Editors and Editorial Board

This journal is indexed in MEDLINE, Current Contents (CM), Science Citation Index and in EMBASE/Excerpta Medical Abstract Journals

Impact Factor: 2.389

www.thieme-connect.de
www.thieme.de/ultraschall

manuscript submission at
http://mc.manuscriptcentral.com/eju

Evaluation of an OSCE Assessment Tool for Abdominal Ultrasound Courses

Evaluation eines standardisierten, praktischen Prüfungsparcours im OSCE-Format für Ultraschallkurse des Abdomens

Authors M. Hofer[1], L. Kamper[2], M. Sadlo[3], K. Sievers[1], N. Heussen[4]

Affiliations
[1] AG Medizindidaktik, University Clinic, H.-Heine-University Düsseldorf
[2] Dept. of Diagnostic and Interventional Radiology, Helios Klinikum Wuppertal, University Hospital Witten/Herdecke
[3] General/Family Medicine, General Practitioner
[4] Institute for medical statistics, University Clinic UKA

Key words
- ultrasound
- QA/QC
- technology assessment
- teaching
- assessment

received 21.10.2010
accepted 3.1.2011

Bibliography
DOI http://dx.doi.org/10.1055/s-0029-1246049
Published online February 14, 2011
Ultraschall in Med 2011; 32: 184 – 190 © Georg Thieme Verlag KG Stuttgart · New York · ISSN 0172-4614

Correspondence
Dr. Matthias Hofer, MME
AG Medizindidaktik,
Univ.-Klinikum
Moorenstr. 5
40225 Düsseldorf
Germany
Tel.: ++49/2 11/8 11 07 43
Fax: ++49/2 11/8 11 07 46
matthias.hofer@uni-duesseldorf.de

Zusammenfassung

Ziel: Die Zielsetzung dieser Studie lag in der Konzeption und Evaluation eines standardisierten Prüfungsinstruments zur Beurteilung praktischer, sonografischer Untersuchungskompetenzen des Abdomens im OSCE-Format.

Material und Methoden: Vorgestellt werden der inhaltliche Aufbau, die Logistik und der zeitliche Ablauf der OSCE-Rotationsprüfung sowie mögliche Ansätze zur Qualitätssicherung durch den Einsatz detaillierter Checklisten und einer systematischen Prüferschulung. Der Parcours wurde mit über 5000 Studierenden und 2000 Ärzten im Laufe von 15 Jahren stufenweise bis zur aktuellen Version entwickelt. Als Qualitätsparameter wurden von 626 Prüfungen die Itemschwierigkeiten und Trennschärfen der Praxisstationen und die Reliabilität der Gesamtprüfung ermittelt.

Ergebnisse: Die Trennschärfen der insgesamt 14 Praxis- und 13 Zeichenstationen erreichen Werte von 0,31 – 0,65 (Praxis: 0,30 – 0,59 und Zeichnungen 0,35 – 0,65, im Mittel 0,48 bzw. 0,50) und weisen mittlere homogene Itemschwierigkeiten von 0,78 (SD 0,02; Praxis) und 0,62 (SD 0,04; Zeichnen) auf. Cronbach's alpha betrug bei 5 Prüfungsstationen 0,69 und überschreitet ab einer Stationszahl von 9 die Grenze von 0,8.

Schlussfolgerung: Die homogene Verteilung der Itemschwierigkeiten ermöglicht einen flexiblen Austausch der OSCE-Stationen zu Parcoursversionen unterschiedlicher Länge und Reliabilitätswerte. Mögliche Adjustierungen der Bestehensgrenze sowie Einflussfaktoren auf die Akzeptanz eines solchen Prüfungsinstruments werden diskutiert. Weltweit erstmalig steht ein standardisierter OSCE-Prüfungsparcours für die Abdomensonografie zur Verfügung, der auch für summative (karrierewirksame) Prüfungen rechtssicher eingesetzt werden kann.

Abstract

Purpose: The purpose of this study was the conception and evaluation of a standardized and reliable assessment tool in the OSCE format to measure the performance and practical skills of abdominal ultrasound users in PGME.

Materials and Methods: The design, logistics, pacing and the choice of tested competencies of a rotating OSCE parcours, as well as the options for quality control using detailed checklists versus global rating scales and different approaches to the training of the involved raters are described. Over the last 15 years the parcours has undergone incremental improvement and has been used in final examinations of abdominal ultrasound courses with approximately 5000 medical students and 2000 residents and fellows. For evaluation, all item difficulties and discrimination coefficients of the individual stations and the reliability (Cronbach's alpha) were calculated for the last 626 assessments.

Results: All 14 hands-on stations showed discrimination coefficients from 0.31 to 0.65 (mean 0.48; SD 0.09). The 13 diagram stations showed mean values of 0.50 (SD 0.16). Data analysis revealed mean homogeneous item difficulties of 0.78 (SD 0.02) and 0.62 (SD 0.04), respectively. Cronbach's alpha was 0.69 with five stations and reached values above 0.8 when more than 8 stations are combined in one parcours.

Conclusion: The homogeneous distribution of item difficulties provides an opportunity for designing different OSCE versions with different levels of reliability. Several options to adjust the cut-off values, the choice of the examined contents and factors that influence the examinees' acceptance of this assessment tool for PGME or CME ultrasound courses are discussed. Overall, the values of reliability and accuracy of this assess-

ment tool are high enough to be used also for high-stakes examinations in the field of abdominal ultrasound.

Introduction

In the opinion of over 1000 German medical board examinees, sonographic examination skills are among to the most relevant core competencies in daily diagnostic algorithms of multiple disciplines. At the same time, the physicians feel that the general training in ultrasound techniques requires definitive approval [1]. Since ultrasound techniques are among the most frequently used diagnostic procedures and result in costs of approximately two billion Euros in Germany, the quality assurance of ultrasound courses for physicians is an important public health issue [2]. Therefore, a professional cost-benefit analysis requires a reliable and feasible measurement instrument for assessing the competence levels of trained physicians, in order to estimate and compare the effectiveness of different approaches to the design of ultrasound courses in postgraduate medical education (PGME). Standardized, hands-on assessment parcours in the format of OSCE ("objective structured clinical examination") are widely established internationally [3–5] and also used for final examinations [6, 7].

Examination Protocol — Examinee: (last name, first name):
Sono OSCE date: — Examiner:

Station 10 (Retroperitoneal space with aorta)

"A patient presents with suspected lymphoma: Please **1.)** Examine the entire retroperitoneal space to measure or exclude enlarged lymph nodes (LN) and **2.)** Determine the diameter of the aortic lumen **3.)** Please identify five hypoechoic/anechoic egg-shaped structures in a frozen image, which might be changed by mistake with LN." *Systematic scanning of RP space from L paraaortic to R paracaval sections, following the aorta to its bifurcation, AO-∅, shows 5 dark/black egg-shaped DDs.*

Initial handling of the transducer

Orientation:
- Correct, initial check by disconnecting the cranial part of the transducer, done by himself/herself
- Initial problems or forgotten, corrected after examiner's reminder (see: guideline) 1
- Needs assessor's help to find the adequate orientation 0

Positioning:
- Correct, puts transducer in place right away 2
- Initial problems, but can adjust to examiner's reminder (see: guideline) 1
- Needs assessor's help to position the transducer at the adequate place 0

Connection:
- Connects the transducer adequately to the skin, adjusts pressure if necessary 2
- Initial problems, but can adjust to examiner's reminder (see: guideline) 1
- Insufficient pressure, does not succeed in connecting the transducer to the skin without help 0

Adequate zoom factors:
- Adjusts the zoom factor by himself/herself immediately if necessary for the present task 2
- Initial problems, but can adjust to examiner's reminder (see: guideline) 1
- Does not achieve an adequate zoom factor even after examiner's feedback / 2° to time limits 0

Communication with the patient/model

Breathing commands:
- Correct: "Please take a deep breath – and hold it." 4
- Incomplete/initial problems (forgotten)/reminder from examiner necessary (see guideline) 2
- Even after reminder incomplete or forgets several times to ask the patient to inhale 0
- Immediately asks the patient to breathe again after freezing the image 2

Scanning performance

Scanning/following the blood vessels:
- RP space completely scanned – even with left paraaortic and right paracaval spaces 8
- RP space scanned down to the bifurcation without scanning the paravascular areas 4
- Scanning of the RP space only possible with examiner's help (see guideline) 2
- Even with examiner's support no adequate scanning possible 0

Measurement:
- Correct: perpendicular to the long axis of the vessel, vessel walls not included 5
- Wrong endpoints of the diameter/corrective feedback from examiner necessary (see guideline)
- Wrong or missing measurement – even after examiner's feedback / 2° to time limits 0

Explanation of frozen image

- Identifies five hypoechoic/anechoic egg-shaped structures (as DD to LN) correctly: Esophagus **(1)** Crura of the diaphragm **(1)** Confluens PV **(1)** Left renal vein **(1)** Duodenum **(1)** 0 - 5

Overall performance (global rating scale)

Outstanding **8 - 7 - 6 - 5 - 4 - 3 - 2 - 1** Poor 0 - 8

Theoretical background

- Lymph nodes are nodular structures – how can you differentiate them from tubular structures? 0 - 4
  Please explain two options.
  *Continuous scanning **(1)** => LN: sudden (dis-)appearance **(1)**; rotation of the transducer by 90° around its cable **(1)** LN: keeps round shape and does not become tubular **(1)***
- Please use a diagram to explain the increased risk for rupture of a partially thrombosed aneurysm with eccentric lumen in contrast to an aneurysm with concentric lumen: 0 - 6
  *Diagram of aortic aneurysm with concentric **(1)** versus eccentric **(1)** lumina: protective thrombotic ring in case of concentric lumina **(2)** / correct predilection site for rupture in eccentric cases **(2)***

Station score (max. 50):

Version 2010

Fig. 1 Exemplary examination protocol for hands-on station 10 (retroperitoneal space and aorta).

Abb. 1 Beispielhafter Bewertungsbogen für Praxisstation 10 (Retroperitoneum und Aorta).

Goals

The main goal of this study was to incrementally design a valid and reliable assessment tool allowing measurement of practical ultrasound performance skills with regard to the abdomen. Every single station of the chosen OSCE parcours should allow the differentiation of overall good examinees from overall poor examinees. They should also have a similar "item difficulty," in order to be easily interchangeable when designing subsequent assessments. The assessment should measure the handling of the ultrasound unit and the transducer, the communication with the patient during the examination and the establishment of pattern recognition of abdominal cross-sectional anatomy with predetermined time limits. The minimal number of single tasks should be elaborated to achieve an overall reliability above 0.8 in order to be able to use the OSCE assessment in high stakes examinations.

Methods

Since 1992, the OSCE assessment tool for abdominal ultrasound courses presented here has been tested and used in ten-week UGME ultrasound courses for over 5000 medical students. Since 1995, it has been used as the final examination in three-day PGME ultrasound courses for over 2000 colleagues [8], mostly residents (80%) and attendings (15%). The course participants were informed in advance by website and sign-in procedure that they were going to be assessed by an OSCE assessment at the end of their course (informed consent). To date, the training program for instructors and examiners has been completed by 116 teachers, 12 to 18 of which belong to the present team with an annual change rate between 2 and 6. The first 20 ultra-

sound teachers wrote a catalog of goals and objectives, specifying which hands-on skills and which background knowledge should be covered by the OSCE parcours in order to assess the basic skills and pattern recognition in abdominal ultrasound in a broad range of single tasks. Based on these objectives, task sheets for hands-on procedures, checklists and scoring instructions for the rates were designed. This initial material has been tested several times for complete overlap between expected and requested actions and adjusted to a level that could be mastered by 90% of 800 medical students within the given time of five minutes per task. In the first years of the study, the initial 11 hands-on stations took 3 minutes each, combined with 2 minutes for additional questions concerning background knowledge and reference values. In the last three years, the focus of 14 hands-on stations has shifted more towards the hands-on portion (4 minutes), combined with the identification of defined organs or blood vessels in frozen images and only one minute for additional questions or interpretations. Furthermore, immediate feedback lasting 1.5 minutes was introduced after each station. The maximal score remained stable at 50 credits per station. ▶ **Fig. 1** shows an example of an examiner scoring protocol for one task. In order to improve three-dimensional topographic pattern recognition, twelve standard image levels were taken from the preparation literature as diagrams to be drawn from memory and labeled by the examinees in pairs of two diagrams within five minutes. For each diagram, a checklist and scoring guideline were designed to support the raters' immediate feedback. The present version of the OSCE parcours contains immediate feedback lasting approximately 1.5 minutes after each task, so that each task requires 7.5 minutes including the time to walk to the next OSCE station. All examinees alternate between the role of sonographer, the role of supine patient and the diagram drawer (▶ **Fig. 2**). Of course, the sequence for each examinee takes into consideration that nobody in the role of patient can watch or listen to the performance of another sonographer and will have to perform the same task later in the parcours. To keep the assessment tool feasible, each course participant rotates through only 3 out of 11 (later 14) hands-on stations and 2 out of 13 diagram stations and will serve three times as a patient, without knowing which of the stations in the catalog he or she will have to master. Thus, a group of seven course participants can be tested within an examining time of 53 minutes, so that four examiners can assess seven examinees every hour. In our CME ultrasound courses, three to four parcours with 12 to 16 examiners take place simultaneously, while a student takes care of the acoustic time and change signals. Once per year, our examiners complete a special training program with video-supported role-playing, in which they act as examinees and ex-

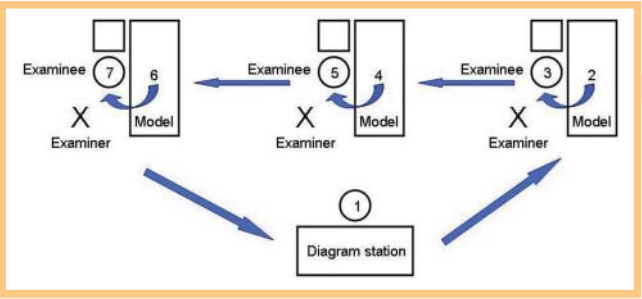
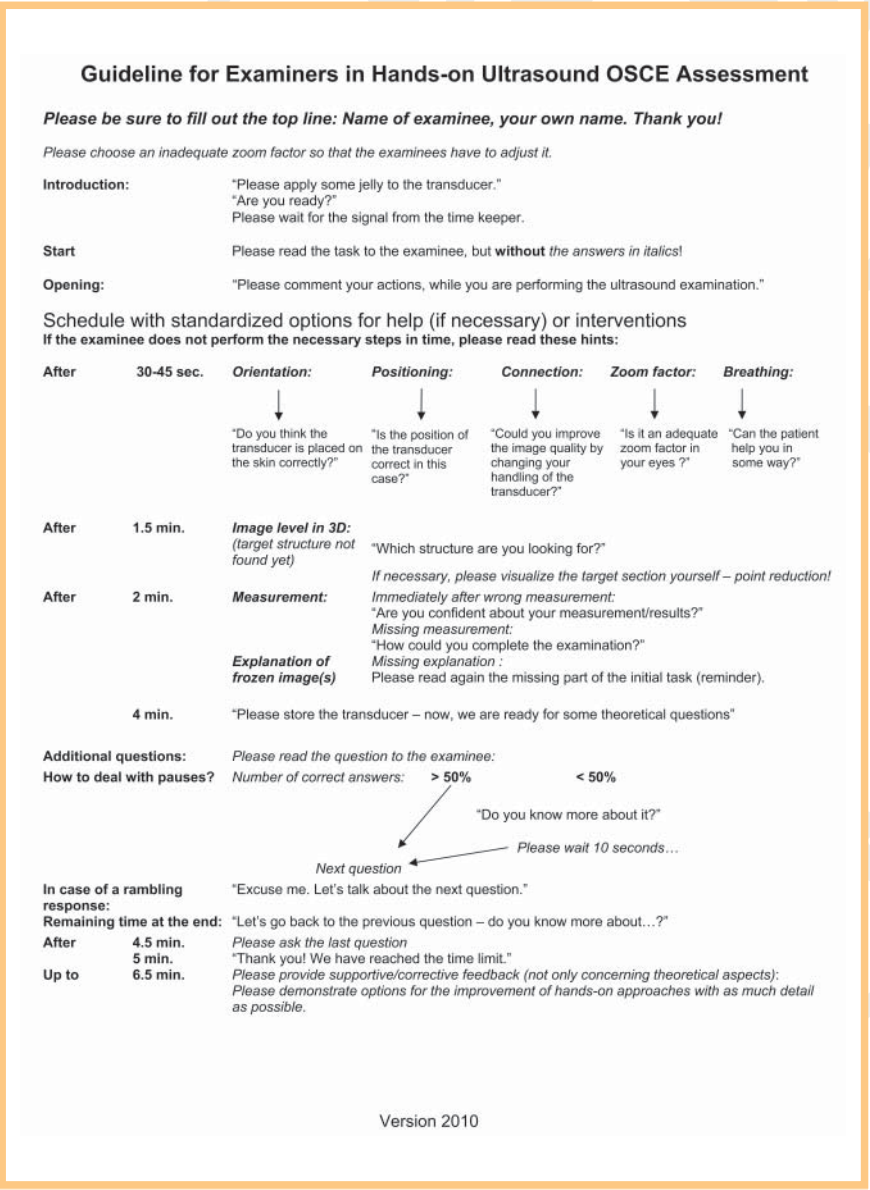


Fig. 2 Rotation path in abdominal OSCE Parcours.

Abb. 2 Rotationsmodus im OSCE-Parcours.



Guideline for Examiners in Hands-on Ultrasound OSCE Assessment

Please be sure to fill out the top line: Name of examinee, your own name. Thank you!

Please choose an inadequate zoom factor so that the examinees have to adjust it.

Introduction: "Please apply some jelly to the transducer."
"Are you ready?"
Please wait for the signal from the time keeper.

Start Please read the task to the examinee, but **without** *the answers in italics*!

Opening: "Please comment your actions, while you are performing the ultrasound examination."

Schedule with standardized options for help (if necessary) or interventions
If the examinee does not perform the necessary steps in time, please read these hints:

After	**30-45 sec.**	***Orientation:***	***Positioning:***	***Connection:***	***Zoom factor:***	***Breathing:***
		"Do you think the transducer is placed on the skin correctly?"	"Is the position of the transducer correct in this case?"	"Could you improve the image quality by changing your handling of the transducer?"	"Is it an adequate zoom factor in your eyes ?"	"Can the patient help you in some way?"

After **1.5 min.** ***Image level in 3D:*** *(target structure not found yet)* "Which structure are you looking for?"
If necessary, please visualize the target section yourself – point reduction!

After **2 min.** ***Measurement:*** *Immediately after wrong measurement:*
"Are you confident about your measurement/results?"
Missing measurement:
"How could you complete the examination?"

Explanation of frozen image(s) *Missing explanation :*
Please read again the missing part of the initial task (reminder).

4 min. "Please store the transducer – now, we are ready for some theoretical questions"

Additional questions: *Please read the question to the examinee:*
How to deal with pauses? *Number of correct answers:* **> 50%** **< 50%**
"Do you know more about it?"
Please wait 10 seconds…
Next question

In case of a rambling response: "Excuse me. Let's talk about the next question."

Remaining time at the end: "Let's go back to the previous question – do you know more about…?"

After **4.5 min.** *Please ask the last question*
5 min. "Thank you! We have reached the time limit."
Up to **6.5 min.** *Please provide supportive/corrective feedback (not only concerning theoretical aspects):*
Please demonstrate options for the improvement of hands-on approaches with as much detail as possible.

Version 2010

Fig. 3 Guideline for possible interventions of the examiner in the hands-on stations.

Abb. 3 Leitfaden für mögliche Eingriffe durch den Prüfer bei Praxisaufgaben.

aminers and also provide feedback. The scores of all examiners are compared and discussed, if necessary, until the scores do not vary by more than 8% on average. The examiners also train to provide supporting feedback to excellent participants as well as to provide corrective feedback with constructive criticism to poor performers. A special guideline has been developed for the timing of possible examiner intervention (**Fig. 3**). Furthermore, all examiners are tested since they are expected to be able to accurately draw each standard image diagram with all labels within 90 seconds and to provide feedback on the diagrams from the 180° perspective, so that they can supervise and support several course participants at once during exercises throughout the course.

Based on intermittent evaluation data, 14 hands-on skills (**Table 1**) and 13 standard images (to be drawn from memory, **Table 2** [9]) have been combined to form the present assessment catalog.

Evaluation

Each protocol has been labeled with the names of both the rater and the examinee to allow subsequent comparison with the total scores achieved by each individual and to determine the inter-rater reliability. Each station was analyzed annually with respect to the item difficulty and discrimination coefficient and has been modified or erased if the grade of difficulty differed significantly from the average of all stations or a discrimination value above 0.3 could not be achieved. This was performed annually at first, and then every 2 to 3 years. All analyses were performed with SAS® statistical analysis software, V9.1.3 (SAS Institute Inc., Cary, NC, USA): The standardized item difficulty (average percent correct across all stations contributing to the test) and discrimination (correlation between the number of points and the sum of the points in all other stations) were derived for each task. To evaluate the reliability of the OSCE assessment, Cronbach's alpha was calculated. An increase in reliability was assessed by Spearman-Brown correction [10]. For the assessment of the agreement between the global rating scale and detailed checklist, Lin's concordance coefficient [11] was calculated. This coefficient combines the Pearson correlation coefficient as a measure of precision and the bias correction factor as a measure of accuracy and results in values between –1 (perfect discordance) and +1 (perfect concordance between the two scoring methods).

Exclusion criteria

For statistical evaluation, only the last 626 examination protocols (300 PGME examinees and 326 UGME examinees) from the years 2007 to 2010 were used, all of which had been filled out completely by the examiners. Stations with discrimination coefficients below 0.3 were regarded as inadequate and will not be used for future OSCE parcours.

Results

623 of 626 examination protocols were filled out completely and could be used for statistical analysis. The item difficulties among the hands-on stations showed a mean homogeneous distribution of 77.98% with a standard deviation (SD) of 2.43% and mean discrimination coefficients of 0.48 (SD 0.09, **Fig. 4**). The values of the diagram stations were 61.83% (4.46%) and 0.50 (SD 0.16), respectively (**Fig. 5**, **Table 3**).

Table 1 Main topics of hands-on stations in abdominal ultrasound OSCE.

1	Systematic scanning of entire right kidney in long and short axis, measurement of organ size and PPI, identification of pyramids und borderline between parenchyma and pelvis.
2	Performance of collapse test of IVC under enhanced inspiration, measurement of maximal and minimal vessel diameter of IVC, identification and measurement of caudate lobe in frozen image.
3	Systematic scanning of entire left lobe of thyroid gland in long and short axis, volumetry of thyroid gland, performance of Valsalva and identification of CCA and IJV in frozen image.
4	Systematic scanning of hepatic hilum and portal vein, measurement of luminal diameter of portal vein with interpretation/normal values, identification of common bile duct and hepatic artery in frozen image.
5	Systematic scanning of entire gall bladder in sagittal and transverse section, measurement of wall thickness with interpretation and normal values, differentiation between NPO and postprandial status.
6	Systematic scanning of entire left hepatic lobe in sagittal and axial views, measurement of caudate lobe, comparison with normal values identification of one branch of hepatic artery and vein.
7	Systematic scanning of the spleen, measurement of organ size and comparison with normal values/interpretation, identification of predilection sites for accessory spleens.
8	Systematic scanning of retroperitoneal space in axial (transverse) section, systematic scanning of the entire pancreas; measurement of organ size and pancreatic duct, identification of splenic vein in frozen image.
9	Systematic scanning of the entire urinary bladder in sagittal and transverse section, measurement of wall thickness and volume with comparison to normal values; identification of prostate gland versus uterus.
10	Systematic scanning of sagittal retroperitoneal space, measurement of suprarenal and infrarenal aortic luminal distances, identification of five hypoechoic, egg-shaped differential diagnoses to LN.
11	Performance of FAST algorithm for trauma patients, identification of 8 predilection sites for free blood or hematoma in 4 frozen images.
12	Systematic scanning of right hepatic lobe in sagittal and axial views, measurement of organ size in right MCL.
13	Scanning of right hepatic lobe, measurement of luminal diameter of peripheral hepatic veins interpretation with normal values and comparison in case of acute RVF.
14	Systematic scanning of entire left kidney in long and short axis, measurement of organ size, identification of pyramids und borderline between parenchyma and pelvis.

Table 2 Abdominal standard sections for diagram-stations [9].

1	Cervical transverse view of the thyroid gland and cervical vessels (20 credit points).
2	Left paramedian sagittal view of upper abdomen of the aorta (20).
3	Right paramedian sagittal view of upper abdomen of the IVC (20).
4	Oblique view of lower abdomen parallel to iliac blood vessels (10).
5	Axial view of upper abdomen at origin of celiac trunk (20).
6	Right subcostal oblique view of hepatic veins and IVC (10).
7	Axial view of left renal vein crossing into IVC (20).
8	Right oblique view of upper abdomen parallel to portal vein (20).
9	Right sagittal and oblique view of lateral mid-abdomen, long axis of right kidney (10).
10	Right axial view of lateral mid-abdomen, transverse axis of right renal hilum (20).
11	High lateral oblique view of the spleen (10).
12	Median sagittal suprapubic view of urinary bladder and uterus or prostate gland (10).
13	Axial and transverse suprapubic view of urinary bladder, prostate gland, rectum (10).

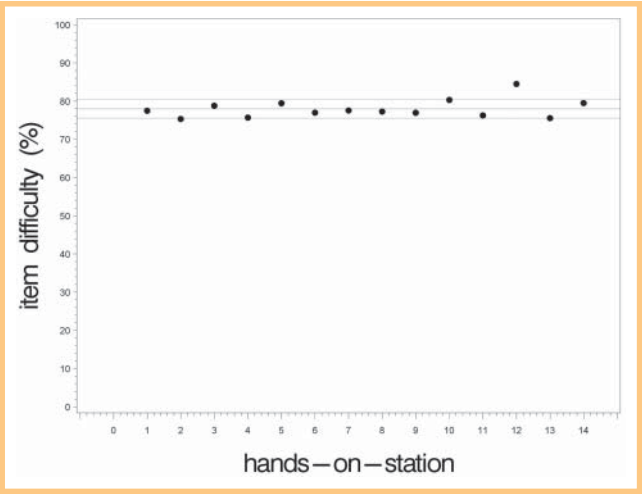


Fig. 4 Mean item difficulties of all 14 hands-on stations revealed very low standard deviations, which allows for flexible exchangeability in random sampling.

Abb. 4 Die mittleren Item-Schwierigkeiten für alle 14 Praxisaufgaben weisen nur eine sehr geringe Standardabweichung auf. Dadurch ist ein flexibler und randomisierter Austausch möglich.

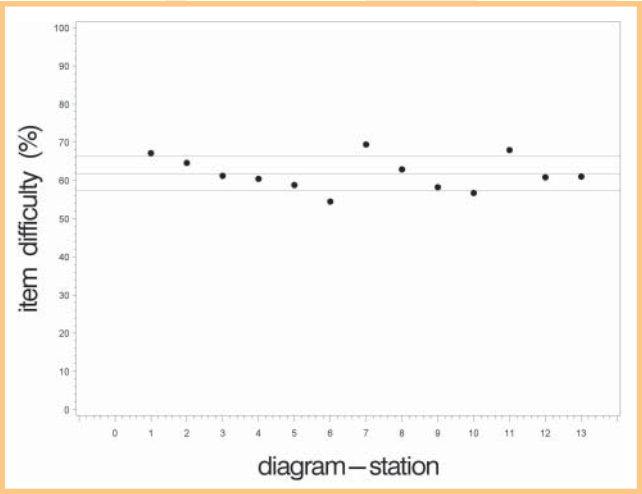


Fig. 5 Mean item difficulties of all 13 diagram stations also show a homogeneous pattern, which allows for high exchangeability among the stations to design different random samples.

Abb. 5 Die mittleren Item-Schwierigkeiten für alle 13 Zeichenaufgaben zeigen ebenfalls eine homogene Verteilung, dies erlaubt eine hohe Austauschbarkeit der Stationen im Design verschiedener Gesamtprüfungen.

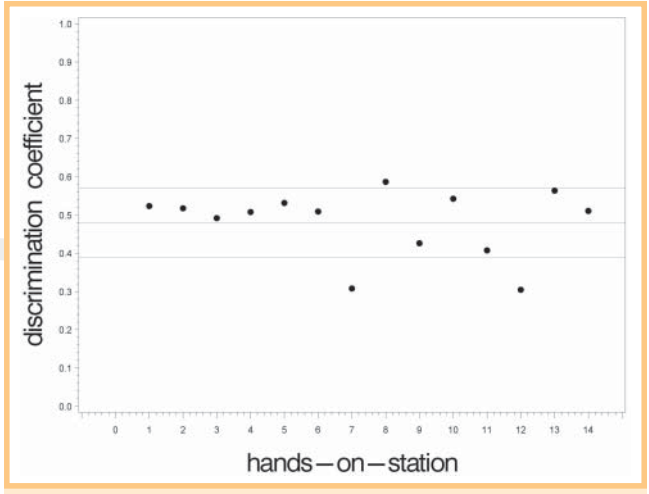


Fig. 6 High values of discrimination coefficients of 14 hands-on stations: mean value of 0.48 (horizontal center line) with a standard deviation of 0.09.

Abb. 6 Hohe Trennschärfen der 14 Praxisaufgaben mit einem Mittelwert von 0,48 (horizontale Linie) und einer Standardabweichung von 0,09.

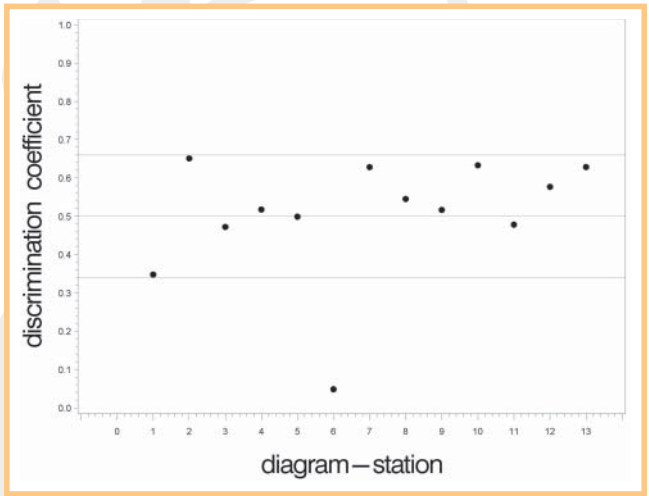


Fig. 7 Discrimination coefficients of 13 diagram stations with an SD of 0.16 around the mean value of 0.50. Station no. 6 had to be excluded from the parcours due to its insufficient validity.

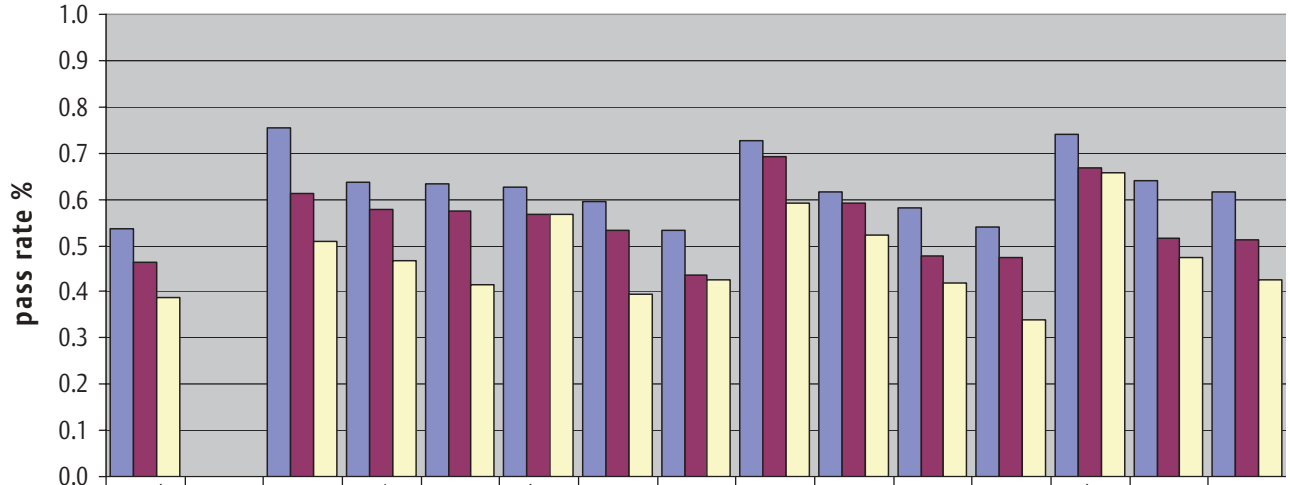
Abb. 7 Trennschärfen der 13 Zeichenaufgaben mit einer Standardabweichung von 0,16 und einem Mittelwert von 0,50. Aufgabe 6 wurde aufgrund der unzureichenden Validität aus dem Prüfungsparcours entfernt.

Table 3 Standardized item difficulties (average percent correct across all stations contributing to the test) and discrimination (correlation between the number of points with the sum of the points in all other tasks) derived for each station.

	item difficulty: mean (SD)	discrimination: mean (SD)
hands-on stations	77.98 (2.43)	0.48 (0.09)
diagram stations	61.83 (4.46)	0.50 (0.16)
diagrams without no. 6	62.44 (4.05)	0.57 (0.10)

Discrimination coefficients for the different hands-on stations and diagram stations are shown in **Fig. 6** and **Fig. 7**, respectively. One diagram station (no. 6) revealed an insufficient discrimination coefficient of 0.05 even after modification and has been excluded from the parcours.

With the given number of 3 hands-on stations and 2 diagram stations, Cronbach's alpha of the OSCE parcours was 0.69. With one additional hands-on station, this indicator for overall reliability can be expected to be 0.73 and reaches values above 0.8 when more than 8 stations are combined in one parcours. The scores of eight-step global rating scales resulted in a Spearman correlation coefficient of 0.70 (95% limits of confidence: 0.20 – 0.91) with the scores determined by detailed checklists.



	Total		TG	AO / LN	IVC	Iliac	Trunc	HVv.	LRV	PV	K sag	K tr	Spleen	UB sag	UB tr
60%	0.54		0.76	0.64	0.63	0.63	0.59	0.53	0.73	0.62	0.58	0.54	0.74	0.64	0.62
65%	0.46		0.61	0.58	0.57	0.57	0.53	0.44	0.69	0.59	0.48	0.47	0.67	0.52	0.51
70%	0.39		0.51	0.47	0.42	0.57	0.40	0.43	0.59	0.52	0.42	0.34	0.66	0.47	0.43

Fig. 8 Overview of anticipated changes in pass/fail ratios depending on different predefined cut-off values of 60%, 65% and 70%, respectively.

Abb. 8 Übersicht der zu erwartenden Änderungen der Bestehens-Rate abhängig von vordefinierten Grenzwerten jeweils für 60%, 65% und 70%.

Discussion

To date, only one standardized OSCE tool was reported for the assessment of the "FAST" algorithm in trauma patients [12], but not yet for a comprehensive and reliable OSCE parcours for broader assessment of sonographer competencies in sonography of the entire abdomen. In critical evaluations of multiple choice questions, discrimination values above 0.2 or 0.3 and reliability values above 0.8 are usually demanded for high-stakes assessments in medical education [13–15]. Compared with written tests in MCQ format, the assessment of hands-on skills by different raters is of course influenced by more and stronger confounders. Therefore, detailed checklists have been introduced in order to minimize rater-dependent deviations [16]. With this difference between both assessment formats in mind, the high discrimination values achieved here allow excellent discrimination between overall outstanding examinees and poor performers. Because of low deviations in item difficulty among the stations, there is a high degree of flexibility, and different sets of hands-on stations can easily be combined to produce new OSCE parcours without changing the pass/fail ratio of the test. Applied to the same or similar target group of course participants, the ratio can also be anticipated and changed as needed to define variable cut-off scores such as 60%, 65% or 70% (**Fig. 8**). Previous studies [16, 17] showed that a scoring method with global rating scales can produce the same results as the use of detailed checklists, but with less effort invested in the training of the examiners. However, the high correlation between both methods in our study might be caused by the fact that both methods had been applied in parallel by the same, non-blinded examiners. Therefore, a Hawthorne effect has probably confounded this correlation, so that we cannot draw any conclusions about this issue. Unsurprisingly, experienced examiners showed higher levels of inter-rater reliability than inexperienced colleagues, who underwent only one examiner training with video-supported role-playing. We developed a set of "survival guides" or manuals for the examiners, which provided hints and rules on one page with respect to how and when to verbalize supporting commands or corrections (e.g. of improper positioning of the transducer). Thus, we could establish a lower variation in the time span in which the examiners intervene and in the extent of their corrections during the tests. Consequently, the possibility of the trainee gaining further credit points in the remaining test time at one specific station could be kept closer together. In addition, the checklists set specific rules for each hands-on station with respect to which action or forgotten step would result in how many credit points or reductions (**Fig. 1**). All course participants knew the tested competencies of all hands-on stations and diagram stations in advance in order to enhance their motivation to train all necessary steps in depth throughout the course: A mild stress by "short-time challenge" in the duration of the hands-on station was also intended to force a higher level of training motivation. A short-term repetition of OSCE series with 800 medical students resulted in only small, but insignificant improvements in the final scores. A major modification in the ratio of hands-on parts versus theoretical parts of the stations (from 3/2 to 4/1) after several years was intended to reduce the testing of factual knowledge in favor of testing hands-on skills. From the educators' point of view, this modification was successful because it enhanced the participants' motivation to develop their skill with transducers and instructors, instead of trying to learn by heart only normal values and theoretical lists of symptoms, sometimes without attempting to apply this knowledge to "real" diagnostic situations. Thus, the approach of concluding CME abdominal ultrasound courses with a standardized OSCE with immediate feedback offers an effective opportunity to influence trainee behavior with respect to preparation efforts and training motivation during the courses. In our experience with the target

group of colleagues in residency, negative reactions or problems regarding the acceptance of this tool are not expected.

Limitations of this study

We could not assess the individual examinees by several blinded examiners in order to test for inter-observer reliability in the real course setting. This could only be done during the previous training sessions for the examiners, because there were not enough trained examiners available at the end of the courses. Otherwise, the time spent for the OSCEs would have doubled, which would have led to feasibility problems. In addition, one might consider using real patients or ultrasound models [18], especially for advanced ultrasound courses. Here, the authors dispensed with that for logistical and economical reasons. Further discussion will be necessary to determine whether the choice of examination topics adequately assesses the core competencies for abdominal ultrasound evaluations. One next step from the authors' point of view could be a kind of Delphi process to elaborate a broader agreement on the validity of the selected stations. An initial blueprint is now available for this process. Besides this, we covered all criteria for the evaluation of OSCE assessment tools that are required by the AMEE recommendation [19].

In conclusion, this study presents for the first time a comprehensive ultrasound OSCE parcours that shows a homogeneous pattern of item difficulty among its stations and very high values of discrimination coefficients. It is thus suitable for high-stakes examinations (Cronbach's alpha above 0.8 with 9 stations or more), in areas such as general medicine, internal medicine, surgery or final UGME examinations. One prerequisite is intensive training of all involved examiners in order to maintain adequate inter-observer reliability. Another option would be to use this OSCE parcours in ultrasound courses for medical students and technicians and in CME courses as well. Interested colleagues are invited to visit one of our ultrasound courses in Duesseldorf on five OSCE days annually [20].

Abbreviations:

AMEE = Association of Medical Education in Europe
CME = Continuing Medical Education
MCQ = Multiple Choice Questions
NPO = nil by mouth, fasting
OSCE = Objective Structured Clinical Examination
PGME = Postgraduate Medical Education
SD = Standard Deviation
UGME = Undergraduate Medical Education

References

1 *Hofer M, Jansen M, Soboll S.* Potential improvements in medical education as retrospectively evaluated by candidates for specialist examinations. Dtsch med Wschr 2006; 131: 373–378
2 *Nickenig N.* Kosten B-Bild sonographischer Verfahren in 1998. Köln: Statistik der KBV (Kassenärztliche Bundesvereinigung); 2001
3 *Harden RM, Stevenson M, Downie WW et al.* Assessment of clinical competence using objective structured clinical examination (OSCE). Br Med J 1975; 1: 447–451
4 *Petrusa ER, Blackwell TA, Rogers MS et al.* An objective measure of clinical performance. Am J Med 1987; 83: 34–42
5 *Carpenter JL.* Cost analysis of objective structured clinical examinations. Acad Med 1995; 70: 828–833
6 *Townsend AH, McIlvenny S, Miller CJ et al.* The use of OSCE for formative and summative assessment in a general practice clinical attachment and its relationship to final medical school examination performance. Med Educ 2001; 35: 841–846
7 *Reznick R, Smee S, Rothmann A et al.* An objective structured clinical examination for the licentiate. Report of the pilot project of the Medical Council of Canada. Acad Med 1992; 67: 487–494
8 *Hofer M, Mey N, Metten J et al.* Quality control of sonography courses in advanced training of physicians: analysis of present status and potential for improvement. Ultraschall in Med 2002; 23: 189–197
9 *Hofer M.* Sono-Grundkurs. Stuttgart, New York: Thieme; 2009; 6th ed: 114–115
10 *Sachs L.* Angewandte Statistik. Berlin, Heidelberg, New York: Springer; 2002: 598
11 *Lin LIK.* A concordance correlation coefficient to evaluate reproducibility. Biometrics 1989; 45: 255–268
12 *Sisley C, Johnson SB, Erickson W et al.* Use of an objective structured clinical examination for the assessment of physician performance in the ultrasound evaluation of trauma. J Trauma 1999; 47: 627–631
13 *Briggs SR, Cheek JM.* The role of factor analysis in the development and evaluation of personality scales. J of Personality 1986; 54: 106–148
14 *DeVellis RF.* Factor analysis. In Scale developement: Theory and applications. Thousand Oaks: Sage, 2003: 102–137
15 *Schultz JH, Nikendei C, Weyrich P et al.* Qualitätssicherung von Prüfungen am Beispiel des OSCE-Prüfungsformats: Erfahrungen der Medizinischen Fakultät der Univ. Heidelberg. Z Evid Fortb Qual Gesundh Wesen (ZEFQ) 2008; 102: 668–672
16 *Hodges B, Regehr G, McNaughton N et al.* OSCE checklists do not capture increasing Level of expertise. Acad Med 1999; 10: 1129–1134
17 *Regehr G, MacRae H, Reznick RK et al.* Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med 1998; 73: 993–997
18 *Terkamp C, Kirchner G, Wedemeyer J et al.* Simulation of abdomen sonography Evaluation of a new ultrasound simulator. Ultraschall in Med 2003; 24: 231–232
19 *Patricio M, Juliao M, Fareleira F et al.* A comprehensive checklist for reporting the use of OSCEs. Medical Teach 2009; 31: 112–124
20 *www.medidak.de/intensiv/sono/?page=osce last visit on Oct. 15th 2010*