scoring.³² Further, there is a strong and growing validity argument supporting the use of the OSCORE across multiple procedural contexts.^{24,33} The domain anchors and three-point scales are intended to act as a simple framework for entrustment score justification, but also to provide formative feedback to trainees that is grounded in key aspects of the competency in question. By breaking down the skill to its component parts in each domain, faculty can

provide specific feedback to trainees regarding areas for improvement, thus moving trainees forward on the path to competence and independence. The above rationale supports response process validity pending additional studies measuring assessor's process and perspectives while using the tool.

Pilot Evaluation and Validity

Following tool derivation, the aim of our pilot process was to evaluate how the tool performed in practice. As per the ICC guidelines of Koo and Li,²⁷ the UCAT domains demonstrated moderate to excellent inter-rater reliability (ICCs = 0.69–0.89). This level of agreement was achieved with one local rater who was aware of trainees' level and general ability and one external rater who was unaware of such information, which speaks to the robustness of the tool. Consistent with previous literature,^{24,33} the entrustment category had one of the highest ICC coefficients (ICC = 0.85), which reaffirms our decision to utilize the OSCORE anchors. Our results also showed that the UCAT demonstrated high internal consistency ($\alpha = 0.91$), suggesting that all four domains (preparation, image acquisition, image optimization, and clinical integration) and the entrustment rating captured related, but not redundant, information on POCUS competence. These measures of inter-rater reliability and internal consistency both further support the validity argument for the UCAT.

Additional bolstering of the argument for validity comes from the correlation of the UCAT's performance score with other relevant variables. We compared UCAT performance with previous ultrasound experience (i.e., number of previous FAST studies completed, and number of total POCUS studies completed), level of training, and self-rated POCUS confidence. All senior trainees (PGY3–5) participating in this study had reported having completed at least 50 previous FASTs and 200 total scans. Not all of the participating junior (PGY1–2) residents and CCFP-EM residents had reached this level of experience (Table 3). While there was no significant difference in UCAT performance between senior, junior, and CCFP-EM trainees, those who had completed more than 50 FAST scans (or more than 200 total scans) within the junior and CCFP-EM groups scored significantly higher on the UCAT than their peers who had less experience with POCUS examinations (Figure 2). Our results also suggest that while the previous number of scans performed was highly correlated with self-

rated POCUS confidence ($r = 0.91$), self-rated confidence did not always translate to higher UCAT performance ($r = 0.45$). These findings were consistent when evaluating UCAT composite scores or UCAT entrustment scores.

We propose that formal assessment using the UCAT may be a valuable addition to the assessment of trainee POCUS competence, supplementing direct observation, numeric assessments, and written examinations. It was not surprising that even early in training, most trainees were assessed as competent to independently perform FAST. There is POCUS literature that demonstrates a rapid rise to competent performance with an early technical skill plateau for the FAST in trainees.⁵ This plateau may occur because the FAST typically has large, accessible sonographic windows and organs with reflective echotextures that allow for broad angles of insonation. Therefore, regions of interest in the FAST may be adequately visualized from multiple approaches, greatly decreasing the technical difficulty. The early plateau followed by gradual skill improvement in ultrasound as a technical skill has been demonstrated across multiple POCUS study types in EM trainees.⁵ Conversely, technical skill in physician transthoracic echocardiography (TTE) has been shown to weakly correlate with the number of examinations performed and interpreted and months of training.⁶ This may be because TTE is a more technical study with small sonographic windows and because the heart possesses an echotexture requiring narrow angles of insonation for adequate visualization. Both of these features make TTE more technically challenging than the FAST. Future studies would be beneficial to evaluate the UCATs performance across different types of POCUS studies with varying degrees of technical difficulty.

Recently medical education has seen a movement away from checklist-type assessments and toward entrustment-based ratings.²⁰ In our study, our composite score and entrustment score alone had very similar results. This raises the question as to whether the scores related to the four domains of the UCAT are needed for competency assessment decisions or if one entrustment score is enough. Recent work demonstrates strong correlation between the original OSCORE assessment tool (in fact involving nine component scores using the entrustment scale) to a single-item performance measure.³³ It may be that the domain anchors can truly be used to inform the entrustment decision and act as a stimulus for the

provision of formative feedback, but are not needed for specific scoring. This would be analogous to similar concepts in the Royal College of Canada's model of CBME, in which entrustable professional activities are broken down into milestones to provide feedback, while only the single entrustment scores for EPAs are typically used to inform higher stakes progression decisions. Alternatively, one could argue that each domain would benefit from its own entrustment score. This would mirror the original OSCORE²³ and a recently published resuscitation assessment tool for use in simulation.³⁴ We felt it more important to focus on simplicity and usability and chose to mirror existing workplace-based assessment tools in Canadian EM that have a single entrustment score. Further evaluation work will help determine the utility of specific domain scoring.

LIMITATIONS

While we employed the CAEP EUC as our Delphi sample population, we acknowledge that there may have been other experts who were not included. Second, the lack of face-to-face interaction during the Delphi process may have been a limitation. The electronic survey distribution method was chosen for ease of distribution and to maintain anonymity. While the anonymous electronic survey gave equal weight to every respondent, the absence of a real-time interaction may have deprived experts the opportunity to exchange the reasons behind their viewpoints and prevented the group from reaching a greater consensus.³⁵ Furthermore, with the electronic surveys, we are unable to guarantee the person completing the Delphi was the intended CAEP EUC member. In an effort to maximize survey completion, we initiated our Delphi drawing on the domains and anchors from the tool of Dashi et al.¹² Accordingly, there was limited opportunity for revision on the domain titles. Finally, while the Delphi was distributed to all CAEP EUC members, only 37% to 43% participated in each round. Members who did not participate may have possessed a valuable perspective on the proposed anchors.

For the validation component, only a single center was used with a population of residents who have largely been exposed to and completed their core ultrasound curriculum.⁴ There were very few learners who had not completed their core ultrasound curriculum. Because there is little information available on measuring ultrasound competence, we were unable to

calculate an a priori sample size. With our small sample, there is a risk that the study was underpowered. Finally, because the FAST is typically one of the initial POCUS studies taught in residency programs, it is also known to be an examination with a flat learning curve.⁵ Thus, a more technically challenging POCUS examination may have revealed a greater variation in scores among the learners.

CONCLUSIONS

We have developed a novel point-of-care ultrasound competency assessment tool using a modified Delphi process involving members of a national point-of-care ultrasound organization and engaged in an initial evaluation of its performance in an assessment of trainee FAST competence. The Ultrasound Competency Assessment Tool is a simple and adaptable tool for use in formative assessments at the bedside or in higher stakes evaluative assessments. The tool has early evidence of validity in the assessment of trainee point-of-care ultrasound competence. While Ultrasound Competency Assessment Tool scores did not correlate with level of training, performance did correlate with the number of previous point-of-care ultrasound studies completed. Additional studies are needed to assess the performance of the Ultrasound Competency Assessment Tool in different populations performing a variety of point-of-care ultrasound examinations.

The authors acknowledge the CAEP EUC members for their engagement and contributions in the Delphi derivation process.

References

1. Solomon SD, Saldana F. Point-of-care ultrasound in medical education—stop listening and look. *N Engl J Med* 2014;370:1083–5.
2. Leschyna M, Hatam E, Britton S, et al. Current state of point-of-care ultrasound usage in Canadian emergency departments. *Cureus* 2019;11:e4246.
3. American College of Emergency Physicians. Ultrasound guidelines: emergency, point-of-care and clinical ultrasound guidelines in medicine. *Ann Emerg Med* 2017;69:e27–54.
4. Olszynski P, Kim DJ, Chenkin J, Rang L. The CAEP emergency ultrasound curriculum - objectives and recommendations for implementation in postgraduate training: executive summary. *CJEM* 2018;20:736–8.
5. Blehar DJ, Barton B, Gaspari RJ. Learning curves in emergency ultrasound education. *Acad Emerg Med* 2015;22:574–82.
6. Nair P, Siu SC, Sloggett CE, Bicular L, Sidhu RS, Eric HC. The assessment of technical and interpretative proficiency in echocardiography. *J Am Soc Echocardiogr* 2006;19:924–31.
7. Kim DJ, Theoret J, Liao MM, Hopkins E, Woolfrey K, Kendall JL. The current state of ultrasound training in Canadian emergency medicine programs: perspectives from program directors. *Acad Emerg Med* 2012;19:E1073–8.
8. Amini R, Adhikari S, Fiorello A. Ultrasound competency assessment in emergency medicine residency programs. *Acad Emerg Med* 2014;21:799–801.
9. Damewood SC, Leo M, Bailitz J, et al. Tools for measuring clinical ultrasound competency: recommendations from the ultrasound competency work group. *AEM Educ Train* 2020;4:S106–12.
10. Lewiss RE, Pearl M, Nomura JT, et al. CORD-AEUS: consensus document for the emergency ultrasound milestone project. *Acad Emerg Med* 2013;20:740–5.
11. Hofer M, Kamper L, Sadlo M, Sievers K, Heussen N. Evaluation of an OSCE assessment tool for abdominal ultrasound courses. *Eur J Ultrasound* 2011;32:184–90.
12. Dashi G, Hall A, Newbigging J, Rang L, La Rocque C, McKaigney C. The development and evaluation of an assessment tool for competency in point-of-care ultrasound in emergency medicine [abstract 636]. *Acad Emerg Med* 2016;23:S262.
13. Ma IW, Desy J, Woo MY, Kirkpatrick AW, Noble VE. Consensus-based expert development of critical items for direct observation of point-of-care ultrasound skills. *J Grad Med Educ* 2020;12:176–84.
14. Boniface KS, Ogle K, Aalam A, et al. Direct observation assessment of ultrasound competency using a mobile standardized direct observation tool application with comparison to asynchronous quality assurance evaluation. *AEM Educ Train* 2019;3:172–8.
15. Norcini J, Anderson B, Bollela V, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach* 2011;33:206–14.
16. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. *Med Teach* 2010;32:676–82.
17. Ten Cate O. Entrustment as assessment: recognizing the ability, the right, and the duty to act. *J Grad Med Educ* 2016;8:261–2.
18. Rekman J, Gofton W, Dudek N, Gofton T, Hamstra SJ. Entrustability scales: outlining their usefulness for competency-based clinical assessment. *Acad Med* 2016;91:186–90.
19. Weller JM, Misur M, Nicolson S, et al. Can I leave the theatre? A key to more reliable workplace-based assessment. *Br J Anaesth* 2014;112:1083–91.

20. Ten Cate O, Hart D, Ankel F, et al. Entrustment decision making in clinical training. *Acad Med* 2016;91:191–8.
21. Messick S. Validity. In: Linn RL, editor. *Educational Measurement*. New York: Macmillan, 1989:13–103.
22. Humphrey-Murto S, Varpio L, Gonsalves C, Wood TJ. Using consensus group methods such as Delphi and nominal group in medical education research. *Med Teach* 2017;39:14–9.
23. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med* 2012;87:1401–7.
24. MacEwan MJ, Dudek NL, Wood TJ, Gofton WT. Continued validation of the O-SCORE (Ottawa Surgical Competency Operating Room Evaluation): use in the simulated environment. *Teach Learn Med* 2016;28:72–9.
25. Sherbino J, Bandiera G, Doyle K, et al. The competency-based medical education evolution of Canadian emergency medicine specialist training. *CJEM* 2020;22:95–102.
26. Hagel CM, Hall AK, Dagnone JD. Queen's university emergency medicine simulation OSCE: an advance in competency-based assessment. *CJEM* 2016;18:230–3.
27. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 2016;15:155–63.
28. Lockyer J, Carraccio C, Chan MK, et al. Core principles of assessment in competency-based medical education. *Med Teach* 2017;39:609–16.
29. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. *Med Teach* 2010;32:638–45.
30. Gruppen LD, Ten Cate O, Lingard LA, Teunissen PW, Kogan JR. Enhanced requirements for assessment in a competency-based, time-variable medical education system. *Acad Med* 2018;93:S17–21.
31. Wagner N, Acai A, McQueen SA, et al. Enhancing formative feedback in orthopaedic training: development and implementation of a competency-based assessment framework. *J Surg Educ* 2019;76:1376–401.
32. Ilgen JS, Ma IW, Hatala R, Cook DA. A systematic review of validity evidence for checklists versus global rating scales in simulation-based assessment. *Med Educ* 2015;49:161–73.
33. Saliken D, Dudek N, Wood TJ, MacEwan M, Gofton WT. Comparison of the Ottawa Surgical Competency Operating Room Evaluation (O-SCORE) to a single-item performance score. *Teach Learn Med* 2019;31:146–53.
34. Weersink K, Hall AK, Rich J, Szulewski A, Dagnone JD. Simulation versus real-world performance: a direct comparison of emergency medicine resident resuscitation entrustment scoring. *Adv Simul* 2019;4:9.
35. Walker AM, Selfe J. The Delphi method: a useful tool for the allied health researcher. *Br J Occup Ther* 1996;3:677–81.

Supporting Information

The following supporting information is available in the online version of this paper available at <http://onlinelibrary.wiley.com/doi/10.1002/aet2.10520/full>

Data Supplement S1. Supplementary material.