Ultrasound Obstet Gynecol 2014; **44**: 693–699
Published online 9 November 2014 in Wiley Online Library (wileyonlinelibrary.com). **DOI:** 10.1002/uog.13400

Assessment of performance measures and learning curves for use of a virtual-reality ultrasound simulator in transvaginal ultrasound examination

M. E. MADSEN*, L. KONGE†, L. N. NØRGAARD‡, A. TABOR*, C. RINGSTED§, Å. K. KLEMMENSEN*, B. OTTESEN* and M. G. TOLSGAARD*†

**Department of Obstetrics, Juliane Marie Centre, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark; †Centre for Clinical Education, University of Copenhagen and The Capital Region of Denmark, Copenhagen, Denmark; ‡Department of Gynaecology and Obstetrics, Nordsjælland Hospital Hillerød, University of Copenhagen, Hillerød, Denmark; §Department of Anesthesia and The Wilson Centre, University of Toronto and University Health Network, Toronto, Canada*

KEYWORDS: criterion-based training; learning curves; simulation-based training; standard setting; ultrasound simulation

ABSTRACT

Objective *To assess the validity and reliability of performance measures, develop credible performance standards and explore learning curves for a virtual-reality simulator designed for transvaginal gynecological ultrasound examination.*

Methods *A group of 16 ultrasound novices, along with a group of 12 obstetrics/gynecology (Ob/Gyn) consultants, were included in this experimental study. The first two performances of the two groups on seven selected modules on a high-fidelity ultrasound simulator were used to identify valid and reliable metrics. Performance standards were determined and novices were instructed to continue practicing until they attained the performance level of an expert subgroup* (n = *4).*

Results *All 28 participants completed the selected modules twice and all novices reached the expert performance level. Of 153 metrics, 48 were able to be used to discriminate between the two groups' performance. The ultrasound novices scored a median of 43.8% (range, 17.9–68.9%) and the Ob/Gyn consultants scored a median of 82.8% (range, 60.4–91.7%) of the maximum sum score* ($P < 0.001$). *The ultrasound novices reached the expert level (88.4%) within a median of five iterations (range, 5–6), corresponding to an average of 219 min (range, 150–251 min) of training. The test/retest reliability was high, with an intraclass correlation coefficient of 0.93.*

Conclusions *Competence in the performance of gynecological ultrasound examination can be assessed in a valid and reliable way using virtual-reality simulation. The novices' performance improved with practice and their learning curves plateaued at the level of expert performance, following between 3 and 4 h of simulator training.* *Copyright © 2014 ISUOG. Published by John Wiley & Sons Ltd.*

INTRODUCTION

Ultrasonography has become used increasingly in many specialties, including obstetrics and gynecology (Ob/Gyn), for which it has become an integral part of the gynecological examination. Ultrasound is traditionally considered safe, but lack of operator skills may lead to diagnostic errors[1]. Little research has been carried out on how to provide the most effective initial training for healthcare providers learning ultrasonography, although hands-on experience and supervised practice are considered key elements. However, the level of supervised clinical practice has been reported to be low, and clinical training is often unstructured and without clear educational goals[2]. A recent study demonstrated large discrepancies between Ob/Gyn trainees' confidence in performing ultrasound and their expected levels of performance, which further adds to the concerns raised about the adequacy of current ultrasound training programs[3].

The emerging field of simulation-based education has been suggested for improving basic ultrasound training[2,4,5]. Simulation provides a safe, controlled and learner-centered environment, which allows repeated practice without patient discomfort or harm[6,7]. Simulation-based training may enable trainees to become familiar with image optimization and probe orientation, and to practice a systematic approach to ultrasonography

Correspondence to: Dr M. E. Madsen, Juliane Marie Centre, Rigshospitalet, University of Copenhagen, Blegdamsvej 9, 4073, DK-2100 OE, Copenhagen, Denmark (e-mail: metteelkjaer@yahoo.dk)

Accepted: 20 April 2014

Copyright © 2014 ISUOG. Published by John Wiley & Sons Ltd.

before beginning clinical training[5,8]. However, there is limited evidence on how to assess simulated performance, what elements the training should include and how much practice is needed. The first step towards answering these questions is to establish valid and reliable performance measures. Moreover, credible performance standards should be used to ensure that trainees have acquired well-defined levels of competence before entering clinical practice[9]. Such performance standards may furthermore be used for certification and remediation purposes, as well as quality assurance to answer to public accountability[10].

This study focused on transvaginal ultrasound, since potential patient discomfort and the examination's intimate nature make simulation-based training particularly advantageous. The aims were: (1) to assess the validity of simulator metrics for discriminating between different levels of competence when performing a gynecological ultrasound scan (that is, construct validity); (2) to determine a pass/fail level and an expert level of performance on an ultrasound simulator; and (3) to assess how much simulation training is needed for novice trainees to attain expert levels of performance (that is, learning curves).

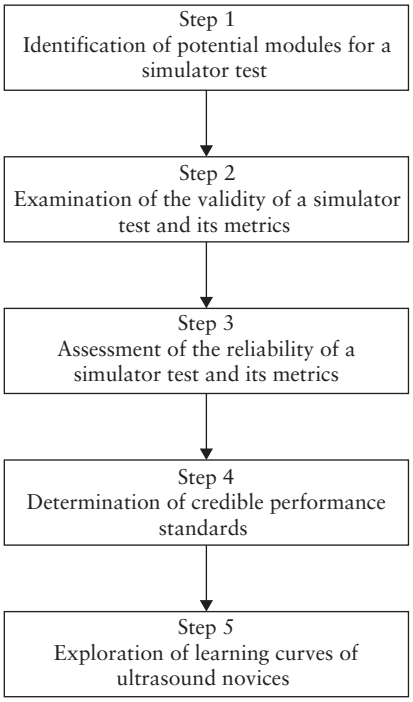

Figure 1 Flow-chart summarizing the five-step approach to assessing peformance and learning curves using a virtual-reality ultrasound simulator.

MATERIALS AND METHODS

In an experimental set-up, the validity and reliability of selected metrics on a high-fidelity transvaginal ultrasound simulator were assessed, performance standards were established and learning curves of novices were examined. The study was conducted at the Departments of Gynecology and Obstetrics at the Juliane Marie Centre, Rigshospitalet, University of Copenhagen and at Nordsjælland Hospital Hillerød, University of Copenhagen. All training and assessment were carried out in an undisturbed environment with optimal lighting conditions. The participants were recruited in March 2013 and the study was conducted between 1 April and 1 July 2013. Approval was obtained from the regional ethics committee of the Capital Region of Denmark before undertaking this study (Protocol No. H-2-2013-FSP28). A five-step approach to validity testing, assessment of reliability, standard setting and exploration of learning curves was used (Figure 1).

The participants included 16 final-year medical students who were ultrasound novices and 12 Ob/Gyn consultants who were all experienced ultrasound practitioners. All participants provided written informed consent. Participants with any kind of virtual-reality simulation experience were excluded. The inclusion criteria for the novices were no previous practical gynecological training experience and fewer than 12 months until graduation as medical doctors. The medical program at the University of Copenhagen is a 6-year traditional curriculum, and the gynecology rotations are completed during the final 6 months. The novices were recruited through the University of Copenhagen student newspaper; all students who met the inclusion criteria were enrolled. During their prior medical training, the students had completed courses in pelvic anatomy, as well as a 3-h course in abdominal ultrasound, which included ultrasound theory and hands-on training. Students with any additional ultrasound experience besides these mandatory elements were excluded.

The group of Ob/Gyn consultants included gynecologists who used ultrasound on a daily basis ($n = 8$) and an expert subgroup of fetal medicine consultants ($n = 4$). The participants were recruited locally at the departments of Gynecology and Obstetrics at Rigshospitalet and Nordsjælland Hospital Hillerød.

Training and assessment were performed using a high-fidelity simulator (Scantrainer; Medaphor™, Cardiff, UK) designed for transvaginal ultrasonography and consisting of a monitor and a transvaginal probe docked into a haptic device that provides realistic force-feedback when the probe is moved. The monitor provides B-mode ultrasound pictures obtained from real patients as well as a three-dimensional (3D) animated illustration of the probe's anatomical scan position (Figure 2). The system includes various training modules ranging from basic to advanced gynecological and early pregnancy modules.

After completing a module, the simulator provides feedback using dichotomous metrics in a number of task-specific areas (e.g. scanning through the entire uterus), as well as general performance aspects (e.g. optimizing the image sufficiently).

Step 1: identification of modules

A pilot study involving three medical students, three first-year Ob/Gyn residents and one Ob/Gyn consultant was conducted to identify modules that potentially

Copyright © 2014 ISUOG. Published by John Wiley & Sons Ltd.

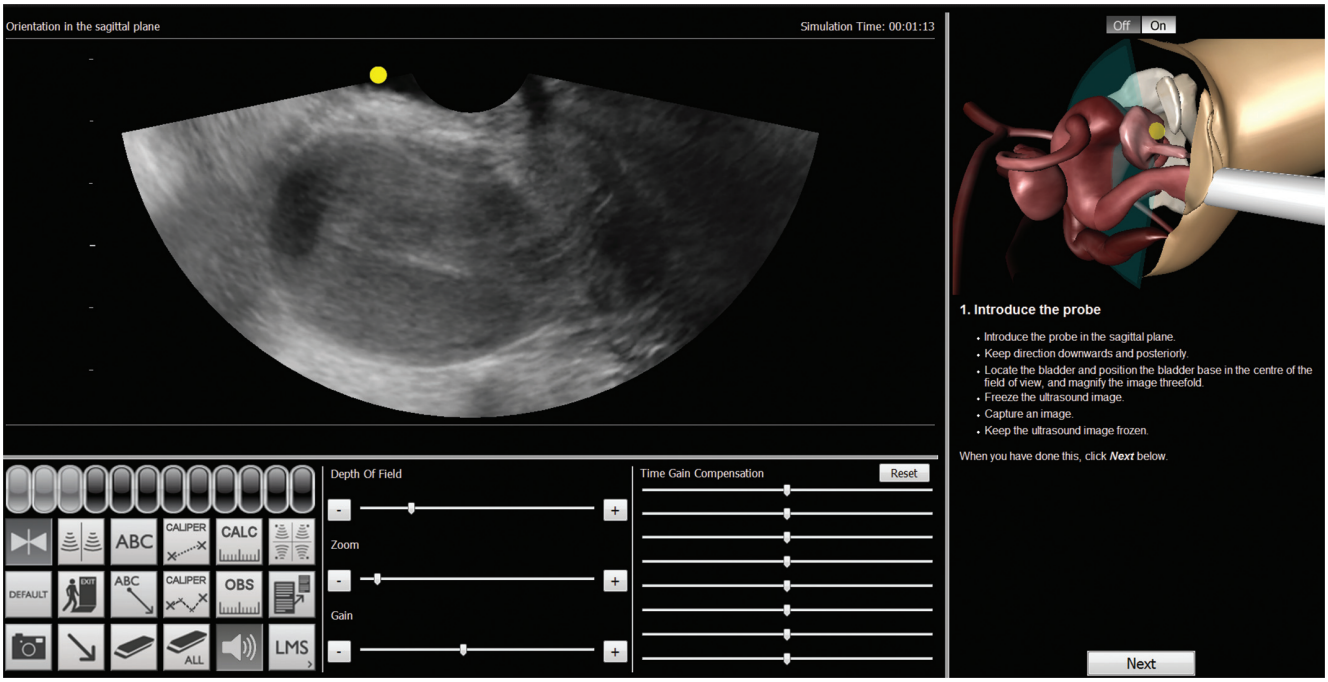

Figure 2 Monitor output on the simulator in a virtual-reality ultrasound training program. Two-dimensional ultrasound image is displayed in upper left of monitor. On right-hand side a three-dimensional animated illustration and module assignments are shown. Buttons for image optimization and documentation are on bottom left of monitor.

reflected differences in ultrasound competence. To sample broadly across different cases, modules involving different scans were included. Of these modules, those providing multiple metrics (that is, performance markers) were selected. The pilot study included 14 modules, of which seven were selected based on their face validity (i.e. participant comments during the pilot study). Performance on these seven modules was regarded as one simulator test for all future validity testing.

Step 2: validity of simulator test and its metrics

All participants received a short introduction to the simulated setting, including how to operate the simulator and its functions. Participants were then allowed to practice on a basic module without any feedback. A maximum of 15 min was allowed for participants to become familiar with the equipment, and they were then asked to complete the simulator test twice in sequence. The 3D animated illustration was concealed for all participants to minimize learning-by-testing effect. The participants had a short break (10 min) between the two rounds to prevent fatigue. Technical assistance was provided during the simulator test, but no instructions or feedback were provided.

Each simulator metric was marked either pass or fail. Metrics that were passed by fewer than 50% of the Ob/Gyn consultants were excluded from further analysis, as these were not considered stable performance measures. Only metrics that significantly discriminated between novice and Ob/Gyn consultant performance were included in the final simulator test. A sum score for all metrics with validity evidence in the seven modules was calculated by adding the scores for each (0, fail; 1, pass).

Step 3: reliability of simulator test and its metrics

The internal consistency of the metrics included in the final test was assessed for each of the initial two attempts undertaken by the participants. The test/retest reliability of the simulator test was assessed for the first and the second iteration of the test.

Step 4: determination of performance standards

Two performance standards were used in this study, a pass/fail level and a level of expert performance. The pass/fail level was determined using the contrasting-groups method[11], which uses the intersection between the sum score distribution of a group of competent performers (the Ob/Gyn consultants) and a group of non-competent performers (the ultrasound novices) to determine a level that ensures as few false positives (that is, passing novices) and false negatives (that is, failing consultants) as possible. The second performance standard, the 'expert level', was determined according to the median score of the subgroup of fetal medicine consultants. All scores were calculated as a percentage of the maximum score.

Step 5: learning curves of novices

After 2 months, all participants had completed the simulator test twice and metrics with validity evidence were determined and included in the simulator test. The

Copyright © 2014 ISUOG. Published by John Wiley & Sons Ltd.

16 novices were then asked to resume simulation-based training until they had reached the same level as, or a higher level than, the expert subgroup of fetal medicine consultants. This time, however, the 3D animated illustration and the feedback from simulator metrics were enabled and an instructor (M.E.M.) provided feedback upon completing each module. Only metrics with validity evidence were included for feedback purposes. The simulator test was repeated until the expert level was attained on two consecutive trials, and the participants' sum scores were recorded for each trial. Participants were instructed to attempt a maximum of seven iterations for all the modules in order to reach the expert level. The maximum duration of each training session was limited to 2 h to avoid trainer fatigue. The novice learning curves were examined by recording the sum score for each simulator test attempt.

Statistical analysis

Simulator-test sum scores were calculated as a percentage of the maximum possible score. Pass/fail rates were compared between the two groups using Pearson's chi-square or Fisher's exact test. If the expected count in one or more of the cells in the crosstabs was under five, Fisher's exact test was used, while if the expected count was above five in every cell, Pearson's chi-square test was used. The sum scores were compared between the two groups using the Mann–Whitney *U*-test. Time spent on the first two simulator test trials was also compared between groups using the latter test.

Finally, test/retest reliability of the first and second trials was examined using intraclass correlation coefficients (ICC), and internal consistency was assessed using Cronbach's alpha. Missing values occurred whenever participants failed to follow the instructions for the simulated task, such as saving an image or measuring a particular structure. These values were replaced by imputing group means for the relevant metrics. Participants who were lost to follow-up were not included in the calculation of total time spent on the simulator or number of attempts needed to reach expert level.

RESULTS

The demographics of the 28 participants are shown in Table 1. All participants completed the two initial validity study iterations. Three novices were lost to follow-up and passed the simulator test only once, not twice, which was the completed training criterion. These three participants were unable to attend a final training session owing to forthcoming exams.

The seven pilot study modules that constituted the simulator test included 153 metrics in total. There were significant differences between the two groups below the $P = 0.05$ level on 50 of these metrics. Two of the metrics were excluded, as only 46% of the Ob/Gyn consultants passed the first, and a higher number of ultrasound novices than Ob/Gyn consultants passed the second. Hence, the final simulator test consisted of 48 metrics (Appendix S1). The novice group had a median sum score of 43.8% (range, 17.9–68.9%) and the Ob/Gyn consultants scored significantly higher, with a median sum score of 82.8% (range, 60.4–91.7%; $P < 0.001$) (Figure 3).

Table 1 Demographics of study participants

		Experienced practitioners (n = 12)	
Parameter	*Novices* (n = 16)	*Ob/gyn consultants* (n = 8)	*Fetal medicine consultants* (n = 4)
Age (years)	26 (24–32)	45 (39–48)	51.5 (42–64)
Gender			
Female	12 (75)	5 (62.5)	3 (75)
Male	4 (25)	3 (37.5)	1 (25)
Years of clinical experience	—	10 (7–14)	23.5 (10–35)

Data are given as median (range) or *n* (%). Ob/gyn, obstetrics/gynecology.

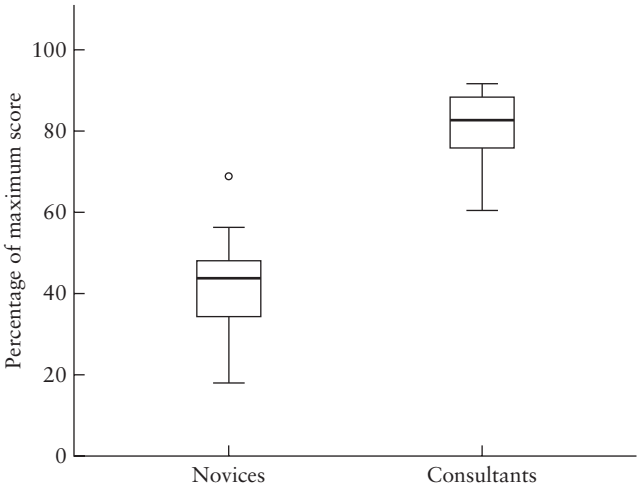

Figure 3 Box-and-whisker plots showing median sum score distribution of novices and consultants in first two iterations in a test of a virtual-reality ultrasound training program. Boxes show median and upper and lower quartiles, whiskers show range excluding outliers and the open circle is an outlier.

The novices needed a median of 81.5 (range, 51–105) min, whereas the Ob/Gyn consultants needed a median of only 53.0 (range, 49–81) min to complete the first two simulator test iterations ($P < 0.001$). Of all the metrics examined, the result of five were not recorded on the simulator. These missing data were equally distributed between the two groups.

The test/retest reliability of the simulator was high, with an ICC of 0.93. Single-measure ICC was also high (0.87) as a measure of the reliability, if the simulator test had only been attempted once. Finally, internal consistency was very high, with Cronbach's alpha of 0.95 on the first iteration and 0.92 on the second iteration of the simulator test.

The pass/fail level determined via the contrasting-groups method corresponded to a sum score of 62.9% (Figure 4). The fetal medicine consultants' median score

Copyright © 2014 ISUOG. Published by John Wiley & Sons Ltd.

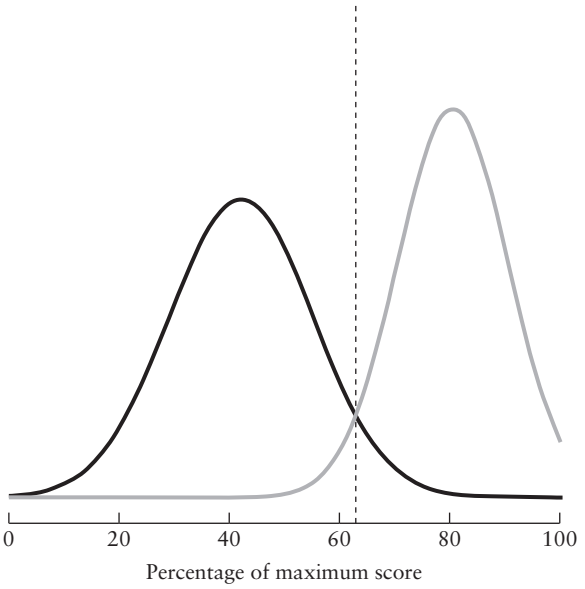

Figure 4 Pass/fail level determined using contrasting-groups method of novices (——) and obstetrics/gynecology consultants (——) in a virtual-reality ultrasound training program. Dashed line, pass/fail level.

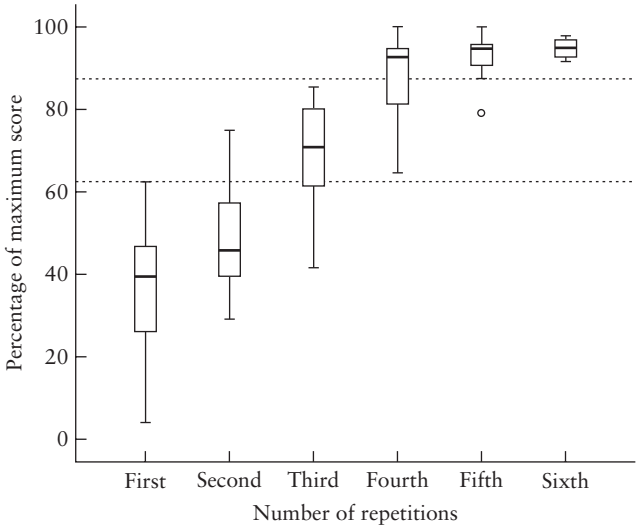

Figure 5 Box-and-whisker plots showing learning curve of novices on ultrasound simulator. Dashed horizontal lines illustrate performance standards: upper line, expert level (88.4% of maximum possible score); lower line, pass/fail level (62.9% of maximum possible score). Boxes show median and upper and lower quartiles, whiskers show range excluding outliers and the open circle is an outlier.

was 88.4% (range, 80.2–91.7%), which was used as the expert performance level. This was slightly higher than that of the Ob/Gyn consultants, who had a median score of 77.6% (range, 60.4–89.5%) ($P = 0.05$). One novice reached the pass/fail level in the first two attempts (false positive) and one Ob/Gyn consultant did not pass this level (false negative). None of the novices reached the expert level within the first two attempts.

The ultrasound novice learning curve is illustrated in Figure 5. One novice reached the pass/fail level on both the first and second attempts without any kind of feedback. After a median of three (range, one to four) attempts, the novices reached the pass/fail level. The learning curve plateaued at the expert level after a median of five (range, five to six) attempts. Three novices did not reach the expert level twice and ended the training after five ($n = 2$) or six ($n = 1$) repetitions. The sum score improved progressively for each trial, except in four cases, and the performances became more consistent throughout each trial (Figure 5). The median time spent on the simulator before the novices reached the expert level twice was 219 (range, 150–251) min.

DISCUSSION

In this study we examined the validity and reliability of a simulator test, established credible standards of performance and explored ultrasound novice learning curves on a high-fidelity transvaginal simulator. Of the 153 metrics, only 48 reliably discriminated between levels of competence and demonstrated evidence of construct validity. The ultrasound novices needed a median of five iterations of the simulator test to reach the 'expert level', which corresponded to a median of 3 h 39 min of training. To our knowledge, this is the first study to explore ultrasound novice learning curves on a simulator using valid and reliable metrics and credible performance standards.

The lack of rigorous research on simulator metric validity and learning curve characteristics for ultrasound performance may explain why no studies to date have demonstrated skills transfer from simulator to clinical performance[12]. This study examined the use of criterion-based training, in which the ultrasound novices were required to attain a predefined performance level before training was complete. Having clear educational goals and providing valid and reliable performance assessments are essential to learning in the simulated environment[6,7]. However, reliance on manufacturers' choice of metrics may not provide a valid performance assessment, as demonstrated in this study, in which only one-third of the examined metrics were valid markers. This does not necessarily mean that the remaining metrics were useless, as they may draw attention to important elements of the procedure, but they should not be included in criterion-based training or performance assessment. The majority of the excluded metrics were too easy to pass for both groups and hence provided little information. Furthermore, some of them may have failed to discriminate between different performances owing to the inherent data loss associated with dichotomous items rather than continuous scales.

Simulator-performance standards are often determined arbitrarily[13], but should be established based on their potential consequences, such as passing non-competent performers or failing competent performers[11]. Concerns regarding patient safety as well as discomfort during the initial learning phase indicate that participants should be trained to the highest possible performance level before

Copyright © 2014 ISUOG. Published by John Wiley & Sons Ltd.

ending simulation-based training and beginning clinical practice[14]. In this study, the expert subgroup of fetal medicine consultants performed slightly better than did the remaining Ob/Gyn consultants, which is consistent with a recent study on ultrasound-performance quality of Ob/Gyn consultants with a different subspecialty. In that study, fetal medicine consultants outperformed fertility consultants with respect to overall performance score[15], which supports the use of fetal medicine consultants as expert performers in the present study. Once the ultrasound novices had attained the expert level, the learning curves also plateaued; therefore, lower standards than these should not be used. This is supported by recent research suggesting that trainees should be slightly 'overtrained' to allow movement automaticity, which has been shown to improve clinical performance[16]. The contrasting-groups method has been used to determine performance standards in several recent studies[15,17], but our results indicate that this approach would result in prematurely terminating training before trainees reached a learning plateau.

This study has some limitations. Although the learning curves plateaued after 3–4 h of practice, this does not imply that the ultrasound novices were proficient operators, but rather that they were fit for supervised clinical practice. The learning curves nonetheless suggest substantial improvement in the novice participants' ultrasound skills[18].

We originally assumed that differences in years of experience were reflected in differences in competence, although recent studies have shown that number of procedures may be a better marker of competence than years of experience[19]. Furthermore, some of the Ob/Gyn consultants may actually have performed worse than expected owing to a lack of familiarity with the simulated environment and the equipment used. Finally, the study participants may not be representative of the general population of ultrasound novices, as they were all highly motivated volunteers. This may suggest that more practice is needed in the trainee population in general and hence an increase in the time needed to reach expert levels. However, having established a proficiency level will be of help in tailoring future training programs to individual learning curves.

Simulation-based ultrasound training should not be viewed as a replacement for traditional clinical training, but rather as a preparation before entering clinical practice[6]. Nonetheless, simulation-based training may be used in basic training programs to meet the calls for increased hands-on training of basic technical aspects and of a systematic approach to transvaginal ultrasonography[15]. These aspects may be taught effectively on an ultrasound simulator, which, in turn, may shorten the learning curve for trainees, decrease patient discomfort during initial clinical training and potentially reduce the number of supervised scans needed. To date, however, there have been few studies assessing the effectiveness of ultrasound simulation on clinical performance[12]. One study on ultrasound-guided central venous catheter placement found that trainees who underwent simulation-based training were more successful in placing catheters in real patients than were a control group[20], but evidence of skills transfer has not yet been demonstrated for diagnostic ultrasound simulation.

In conclusion, performance on a transvaginal ultrasound simulator can be assessed in a reliable and valid way. Credible performance standards should be used to determine when trainees are fit for clinical practice, which may be after a total of 3–4 h of simulation and feedback.

ACKNOWLEDGMENT

The Tryg Foundation funded the project.

REFERENCES

1. Moore CL, Copel JA. Point-of-care ultrasonography. *N Engl J Med* 2011; **364**: 749–757.
2. Salvesen KÅ, Lees C, Tutschek B. Basic European ultrasound training in obstetrics and gynecology: where are we and where do we go from here? *Ultrasound Obstet Gynecol* 2010; **36**: 525–529.
3. Tolsgaard MG, Rasmussen MB, Tappert C, Sundler M, Sorensen JL, Ottesen B, Ringsted C, Tabor A. Which factors are associated with trainees' confidence in performing obstetric and gynecological ultrasound examinations? *Ultrasound Obstet Gynecol* 2014; **43**: 444–451.
4. Burden C, Preshaw J, White P, Draycott TJ, Grant S, Fox R. Validation of virtual reality simulation for obstetric ultrasonography: a prospective cross-sectional study. *Simul Healthc* 2012; **7**: 269–273.
5. Nitsche JF, Brost BC. Obstetric ultrasound simulation. *Semin Perinatol* 2013; **37**: 199–204.
6. Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Med Teach* 2005; **27**: 10–28.
7. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003–2009. *Med Educ* 2010; **44**: 50–63.
8. Burden C, Preshaw J, White P, Draycott TJ, Grant S, Fox R. Usability of virtual-reality simulation training in obstetric ultrasonography: a prospective cohort study. *Ultrasound Obstet Gynecol* 2013; **42**: 213–217.
9. Gallagher AG, Ritter ME, Champion H, Higgins G, Fried MP, Moses G, Smith CD, Satava RM. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Ann Surg* 2005; **241**: 364–372.
10. Scalese RJ, Obeso VT, Issenberg SB. Simulation technology for skills training and competency assessment in medical education. *J Gen Intern Med* 2008; **23**: 46–49.
11. Zieky M, Perie M. *A Primer on Setting Cut Scores on Tests of Educational Achievement*. Educational Testing Service: Princeton, NJ, USA, 2006.
12. Sidhu HS, Olubaniyi BO, Bhatnagar G, Shuen V, Dubbins P. Role of simulation-based education in ultrasound practice training. *J Ultrasound Med* 2012; **31**: 785–791.
13. Stefanidis D. Optimal acquisition and assessment of proficiency on simulators in surgery. *Surg Clin North Am* 2010; **90**: 475–489.
14. Ziv A, Wolpe PR, Small SD, Glick S. Simulation-based medical education: an ethical imperative. *Simul Healthc* 2006; **1**: 252–256.
15. Tolsgaard M, Ringsted C, Dreisler E, Klemmensen A, Loft A, Sorensen JL, Ottesen B, Tabor A. Reliable and valid

Copyright © 2014 ISUOG. Published by John Wiley & Sons Ltd.

assessment of ultrasound operator competence in obstetrics and gynecology. *Ultrasound Obstet Gynecol* 2014; **43**: 437–443.
16. Stefanidis D, Scerbo MW, Montero PN, Acker CE, Smith WD. Simulator training to automaticity leads to improved skill transfer compared with traditional proficiency-based training: a randomized controlled trial. *Ann Surg* 2012; **255**: 30–37.
17. Konge L, Annema J, Clementsen P, Minddal V, Vilmann P, Ringsted C. Using virtual-reality simulation to assess performance in endobronchial ultrasound. *Respiration* 2013; **86**: 59–65.
18. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, Erwin PJ, Hamstra SJ. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. *JAMA* 2011; **306**: 978–988.
19. Birkmeyer JD, Finks JF, O'Reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJ; Michigan Bariatric Surgery Collaborative. Surgical skill and complication rates after bariatric surgery. *N Engl J Med* 2013; **369**: 1434–1442.
20. Barsuk JH, McGaghie WC, Cohen ER, Balachandran JS, Wayne DB. Use of simulation-based mastery learning to improve the quality of central venous catheter placement in a medical intensive care unit. *J Hosp Med* 2009; **4**: 397–403.

SUPPORTING INFORMATION ON THE INTERNET

The following supporting information may be found in the online version of this article:

Appendix S1 Simulator metrics containing construct validity.

Copyright © 2014 ISUOG. Published by John Wiley & Sons Ltd.