Reliable and Valid Assessment of Point-of-care Ultrasonography

Tobias Todsen, MD,* Martin Gronnebæk Tolsgaard, MD, PhD,†† Beth Hærstedt Olsen, MD,‡ Birthe Merete Henriksen, MD,§ Jens Georg Hillingsø, MD, PhD,¶ Lars Konge, MD, PhD,* Morten Lind Jensen, MD, PhD,* and Charlotte Ringsted, MD, PhD∥

Objective: To explore the reliability and validity of the Objective Structured Assessment of Ultrasound Skills (OSAUS) scale for point-of-care ultrasonography (POC US) performance.

Background: POC US is increasingly used by clinicians and is an essential part of the management of acute surgical conditions. However, the quality of performance is highly operator-dependent. Therefore, reliable and valid assessment of trainees’ ultrasonography competence is needed to ensure patient safety.

Methods: Twenty-four physicians, representing novices, intermediates, and experts in POC US, scanned 4 different surgical patient cases in a controlled set-up. All ultrasound examinations were video-recorded and assessed by 2 blinded radiologists using OSAUS. Reliability was examined using generalizability theory. Construct validity was examined by comparing performance scores between the groups and by correlating physicians’ OSAUS scores with diagnostic accuracy.

Results: The generalizability coefficient was high (0.81) and a D-study demonstrated that 1 assessor and 5 cases would result in similar reliability. The construct validity of the OSAUS scale was supported by a significant correlation between OSAUS scores and diagnostic accuracy. The generalizability coefficient was high (0.81) and a D-study demonstrated that 1 assessor and 5 cases would result in similar reliability. The construct validity of the OSAUS scale was supported by a significant correlation between OSAUS scores and diagnostic accuracy (Spearman ρ correlation coefficient = 0.76; P < 0.001).

Conclusions: This study demonstrates high reliability as well as evidence of construct validity of the OSAUS scale for assessment of POC US competence. Hence, the OSAUS scale may be suitable for both in-training as well as end-of-training assessment.

Keywords: abdominal ultrasound, assessment, point-of-care ultrasonography, surgical training. (Ann Surg 2015;261:309–315)
was conducted at the Centre for Clinical Education at Copenhagen University Hospital, Rigshospitalet, Denmark.

Participants

Twenty-four volunteer physicians were recruited as a strategic sample for this explorative study. The participants were distributed across 3 groups (novice, intermediate and expert) on the basis of their experience with abdominal POC US:

The novice group consisted of physicians enlisted for a basic hands-on abdominal ultrasonography course. Inclusion criteria were (1) no prior formalized ultrasonography course and (2) fewer than 25 prior abdominal ultrasound examinations. The learning curve for right upper quadrant POC US has been reported to level out after 25 examinations. Therefore, we considered physicians as intermediates based on experience of a minimum of 25 abdominal POC US examinations and a formal ultrasonography course.

The intermediate group consisted of residents and consultants recruited from departments of general surgery from 2 different University Hospitals in the Capital Region of Denmark. Inclusion criteria were (1) completion of a formal ultrasonography course, and (2) prior experience of more than 25 abdominal ultrasound examinations on patients.

The expert group consisted of 1 consultant in general surgery and 1 consultant in emergency medicine. Inclusion criteria were (1) practicing ultrasonography on a regular basis; (2) experience with more than 500 abdominal ultrasonography examinations; and (3) involvement in postgraduate teaching of ultrasonography. The expert group criteria were inspired by recommendations of the European Federation of Societies for Ultrasound in Medicine and Biology.

Data from all the participants were used in the analyses of reliability and to calculate the association between OSAUS scores and diagnostic accuracy.

Participants who did not represent the predefined group criteria of ultrasound experience (ie, a novice-, intermediate- and expert group) were not included in the construct validity analyses concerning the ability of OSAUS to discriminate between levels of competence (see Fig. 1).

Patient Cases

To assess competence in acute abdominal POC US, we chose cases where general surgeons would potentially perform POC US in the emergency department. Four volunteer individuals were selected to portray cases of cholecystolithiasis, hydronephrosis, hemoperitoneum, and aorta aneurysm. Three of the 4 were patients with sonographic verifiable pathology findings: a 1-cm stone in a normal gallbladder, moderate hydronephrosis, and 1.5 L free intra-peritoneal fluid, respectively. The patients were recruited from the Department of Surgery, Copenhagen University Hospital Rigshospitalet. The fourth person had a normal caliber of the abdominal aorta. The physicians were given a short written description about the patients’ symptoms before entering the room with the patient (Table 1).

After the written case descriptions, the physicians performed 4 POC US examinations consecutively to confirm or reject the possible diagnoses (Table 1). The physicians were allowed 5 minutes to complete each of the ultrasound examinations. After 4½ minutes, the physicians were reminded to make a short dictation of their findings to the medical record. The examinations and dictations were videorecorded. Screen output from the ultrasound equipment was recorded with an external hard-disk recorder.

Video Processing Before Performance Rating

The technical performance video recordings were merged with the ultrasonography screen recordings to form 1 anonymized clip (see Fig. 2). Thereby, the assessors could review the ultrasound image and how the physician handled the equipment at the same time. The specific equipment used was as follows: a portable ultrasound machine (GE LOGIQ e with a Convex Probe 2.0–5.5 MHz), a Sony Exmor Handycam video recorder, a MediCapTM USB200 medical recorder, and Final Cut Pro 7 video-editing software.

**FIGURE 1.** Flowchart: The inclusion of physicians to the construct validity analyses concerning OSAUS’ ability to discriminate between levels of competence.
TABLE 1. The Written Case Descriptions Given to the Physicians Before They Performed Ultrasonography of the 4 Cases

Case 1: A patient has pain from the right costal margin.
Use ultrasonography to do the following:
1. Assess the gallbladder
2. Measure the wall thickness
3. Assess whether there are gallstones
Dictate a summary of your findings for the medical record.

Case 2: A patient has colicky flank-pain.
Use ultrasonography to do the following:
1. Assess if there is hydronephrosis on the affected side and compare with the other side.
2. Measure length and diameter of the kidney
3. Measure the parenchyma cortex
Dictate a summary of your findings for the medical record.

Case 3: A patient develops strong abdominal pain 2 days after an operation.
Use Focused Assessment with Sonography for Trauma (FAST) to examine the areas in which free fluid is most likely to accumulate.
Dictate a summary of your findings for the medical record.

Case 4: A patient has strong constant abdominal pain radiating to the back.
Use ultrasonography to do the following:
1. Assess if there is an abdominal-aortic aneurysm
2. Examine the full length of the abdominal aorta
3. Measure a transverse section of the aorta
Dictate a summary of your findings for the medical record.

All videos were given a randomized number and embedded into a password-protected Web page where performances could be assessed using an incorporated OSAUS rating scale.

Assessment of Performance Using OSAUS

The video-recorded performances were assessed by 2 consultant radiologists subspecialized in ultrasound examinations and interventions (B.H.O. & B.M.H.).

After the data acquisition, our 2 assessors participated in a 90-minute training session. Here they received a short introduction to the OSAUS rating scale; thereafter, they assessed 5 pilot videos independently and discussed their ratings until consensus was reached. Participants portrayed in the pilot videos; that is, 2 medical students and 3 physicians conducted the same POC US examinations as in the main study; however, these 5 subjects were not included in the main study sample. E-mails with Web access to the video-recorded performances were then sent to the 2 assessors, who independently assessed all the ultrasound performances, blinded to the identity and level of ultrasonography experience of the physicians.

All individual data points, consisting of a combination of rating value (1–5), OSAUS element (1–5), case (1–4), participant (1–24) and assessor (1–2), were electronically transferred from the Web page to an electronic database.

The 2 OSAUS elements concerning “indication for the examination” and “medical decision making” are described as elements that should only be included if applicable. These 2 elements were not applicable in our controlled setting and were therefore not included (Fig. 3).

Diagnostic Accuracy

The physicians’ dictations of their ultrasonography findings were transcribed into written text. The medical transcripts were evaluated by a consultant in general surgery who was responsible for postgraduate surgical ultrasonography training (J.G.H.). The consultant...
TABLE 1. The Objective Structured Assessment of Ultrasound Skills (OSAUS) used in this study of point-of-care ultrasonography skills.

<table>
<thead>
<tr>
<th>Category</th>
<th>Level 1: Unable to operate equipment</th>
<th>Level 2: Operates the equipment with some experience</th>
<th>Level 3: Competent image optimization but not done consistently</th>
<th>Level 4: Consistent optimization of images</th>
<th>Level 5: Consistently displays systematic approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applied knowledge of ultrasound equipment</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Familiarity with the equipment and its functions, i.e. selecting probe, using buttons and application of gel.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image optimization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistently ensuring optimal image quality by adjusting gain, depth, focus, frequency etc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Systematic examination</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consistently displaying systematic approach to the examination and presentation of relevant structures according to guidelines.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interpretation of images</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recognition of image pattern and interpretation of findings.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Documentation of examination</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Image recording and focused verbal/written documentation.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FIGURE 3. The Objective Structured Assessment of Ultrasound Skills (OSAUS) used in this study of point-of-care ultrasonography skills.

surgeon was aware of the correct diagnoses of the patients but was blinded to the physicians’ identity, level of ultrasonography experience, and ultrasonography performance. The diagnostic accuracy of the medical record was scored as correct diagnosis, inconclusive, or false diagnosis.

Reliability

Generalizability theory was used to estimate the size of the relevant variables that influenced the reliability of the OSAUS scores. All individual data points were electronically transferred to the G String IV statistical software package, based on urGENOVA. To estimate all variance components (VC), we used a fully crossed design of every facet of the assessment (case, participants, assessors, and their interactions). We assumed that “true variance” was the differences in OSAUS score because of different competence between physicians. “Error variance” was attributed to differences in OSAUS score affected by different assessors, cases, and interactions between them. The estimated VCs were subsequently used in Cronbach’s formula to calculate the generalizability coefficient. Reliability is presented from 0 to 1, expressing the percentage of the OSAUS score attributable to true score, not to the error of measurement. Furthermore, a D-study was conducted to determine the number of ultrasonic cases and assessors needed to make a reliable assessment of physician competence. A generalizability coefficient greater than 0.8 is generally accepted as sufficient for high-stakes decisions, whereas greater than 0.6 is sufficient for formative evaluations.

Construct Validity

Individual participants’ OSAUS scores were calculated as a percentage of the maximum possible score from all 4 cases. The primary indicator of construct validity was whether the OSAUS rating scale could discriminate between the groups with different experience. Therefore, the OSAUS scores between the novice, intermediate, and expert groups were compared using the Kruskal-Wallis test. Comparisons were made using the Mann-Whitney test between the novice and intermediate groups, and between the intermediate and expert groups. The results were considered statistically significant when \( P < 0.05 \). A secondary indicator of construct validity was determined to be a correlation between the OSAUS rating scale scores and the diagnostic accuracy. Therefore, Spearman \( \rho \) was used to explore the correlation between the OSAUS scores and the number of correct ultrasonography diagnoses. The statistical analyses were performed using PASW, version 20.0 (IBM Corp., Armonk, NY).

RESULTS

From the 24 physicians who volunteered to participate, OSAUS scores from 95 of 96 possible cases were obtained. One case was missing due to technical problems with the video equipment. Two physicians did not meet the inclusion criteria for either the novice or the intermediate group in the construct validity analyses (see Fig. 1). One of these physicians had experience with 30 abdominal ultrasound examinations, but no formalized ultrasonography course experience.
Annals of Surgery Volume 261, Number 2, February 2015
Assessment of Point-of-care Ultrasonography

The other had completed a formalized ultrasonography course but had no clinical experience with abdominal ultrasonography.

Reliability
All of the 24 physicians were included in the analyses of reliability. Table 2 summarizes the contribution of each source of variance to the OSUAS score. The generalizability coefficient in our test setup was 0.81. A D-study predicted that a test setup with 5 ratings from 1 assessor would also ensure a generalizability coefficient greater than 0.8 (see Fig. 4).

Construct Validity
The analysis of the primary indicator of construct validity included 12 novices, 8 intermediates, and 2 experts. The mean OSUUS scores were 17.0 (SD 8.4) for novices, 30.1 (SD 10.0) for intermediates, and 71.9 (SD 4.4) for experts (see Fig. 5). The Kruskal-Wallis test demonstrated a significant difference between the groups \(P = 0.03\), whereas the Mann-Whitney Test demonstrated a significant difference between novices and intermediates \(P = 0.007\) and between intermediates and experts \(P = 0.04\). The analysis of the secondary indicator of construct validity included all 24 participants and Table 3 summarizes the diagnostic accuracy of the different experience groups. A strong correlation between the physicians’ OSUUS score and the number of correct diagnoses was demonstrated by a Spearman \(\rho\) of 0.76 \((P < 0.001)\).

DISCUSSION
This study strongly supports the reliability and validity of using a generic scale (OSUAS) to assess abdominal POC US competence. The generalizability coefficient was high (0.81) in our simulated test setup using 2 trained assessors and 4 different patient cases. This reflects the fact that 81% of our OSUUS score was attributable to true score and not to the error of measurement, which is considered sufficient for high-stake assessments.\(^1\)\(^6\) The construct validity of the OSUUS scale was supported by both the significant differences in scores between physicians with different levels of ultrasound experience, and by the strong correlation between OSUUS score and the number of correct diagnoses. These results contribute to the validity evidence of the OSUUS scale for an objective measure of POC US performance.\(^1\)\(^7\)

Reliability
As the assessment literature recommends, we conducted a generalizability study with a fully crossed design in a controlled setting.\(^1\)\(^8\)

<table>
<thead>
<tr>
<th>Source of Variance (V)</th>
<th>Description</th>
<th>Estimated VC</th>
<th>Relative Contribution</th>
<th>Interpretation of Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physicians, V</td>
<td>Systematic variation among physicians</td>
<td>9.91</td>
<td>43.7%</td>
<td>Most of the measured variance derives from different competence between the physicians</td>
</tr>
<tr>
<td>Assessor, Vr</td>
<td>Systematic variability among assessors</td>
<td>1.30</td>
<td>5.75%</td>
<td>The assessors had an overall equal level of stringency</td>
</tr>
<tr>
<td>Case, Vc</td>
<td>Systematic variability among cases</td>
<td>0.201</td>
<td>0.887%</td>
<td>The cases were almost equally difficult</td>
</tr>
<tr>
<td>Interaction between physician and assessor, Vpr</td>
<td>Consistent trend for an assessor to assess a particular physician differently</td>
<td>0.294</td>
<td>1.30%</td>
<td>There was no bias between assessor and physician due to effectively blinding</td>
</tr>
<tr>
<td>Interaction between case and assessor, Vcr</td>
<td>Consistent trend for a assessor to assess differently on a particular case</td>
<td>5.96</td>
<td>23.6%</td>
<td>The assessors vary in their perceptions of the challenge of an ultrasound case.</td>
</tr>
<tr>
<td>Interaction between physician, case and assessor, Vpcr</td>
<td>All remaining variability</td>
<td>4.99</td>
<td>22.0%</td>
<td>Expected unexplained error</td>
</tr>
</tbody>
</table>

© 2014 Wolters Kluwer Health, Inc. All rights reserved.

FIGURE 4. D-study for increasing numbers of assessors and observed cases. A critical generalizability coefficient of 0.8 (represented by the dotted line) is reached with 1 assessor observing 5 cases or 2 assessors observing 4 cases.

This approach allows for a combined measure of interrater and test-retest reliability and is intended to be a more accurate exploration of reliability than just calculating interrater reliability. This approach also minimizes unknown bias in the test score and explores facets that could be a threat to reliability. Our results rely on carefully trained assessors. This is confirmed by the high interrater agreement, which only contributes with 5.75% of the overall variation in test score (Table 2). Hence, rater training of clinicians will probably be needed if the instrument is to be used in a clinical setting. However, rater training is unlikely to remove all variations due to the existence of different raters, and in this study we found that 23.6% of the variation in test score was due to the interaction between assessor and case (Table 2). This reflects the fact that the assessors vary in their
perceptions of the challenge of a specific ultrasound case. However, this will equally affect the score of all participants and, therefore, not impact their order of ranking in this study.14 In light of our results, it seems important to use more assessors and cases to ensure test scores with high reliability. For example, using only 1 assessor rating a single case would result in poor reliability (Fig. 4), which corresponds well to results from other similar studies.19,20

Construct Validity

In our setting, the OSAUS score demonstrated a significant ability to discriminate between increasing levels of POC US experience (Fig. 5). However, even though defining expertise according to the number of examinations performed is a feasible and pragmatic approach to measuring the complex construct of expertise development,21 it is also a very simplified approach. Therefore, we chose a test setup in which the sonographically verifiable findings were known beforehand to estimate diagnostic accuracy. By correlating the diagnostic accuracy and the OSAUS rating scale score, we found a strong correlation (correlation coefficient 0.76) to support our claim of construct validity; that is, that the OSAUS rating scale score is a good measure of competence.

TABLE 3. Diagnostic Accuracy of the Different Experience Groups Based on the Physicians’ Medical Record Dictates After the Ultrasound Examinations

<table>
<thead>
<tr>
<th>Interpretation of the Ultrasound Examinations</th>
<th>No. False Diagnoses (%)</th>
<th>No. Inconclusive Examinations (%)</th>
<th>No. Correct Diagnoses (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novices</td>
<td>17 (30%)</td>
<td>21 (38%)</td>
<td>18 (32%)</td>
</tr>
<tr>
<td>Intermediates</td>
<td>7 (22%)</td>
<td>3 (9%)</td>
<td>22 (69%)</td>
</tr>
<tr>
<td>Experts</td>
<td>1 (12%)</td>
<td>0</td>
<td>7 (88%)</td>
</tr>
</tbody>
</table>

STRENGTHS AND LIMITATIONS

Simulated Cases

One limitation of our study is the use of simulated patient cases. However, we did use real patients with sonographically verifiable findings, which supports the generalizability of our findings to clinical settings. Furthermore, video recordings of both the technical performance and the ultrasonography screen recordings provided the assessors with similar information as direct observation in a clinical setting. Therefore, we believe that the OSAUS rating scale is valuable for clinical use.

Classification of Experience Levels

The criteria for the groups were based on studies about the initial learning curve for abdominal POC US11,12 and international training recommendations for the practice of ultrasonography.13 The intermediate group’s mean OSAUS score of 30.1 was sizable smaller compared with the expert group, which was 71.9. This may indicate that some of the surgeons in the intermediate group were still at the beginning of their abdominal ultrasonography learning curve. We only had 2 physicians in the expert group, which makes it difficult to determine the variation in the OSAUS score among different experts. Despite this limitation, the large difference between the performance of intermediates and experts on 4 different cases assessed by 2 raters supports the validity of the results. This indicates that, in our setup, the OSAUS scale can discriminate between competence levels.

Case Design

In our simulated setting, there is a risk that the physicians’ performance will be artificially inflated because of the written case description that points out which POC US examination to perform (Table 1). In contrast, in a real clinical scenario the physician will have to decide which examination they will perform on the basis of the medical history. However, to standardize the reliability analyses, we needed to ensure that all physicians performed the same POC US examinations. For future studies on the OSAUS scale, it would improve realism if the trainee had to decide which POC US to apply.

The availability of volunteer patients admitted to the Department of Surgery at the time of the study restricted the case combination. This was primarily due to the need for nonacute cases with findings that were sonographically identical, or very similar, to the pathological findings of an acute case. However, other common acute abdominal cases, like appendicitis and intussusception, could potentially be used in future assessments in a different setup.

Generic Rating Scale

We chose a generic scale in this study because it is easy to use in a clinical setting.18 A procedural-specific rating scale for different types of abdominal scans has previously been developed for an end-of-training objective structured clinical examination (OSCE).22 However, the use of such a comprehensive instrument might not be feasible in a clinical setting. Moreover, procedure-specific rating scales can have poorer validity and reliability than generic rating scales.23 In this study, we used a generic rating scale in combination with procedure-specific cases.

Implications

The primary implication of this study is that the OSAUS rating scale can be a useful tool for objectively measuring POC US performance. It has the potential to be an integrated part of a POC US competence-based training curriculum to determine advancement in training stages and certification of physicians. We used editing software to merge the technical performance video recordings and the ultrasonography screen recordings into 1 clip (see Fig. 2). However,
it may be easier to use screen recordings only for in-training assessment in a clinical setting. Although this would be simpler and less time-consuming to generate the videos, the assessor would lose some information. We believe that for ultrasound performance assessment, it is important to see both the screen output and the person performing with hand movements. A topic for future research could be to examine both sources in terms of their individual contribution to reliability and validity.

Another way forward could be to eliminate the need for video recordings in the clinical setting. Then the assessor would need to be available and could observe the ultrasound examinations directly. However, this strategy includes the risk of bias due to the assessor-trainee relationship. Overall, additional insights are needed from further research to optimize using the OSAUS in a clinical setting.

CONCLUSIONS

Our findings show that assessment of POC US competence using the generic OSAUS rating scale has high reliability and strong construct validity. The primary implication of this study is that the OSAUS rating scale can be a useful tool during training and certification of physicians using POC US.

ACKNOWLEDGMENTS

The authors are grateful to Matthias Hofer for the opportunity to see his specific OSCE assessment tool, and to Anke Katrin Hofmann for support with the translation of the assessment tool. We would also like to thank Yousif Subhi for the development of the video assessment Web page.

REFERENCES