Assessment and learning of ultrasound skills in

Obstetrics & Gynecology

Martin Grønnebæk Tolsgaard

Table of contents

List of publications

1. 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 Apr;43(4):444-51

2. Tolsgaard MG, Todsen T, Sorensen JL, Ringsted C, Lorentzen T, Ottesen B, Tabor A. International multispecialty consensus on how to evaluate ultrasound competence: a Delphi consensus survey. PLoS One. 2013;8(2):e57687.

3. Tolsgaard MG, Ringsted C, Dreisler E, Klemmensen A, Loft A, Sorensen JL, Ottesen B, Tabor A. Reliable and valid assessment of ultrasound operator competence in obstetrics and gynecology. Ultrasound Obstet Gynecol. 2014 Apr;43(4):437-43.

4. Madsen ME, Konge L, Nørgaard LN, Tabor A, Ringsted C, Klemmensen AK, Ottesen B, Tolsgaard MG. Assessment of performance measures and learning curves for use of a virtual-reality ultrasound simulator in transvaginal ultrasound examination. Ultrasound Obstet Gynecol. 2014 Dec;44(6):693-9.

5. Tolsgaard MG, Madsen ME, Ringsted C, Oxlund BS, Oldenburg A, Sorensen JL, Ottesen B, Tabor A. The effect of dyad versus individual simulation-based ultrasound training on skills transfer. Med Educ. 2015 Mar;49(3):286-95.

6. Tolsgaard MG, Ringsted C, Dreisler E, Nørgaard LN, Petersen JH, Madsen ME, Freiesleben NL, Sørensen JL, Tabor A. Sustained effect of simulation-based ultrasound training on clinical performance: a randomized trial. Ultrasound Obstet Gynecol. 2015 Sep;46(3):312-8.

7. Tolsgaard MG, Ringsted C, Rosthøj S, Nørgaard LN, Møller LMA, Freiesleben NLC, Dyre L, Tabor A, The effects of simulation-based transvaginal ultrasound training on quality and efficiency of care. A multi-centre single-blind randomized trial. Ann Surg. 2016 Jan 25.

8. Tolsgaard MG, Tabor A, Madsen ME, Wulff CB, Dyre L, Ringsted C, Nørgaard LN. Linking quality of care and training costs: cost-effectiveness in health professions education. Med Educ. 2015 Dec;49(12):1263-71.

Preface

Ultrasound has become a core diagnostic examination in multiple medical specialties, including obstetrics-gynecology. Before ultrasound became readily available as a routine examination, clinicians had to rely on their physical examination findings when diagnosing pelvic masses and pathology during pregnancy. Today, almost every clinician in obstetrics-gynecology is using ultrasound, and unceasing technological advances have continued to provide new applications for its clinical use. Despite these developments, one key aspect of ultrasound has not changed much since its introduction, and that is the highly operator-dependent nature of the ultrasound examination. In ultrasound, the quality of the examination in terms of diagnostic accuracy depends not only on the equipment, but also on the skills of the clinician performing the ultrasound scan. Although this aspect has profound implications for patient safety, the role of training and assessment of ultrasound skills has received very limited attention until now.

My interest in health professions education started during my employment as a student teacher at Copenhagen Academy for Medical Education and Simulation (CAMES), Copenhagen University Hospital Rigshospitalet, where I did my first studies within the field of health professions education. These studies were later compiled in a PhD on the subject of undergraduate skills training. When I started my clinical training in obstetrics-gynecology at the Juliane Marie Centre, Copenhagen University Rigshospitalet, I became interested in ultrasound and in the development of ultrasound skills. Over the following years, I had the opportunity to dedicate time and receive financial support to conduct a series of studies on assessment and learning of ultrasound skills in obstetrics-gynecology in collaboration with leading ultrasound experts and medical educators. The aim of these studies, on which the present thesis is based, was to provide evidence of how to assess ultrasound skills and to explore methods to improve the basic training of novice clinicians.

I would like to express my sincere gratitude to my two mentors, Ann Tabor and Charlotte Ringsted, who throughout the years have provided their competent

advice and continuous support. Their combined expertise and guidance has shaped me as a researcher and helped me to design and conduct studies of relevance to clinicians as well as educators.

I owe thanks to all of my co-authors, who helped me conduct the studies included in this thesis – without them, there would be no studies at all. In particular, Lone Nørgaard, Åse Klemmensen, Nina Freiesleben, Eva Dreisler, Tobias Todsen, Liv Dyre, and Mette Madsen have played a significant role in the data collection for several of the studies in this thesis, for which I am very grateful. A special thanks to my colleague Maria Birkvad Rasmussen, who always offers her support and advice on on-going or new projects. Finally, I would like to thank the Juliane Marie Center, Copenhagen Academy for Medical Education and Simulation, Rigshospitalet, the University of Copenhagen, the Tryg Foundation, and the Laerdal Foundation for their financial support for the projects included in this thesis.

Summary in English

Ultrasound is a core skill in obstetrics-gynecology, but is highly operator-dependent. The evidence supporting the use of different methods for assessment and training of ultrasound skills was examined from different perspectives through a series of explorative and experimental studies.

We found that ultrasound performance of trainees in obstetrics-gynecology depended on a combination of motor skills, visual skills, and cognitive skills. We then established international multispecialty consensus on an assessment instrument designed to evaluate ultrasound skills. The validity evidence of assessments made using this instrument was then examined using empirical data on the performances of obstetrician-gynecologists with different levels of clinical experience. There was evidence to suggest that technical aspects of trainee performance may need improvement, and that simulation-based ultrasound training may play a role by allowing trainees to achieve mastery levels prior to their clinical training. We found that the use of simulation-based ultrasound training led to immediate as well as sustained improvements in trainees' performances with patients. Moreover, simulation-based ultrasound training led to improvements in patient-reported discomfort, perceived safety, and confidence in the ultrasound operator. From an organizational perspective, we found evidence that providing initial simulation-based ultrasound training combined with clinical training to reduce the need for supervised practice and repeated patient examinations. This evidence supported the hypothesis that simulation-based medical education can act as preparation for future learning. Finally, by taking a health economics perspective, we examined how ultrasound training could be linked to monetary costs, and demonstrated how training efficiency could be doubled using collaborative learning without negative consequences.

Summary in Danish

Ultralyd er blevet en essentiel del af den gynækologisk-obstetriske undersøgelse, men kvaliteten af ultralydsundersøgelsen afhænger af operatørens kompetence. Vi undersøgte, hvordan kompetencevurdering og oplæring i ultralyd kan belyses ud fra forskellige perspektiver. Fra de uddannelsessøgendes synspunkt afhænger udførelse af ultralydsundersøgelser af tekniske aspekter, evne til billedfortolkning og integrering af undersøgelsesresultater i patient-behandlingen. Vi undersøgte, hvad internationale eksperter mener, der bør inkluderes i vurdering af ultralydskompetence og understøttede disse fund med data fra kompetencevurderinger af læger med forskellige erfaringsniveauer i både den kliniske og simulerede kontekst. Vi fandt, at novicer var i stand til at opnå ekspertniveau igennem simulations-baseret oplæring i ultralyd. Simulations-baseret oplæring i ultralyd førte desuden til forbedrede kliniske præstationer både umiddelbart efter endt træning samt efter flere måneders klinisk oplæring. Derudover førte brugen af simulations-baseret oplæring i transvaginal ultralyd til forbedringer i patienters opfattelse af kvalitet i behandlingen i form af nedsat ubehag, øget tryghed og tillid til operatøren. Behovet for supervision og gentagelse af ultralydsundersøgelsen faldt over tid med den kliniske oplæring for novicer, der havde gennemgået simulations-baseret oplæring i ultralyd forud for deres kliniske træning sammenlignet med dem, der blot gennemførte den kliniske træning. Dermed kunne vi understøtte hypotesen om, at brugen af initial simulations-baseret oplæring virker som 'forberedelse til fremtidig læring'. Vi fandt desuden, at effektiviteten af simulations-baseret oplæring i ultralyd kunne fordobles ved brugen af træning i par uden negative konsekvenser for transfer af færdigheder til klinikken. Ud fra en økonomisk vinkel blev omkostningerne ved oplæring i ultralyd undersøgt og vurderet i forhold til den kliniske effekt i et omkostnings-effekt studie.

1. Background

In 1958, Ian Donald and colleagues published their seminal article on clinical application of diagnostic ultrasound in *The Lancet* (Donald et al. 1958). The authors described how they used ultrasonography in obstetrics and gynecology to visualize abdominal masses and basic fetal anatomy. In subsequent years, ultrasound was used for detection of hydatidiform mole, assessment of cephalic growth, placenta previa, and early pregnancy complications. During the 1970s and 1980s, ultrasound enabled screening for fetal anomaly and assistance during invasive procedures; in addition, the introduction of color Doppler helped identify growth-restricted fetuses and pregnancies at risk for preeclampsia. Technological advances continued during the 1990s and 2000s to include the 3D/4D scan, automated follicle count, assessment of fetal anemia, and ultrasound elastography (Campbell 2013).

The introduction of real-time ultrasound equipment allowed operators to move the probe freely around the abdomen, leading to a revolution not only in the speed of diagnosis, but also in curtailment of costs. Instead of being limited to only a few experts and researchers to use ultrasound in selected centers, ultrasound machines have increasingly been adopted by practicing obstetrician-gynecologists, midwives, residents, and even medical students over the past 50 years (Greenbaum 2003). Today, ultrasound has become as essential to the evaluation of early pregnancy complications and pelvic masses as the clinical examination. Hence, the medical applications for diagnostic ultrasound have expanded rapidly, but often rely on the use of sophisticated equipment by non-expert ultrasound operators (Moore & Copel 2011). This has caused concern because the quality of ultrasound examinations is thought to be highly operator-dependent and because ultrasound learning curves are considered quite long (Salvesen et al. 2010). For these reasons, the International Society for Ultrasound in Obstetrics and Gynecology (ISUOG) has recommended that trainees spend at least 100 hours of supervision and complete a minimum of 100 ultrasound examinations before independent practice is commenced (ISUOG 2014). The European Federation of Societies for Ultrasound in Medicine and Biology (EFSUMB) recommended even stricter criteria by suggesting that trainees should have completed at least 300 scans before performing independent ultrasound examinations (EFSUMB 2006). These recommendations reflect the notion that experience contributes to diagnostic accuracy, which find some support in the literature. For example, a study on antenatal detection of congenital heart disease (CHD) showed that sonographers with extensive experience (more than 2,000 ultrasound examinations) were more accurate in their diagnoses than their less experienced colleagues, which suggested long learning curves for complex ultrasound examinations (Tegnander & Eik-Nes 2006). However, simple tasks such as assessment of the presence of an intrauterine pregnancy may require very few supervised examinations before the operator attains a sufficient level of diagnostic accuracy (Jang et al. 2010). These large differences and the substantial individual variation in performance reported in existing studies on ultrasound learning curves suggest that the number of completed or supervised examinations is a poor predictor of ultrasound competence. However, no international consensus exists on how to assess trainees' ultrasound skills or on the level of competence that should be attained before trainees engage in independent clinical practice.

Experience may not be the only predictor of ultrasound skills (Hertzberg et al. 2000), and skill level may not be the only predictor for quality of care (Cook & West 2013). Multiple factors probably account for diagnostic failures during antenatal ultrasound screening. According to a review of 10 years of maternity claims in the National Health Service (NHS), human errors as well as lack of training and supervision were identified as areas needing further attention (NHS 2012). For intimate examinations such as transvaginal ultrasound, lack of training and supervision may also lead to increased discomfort, prolonged examination time, and repeated ultrasound examinations to address diagnostic uncertainty. Insufficient training is also considered to increase the risk for unnecessary tests and interventions and thereby poses a threat to patient safety (Moore & Copel 2011). However, to improve ultrasound training, a deeper understanding is needed of how complex diagnostic skills are developed, the challenges physicians face during training, and the most effective methods for training.

This thesis focuses on ultrasound skills development, assessment, and training in obstetrics and gynecology. The theoretical aspects of complex diagnostic skills development, training, and assessment are discussed below, and integrated with results from eight of our own studies.

2. Developing ultrasound skills

Ultrasonography may be considered a complex diagnostic skill and is likely to depend on a combination of motor skills and visual-cognitive skills. Motor skills such as hand-eye coordination are needed to operate the ultrasound equipment, which involves matching hand movements to the visual feedback provided on the ultrasound monitor. Visual-cognitive skills are also needed during image search and interpretation, while medical decision-making skills are needed to integrate the scan results into patient care.

2.1 Motor skills development

The development of motor skills described in the model proposed by Fitts and Posner (1967) includes three steps: 1) the cognitive phase; 2) the associative phase; and 3) the autonomous stage. During the cognitive phase, considerable cognitive effort is required in the conscious planning of each movement. Movements are prone to slowness, inconsistency, and error. With practice, the learner gradually moves into the associative stage, characterized by smoother and more reliable movement patterns that require less cognitive effort. After extensive practice, movements become increasingly consistent, efficient, and accurate with little or no cognitive effort required (Fitts & Posner 1967; Wulf 2007).

Research from the field of cognitive psychology on information encoding and retrieval provides an explanatory framework that aids in understanding skills development. According to information-processing theory, stimuli are identified through the sense organs and processed in the working memory (Grierson 2014). The working memory is only able to hold limited amounts of information – approximately seven elements at one time (Miller 1956) – and is thought to be controlled by a central executive function (Baddeley & Hitch 1974). This central executive function controls three types of cognitive processes, including: 1) a phonological loop, related to auditory information; 2) the visuospatial sketchpad, related to visual or spatial information; and lastly 3) the episodic memory system that binds together visual, spatial, and phonological information (Baddeley 2000). When information is processed in the working memory, it is encoded into long-term memory in the form of *schemas* (Sweller et al. 2011). Schemas are cognitive structures that tie related pieces of information together into coherent units that can be accessed during subsequent retrieval (Bruning et al. 2010). Learners as opposed to experts have limited cognitive processing capacities (Miller 1956), and working memory is therefore considered a bottleneck for information processing according to cognitive load theory (Sweller 1988). Cognitive load is divided into three parts: loads caused by the information to be learned (known as *the intrinsic load*), the *germane load,* comprised of processes that are beneficial to the act of learning, or *extraneous load,* defined as ineffective processes and instructional formats (Sweller 1988, 2011; Kirschner 2009). During complex skills learning, there is a risk of cognitive overload due to the combination of high intrinsic load with ineffective learning formats. Cognitive overload is thought to impair learning, which may be the case for novice learners who are practicing a new and complex skill such as ultrasonography. With training, the cognitive load associated with the primary task may decrease as a consequence of schema automation, when larger chunks of information are gathered into schemas and executed with less effort by the working memory. After extensive amounts of practice, the learner may free up additional cognitive resources to manage other related tasks through increasing levels of movement automaticity (Magill 2010). It is therefore reasonable to hypothesize that during the early phases of learning ultrasonography, hand-eye coordination requires substantial cognitive resources in addition to the attentional demands required from image processing and clinical decision-making. However, with extensive training, hand-eye coordination may be automated and the cognitive load required for technical aspects of the task is likely reduced.

2.2 Visual-cognitive skills

Meaningful use of medical imaging may require that users be able to detect distinct features by searching the image, as well as to decide whether a certain feature represents normal anatomy or an abnormal finding. In addition, physicians need to translate two-dimensional images as they appear on the monitor into a three-dimensional representation of the structure or organ of interest. Hence, both visual and cognitive components are responsible for search and interpretation of images (Lesgold et al. 1988; Nodine et al. 1996; Crowley et al. 2003). Visual search is considered to rely on a two-step process, an initial global impression followed by a focal search (Krupinski 2011; Crowley et al. 2003; Kundel & Nodine 1975). During this search, key features including the color, shape, and symmetry of relevant structures are identified. Perceptions of these features are continually compared and evaluated against the operator's past experiences (Kundel & Nodine 1983; Krupinski 2011). Compared to novices, experts tend to search more efficiently, require less information-gathering, and focus less on non-relevant areas (Kundel et al. 1978, 1989; Nodine et al. 1999). Novices, on the other hand, generally exhibit longer viewing times (Nodine et al. 1996), and generate fewer explicit hypotheses than do experts (Crowley et al. 2003).

The change in search patterns that accompany increasing amounts of experience may develop secondary to the acquisition of knowledge and the developments in the cognitive aspects of expertise (Kundel & La Folette 1972). With increasing levels of expertise, physicians are thought to organize past experiences in knowledge-based cognitive schemas representing a number of differential diagnoses (Krupinski 2011; Schmidt et al. 1990). These elaborate memory structures allow experienced physicians to aggregate key features and presentations of a particular medical condition or disease into larger chunks of information (Schmidt et al. 1990). The development of these elaborate chunks of information allows experienced clinicians to rely on fewer pieces of information for some diagnoses (Norman et al. 1992).

The use of chunking allows physicians to use pattern recognition in visual diagnosis, which is considered effortless and fast compared to the slow and laborious hypothetico-deductive process known as analytical reasoning (Schmidt et al. 1990). These two types of reasoning relate well to dual-process theory, which describes two systems of diagnostic processing: System 1 is characterized by unconscious, intuitive, and rapid processing, whereas system 2 is characterized by slow, effortful and analytical processing (Kahneman 2011).

Some researchers have argued that slowing down using the deliberate analytical reasoning characterized by system 2 processing may reduce cognitive bias during clinical decision-making (Kahneman 2011; Croskerry 2013). However, cognitive forcing strategies to promote system 2 reasoning have often failed to improve diagnostic accuracy, and evidence to support the notion that system 2 should be adopted over system 1 processing is conflicting at best (Monteiro & Norman 2013). Moreover, there is evidence to suggest that experts should make use of both types of reasoning processes, since visual expertise development alone is not contingent on the increased use of system 2 reasoning (Norman et al. 1992). This hypothesis is in part supported by the lack of effectiveness of cognitive and visual hinting strategies on the diagnostic accuracy of novices learning to read radiographic images (Boutis et al. 2013). Hence, in ultrasound training, efforts may be best invested in developing a sound theoretical knowledge base for the cognitive aspects of performance, as well as to ensure automation of hand-eye coordination to reduce the cognitive load associated with the technical aspects of performance for novice learners.

2.3 From theory to practice – what challenges do learners face during their ultrasound training?

From the motor-skills learning literature and medical imaging research, we may hypothesize that both motor skills and visual-cognitive skills are needed during learning and performance of ultrasonography. However, the practical challenges to learning ultrasonography in obstetrics-gynecology are less well-described (Blumenfeld et al. 2013). Other factors such as knowledge about relevant differential diagnoses, ultrasound equipment, and communication with staff and patients – as well as the ability to receive and ask for supervision from more experienced operators – may affect performance and learning (EFSUMB 2006, AIUM 2015, ISUOG 2014). Current ultrasound training methods often include apprenticeship teaching, in which learners observe senior clinicians and receive supervision during clinical training, as well as self-directed unsupported learning. Workplace-based learning has been described as *situated learning* and follows the concept of legitimate peripheral participation (Lave & Wenger 1991): Learners first observe experts, and through professional and social interaction, they gradually enter the "community of practice" as they become increasingly proficient and independent (Wenger 1998). Interaction with a senior colleague is therefore central to workplace-based learning; however, previous research has shown that requesting frequent supervision may be perceived by learners as threatening to their credibility and is therefore avoided (Kennedy et al. 2009). Moreover, the opportunistic nature of workplace-based learning and the degree of self-direction that is associated with this type of learning has led some researchers to question its effectiveness for basic clinical skills training (Tolsgaard et al. 2013 A).

Hence, a number of questions regarding ultrasound learning and performance remain unanswered, including determinants of independent practice, availability of supervision, and the role of clinical experience and training in specialized ultrasound units. Given that diagnostic performance is considered content-specific and context-dependent (Elstein 1978; Schmidt et al. 1990), evidence regarding learning and performance of ultrasonography should be compiled across multiple institutions and for several types of ultrasound examinations. In our first study, we therefore aimed at exploring learners' challenges during ultrasound performance in the Scandinavian countries to inform future training programs in obstetric-gynecological ultrasound.

2.4 Factors associated with trainees' confidence in performing ultrasound examinations.

The research questions for Study 1 (Tolsgaard et al. 2014 A) were as follows: (a) "How do clinical experience and the amount of time spent in specialized ultrasound units predict trainees' levels of confidence in performing ultrasound scans independently?" (b) "Which factors explain trainees' levels of confidence in performing ultrasound scans?" (c) "How does confidence in managing selected procedures independently relate to trainee expectations regarding their daily clinical work?" and (d) "How satisfied are trainees with their clinical training?"

We surveyed 973 trainees in obstetrics-gynecology in Denmark, Sweden, and Norway. A total of 621 eligible trainees completed the questionnaire (response rate, 70.1%). We found that clinical experience and the number of days spent in a specialized ultrasound unit were predictors for trainees' confidence in performing transvaginal and transabdominal ultrasound examinations independently (P < 0.001). It took trainees on average more than 24 months of clinical experience to manage ultrasound examinations independently, while only 12 to 24 days in a specialized ultrasound unit were needed to reach the same level. This corresponded well with the reported need for supervised practice, which seldom occurred after 24 months of clinical experience.

Contrary to our initial hypothesis, trainees did not regard requesting supervision as a threat to their professional credibility. Nonetheless, they reported significant gaps between the types of ultrasound examinations that they felt confident in performing independently and the degree to which they were expected to manage these examinations independently (P < 0.001). An exploratory factor analysis was carried out to identify which components affected trainees' confidence in performing ultrasound examinations independently. We identified three factors, including technical aspects of the ultrasound examination, image interpretation, and integration of scan results into patient care.

To date, our study is the only international survey of challenges to ultrasound learning and performance among trainees in obstetrics and gynecology. The large number of respondents and the fact that we sampled data across multiple institutions in the Scandinavian countries support the generalizability of the study results. Although the use of trainees' confidence is not a valid marker of competence on an individual level, it may be used on a group level to assess the quality of training programs (D'Eon & Trinder 2014). Moreover, our intent was not to assess the competence of the trainees, but rather to identify which factors facilitated their progress and which factors served as potential obstacles during their learning and performance.

Some important conclusions arose from this study. First, ultrasound training is a time- and resource-intensive process that requires years of clinical training before supervision is no longer needed. Second, the gaps between expected levels of performance and perceived ability suggest that clinical apprenticeship training may be insufficient, when not combined with dedicated time for basic training. However, trainees' perceptions of adequacy of ultrasound training programs in obstetrics and gynecology have been evaluated in previous and subsequent surveys. The results have varied with respect to trainees' perceptions of the adequacy of training programs, which may suggest a high degree of context-specificity of such evaluations (Lee et al. 2004; Green et al. 2015). In addition, results of the factor analysis support the hypothesis that ultrasound skills are a mix of motor skills (technical aspects of performance), visual skills (image interpretation), and cognitive skills (integration of scan results into patient care). Finally, the relatively low confidence scores on technical aspects of performance indicate that an increased focus on equipment knowledge and motor skills learning may be beneficial during basic training. These findings were supported by a recent study demonstrating that cognitive load imposed by "knobology" negatively affected novice learners' perceived utility of ultrasound for learning physical examination skills (Jamniczky et al. 2015). The load caused by image interpretation, on the other hand, was reported to enhance the perceived utility of ultrasound for learning physical examination skills. Insufficient technical skills may therefore be at odds with the acquisition of image interpretation skills, and may perhaps constitute a bottleneck for information processing when performing ultrasound examinations.

3. Mastery learning and assessment of ultrasound skills

The scientific ultrasound communities have proposed a set of minimum standards for the amount of supervision and number of scans completed before trainees are allowed to commence independent practice (EFSUMB 2006, AIUM 2015, ISUOG 2014). These recommendations do not take into consideration the different rates at which trainees may learn new skills. Consequently, some trainees may be fit for independent practice before completion of the required number of scans, whereas others may need additional training. To ensure that all trainees are at the same level before independent practice, the concept of mastery learning has gained popularity in health professions education during the past decade (McGaghie et al. 2010; Barsuk et al. 2009).

Mastery learning may be defined as the acquisition of essential knowledge and skills until a predefined performance standard is reached, regardless of the time needed to attain this level (Wayne et al. 2006). This concept of mastery learning is appealing for a number of reasons. First, training until attainment of a fixed performance standard ensures that all trainees are at the same level at the completion of training. Therefore, the only variable that differs between trainees is the time to achieve mastery learning levels (McGaghie et al. 2011 A; 2011 B). Second, mastery learning resonates well with the concept of social accountability, as trainees are first allowed to practice independently with patients only after being assessed against rigorous standards. Finally, mastery learning aligns well with the concept of entrustable professional activities (Ten Cate 2013), which describes the entrustment of different clinical tasks to trainees based on competency levels and need for supervised practice. To adopt mastery learning in ultrasound training, credible performance standards and reliable assessment instruments with sufficient validity evidence must be defined and developed. Such instruments may be used to determine which trainees should be allowed to practice ultrasound without direct supervision. In the following sections, the concepts of reliability and validity are discussed from a psychometric perspective.

3.1 Validity and reliability of performance assessment

Validity is a key concept in assessment research in medical education. Validity has been defined as the evidence supporting the interpretation of test scores (Downing 2003; American Educational Research Association 2014). In other words, validity refers to the degree to which test scores actually measure what the test has been designed to measure. Without any evidence of validity, the interpretation of test scores is meaningless, and the consequences of testing cannot be justified. Hence, the concept of validity relates to the interpretation of scores and not to an assessment instrument.

Different conceptual frameworks for validity have been proposed, of which the most recent include the work of Messick and Kane. According to Messick (Messick 1989), validity is considered a unitary concept that includes content, criteria, and consequences. In Kane's (Kane 2006) view, validity evidence is collected through different phases to build the validity argument. In the 2014 version of the *Standards for Educational and Psychological Testing* published by the American Educational Research Association, both views are supported, and validity evidence is divided into five sources. The first of these sources is *content evidence*, which was previously known as content validity. Content evidence is the documentation of the representativeness of the test contents to the achievement domains. Content evidence may be collected through expert review, blueprinting, or stakeholder opinions. *Response process,* the second category, involves the way in which a test is used and administered (Downing 2003). In evaluating response process, instructions provided during test administration and the materials available to test-takers are documented and quality control of final scores is performed. The third source of validity evidence is *internal structure*, which includes the psychometric properties of the test, such as internal consistency, item discrimination, inter-rater reliability, and factor analysis. The term reliability refers to the reproducibility of the test, which in classical test theory is a measure of the amount of error to true score among the observed scores on the test instrument (Streiner & Norman 2008). The fourth validity source is called *relationship to other variables*, previously known as construct validity (Messick 1989). The underlying ability represented by differences in test scores is in this step associated with clinical performance markers such as diagnostic accuracy – or, in the absence of such markers, clinical experience levels. Finally, the *test consequences* are explored by determining credible pass/fail levels of performance and the implications of these standards (Downing & Yudkowsky 2009).

3.2 Improving validity and reliability of test scores

The validity and reliability of performance assessments may be influenced by a number of factors that can be taken into account when designing a new assessment instrument. Experts tend to use shortcuts in both clinical reasoning and performance, whereas novices tend to display rule-bound and checklist-oriented behaviors (Schmidt et al. 1990; Norman et al. 1994). These differences in reasoning and performance may lead to paradoxes during assessment. For example, procedure-specific checklists often fail to discriminate between increasing levels of clinical expertise, and novices are sometimes assigned even higher checklist scores than experts (Hodges et al. 1999). One way to improve the validity of test scores is to use generic rating scales instead of checklists; this practice has been shown to provide better discrimination between different levels of expertise (Hodges et al. 1999; Hodges 2013). The use of excessively detailed and elaborate assessment instruments is thought to interrupt the automatic top-down processing (in other words, moving from general to specific features) of expert raters, resulting in inaccurate test scores and lower reliability (Govaerts et al. 2011). Accordingly, there is some evidence to suggest that expert raters often agree on the overall performance of trainees, but disagree over the interpretation of the scoring format (Ginsburg 2011). In one study, the reliability of test scores was improved by relating the performances of trainees to increasing levels of clinical sophistication and independence (Crossley et al. 2011). This "construct-alignment" of rating scales relates closely to the concept of entrustable professional activities (EPAs), in which trainee progress is evaluated based on the degree of clinical independence (Ten Cate 2013). However, this view assumes that levels of independence and experience reflect the development of competence, a contention that is not always supported by clinical data. For example, studies on thyroid and cardiac surgery have demonstrated surgeons' clinical experience in years correlated positively with the frequency of adverse complications (Duclos et al. 2012; Hickey et al. 2014).

Based on this evidence, multiple sources of validity evidence should be gathered to justify the use of a new assessment instrument for the evaluation of ultrasound skills. The resulting assessment instrument should be designed as a generic scale that provides scores based on the target behavior or on increasing levels of clinical independence.

3.3 Gathering validity evidence for the assessment of ultrasound skills in obstetrics and gynecology.

In studies 2 and 3, we aimed to develop a new generic instrument for the assessment of ultrasound skills (Study 2, Tolsgaard et al. 2013 C) and to collect validity evidence to support its use in obstetrics and gynecology (Study 3, Tolsgaard et al. 2014 B). Finally, we sought to establish credible pass/fail levels of performance for basic transvaginal and transabdominal ultrasound scans.

The objective of Study 2 was to establish international multispecialty consensus on the content of a generic instrument for the assessment of ultrasound skills.

We performed a Delphi study among 60 ultrasound experts from obstetrics and gynecology, radiology, urology, surgery, emergency medicine, rheumatology, and gastroenterology practicing in North America, Australia, and Europe. A list of seven items was drafted for the first Delphi round, based on a synthesis of practice recommendations from the international ultrasound societies as well as from existing imaging and assessment literature. The experts were asked to rate the importance of each of the seven items on five-point Likert scales and were also encouraged to suggest additional items. In the second Delphi round, the experts were informed regarding the distribution of scores and comments made by the expert panel during the first Delphi round. Each expert was asked to reconsider his or her ratings based on the comments from the rest of the expert panel. Two new items resulted from the first Delphi round and these items were also rated during the second Delphi round. Items that were rated important by more than 80% of participants were included in the third and final Delphi round. Descriptive anchors were added to five-point Likert scales for each of the remaining seven items. The expert panel was finally asked to provide any final comments on the outline of the assessment instrument. Of the 60 experts invited, 44 agreed to participate in the first round; out of this sample, 41 responded in the second round, and 37 completed the third round of the Delphi study. The final assessment instrument – the Objective Structured Assessment of Ultrasound Skills (OSAUS) – included seven elements; the first and last of these (*indication for the examination* and *medical decision- making*) were marked "if applicable," depending on the context of use (see Table 1). There were no statistically significant differences between countries in the ratings. Differences between raters were only observed for one item in the second Delphi round (*documentation of examination*), but this difference had no implication for the inclusion or exclusion of the item.

Our study was the first study to generate international, multispecialty consensus on the contents of a generic assessment instrument for the evaluation of ultrasound skills. The study served to establish content evidence for the use of OSAUS as an assessment instrument. The choice of including experts from multiple specialties ensured that the content of the OSAUS scale was context-independent and that more general aspects of competence were evaluated rather than just procedure-specific skills. We therefore hypothesized that the instrument could be used for assessment of both gynecological and obstetric ultrasound skills.

Table 1. The Objective Structured Assessment of Ultrasound Skills (OSAUS) scale.

1. Indication for the examination If applicable. Reviewing patient history and knowing why the examination is indicated.	1 Displays poor knowledge of the indication for the examination	2	3 Displays some knowledge of the indication for the examination	4	5 Displays ample knowledge of the indication for the examination
2. Applied knowledge of ultrasound equipment Familiarity with the equipment and its functions, i.e. selecting probe, using buttons and application of gel.	1 Unable to operate equipment	2	3 Operates the equipment with some experience	4	5 Familiar with operating the equipment
3. Image optimization Consistently ensuring optimal image quality by adjusting gain, depth, focus, frequency etc.	1 Fails to optimize images	2	3 Competent image optimization but not done consistently	4	5 Consistent optimization of images
4. Systematic examination Consistently displaying systematic approach to the examination and presentation of relevant structures according to guidelines.	1 Unsystematic approach	2	3 Displays some systematic approach	4	5 Consistently displays systematic approach
5. Interpretation of images Recognition of image pattern and interpretation of findings.	1 Unable to interpret any findings	2	3 Does not consistently interpret findings correctly	4	5 Consistently interprets findings correctly
6. Documentation of examination Image recording and focused verbal/written documentation.	1 Does not document any images	2	3 Documents most relevant images	4	5 Consistently documents relevant images
7. Medical decision making If applicable. Ability to integrate scan results into the care of the patient and medical decision making.	1 Unable to integrate findings into medical decision making	2	3 Able to integrate findings into a clinical context	4	5 Excellent integration of findings into medical decision making

In Study 3, we aimed to: 1) gather validity evidence for the clinical use of the OSAUS scale in obstetrics and gynecology; 2) determine the reliability of OSAUS ratings; and finally 3) establish credible pass/fail standards of performance.

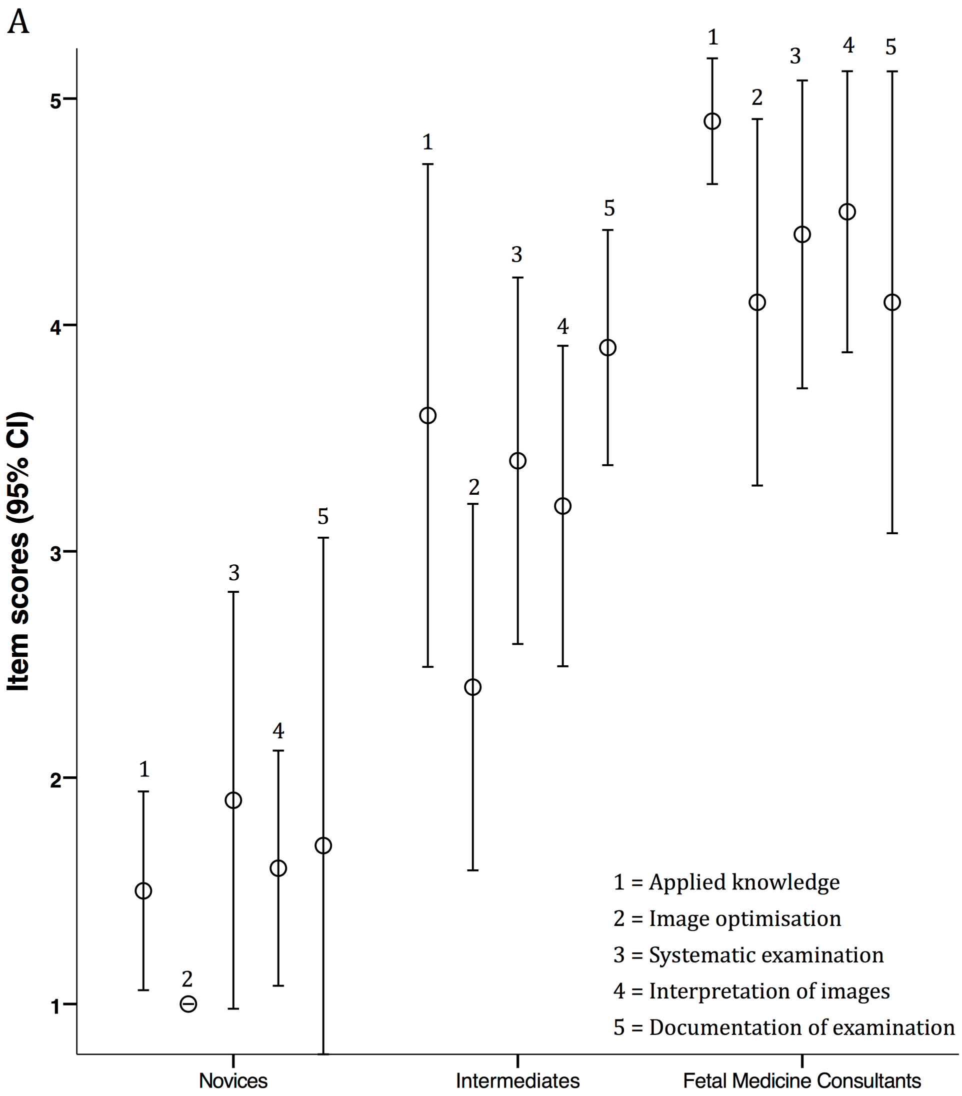
To gather data on validity evidence and reliability of the OSAUS ratings in a clinical context, we collected data on ultrasound scans performed by three groups of gynecologists with different levels of clinical experience (N=30). We included a group of novices with less than one month of clinical experience, a group of intermediates who had between 12 and 60 months of clinical experience, and a senior group consisting of consultant obstetrician-gynecologists.

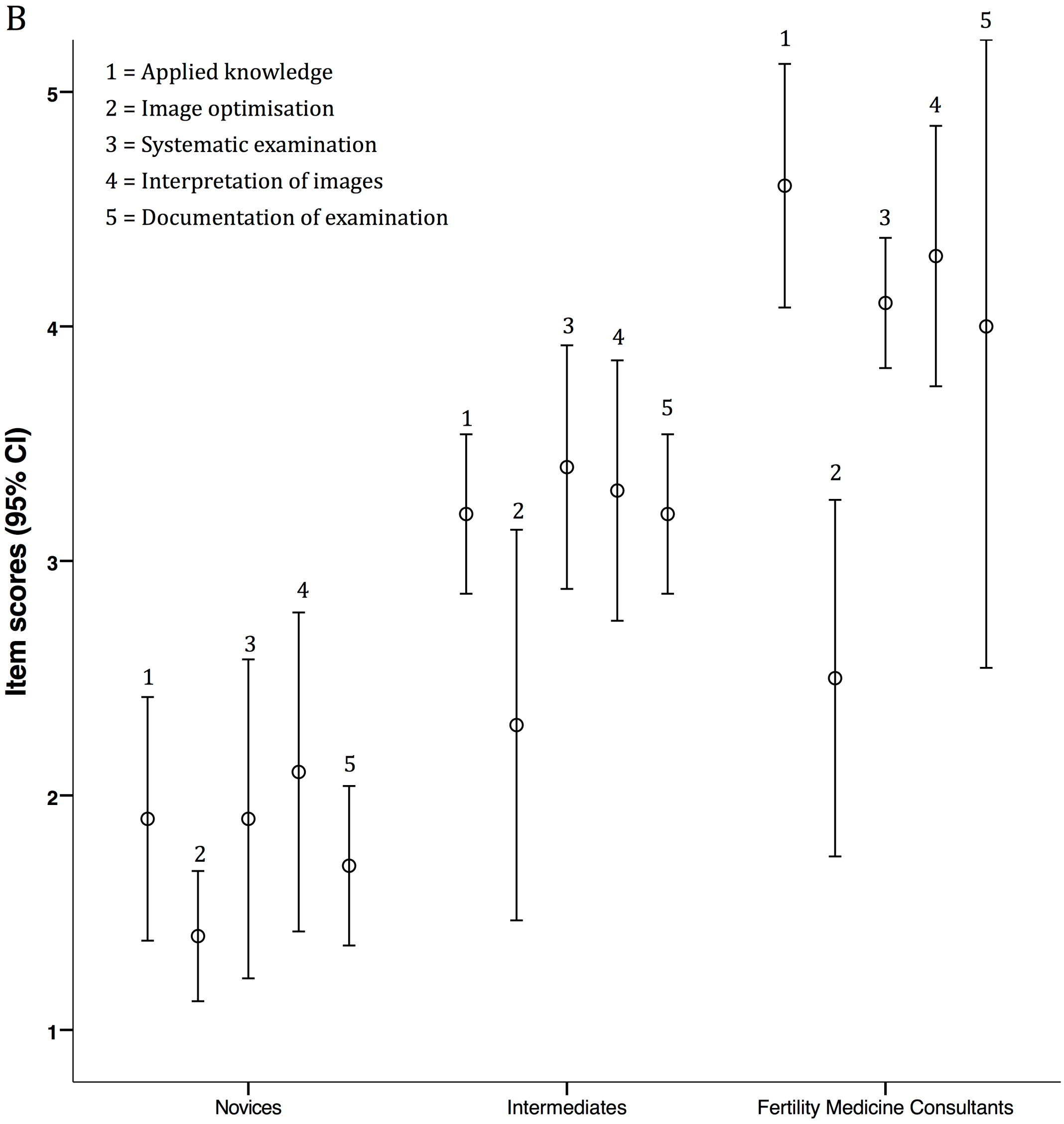
Participants were instructed to perform either a systematic transvaginal ultrasound scan or a transabdominal fetal biometry scan. The senior participants who performed the transvaginal scans were fertility medicine consultants, whereas fetal medicine consultants performed the transabdominal fetal biometry scans. Hand movements were video recorded and paired with the ultrasound output. Finally, two consultant obstetrician-gynecologists with research backgrounds in ultrasound rated the performances using the OSAUS scale.

The results of Study 3 provide validity evidence for OSAUS test scores in terms of *response process*, *internal structure*, *relationship to other variables*, and *test consequences*. The *response process* was examined through the rater training and calibration that was performed prior to the actual assessments. This calibration was performed to ensure that the raters agreed on the interpretation of test scores as well as on the expected levels of performance. We found that four videos were sufficient to reach consensus on ratings through discussion. The *internal structure* of the OSAUS item scores were supported by the high internal consistency and inter-rater reliability coefficients demonstrated, through Cronbach's alpha of 0.96 and Intraclass Correlation Coefficient of 0.89, respectively. We used clinical experience levels and use of time as proxy measures for *relationship to other variables.* There were significant differences between scores in the three groups for both the transvaginal (P = 0.003) and transabdominal scans (P = 0.003). Post hoc comparisons showed significant differences across all three experience levels. There were significant differences between fetal medicine consultants and fertility medicine consultants on their image optimization scores (P = 0.014), but no differences for the remaining items. Time to complete the ultrasound examination was not associated with

OSAUS scores (P > 0.05). *Consequences of testing* were determined using the contrasting groups method, which resulted in a pass/fail level of 50% and 60% of maximum total OSAUS score for the basic transvaginal and transabdominal scans, respectively. There were no false positives in terms of failing consultants; however, 40% of participants in the intermediate group failed the transabdominal scans when using these criteria.

Figure 1. Distribution of OSAUS scores for transabdominal ultrasound (A) and for transvaginal ultrasound (B).





Studies 2 and 3 were the first studies to establish multi-source validity evidence for the assessment of ultrasound skills in obstetrics and gynecology. According to the *Standards for Educational and Psychological Testing*, performance assessment using OSAUS scores is supported by all five sources of validity. This evidence has received further support by a subsequent validation study involving the use of OSAUS scores for assessment of transabdominal point-of-care ultrasound competence (Todsen et al. 2015). Participants in the intermediate group of our study received poor scores for their image optimization skills, which may warrant a heightened focus on technical aspects of performance during basic training. These findings are in accordance with Study 1 (Tolsgaard et al. 2014 A), in which trainees scored image optimization as the most difficult part of the examination. Interestingly, we found that fertility medicine consultants received relatively low scores on their image optimization skills compared to fetal medicine consultants. This may in part be attributed to the type of scans performed (transvaginal versus transabdominal), but may also reflect differences in the use of ultrasound for point-of-care examination versus for diagnostic purposes. Although the fertility medicine consultants were all senior clinicians, these findings may also suggest that insufficient basic skills are not automatically corrected with increasing levels of clinical experience.

We did not find that the length of time per examination was associated with OSAUS scores or with experience. While a true non-association between diagnostic performance and use of time may exist, this would be contrary to the diagnostic reasoning literature reviewed above (Schmidt et al. 1990; Krupinski 2011). Participants in the novice group were very inexperienced, which may have made them unable to complete the scan and abandon the procedure after having tried for some time. Therefore, the importance of time expenditure for ultrasound performance and quality of care needs to be addressed in larger populations of trainees with increasing levels of clinical experience.

Based on the findings in studies 1–3, we hypothesized that technical aspects of performance may be improved during basic training but that clinical training alone was insufficient to achieve mastery learning. Simulation-based medical education may be a useful method for training basic aspects of the ultrasound examination and a valuable adjunct to clinical training. In the following sections, we will review the arguments for the use of simulation-based medical education and present data for its use in basic ultrasound training (studies 4–8).

4. Simulation-based ultrasound training

Simulation can be defined as a technique “to replace or amplify real experiences with guided experiences that evoke or replicate substantial aspects of the real world in a fully interactive manner” (Gaba 2004). The use of simulators for skills learning in medical education dates back to the 17th century, when midwives practiced obstetric skills on physical mannequins to reduce maternal mortality (Buck 1991). During the 1960s, more sophisticated medical simulators were developed for resuscitation, anesthesia, and cardiopulmonary auscultation training (Cooper & Taqueti 2004). The use of simulation as a method for improving patient safety through team training increased dramatically during the 1980s and 1990s and involved the use of interactive simulators and complex simulated settings (Aggarwal et al. 2010). Training concepts and theories in simulation-based medical education (SBME) have often been inspired by the use of simulation-based training in aviation, nuclear energy, the oil industry, and the military (Page 2000). In these high-risk and high-stakes industries, simulation-based training is being used to improve safety and performance through improved communication, leadership, and decision-making skills (Aggarwal et al. 2004). In aviation, simulation-based training and assessment is now relied upon to such a great extent that in some cases, the first time a pilot takes off with a new airplane type, there are passengers on board (Page 2000).

During the past 15 years, the use of virtual reality simulators has become a key element in many surgical training programs, and considerable amounts of time and monetary resources are now invested in SBME for technical skills training (Zendejas et al. 2013 B). Several reviews have examined the effectiveness of SBME for technical skills training and have found that, compared to nothing, SBME produces superior learning outcomes (McGaghie et al. 2010, 2011 A; Teteris et al. 2012). A large meta-analysis involving 609 studies demonstrated large effects of SBME on knowledge, skills, and behaviors, and moderate effects on patient outcomes when compared to nothing (Cook et al. 2011). The potential benefits associated with SBME in terms of increasing quality and safety in care has therefore led some researchers to regard SBME as an ethical imperative in health professions education (Ziv et al. 2003). For these reasons, the World Health Organization (WHO) now strongly recommends that educational institutions use SBME in training future health professionals (WHO 2013).

4.1 Theoretical foundations of SBME

There are several purported advantages associated with SBME. The opportunity for repeated practice in a safe environment, in which there is no risk of patient harm, is often highlighted as an important factor (Issenberg et al. 2005). However, repeated practice alone is not always enough to attain high levels of performance but deliberate strategies and methods are often required to improve performance under the guidance from expert teachers (Ericsson et al. 1993). The combination of repeated practice and expert supervision enables what in the expertise literature is referred to as *deliberate practice*, which is thought to be a determinant for the acquisition of expert levels of performance in virtually any domain of expertise (Ericsson et al. 1993). According to Ericsson's concept of deliberate practice, expert performance is attained through deliberate efforts to improve and extended periods of practice over several years. Prolonged practice beyond achieving a set training criterion – also known as *overlearning* or *automaticity training* – has been shown to improve long-term retention as a function of the amount of additional practice (Driskell et al. 1992), as well as skills transfer (Stefanidis et al. 2012). This again resonates well with cognitive load theory, as the cognitive load associated with the task at hand is thought to decrease with increasing levels of schema automation in long-term memory (Sweller et al. 1988). In this view, expertise is thought to develop through deliberate and extended periods of practice rather than as a result of innate ability. However, whether learners engage in deliberate practice depends on their motivation, the available amount of monetary and time resources, as well as their access to expert supervision and feedback (Ericsson et al. 2006).

SBME allows repeated practice in an authentic environment that mimics the clinical setting, while allowing educators to control and direct training in ways that would not be possible during clinical training (Gaba 2004; Issenberg et al. 2005). The use of SBME is therefore thought to provide optimal conditions for deliberate practice, and deliberate practice is considered by many to be a keystone for effective learning in the simulated setting (McGaghie et al. 2010). However, the specific requirements for practice to become deliberate are usually not described in greater detail in the SBME literature, and there is limited evidence that trainees automatically engage in deliberate practice when presented with optimal training conditions. A second proposed keystone for effective learning in SBME is the use of mastery learning (McGaghie et al. 2011 B). According to a recent meta-analysis, there is some evidence to support the adoption of mastery over non-mastery learning, although the number of available studies is limited and the authors did not demonstrate significant effects of mastery learning on patient-related outcomes (Cook et al. 2013 A). This may in part be explained by the ill-defined mastery learning levels, as there is no consensus on which standards should be used for the assessment of mastery (Cook et al. 2013 A).

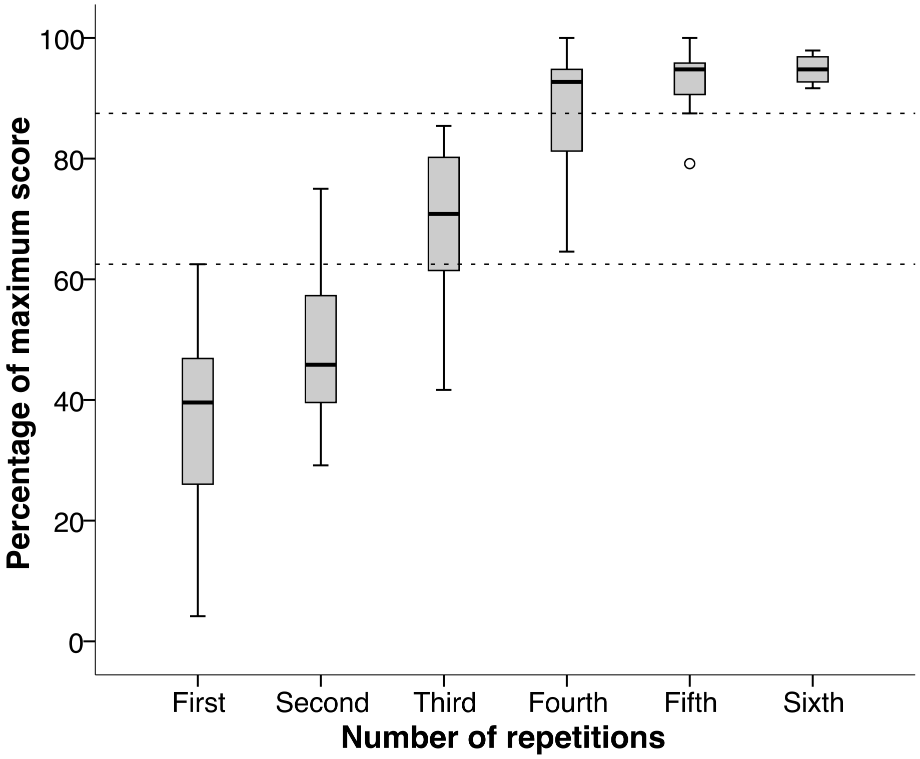
There are several indications that SBME may be a useful adjunct to basic ultrasound training in obstetrics and gynecology. However, there is limited evidence of the effectiveness of SBME on complex diagnostic skills (Teteris et al. 2012) such as ultrasonography, which requires a combination of motor skills as well as visual-cognitive skills. We hypothesized that mastery learning using SBME may be a useful adjunct to clinical training by improving technical aspects of performance. As discussed above, mastery learning relies on the achievement of pre-specified learning goals using reliable and valid performance assessments. Performance assessment in the simulated setting may be done through expert supervision or through built-in automated simulator data on performance (i.e. simulator metrics), which is available with most virtual reality (VR) simulators (Aggerwal et al. 2010; Issenberg 2005). A variety of performance standards may be used, and may include pass/fail levels that discriminate between competent and non-competent performers as well as expert levels of performance (Downing & Yudkowsky 2009). In Study 4, we aimed to develop reliable and valid performance assessments in the simulated setting and determine credible performance standards that may be used for the adoption of mastery learning.

4.2 Assessment of performances in the simulated setting

The objective of Study 4 (Madsen et al. 2014) was to: 1) determine the validity evidence supporting the use of automated simulator metrics for the assessment of transvaginal ultrasound skills in obstetrics and gynecology; 2) establish credible performance standards; and 3) assess learning curves for transvaginal ultrasound in the simulated setting.

We conducted a pilot study to identify training modules on a VR simulator designed for training transvaginal ultrasound skills (Medaphor, Cardiff, UK). Seven modules were selected, based on their capabilities for representing different types of cases and on the responses elicited by pilot group participants. To examine the simulator metrics' relationship to other variables, 16 ultrasound novices and 12 OB/GYN consultants (eight gynecologists and four fetal medicine consultants) were asked to complete the seven training modules twice. Simulator metrics that significantly discriminated between novices and OB/GYN consultants were selected for a simulator test. Finally, performance standards were established using the contrasting groups method as described in Study 3 (Tolsgaard et al. 2014 B), and an expert performance level was determined according to the scores of the sub-group of fetal medicine consultants. The novice participants were then instructed to continue training on the seven modules until they scored at the expert performance level twice.

Figure 2. Learning curves and performance standards on a virtual reality ultrasound simulator. The lower dotted line represent the pass/fail criterion and the upper dotted line represents the expert performance level.



The seven training modules identified from the pilot test included 153 simulator metrics, of which 50 metrics discriminated between novices and OB/GYN consultants below a significance level of 0.05. On the simulator test that included these simulator metrics, the median scores of the novices and OB/GYN consultants were 43.8% (range, 17.9–68.9%) and 82.8% (range, 60.4–91.7%; P <

0.001), respectively. The test-retest reliability was high (ICC = 0.93), and the internal consistency was Cronbach's alpha = 0.95 on the first iteration of the test. A pass/fail level of 62.9% of maximum simulator score was estimated using the contrasting groups method, and the expert performance level demonstrated by the fetal medicine consultants was determined at 88.4% (range, 80.2–91.7%). This was slightly higher than the consultant gynecologists, whose median score was 77.6% (range, 60.4–89.5%; P = 0.05). The novices needed a median time of 3 hours 39 minutes (range, 150–251 minutes) to attain the expert performance level.

Study 4 demonstrated that performance could be assessed in a reliable and valid way using a VR ultrasound simulator and that novice trainees could attain expert levels of performance at selected tasks in the simulated setting within an average of three to four hours of hands-on practice. To support the use of mastery learning, we adopted the expert performance level as the training criterion for the novice participants. The mastery learning approach was supported by the findings that the novice participants continued improving beyond the pass/fail level, and that their performances first plateaued after surpassing the expert performance level. Interestingly, we found significant performance differences between consultant gynecologists and fetal medicine consultants on their simulator scores. This relates well with the findings from Study 3 (Tolsgaard et al. 2014 B), where fertility medicine consultants scored significantly lower on their image optimization skills compared to the fetal medicine consultants. The fact that the clinicians included were subject matter experts in different domains of practice (gynecology, fertility medicine, and fetal medicine) may well explain the observed differences. The findings also resonate well with research in diagnostic reasoning, demonstrating differences in the methods used by generalists and specialists during their diagnostic processes (Simpson et al. 1987). In particular, the use of clinical information (van der Gijp et al. 2014) and knowledge of anatomy and image acquisition are thought to influence medical imaging diagnosis and decision-making (Lesgold et al. 1988).

Study 4 demonstrated that novice learners *can* attain expert performance levels during simulation-based ultrasound training. However, the extent to which the large performance improvements observed in the simulated setting in fact *do* translate into improved ultrasound performances with patients is not known. In the following section, the concept of transfer of learning is reviewed in relation to its theoretical foundations, and methods for improving transfer are discussed in relation to SBME.

4.3 Transfer of learning

Transfer of learning can be defined as application of previously learned knowledge or skills to a new problem, context, or domain (Kulasegaram 2013). The concept of transfer can be traced back to Plato (Plato 380 BC) and his descriptions of how mathematics and geometry may help the development of higher-order thinking skills. In the early 1900s, Thorndike and Woodworth conducted their seminal studies on transfer of learning that led to the identical elements theory. According to identical elements theory, transfer of learning is dependent on the degree to which two tasks contain identical key elements; therefore, training in one function rarely leads to improvements in another function (Thorndike & Woodworth 1901). The behaviorist view that transfer is a specific response to certain stimuli has led to some disappointing conclusions regarding transfer (Detterman 1993), which may call into question the effectiveness of any type of training. However, learners are often exclusively assessed based on their ability to repeat the learned information (replicative knowledge or "knowing that") or on their direct application of skills in a new context (applicative knowledge or "knowing how") (Broudy 1977). Educational interventions may be considered ineffective if learners are measured only on "knowing that" or "knowing how". By contrast, the concept of "knowing with" proposed by Broudy (Broudy 1977) provides a way to appreciate how learners use prior knowledge to improve their interpretation, perception, and judgment of new situations. Bransford and Schwartz built on Broudy's notion of knowing with by arguing that transfer should be evaluated based on how educational activities prepare learners to learn from new experiences, rather than on how learners perform immediately after training. Accordingly, the purpose of training is not to make people experts, but to "place them on a trajectory towards expertise" by acting as *preparation for future learning*(PFL) (Bransford & Schwartz 1999).

With regard to health professions education, most studies involving SBME have focused on immediate transfer outcomes (Grantcharov et al. 2004; Stefanidis et al. 2012; Larsen et al. 2009) and only a few studies have examined the long-term consequences of training interventions for performance, learning, and transfer (Barsuk et al. 2009, 2010; Curtis et al. 2013). Hence, the majority of existing studies of SBME have focused on transfer as direct application rather than from a PFL perspective, and the implications of SBME for subsequent clinical training are therefore largely unknown. Given that most educational interventions produce an effect on learning (Cook 2012; Norman 2014), it may come as no surprise that *some* degree of transfer follows the use of SBME. The real question is rather how learners are instructed most effectively during simulation-based ultrasound training to facilitate transfer, as well as how structured initial training using simulation may act as preparation for future learning in the clinical workplace. To answer these clarification questions (Cook et al. 2008), we examined methods for improving learning and transfer in the controlled experimental setting, in addition to the role of simulation-based ultrasound training as preparation for future learning in the clinical setting.

4.4 Improving learning and transfer following simulation-based ultrasound training

A prerequisite for any transfer is that some learning has occurred, although improvements in learning are only moderately correlated with transfer (Colquitt et al. 2000). Several factors may affect learning and thereby transfer, including factors relating to the *individual, context,* and *task* (Ringsted et al. 2006). Individual factors related to learning and transfer include general cognitive skills, motivation, and self-efficacy (Burke & Hutchins 2007), of which SBME is thought to stimulate the latter two (Issenberg et al. 2005). Contextual factors may involve supervision, the opportunity to perform the task, and support from supervisors and peers (Burke & Hutchins 2007; Lave & Wenger 1991). Finally, instructional strategies for learning new *tasks*, such as distributed learning, mixed practice, and automaticity training, have also been shown to benefit learning and transfer (Druckman & Bjork 1994; Burke & Hutchins 2007), and have received empirical support in the SBME literature (Stefanidis et al. 2012, Cook et al. 2013 B, Hatala et al. 2003). From a constructivist point of view, instructional strategies that rely on promoting learners' meta-cognition, self-direction, and reflection may also affect learning, although these aspects have received less attention and their effectiveness has been questioned (Kirschner et al. 2006). According to Chi's *active-constructive-interactive* framework, learning is promoted by adoption of certain activities that may be passive, active, constructive, or interactive. *Passive* activities (like observing a demonstration) are thought to be less effective for learning than *active* activities (such as performing an action), which are in turn inferior to *constructive* activities (such as producing an output that contains new ideas). At the top of the hierarchy, Chi placed *interactive* activities, which are dependent on interaction between learners and experts or peers, and allow learners to build on each other's ideas and inputs through sequential construction. Interactive activities are considered to stimulate cognitive co-construction and shared mental models of the to-be-learned information (Chi 2009). Moreover, from a cognitive perspective, interacting with peers may help reduce the cognitive load associated with the task at hand (Kirschner et al. 2009). According to a social learning perspective, instructional strategies that promote collaborative learning may result in improved motivation and self-efficacy through positive interdependence (Johnson & Johnson 2009). Finally, from a motor-skills learning perspective, there may be considerable benefits associated with peer observation but also reduced hands-on time, which may impair the development of skills automaticity (Shea et al. 1999; Granados & Wulf 2007; Rizzolatti & Craighero 2004). There is some evidence in the health professions education literature to support the use of collaborative learning of clinical skills (Tolsgaard et al. 2013 B; Bjerrum et al. 2014; Räder et al. 2014). However, there is no evidence documenting the effects of collaborative learning on transfer of skills. There are several potential advantages associated with the use of collaborative learning during simulation-based ultrasound training. First, collaborative learning increases training efficiency by increasing the number of trainees per simulator as compared with single training. Second, and in accordance with the theoretical advantages outlined above, the use of collaborative learning may also contribute positively during transfer of skills to the clinical setting. In Study 5, we therefore examined how the use of collaborative learning in terms of training in pairs (dyad training) affects learning and transfer to the clinical setting.

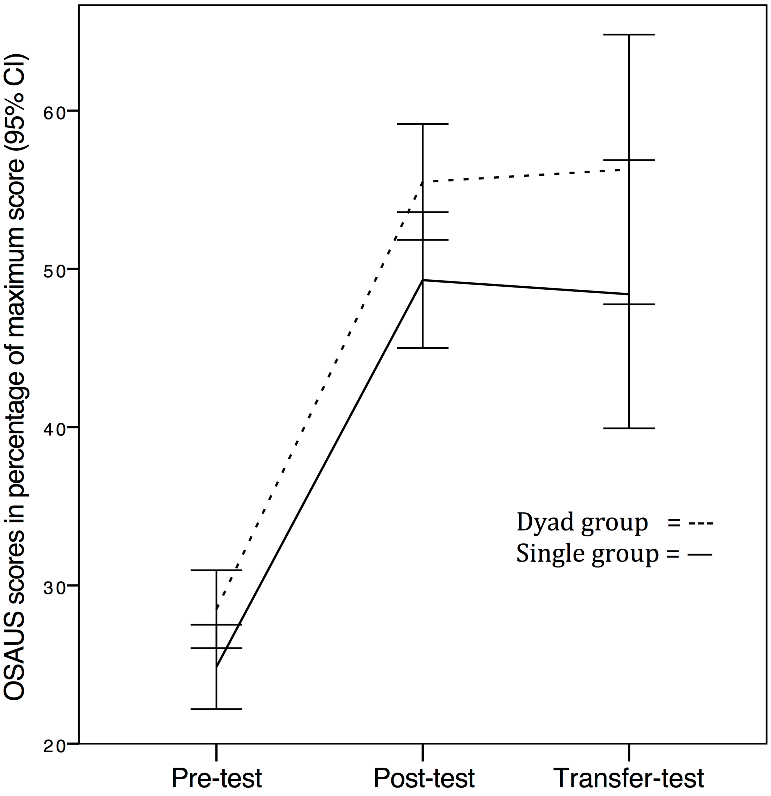
4.5 The effectiveness of dyad training on skills transfer after simulation-based ultrasound training

The objective of Study 5 (Tolsgaard et al. 2015 A) was to determine the effectiveness of dyad compared to individual simulation-based transvaginal ultrasound training on skills transfer to the clinical setting.

We used a randomized non-inferiority design, in which we chose a predefined margin of 4.6% as the least educational meaningful difference, according to findings in Study 3 (Tolsgaard et al. 2014 B). Final-year medical students were randomized to dyad or single practice on a virtual reality transvaginal ultrasound simulator. The students were instructed to practice for two hours on the same modules that were included in the simulator test developed for Study 4 (Madsen et al. 2014). The single practice group practiced alone, whereas the dyad practice group took turns as the active participant and observer, and dialogue between participants was allowed. A pre-test and post-test were performed involving a basic systematic ultrasound scan of a normal female pelvis. On the following day, participants were instructed to perform a systematic transvaginal scan on an actual patient in the gynecological ambulatory unit. Pre-test, post-test, and transfer-test performances were scored by one of two blinded raters using the OSAUS scale.

Thirty participants were randomized and 24 completed the transfer test. The dyad group scored 7.8% (95% CI, -3.8 to 19.6%) higher on their transfer-test OSAUS scores than the single group. This difference was significantly above the non-inferiority limit (P = 0.04) but included zero. When using the pass/fail standards that were developed in Study 3 (Tolsgaard et al. 2014 B), there were significantly more dyad participants who passed the transfer test compared to single group participants (dyad group, 71.4%; single group 30.0%; P < 0.05). There were no interaction effects between the intervention and the simulation-based training with respect to pre-test and post-test performances. However, the dyad group had a higher training efficiency when compared to the single group, with a mean simulator test score of 5.8 (SD 1.13) points/attempt compared to 2.8 (SD 0.92) points/attempt (P < 0.01), as shown in Figure 3. Large effects of training (Cohen's d = 3.85) were demonstrated for both groups, and mean transfer-test scores differed by less than one percentage point from the post-test scores.

Figure 3. Pre-test, post-test, and transfer-test performances of participants randomized to dyad training or single training.



Study 5 was the first study to demonstrate skills transfer following simulation-based ultrasound training using assessment instruments with established validity evidence. We demonstrated that training efficiency could be doubled without any consequences for transfer of learning after simulation-based ultrasound training. The fact that more dyad participants than single participants passed the pass/fail level on the transfer test may even suggest superiority of the use of collaborative learning, although there were no significant differences in mean scores between groups. Previous studies involving collaborative learning of clinical skills (Tolsgaard et al. 2013 B; Räder et al. 2014; Bjerrum et al. 2014; Shanks et al. 2013) have shown mixed results regarding the effectiveness of dyad training on learning. Whereas some researchers have proposed that the effect of dyad training relies on the experience levels of the learners (Shanks et al. 2013), cognitive load scientists have theorized that task complexity was the main determinant of the effectiveness of collaborative learning (Kirschner et al. 2009). Given the empirical data from the current and other studies involving dyad training, we proposed that the beneficial effects associated with collaborative learning in terms of reduced cognitive load and peer-support are balanced against the potentially negative consequences of reduced hands-on time (Tolsgaard et al. 2016 A). The benefits of dyad training may therefore be dependent on time on task, as cognitive load decreases when learners become more proficient and require increased amounts of hands-on practice to achieve skills automaticity (Tolsgaard et al. 2016 A, Räder et al. 2014). Consequently, we may hypothesize that learners benefit from collaborative learning during early skills acquisition, and that at some point during training, they benefit more from individual training before reaching mastery learning levels.

However, irrespective of instructional method the question regarding how well learning is transferred from the simulated to the clinical setting still remains. In particular, the sustained impact of SBME on learning has received limited attention in the literature. In Study 5, we examined the sustained effect of SBME on ultrasound skills after the first two months of clinical training.

5. The impact of simulation-based ultrasound training on clinical performances and quality of care

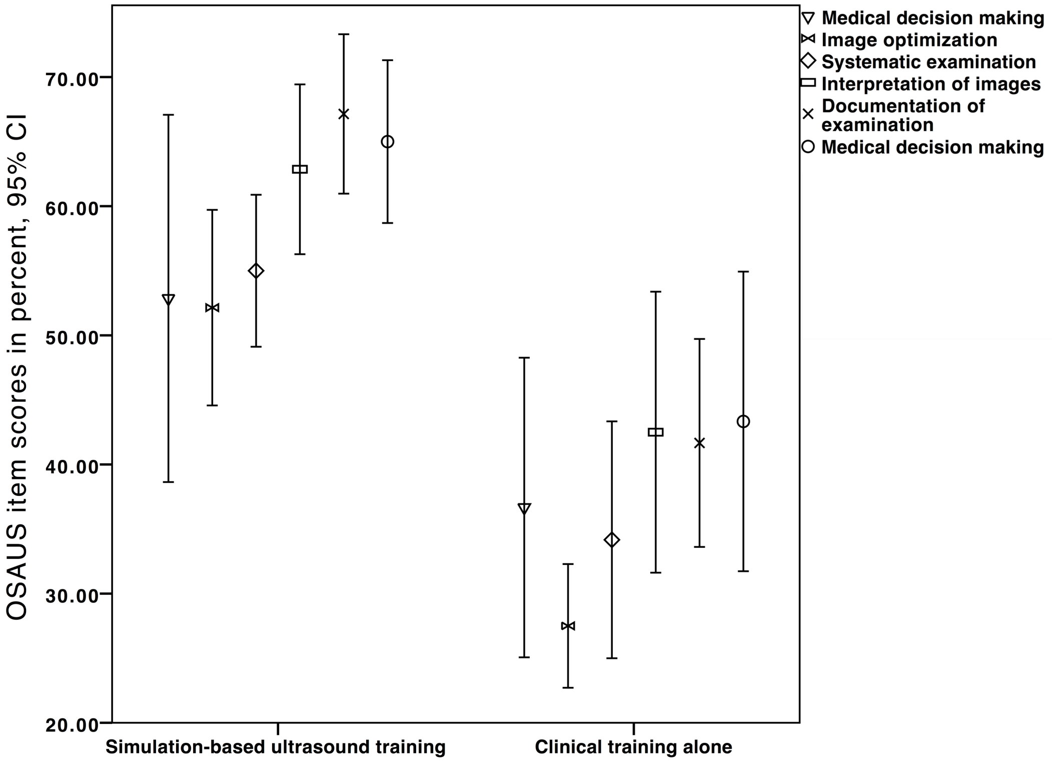
Study 5 (Tolsgaard et al. 2015 A) demonstrated how to improve efficiency of simulation-based ultrasound training with respect to skills transfer immediately after completing training. However, the degree to which these effects are sustained beyond initial clinical training is not known. For training of technical skills such as surgery, there are considerable risks associated with insufficiently trained operators, and it may not be defensible to use patients during the basic training phase (Ziv et al. 2003). For ultrasound training, on the other hand, there is no known patient risk associated with supervised practice. The considerable monetary and time costs associated with SBME may therefore not be justified if the effects of simulation-based ultrasound training only extends to the initial few supervised ultrasound scans during clinical training. However, as we hypothesized in studies 1 and 3, there is some evidence to suggest that clinical training alone is insufficient to ensure adequate ultrasound skills. One explanation may be that trainees are never introduced to basic concepts from which they can build more sophisticated schemas during clinical training. In undergraduate medical education, teaching basic science concepts rather than clinically focused presentations have been shown to improve skills retention as well as to act as PFL (Woods et al. 2006; Mylopoulos & Woods 2014). We may therefore hypothesize that providing systematic basic ultrasound training using SBME could act as preparation for future clinical learning, and consequently that training effects are sustained after several months of clinical training. On the other hand, large initial training effects may decline with time and the effects of short interventions may quickly become engulfed by the vast amounts of time spent on learning in the clinical setting. To examine these hypotheses further, we assessed the impact of simulation-based ultrasound training on trainees' clinical performances after completing the first two months of clinical training.

5.1 Sustained effects of simulation-based ultrasound training on clinical performances

The objective of Study 6 (Tolsgaard et al. 2015 B) was to examine the effects of initial simulation-based transvaginal ultrasound training and clinical training compared with clinical training alone on clinical performances after two months of training.

In a multi-center, randomized design, 33 new residents in obstetrics and gynecology were randomized to initial simulation-based ultrasound training and clinical training (intervention) or only clinical training (control) groups. The intervention group practiced on a transvaginal VR simulator (Medaphor, Cardiff, UK) until they attained the mastery learning level described in Study 4 (Madsen et al. 2014). Subsequently, they practiced equipment handling ("knobology") on a physical mannequin (BluePhantom, CAE Healthcare, Redmond, WA, USA) until they demonstrated the pass/fail level on the OSAUS scale as described in Study 3 (Tolsgaard et al. 2014 B). After two months of clinical training, the participants were assessed on transvaginal ultrasound scans performed on emergency gynecological patients. The scans were recorded and assessed by two blinded expert raters using the OSAUS scale.

Figure 4. OSAUS-scores of participants that received simulation-based ultrasound training (intervention) or clinical training alone (control).



Of the 33 randomized, 26 participants completed the clinical performance test. The intervention and control group participants were assessed after they had completed an average of 57.6 and 62.5 scans, of which means of 43.9 and 45.0 scans had been supervised, respectively. The intervention group participants had significantly higher OSAUS scores compared to the control group (mean, 59.1 ± 9.3% vs. 37.6 ± 11.8%; P < 0.001). A significantly higher number of intervention group participants passed the pass/fail performance level established in Study 3 (Tolsgaard et al. 2014 B) compared to control group participants (85.7% vs. 8.3%, respectively; P < 0.001). There was no main effect of hospital allocation or interaction effect between hospital allocation and the intervention. Finally, there were no significant correlations between performance measures in terms of simulator scores and time used to attain the mastery learning level and the clinical performance scores.

Study 6 demonstrated that initial simulation-based ultrasound training led to performance improvements in the clinical setting that were sustained after more than two months of clinical training and more than 40 supervised scans. Interestingly, participants in both groups reported that they had completed several unsupervised scans, but only 8.3% of control group participants were able to pass a pre-defined pass/fail level that defined the minimally acceptable level of performance. Again, these findings support the notion from Study 1 (Tolsgaard et al. 2014 A) that apprenticeship teaching during clinical training failed to ensure acceptable clinical performances across multiple institutions. However, this does not imply that SBME as a teaching method is superior to clinical training. On the contrary, Moak et al. found that students who practiced ultrasound skills on a pelvic mannequin had lower performances scores than students who practiced on live models, when they were assessed on standardized patients (Moak et al. 2014). These findings relate well to transfer theory, which highlights the importance of contextual similarity in facilitating near-transfer (Gentner et al. 1993) and may at first seem contradictory to our findings in Study 6. However, the use of live models or standardized patients is not equivalent to clinical training using real patients, who may be bleeding, in pain, or under severe psychological stress. As opposed to SBME, clinical training rarely allows trainees to commit errors deliberately or continue practicing under the supervision of expert instructors. According to situated learning theory, complete novice learners may therefore participate very peripherally in patient care and in the "community of practice" (Tolsgaard et al. 2013 A; Lave & Wenger 1991). Providing trainees with some basic skills may enable them to participate more actively in patient management and care through "legitimate peripheral participation" (Wenger 1998). The extremely low number of control group participants who passed the pass/fail level despite being supervised multiple times suggests that supervision in itself was ineffective if not preceded by some systematic form of basic training.

5.2 Quality and efficiency of care

Diagnostic accuracy has been correlated to the amount of operator experience in retrospective studies (Tegnander & Eik-Nes 2006; NHS 2012), and a recent study has also linked diagnostic accuracy to OSAUS scores for abdominal ultrasound scans (Todsen et al. 2014). The finding that simulation-based ultrasound training leads to sustained improvements in ultrasound skills in the clinical setting is therefore promising in terms of diagnostic accuracy and thereby patient safety. However, there is little evidence regarding the implications such performance improvements may have for patient-perceived quality of care as well as for efficiency of care. In terms of public accountability, patients' experiences of care quality are important outcomes that nonetheless have received very limited attention in the health professions education literature. The few studies conducted in this area have shown mixed results (Sedlack et al. 2004; Ahlberg et al. 2005; Curtis et al. 2013; Zendejas et al. 2013 A). In colonoscopy training, for example, SBME has been associated with decreased patient discomfort (Sedlack et al. 2004 A&B; Ahlberg et al. 2005). In other areas, such as communication training, the use of SBME led to higher performance scores of intervention group participants, but no differences in patients' and stakeholders' ratings of residents' performances. In fact, patients being cared for by participants who completed simulation-based communication training had higher depression scores than those cared for by control group participants (Curtis et al. 2013). Hence, correlations between clinical skills and patient-reported outcomes may in some instances be absent or even inverse. Transvaginal ultrasound is generally well-tolerated by patients but may cause some discomfort, and for patients with early pregnancy complications, considerable psychological distress can be expected (Dutta & Economides 2003). Therefore, we hypothesized that simulation-based ultrasound training would decrease patients' discomfort during transvaginal ultrasound examinations as well as patient-reported safety. With respect to factors that are not directly procedure-related, such as general satisfaction with the care provided, we expected little to no effect of simulation-based ultrasound training.

Efficiency of care has also received limited attention in the SBME literature, although factors such as need for supervised practice or repeated patient examinations are of paramount importance to the costs of training and medical care. According to the apprenticeship model of clinical training, trainees gradually become more and more independent with increasing expertise. Study 1 (Tolsgaard et al. 2014 A) confirmed this model, which enables us to hypothesize that improvement in ultrasound skills following simulation-based ultrasound training leads to decreased need for supervision and repeated patient examinations. Hence, the relationship between simulation-based ultrasound training and quality and efficiency of care was the focus of Study 7.

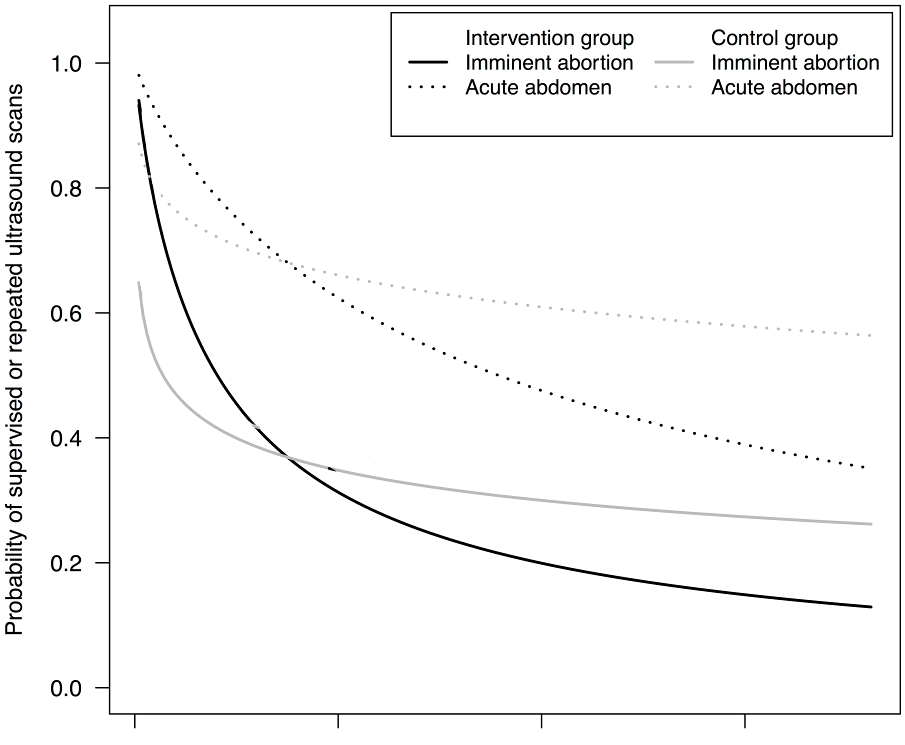
5.3 The effects of simulation-based transvaginal ultrasound training on quality and efficiency of care

The research question of Study 7 (Tolsgaard et al. 2016 B) was as follows: "What is the effect of adding initial simulation-based transvaginal ultrasound training to new trainees' clinical training on quality and efficiency of care measured during the first six months of clinical training, as compared to clinical training only?"

In a multi-center randomized study, 54 new OB/GYN residents were included and randomized to initial simulation-based ultrasound training and clinical training (intervention) or clinical training only (control) groups. The simulation-based ultrasound training followed the mastery learning model described above, and included a VR simulator and a physical mannequin. All emergency gynecological patients, for whom a transvaginal ultrasound examination was performed by study participants, were invited to fill out a standardized scoring form, in which they were asked to rate discomfort, perceived safety, confidence in their ultrasound provider, and satisfaction. The assisting nurse recorded the time spent per ultrasound scan and the need for supervision from a senior colleague or the need for repeated patient examinations.

In total, 1,150 patient ratings were completed for 52 participants from four different departments. Intervention group participants had 18.5% (95% CI, 10.7 to 25.5; $P < 0.001$) lower patient discomfort scores compared with control group participants. Patients rated intervention group participants 7.9% (95% CI, 0.5 to 14.7; $P = 0.04$) higher on safety compared to the control group. Patients scored intervention group participants 11.1% (95% CI, 2.5 to 18.9; $P=0.01$) higher on confidence compared to control group participants. However, there were no differences with regard to overall patient satisfaction (P = 0.61). There were minimal effects of clinical training length on patients' confidence in their ultrasound providers (P = 0.001), and no effects on discomfort, perceived safety, or overall satisfaction. Intervention group participants used 1 minute 32 seconds (95% CI, 7 seconds to 3 minutes 6 seconds; P = 0.03) less per ultrasound examination compared to control group participants. Finally, there was a significant interaction effect between clinical training time and the intervention on the need for supervision or repeated patient examination (P = 0.005). The odds for supervision or repeated patient examination were reduced by 45.3% (95% CI, 33.5 to 55.1) in the intervention group and by 19.8% (95% CI, 4.1 to 32.9) in the control group, when clinical training time was doubled. There were no interaction effects between the intervention and hospital allocation on any of the outcomes.

Figure 5. Need for supervision and repeated patient examination as a function of clinical training time.



Study 7 demonstrated that simulation-based ultrasound training led to improvements in some – but not all – patient-relevant outcomes. These findings align well with previous meta-analyses demonstrating that SBME is associated with large effects on knowledge, skills, and behavior, but only moderate effects on patient outcomes (Cook et al. 2011). As the outcome of interest moves closer to patient care and further away from the controlled simulated setting, it becomes increasingly difficult to prove causality. Some studies have completely failed to demonstrate any effect of SBME on patient outcomes, which may be explained by dilution of training effects, inadequate sample sizes, and failure to establish causal links between intervention and outcome (Cook & West 2013). Although Study 7 was sufficiently powered and among the largest studies conducted on the role of SBME for patient care, we had no a priori indications that improvements in skills actually led to improvements in patients' perceptions of care quality. Reduced examination time combined with improved operator self-efficacy as a result of systematic initial training may have reflected upon patients' ratings. However, we did not attempt to establish which components of simulation-based ultrasound training were responsible for the observed effects. In terms of dilution effects, we found no negative interaction effects between the intervention and length of clinical training for any of the outcomes examined. In fact, we found the opposite for efficiency of care. The large interaction effects between simulation-based ultrasound training and length of clinical training on the need for supervision or repeated patient examinations support the use of SBME as PFL. These results stress the importance of the use of long-term follow-up and large patient populations to determine the link between training interventions and quality of care.

Most previous studies on patient outcomes have focused on immediate main effects of training. Study 7 demonstrates that SBME is not only a method for improving immediate outcomes, but that it also enables trainees to benefit more from their subsequent clinical training. Given the considerable time and monetary costs associated with SBME, improvements in efficiency of care have profound implications for the justification of simulation-based ultrasound training. Nonetheless, there is always a cost of training, regardless of the positive effects on quality and efficiency of care associated with simulation-based ultrasound training. Whether to adopt a new method of training or not is therefore a choice that depends on how much stakeholders are willing to pay for a given change in the outcome of interest. However, there is little research in the costs of SBME (Zendejas et al. 2013 B), and no consensus exists on how to determine cost-effectiveness of training interventions in health professions education. In the final study, we examined how to provide defensible and evidence-based recommendations to decision-makers regarding the adoption of new training methods such as simulation-based ultrasound training.

6. Cost-effectiveness of simulation-based ultrasound training

Medical education is estimated to cost around €80 billion per year worldwide (Frenk et al. 2010). Despite this enormous amount, the cost of medical education is generally underreported in the health professions education literature (Zendejas et al. 2013 B). Decision-makers and leaders in medical education need to prioritize between different educational interventions. However, most educational interventions result in some kind of learning (Cook 2012), although at very different costs (Walsh et al. 2013). Given that decision-makers often have to make decisions based on *costs* and medical education researchers only provide evidence on *effectiveness*, there is a risk of the development of gaps between *actual* practices and *best* practices in health professions education (van der Vleuten & Driessen 2014). A more informative study focus may therefore be the cost-effectiveness of educational interventions.

However, estimation of the cost and effect of educational interventions is not straightforward. Costs of educational interventions may vary between institutions and countries, and there is no general consensus on what should be included in cost estimates. As demonstrated in studies 4–7, effects of an educational intervention may be estimated very differently. Only a small number of experimental trials in medical education have attempted to estimate long-term effects of educational interventions, and most often only the immediate effects on knowledge or skills are examined (Cook et al. 2011). However, in some cases, training effects may be sustained throughout extended periods of clinical practice and training (studies 6 and 7), whereas skills decay and in other cases participant attrition may complicate a meaningful estimation of the effects of an intervention.

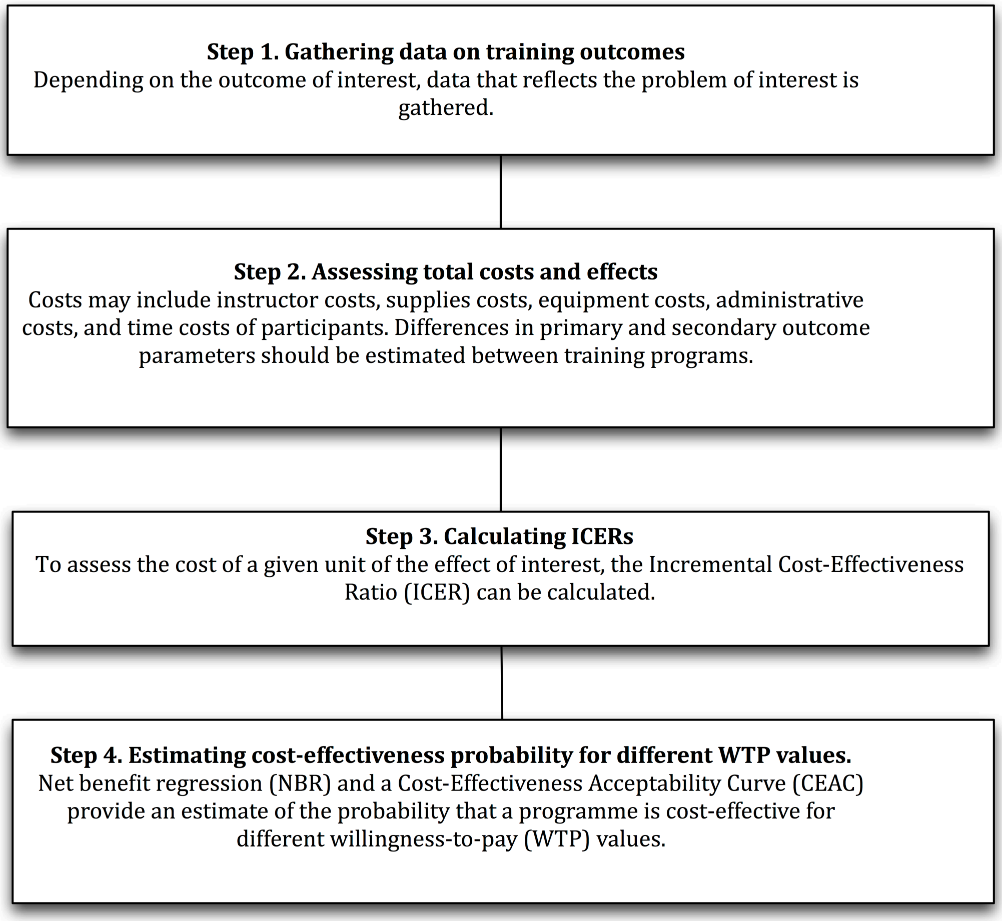
In ultrasound education, the question of cost-effectiveness is highly relevant but also difficult to answer. Although the purpose of ultrasound education is ultimately to improve patient care and safety, other outcomes may also be of interest, including the effects of training on operator skills or efficiency of care. Consequently, the costs that decision-makers are willing to pay for basic ultrasound training may also vary depending on the outcome of interest. The question is therefore not *whether* simulation-based or clinical ultrasound training is cost-effective, but rather *how much* the outcome of interest is changed *relative* to its costs. To provide such estimates, there is a need for models that take decision-makers' willingness to pay into consideration, as well as the uncertainty associated with cost and effect estimates. In Study 8, we attempted to develop a model for cost-effectiveness studies in health professions education using an example from a randomized trial involving simulation-based ultrasound training.

6.1 Linking quality of care and ultrasound training costs

The aim of Study 8 (Tolsgaard et al. 2015 C) was to develop a model for conducting cost-effectiveness studies in health professions education. The research question for the example study that the model was based on was: "What is the cost-effectiveness of training midwives in performing cervical length scans compared with obstetrician-performed cervical scans with respect to patient waiting time?"

A literature review of health economics theory (Drummond et al. 2005; Gold et al. 1996; Hoch et al. 2006; O'Brien et al. 1994; Van Hout et al. 1994) and cost-effectiveness studies in health professions education (Isaranuwatchai et al. 2013; Magee et al. 2013; Zendejas et al. 2013 B; Fletcher & Wind 2013; Wynn et al. 2013; Cohen et al. 2010; Stefanidis et al. 2010; Iribarne et al. 2011) was conducted. Based on this review, we proposed a model that included four steps: 1) gathering data on training outcomes; 2) assessing total costs; 3) calculating incremental cost-effectiveness ratios; and 4) estimating cost-effectiveness probability.

Figure 6. The four steps of the Programme Effectiveness and Cost Generalization (PRECOG) model.



Step 1. Gathering data on training outcomes
Depending on the outcome of interest, data that reflects the problem of interest is gathered.

Step 2. Assessing total costs and effects
Costs may include instructor costs, supplies costs, equipment costs, administrative costs, and time costs of participants. Differences in primary and secondary outcome parameters should be estimated between training programs.

Step 3. Calculating ICERs
To assess the cost of a given unit of the effect of interest, the Incremental Cost-Effectiveness Ratio (ICER) can be calculated.

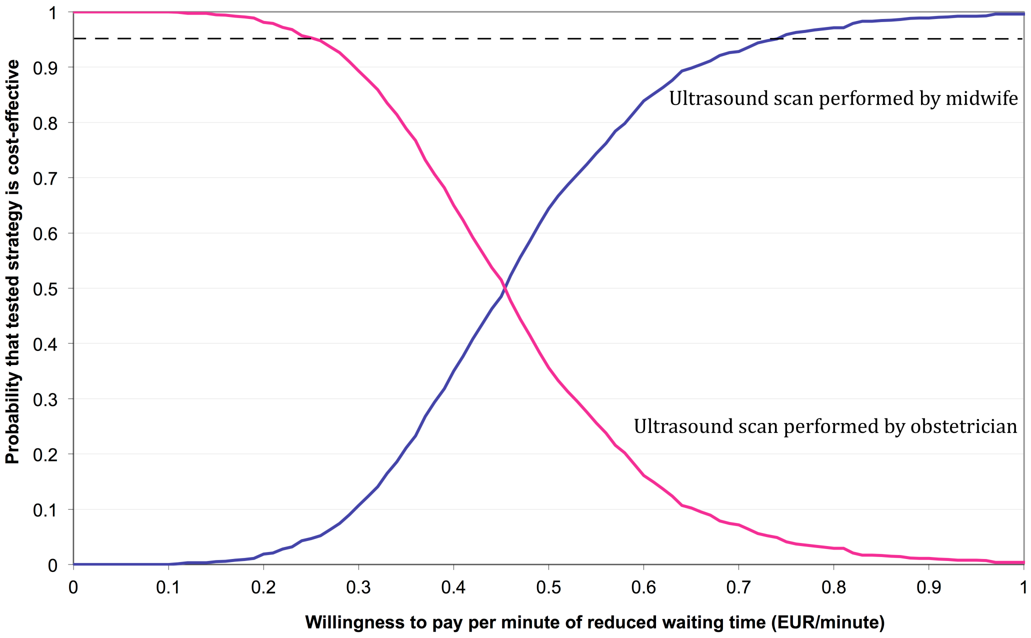
Step 4. Estimating cost-effectiveness probability for different WTP values.
Net benefit regression (NBR) and a Cost-Effectiveness Acceptability Curve (CEAC) provide an estimate of the probability that a programme is cost-effective for different willingness-to-pay (WTP) values.

In the first step, we conducted a randomized trial to examine the effects of training a group of midwives to perform cervical length scans compared to obstetrician-performed cervical length scans. The rationale for conducting the study was that we observed long waiting time for women who presented with signs of preterm onset of labor, which represented a potential threat to their safety. In total, 12 midwives were randomized to simulation-based and clinical training in cervical length measurement (intervention group) or no training (control group). The simulation-based ultrasound training included mastery learning using first a VR simulator and then a physical mannequin as described in studies 6 and 7. The simulation-based ultrasound training was followed by clinical training, in which the participants were required to pass the pass/fail level on the OSAUS scale as described in Study 3 (Tolsgaard et al. 2014 B). Over the next six months, waiting time (primary outcome) and number of shifts for the responsible health care provider were recorded, for women who presented with signs of preterm labor and were cared for by intervention or control group participants. The effects were extrapolated to the first 60 months after completion of training to account for residual training effects and participant attrition.

In step two, training costs were estimated, including implementation costs and equipment costs. Implementation costs were calculated by measuring the amount of time used for training by study participants, the simulator instructor, and the clinician teacher. Step three involved determining the incremental cost-effectiveness ratio (ICER), which is defined as the ratio between differences in cost and effects between the two groups. In step four, the uncertainty represented by each of the preceding steps was taken into account when calculating the cost-effectiveness probability for different willingness-to-pay values.

There was a significant reduction in patient waiting time for patients being cared for by the intervention group (n = 50), compared with patients cared for by the control group (n=65); the mean difference between groups was 36.6 minutes (95% CI, 7.3–65.8; P = 0.008). Intervention group participants were able to discharge the majority of patients (86%) and needed second opinions by an obstetrician in 16% of the cases, compared to 100% in the control group (P < 0.001). The total cost for all participants was €2,688.3 over the study period. The total reduction in patient waiting time over a 60-month period was estimated at 99 hours 50 minutes for 164 patients. This corresponded to an ICER for time saved of €0.45 per minute and an ICER for shifts in responsible health care provider of €19.51/shift. A graphical presentation – a cost-effectiveness acceptability curve (CEAC) – was created to illustrate the probability that the intervention was cost-effective for different willingness-to-pay values (Figure 6). For willingness-to-pay values below €0.26 per minute saved of waiting time, there was a 95% probability that obstetrician-performed cervical scans was the most cost-effective strategy. On the other hand, if decision-makers were willing to pay €0.73 or more per saved minute of waiting time, there was a 95% chance that training midwives in performing cervical scans was the most cost-effective approach.

Figure 7. The Cost-Effectiveness Acceptability Curve (CEAC) demonstrates the probability than the two interventions were cost-effective depending on stakeholders' willingness-to-pay.



Study 8 demonstrated that although educational interventions may result in large effects on the outcome of interest, it does not imply that they are cost-effective. Depending on how much decision-makers were willing to pay, the intervention in Study 8 could be regarded as both cost-effective and cost-ineffective. Hence, the choice of whether or not to adopt a new training intervention rests equally on the effectiveness of the intervention and its costs as well as willingness-to-pay. Although this may seem intuitive, the evaluation of cost-effectiveness and its interpretations are not. The results of previous cost-effectiveness studies that did not make use of probabilistic cost-effectiveness estimates are difficult to interpret if only cost savings or raw cost-effectiveness ratios are presented, which may limit the generalizability of the results (Stefainidis et al. 2010; Cohen et al. 2010). Moreover, the use of immediate outcomes such as skills improvements following SBME may fail to inform educators as to the real outcomes of interest, such as skills retention, transfer, or improvements in PFL.

Study 8 demonstrated that long-term follow-up, the use of patient-related outcomes, and probabilistic models are all needed to provide meaningful cost-effectiveness studies in medical education. Hence, the direct comparison of immediate training effects and the use of outcomes that are not directly related to quality of care may over- or underestimate the true educational and clinical impact of the intervention being studied. Finally, even when all factors are taken into account, the use of cost-effectiveness studies in health professions education is largely limited by the low generalizability of cost estimates across institutions. One way forward could be a more detailed account of the use of participant and training time during future experimental trials, which would allow meta-analysis to generalize cost-effectiveness estimates across countries and institutions.

7. General discussion

This thesis examined the evidence supporting the use of various methods for assessment and training ultrasound skills from multiple perspectives. From the theories on learning and from the perspectives of trainees, ultrasound performance depended on a combination of motor skills, visual skills, and cognitive skills. These factors were also reflected in experts' opinions on what should be evaluated when assessing ultrasound skills and is supported by empirical data on the performances of obstetrician-gynecologists with different levels of clinical experience. We found some evidence to suggest that the technical aspects of trainees' performance need improvement and that simulation-based ultrasound training may be used to achieve mastery levels prior to clinical training. Ultrasound skills were transferred from the simulated to the clinical setting immediately after training as well as after completion of two months of clinical training. From the patients' perspective, the observed improvements in skills were accompanied by reduced patient-reported discomfort, improved safety, and increased confidence in the ultrasound operator. From an organizational perspective, we found that the provision of initial systematic simulation-based ultrasound training interacted with clinical training by reducing the need for supervised practice and repeated patient examinations. From a health economics point of view, training efficiency and costs were evaluated. We demonstrated that training efficiency could be doubled using collaborative learning without any negative consequences for transfer. Finally, a model for linking quality of care to training costs was developed to assess the cost-effectiveness of educational interventions such as ultrasound training.

7.1 Generalization of findings and critiques

The studies included in this thesis make use of a variety of quantitative methods such as cross-sectional studies, validation studies, and randomized controlled trials. The choice of study designs was guided by the research questions, as some studies were explorative (such as studies 1–4) and others explanatory (such as studies 5–8). Weaknesses of the cross-sectional and validation studies include the notion that association does not imply causation. This is particularly problematic in the studies that attempted to establish validity evidence to support the use of assessment scores in the simulated and clinical settings. We chose to use different levels of clinical experience as indicators of competence in studies 3 and 4, which may be confounded by other factors than those relating to competence. As argued by Cook (Cook 2015), the finding that groups with very different experience levels achieve different assessment scores is in itself not particularly informative. Most validation studies – including ours – make use of groups of learners that do not sufficiently represent the target population but rather the extreme ends of the performance spectrum. One challenge in this regard is that large sample sizes are often needed to detect small but relevant performance differences in the target population (Norcini et al. 2003). Another problem with using experience as marker of competence is that experience does not always correlate to the development of expertise (Ericsson et al. 1993, 2006), and even when experience is associated with skills development, it does not imply causation (Cook 2015). If competence is defined according to how experienced clinicians perform, educators may be teaching to the test without any evidence that assessment scores in fact relate to improvements in quality of care or patient outcomes.

In the worst case, the use of expert-novice differences to support the use of assessment scores as relevant outcomes in health professions education may be an exercise of chasing one's own tail. The experience-expertise inconsistency is in part reflected in the results from studies 3 and 4, which demonstrated that even experienced gynecologists failed to display expert behavior when compared to the performances of fetal medicine consultants. However, the assumption that assessment scores reflect skills and that skill levels affect quality of care is to some extent supported by the findings relating to effects of simulation-based ultrasound training in studies 4–7. In these studies, we demonstrated how the use of SBME led to improvements in skills as well as quality of care, which may be considered both multiple independent outcomes and also an interrelated chain of outcomes (Cook & West 2013). Accordingly, similar associations between skills and patient outcomes have been demonstrated in other areas of medical education, such as in surgical education (Zendejas et al. 2011; Birkmeyer et al. 2013). Nonetheless, overreliance on assessment scores that are not supported by correlations with clinically meaningful outcomes remains problematic, but is still widely used in medical education research. The central role of patient outcomes in clinical medicine may be self-evident from a clinician's point of view, but in medical education, the role of patient outcomes has been debated (Cook & West 2013). Among the concerns regarding the use of patient outcomes are that relatively large sample sizes are needed to demonstrate differences in effectiveness between interventions because of dilution of training effects. Non-clinical outcomes such as behavior or trainee reactions are important to the development of education theory and practice, but their implications for and relationship to health outcomes should be evaluated critically. Otherwise, there is a risk that research in medical education will not benefit the stakeholders, which include trainees, supervisors, patients, and policy-makers.

The use of randomized designs in studies 5–8 provided some strength in terms of ability to control for systematic bias but the value of randomized trials in medical education has been debated (Eva 2009). One argument is that although these designs provide unbiased estimates of training effects, their practical implications are limited, as they only explain a fraction of the total variance that derives from competing educational activities (Norman 2003). These concerns seem particularly relevant for the majority of experimental studies in medical education that use short follow-up and narrow focus on changes in behavior under controlled and ideal circumstances. Hence, the use of highly standardized and controlled designs may improve the internal validity of results but at the cost of their external validity. In clinical medicine, this dilemma has resulted in the call for practical clinical trials (PCT) that enable decision-makers to make informed choices regarding clinically relevant alternatives under real-life conditions (Tunis et al. 2003). Traditional explanatory trials may provide evidence on the efficiency of an intervention and answer the question "*can* it work?" under ideal circumstances. Practical clinical trials, on the other hand, are designed to demonstrate the effectiveness of an intervention under real-life conditions and answer the question "*does* it work?" These trials are characterized by the following: 1) having broader inclusion criteria and thereby a more heterogeneous study population; 2) comparing clinically relevant alternatives; 3) recruiting from a variety of contexts; and 4) employing longer follow-up and multiple outcomes including evidence of cost-effectiveness (Tunis et al. 2003; Glasgow et al. 2005). We used a mix of explanatory designs that aimed at determining intervention efficiency (studies 5 and 6) and practical designs that aimed at exploring intervention effectiveness in multiple sites using clinically and educationally relevant outcomes and long-term follow-up (studies 7 and 8). Both types of educational trials are needed in accord with the systematic assessment of evidence in other areas of health care, such as the clinical trial phases in pharmaceutical research (Pocock 1983).

In our studies, the concept of mastery learning during simulation-based ultrasound training demonstrated beneficial effects on near transfer, but it also acted as preparation for future learning. However, as noted previously, there is not always a causal relationship between skills and patient-reported or clinical outcomes (Cook & West 2013), which supports the need for determining whether and how improvements in the first may affect the latter. For example, SBME has shown to benefit trainees' communication skills when assessed according to educational goals by trained raters, but failed to demonstrate similar improvements in patient-reported quality of care and communication

(Fallowfield et al. 2002; Shilling et al. 2003; Curtis et al. 2013). Such differences may reflect the notion that trainees are assessed based on ideas of competence that do not translate into improved performances and a dissonance between different stakeholders' perceptions of competence, which again poses a validity problem for the assessments used. Another explanation is that behavioristic models of skills training fail to acknowledge the importance of preparation for future learning, which according to our findings may include an interaction effect with clinical training, as opposed to a fixed main effect that can be observed immediately after training.

According to recent transfer theory, *adaptive expertise* that relies on an innovative dimension of performance may help explain the role of SBME as preparation for future learning (Schwartz et al. 2005; Bransford & Schwartz 2001). In contrast, traditional apprenticeship training may depend on an effectiveness perspective on performance because of clinical workload, the resulting time pressure, and consideration for the patients being examined. SBME, on the other hand, offers the opportunity for trainees to commit errors and to handle errors during performance. According to literature outside the medical domain, error-management training has been shown to improve transfer of learning by enabling trainees to handle the unexpected when presented with novel situations and cases during subsequent performances (Keith & Frese 2008). However, these hypotheses remain subjects for future research.

7.2 Educational implications

The results of the studies included in this thesis may have implications for the use of SBME in ultrasound training of obstetrician-gynecologists in Denmark and abroad. There is little doubt that simulation-based ultrasound training should be considered early in residency training for the sake of trainees, their clinical supervisors, and their patients. In eastern Denmark, simulation-based training is now mandatory for junior-level trainees in obstetrics-gynecology for a large majority of teaching hospitals (Konge et al. 2015). On an international level, SBME is now increasingly incorporated into basic training courses, and efforts are being made to ensure equal training and assessment standards across different countries and institutions (ISUOG 2014). Emerging new technologies such as online learning platforms (including Massive Open Online Courses and cloud-based simulations) provide new opportunities for ultrasound education by allowing trainees to access large image banks with various types of pathology and anomalies that can be shared through international collaborations. According to our findings, ultrasound competence relies on technical aspects of performance, image interpretation skills, and the ability to integrate scan results into patient care. For now, the use of SBME primarily relates to technical skills training. However, the new advances in technology-enhanced learning may provide the next step in ultrasound education by exposing trainees to large case volumes and thereby stimulate cognitive aspects of performance such as image interpretation skills.

During recent decades, postgraduate medical education has experienced an explosion in the focus on assessment, competency-based education, and the use of SBME. However, postgraduate clinical training is still largely opportunistic, without standardization and systematic use of in-training assessment (RCOG 2012; Ringsted et al. 2004). As a consequence, we found evidence that current postgraduate medical education often fails at ensuring basic skills for trainees entering clinical training. The finding in our studies that even experienced clinicians did not display expert behavior in a core clinical skill that they practice on a daily basis suggests that inadequate basic training has long-term consequences for clinical performances. This notion receives some support from the findings that SBME not only improved trainees' skills following training but also enabled them to benefit more from subsequent clinical training.

We could conclude that SBME should be considered whenever there is sufficient effectiveness evidence to support its use (McGaghie et al. 2014). However, as we demonstrated in our final study, effectiveness of an intervention should be balanced against its costs. We used SBME as a tool to support learning, but it may just as well be replaced by structured clinical training had we applied the same principles and resources for training in the clinical setting that we did in the simulated setting (Moak et al. 2014; Cook et al. 2011). If investments and costs are ignored, there is a risk of being blinded by new technologies that may seem more effective than existing methods for training due to allocation of large time and monetary resources. In other words, clinical training could "look bad" and SBME "look good" merely due to the amount of resources invested in each.

In some respects, this relates to the differences between how interventions are handled in efficiency trials and in practical trials. When evaluating new interventions, researchers often use efficiency designs, in which the intervention is examined under ideal and highly controlled conditions. However, existing methods for training are often evaluated using real-life and less controlled conditions, which usually results in lower effect estimates (Tobler et al. 2000). Accordingly, there is a risk that researchers are overly optimistic toward new interventions when compared against existing practices, which may result in the adoption of new and more costly methods that are not superior to existing educational methods. However, there are remarkably few studies being performed on how to improve clinical training and thereby quality and efficiency of care, compared with the large amounts of studies involving SBME. This is a paradox given the relatively limited time that health professionals spend on SBME compared with the clinical training that spans during the entirety of their careers.

The use of SBME has brought us closer to some level of standardization of postgraduate training; however, SBME needs to be better aligned with subsequent clinical training where serendipitous clinical training remains an accepted practice. Still, we cannot expect that trainees will master clinical skills by random unsupervised clinical practice no matter how much simulation-based training they undergo. Although most clinicians eventually master the skills that are considered essential in their respective specialties, there is sufficient evidence to support the claim that clinicians do not display expert behavior just because they become experienced (Duclos et al. 2012; Hickey et al. 2014; Birkmeyer et al. 2013; Tolsgaard et al. 2014 B).

Striving for clinical independence as the ultimate goal of postgraduate training (Ten Cate et al. 2016) may therefore foster mediocrity more than clinical excellence. Consequently, a change of perspective on the role of education in health care is needed. If our objectives are clinical excellence and high-quality care, we must support their development through systematic allocation of protected training time, supervised practice, and performance assessments that continue throughout clinicians' careers. This inevitably clashes with the workload that may be imposed on trainees and clinicians. Consequently, educational activities such as protected training time or supervised practice may be considered at odds with clinical efficiency and production. However, the time, consequences, and costs of poorly trained clinicians have not been sufficiently investigated. Hence, we need to choose the type of care that we would like to offer patients, and acknowledge that quality of care relies upon educational efforts that may provide long-term rather than short-term returns on investment.

7.3 Future research

Postgraduate medical education in the 21st century still faces some of the same challenges as those observed in undergraduate medical education more than 100 years ago, in terms of lack of standardization and methods for performance assessment (Flexner 1910; Irby et al. 2010). A key question remains regarding how to provide high-quality medical education that results in the production of competent clinicians. Unfortunately, we often tend to substitute difficult questions with questions we can more easily answer or provide a solution for (Kahneman 2011). A challenge for future medical education research is therefore to ask the right questions rather than only to provide tools for their solution (Regehr 2010). Our studies constituted a research program that included needs analysis, development of methods for skills assessment, and the exploration of how to improve learning through SBME and its consequences for learning and performance in the clinical setting. Studies along similar lines are needed in many other areas and disciplines to generate evidence to support the development of clinical excellence.

During our validation studies, we found evidence to support an experience-expertise inconsistency, which may be explored further in future studies. Most studies involving training and assessment of technical and diagnostic skills have focused on novice learners, but other groups of clinicians may also benefit from systematic training. With regard to ultrasound training, this may be particularly interesting given that ultrasound is a relatively new technology and that few senior clinicians have completed systematic competency-based training. However, there is evidence that instructional strategies that are effective for novice learners in terms of facilitating schema formation and automation may not be effective for experienced learners – and in some cases, may even have negative consequences (Sweller et al. 2003). Therefore, future studies are needed to explore how more experienced learners interact with the structured training formats that have been well-described for novice learners, and determine their effectiveness in terms of improvements in quality of care. Experienced practitioners may have acquired undesirable habits that must be unlearned and relearned. How, when, and whether this happens are subjects for future studies to explore.

Our studies suggest that assessment of training effects should be viewed from multiple perspectives, including a focus on the practical implications of training interventions on quality and efficiency of care using long-term follow-up. Although some researchers have argued that patient-relevant outcomes should not be the gold or exclusive standard for the assessment of new innovations in medical education (Cook & West 2013), the time has come to link quality of education with the consequences for quality and efficiency of clinical practice. The less-than-perfect relationship between what *can* work in the controlled setting and what *does* work in clinical practice further supports the notion that evaluation of educational innovations should be assessed based on their consequences for clinical practice. The use of patient-relevant outcomes, however, requires significant funding as well as time investments, which may not always be feasible. Moreover, even when conducting large-scale experimental trials (Curtis et al. 2013; Bilimoria et al. 2016), effects on patient-relevant outcomes are not always present. New methods are therefore needed to bridge the gap between education theory and its relevance to clinicians, patients, and decision-makers. One way of bridging this gap is offered through the use of observational registry-based epidemiological studies that allow researchers to explore the association between educational characteristics relating to the trainee, task, and setting with patient-level data (Norcini et al. 2013, 2014). The use of these methods would allow scholars to advance education research to the next level, where changes in education practice could be linked to small but relevant differences in patient-relevant outcomes (Cook et al. 2010). However, a major challenge remains – although patient registers already exist, there are few countries in which they are linked with care-provider data. Nonetheless, this only suggests that the epidemiological studies are underutilized in medical education, not that they are infeasible.

8. Conclusion

Over the past 50 years of research in ultrasound, multiple clinical applications have been described. With the introduction of new technology to a broader group of clinicians, increasing focus is needed on optimal approaches to ensure its safety through reliable and valid performance assessment and systematic training. We examined the validity evidence supporting the assessment of ultrasound skills in obstetrics-gynecology in the simulated and clinical settings. We then demonstrated how adding initial simulation-based ultrasound training to trainees' clinical training led to large improvements in trainee learning and performances with patients during subsequent clinical training. We demonstrated how the use of simulation-based ultrasound training positively impacted quality and efficiency of care. Finally, we evaluated the economical perspective of ultrasound training by developing a generic model for the assessment of cost-effectiveness of training interventions in health professions education. Our results support the approach of using simulation-based training as preparation for future learning, and stress the importance of applying a multi-level perspective on educational and clinical outcomes over longer periods of time and in multiple institutions.

References

American Educational Research Association: Standards for Educational and Psychological Testing. USA, Washington, DC 2014.

AIUM. American Institute of Ultrasound in Medicine. Training Guidelines for Physicians Who Evaluate and Interpret Diagnostic Ultrasound Examinations of the Female Pelvis 2015: http://www.aium.org/resources/viewStatement.aspx?id=58. Accessed April 24th 2016.

Aggarwal R, Mytton OT, Derbrew M, Hananel D, Heydenburg M, Issenberg B, MacAulay C, Mancini ME, Morimoto T, Soper N, Ziv A, Reznick R. Training and simulation for patient safety. Qual Saf Health Care 2010;19 Suppl 2:i34-43.

Aggarwal R, Undre S, Moorthy K, et al. The simulated operating theatre: comprehensive training for surgical teams. Qual Saf Health Care 2004;13(Suppl 1): i27e32.

Ahlberg G, Hultcrantz R, Jaramillo E, Lindblom A, Arvidsson D. Virtual Reality Colonoscopy Simulation: A Compulsory Practice for the Future Colonoscopist? Endoscopy 2005;37:1198–1204.

Baddeley AD, Hitch GJ. Working memory, in GH. Bower (Ed.): The Psychology of Learning and Motivation: Advances in Research and Theory. New York: Academic Press. 1974;8:47-90

Baddeley AD. The episodic buffer: a new component of working memory? Trends in Cognitive Science 2000;4:417–23.

Barsuk JH, Cohen ER, McGaghie WC, Wayne DB. Long-term retention of central venous catheter insertion skills after simulation-based mastery learning. Acad Med 2010; 85:9 – 12.

Barsuk JH, McGaghie WC, Cohen ER, O'Leary KJ, Wayne DB. Simulation-based mastery learning reduces complications during central venous catheter insertion in a medical intensive care unit. Crit Care Med 2009;37:2697-701.

Bilimoria KY, Chung JW, Hedges LV, Dahlke AR, Love R, Cohen ME, Hoyt DB, Yang AD, Tarpley JL, Mellinger JD, Mahvi DM, Kelz RR, Ko CY, Odell DD, Stulberg JJ, Lewis FR.. National Cluster-Randomized Trial of Duty-Hour Flexibility in Surgical Training. N Engl J Med 2016;25;374:713-27.

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;10;369:1434-42.

Bjerrum AS, Eika B, Charles P, Hilberg O. Dyad practice is efficient practice: a randomised bronchoscopy simulation study. Med Educ 2014;48:705-12.

Blumenfeld YJ, Ness A, Platt LD. Maternal-fetal medicine fellowship obstetrical ultrasound experience: results from a fellowship survey. Prenat Diagn 2013;33:158-61.

Boutis K, Pecaric M, Shiau M, Ridley J, Gladding S, Andrews J, Pusic M. A hinting strategy for online learning of radiograph interpretation by medical students. Med Educ 2013;47:877-87.

Bransford JD, Schwartz DL. Rethinking transfer: a simple proposal with multiple implications. Rev Res Educ 1999;24:61.

Bransford JD, Schwartz DL. Rethinking Transfer: A Simple Proposal With Multiple Implications. Vanderbilt University Review of Research in Education 2001:3:61-100.

Broudy, HS. Types of knowledge and purposes of education. In Anderson RC, Spiro RJ, Montague WE (eds.), Schooling and the acquisition of knowledge. Hillsdale, NJ: Erlbaum. 1977.

Bruning RH, Schraw GJ, Norby MM. Cognitive Psychology and Instruction. Upper Saddle River, NJ. USA. 5th Edition 2010.

Buck GH. Development of simulators in medical education. Gesnerus. 1991;48 Pt 1:7-28.

Burke L, Hutchins HM. Training Transfer: An Integrative Literature Review. Human Resource Development Review 2007;6:263.

Campbell S. A short history of sonography in obstetrics and gynaecology. Facts Views Vis Obgyn 2013;5:213-29.

Chi MT. Active-constructive-interactive: a conceptual framework for differentiating learning activities. Top Cogn Sci 2009;1:73-105.

Cohen ER, Feinglass J, Barsuk JH, Barnard C, O'Donnell A, McGaghie WC et al. Cost savings from reduced catheter-related bloodstream infection after simulation-based education for residents in a medical intensive care unit. Simul Healthc 2010;5:98–102.

Colquitt JA., LePine JA, Noe RA. Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. Journal of Applied Psychology 2000;85:678–707.

Cook DA, Andriole DA, Durning SJ, Roberts NK, Triola MM. Longitudinal research databases in medical education: facilitating the study of educational outcomes over time and across institutions. Acad Med 2010;85:1340-6.

Cook DA, Bordage G, Schmidt HG. Description, justification and clarification: a framework for classifying the purposes of research in medical education. Med Educ 2008;42:128-33.

Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Mastery learning for health professionals using technology-enhanced simulation: a systematic review and meta-analysis. Acad Med.2013;88:1178-86. **(Cook et al. 2013 A)**

Cook DA, Hamstra SJ, Brydges R, Zendejas B, Szostek JH, Wang AT, Erwin PJ, Hatala R. Comparative effectiveness of instructional design features in simulation-based education: systematic review and meta-analysis. Med Teach 2013;35:e867-98. **(Cook et al. 2013 B)**

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;7;306:978-88.

Cook DA, West CP. Perspective: Reconsidering the focus on "outcomes research" in medical education: a cautionary note. Acad Med 2013;88:162-7.

Cook DA. If you teach them, they will learn: why medical education needs comparative effectiveness research. Adv Health Sci Educ Theory Pract 2012;17:305-10.

Cook DA. Much ado about differences: why expert-novice comparisons add little to the validity argument. Adv Health Sci Educ Theory Pract 2015;20:829-34.

Cooper JB, Taqueti VR. A brief history of the development of mannequin simulators for clinical education and training. Qual Saf Health Care 2004;13: i11-i18.

Croskerry P. From mindless to mindful practice - cognitive bias and clinical decision making. N Engl J Med 2013;368:2445-8.

Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: construct alignment improves the performance of workplace-based assessment scales. Med Educ 2011;45:560–9.

Crowley RS, Naus GJ, Stewart J 3rd, Friedman CP. Development of visual diagnostic expertise in pathology - an information-processing study. J Am Med Inform Assoc 2003;10:39-51.

Curtis JR, Back AL, Ford DW, Downey L, Shannon SE, Doorenbos AZ, Kross EK, Reinke LF, Feemster LC, Edlund B, Arnold RW, O'Connor K, Engelberg RA. Effect of communication skills training for residents and nurse practitioners on quality of communication with patients with serious illness: a randomized trial. JAMA 2013;4;310:2271-81.

D'Eon MF, Trinder K. Evidence for the validity of grouped self-assessments in measuring the outcomes of educational programs. Eval Health Prof 2014;37:457-69.

Detterman DL. The case for the prosecution: Transfer as epiphenomenon. In D. K. Detterman & R. J. Sternberg (Eds.), Transfer on trial: Intelligence, cognition, and instruction. Norwood, NJ. 1993.

Donald I, Macvicar J, Brown TG. Investigation of abdominal masses by pulsed ultrasound. Lancet 1958;7;1(7032):1188-95.

Downing SM & Yudkowsky R. Assessment in Health Professions Education. Routledge New York, NY. 2009.

Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ 2003;37:830-7.

Driskell JE, Willis RP, Copper C. Effect of Overlearning on Retention. Journal of Applied Psychology 1992;77:615–22.

Druckman D, Bjork RA (eds). Learning, Remembering, Believing: Enhancing Human Performance; Committee on Techniques for the Enhancement of Human Performance, National Research Council. 1994.

Drummond MF, Sculpher MJ, Torrance GW, O'Brien BJ, Stoddart GL. Methods for the Economic Evaluation of Health Care Programmes. Oxford: Oxford University Press. 3rd edition. 2005.

Duclos A, Peix JL, Colin C, Kraimps JL, Menegaux F, Pattou F, Sebag F, Touzet S, Bourdy S, Voirin N, Lifante JC; CATHY Study Group. Influence of experience on performance of individual surgeons in thyroid surgery: prospective cross sectional multicentre study. BMJ 2012;10;344:d8041.

Dutta RL, Economides DL. Patient acceptance of transvaginal sonography in the early pregnancy unit setting. Ultrasound Obstet Gynecol 2003;22:503-7.

EFSUMB. European Federation of Societies for Ultrasound in Medicine and Biology. Minimum training recommendations for the practice of medical ultrasound. Appendix 3:Gynaecological ultrasound. Education and Practical Standards Committee, Ultraschall Med 2006;27:79-105.

Elstein AS, Shulman LS, Sprafka SA. Medical Problem Solving An Analysis of Clinical Reasoning. Harvard University Press, Cambridge, MA, USA. 1978.

Ericsson KA, Charness N, Feltovich PJ, Hoffman RR (eds). The Cambridge Handbook of Expertise and Expert Performance. Cambridge Handbooks in Psychology. 2006.

Ericsson KA, Krampe R, Tesch-Romer TH. The role of deliberate practice in the acquisition of expert performance. Psychol Rev 1993;100:363–406.

Eva KW. Broadening the debate about quality in medical education research. Med Educ 2009;43:294–6.

Fallowfield L, Jenkins V, Farewell V, Saul J, Duffy A, Eves R. Efficacy of a Cancer Research UK communication skills training model for oncologists: a randomised controlled trial. Lancet 2002;359:650-656.

Fitts PM, Posner MI. Human Performance. Brooks/Cole Pub. Co. Belmont, CA. USA. 1st edition. 1967.

Fletcher JD, Wind AP. Cost considerations in using simulations for medical training. Mil Med 2013;178:37– 46.

Flexner A. Medical Education in the United States and Canada. A Report to the Carnegie Foundation for the Advancement of Teaching. Boston: Updyke, 1910.

Frenk J, Chen L, Bhutta ZA, Cohen J, Crisp N, Evans T, Fineberg H, Garcia P, Ke Y, Kelley P, Kistnasamy B, Meleis A, Naylor D, Pablos-Mendez A, Reddy S, Scrimshaw S, Sepulveda J, Serwadda D, Zurayk H. Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. Lancet 2010;376:1923–58.

Gaba DM. The future vision of simulation in health care. Qual Saf Health Care 2004;13 Suppl 1:i2-10.

Gentner D, Ratterman MJ, Forbus KD. The roles of similarity in transfer: separating retrievability from inferential soundness. Cognitive Psychology 1993;25:524-75.

Ginsburg S. Respecting the expertise of clinician assessors: construct alignment is one good answer. Med Educ 2011;45:546-8.

Glasgow RE, Magid DJ, Beck A, Ritzwoller D, Estabrooks PA. Practical clinical trials for translating research to practice: design and measurement recommendations. Med Care 2005;43:551-7.

Gold MR, Siegel JE, Russell LB, Weinstein MC. Cost- Effectiveness in Health and Medicine: Report of the Panel on Cost-Effectiveness in Health and Medicine. New York: Oxford University Press. 1st edition. 1996.

Govaerts M, Schuwirth L, Van der Vleuten C, Muijtjens A. Workplace-based assessment: effects of rater expertise Adv Health Sci Educ Theory Pract 2011;16: 151–165.

Granados C, Wulf G. Enhancing motor learning through dyad practice: contributions of observation and dialogue. Res Q Exerc Sport 2007;78:197–203.

Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P. Randomized clinical trial of virtual reality simulation for laparoscopic skills training. Br J Surg 2004;91:146-150.

Green J, Kahan M, Wong S. Obstetric and Gynecologic Resident Ultrasound Education Project: Is the Current Level of Gynecologic Ultrasound Training in Canada Meeting the Needs of Residents and Faculty? J Ultrasound Med 2015;34:1583-9.

Greenbaum LD. It is time for the sonoscope. J Ultrasound Med 2003;22:321-2.

Grierson LE. Information processing, specificity of practice, and the transfer of learning: considerations for reconsidering fidelity. Adv Health Sci Educ Theory Pract 2014;19:281-9.

Hatala RM, Brooks LR, Norman GR. Practice makes perfect: the critical role of mixed practice in the acquisition of ECG interpretation skills. Adv Health Sci Educ Theory Pract. 2003;8:17-26.

Hertzberg BS, Kliewer MA, Bowie JD, Carroll BA, DeLong DH, Gray L, Nelson RC. Physician training requirements in sonography: how many cases are needed for competence? AJR Am J Roentgenol 2000;174:1221-7.

Hickey GL, Grant SW, Freemantle N, Cunningham D, Munsch CM, Livesey SA, Roxburgh J, Buchan I, Bridgewater B. Surgeon length of service and risk-adjusted outcomes: linked observational analysis of the UK National Adult Cardiac Surgery Audit Registry and General Medical Council Register. J R Soc Med 2014;107:355-64.

Hoch JS, Rockx MA, Krahn AD. Using the net benefit regression framework to construct cost-effectiveness acceptability curves: an example using data from a trial of external loop recorders versus Holter monitoring for ambulatory monitoring of “community acquired” syncope. BMC Health Serv Res 2006;6:68.

Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med 1999;74:1129-34.

Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. Med Teach 2013;35:564-8.

Irby DM, Cooke M, O'Brien BC. Calls for reform of medical education by the Carnegie Foundation for the Advancement of Teaching: 1910 and 2010. Acad Med 2010;85:220-7.

Iribarne A, Easterwood R, Russo MJ, Wang YC. Integrating economic evaluation methods into clinical and translational science award consortium comparative effectiveness educational goals. Acad Med 2011;86:701–5.

Isaranuwatchai W, Brydges R, Carnahan H, Backstein D, Dubrowski A. Comparing the cost-effectiveness of simulation modalities: a case study of peripheral intravenous catheterization training. Adv Health Sci Educ 2013;19:219–32.

Issenberg SB, McGaghie WC, Petrusa ER, Lee Gordon D, 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.

ISUOG Education Committee recommendations for basic training in obstetric and gynecological ultrasound. Ultrasound Obstet Gynecol 2014;43:113-6.

Jamniczky HA, McLaughlin K, Kaminska ME, Raman M, Somayaji R, Wright B, Ma IW. Cognitive load imposed by knobology may adversely affect learners' perception of utility in using ultrasonography to learn physical examination skills, but not anatomy. Anat Sci Educ 2015;8:197-204.

Jang TB, Ruggeri W, Dyne P, Kaji AH. Learning curve of emergency physicians using emergency bedside sonography for symptomatic first-trimester pregnancy. J Ultrasound Med 2010;29:1423-8.

Johnson DW, Johnson RT. An educational psychology success story: social interdependence theory and cooperative learning. Educ Res 2009;36:365–79.

Kahneman D. Thinking, Fast and Slow. Straus and Giroux, New York, NY, USA. 2011.

Kane MT: Validation; in R.L.Brennan, (ed): Validation. Praeger: Westport, Educational Measurement 2006:17-64.

Keith N, Frese M. Effectiveness of error management training: a meta-analysis. J Appl Psychol 2008;93:59-69.

Kennedy TJ, Regehr G, Baker GR, Lingard L. Preserving professional credibility: grounded theory study of medical trainees' requests for clinical support. BMJ 2009;9;338:b128.

Kirschner F, Paas F, Kirschner PA. A cognitive load approach to collaborative learning: united brains for complex tasks. Educ Psychol Rev 2009;21:31-42.

Kirschner PA, Sweller J, Clark RE. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, exponential and inquiry-based teaching. Educational Psychologist 2006;41:75–86.

Konge L, Ringsted C, Bjerrum F, Tolsgaard MG, Bitsch M, Sørensen JL, Schroeder TV. The Simulation Centre at Rigshospitalet, Copenhagen, Denmark. J Surg Educ 2015;72:362-5.

Krupinski EA. The role of perception in imaging: past and future. Semin Nucl Med 2011;41:392-400.

Kulasegaram KM. PhD thesis: The effect of conceptual and contextual teaching strategies for the transfer of basic science knowledge in medical education. McMaster University. 2013.

Kundel HL, La Follette Jr. PS. Visual search patterns and experience with radiological images. Radiology 1972;103, 523–28.

Kundel HL, Nodine CF, Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. Invest Radiol 1978;13:175-181.

Kundel HL, Nodine CF, Krupinski EA. Searching for lung nodules. Visual dwell indicates locations of false-positive and false-negative decisions. Invest Radiol 1989;24:472-478.

Kundel HL, Nodine CF. A visual concept shapes image perception. Radiology 1983;146: 363-368.

Kundel HL, Nodine CF. Interpreting chest radiographs without visual search. Radiology 1975;116:527–532.

Larsen CR, Soerensen JL, Grantcharov TP, Dalsgaard T, Schouenborg L, Ottosen C, Schroeder TV, Ottesen BS. Effect of virtual reality training on laparoscopic surgery: randomised controlled trial. BMJ 2009; 14; 338: b1802.

Lave J, Wenger E. Situated Learning: Legitimate Peripheral Participation. Cambridge: Cambridge University Press. 1st edition. 1991.

Lee W, Hodges AN, Williams S, Vettraino IM, McNie B. Fetal ultrasound training for obstetrics and gynecology residents. Obstet Gynecol 2004;103:333-8.

Lesgold A, Rubinson H, Feltovitch P, Glasser R, Klopfer D, Wang Y. Expertise in a complex skill: diagnosing x-ray pictures. In: Chi M, Glaser R, Farr M, eds. The nature of expertise. Rillsdale, NJ: Erlbaum, 1988;311- 342.

Madsen ME, Konge L, Nørgaard LN, Tabor A, Ringsted C, Klemmensen AK, Ottesen B, Tolsgaard MG. Assessment of performance measures and learning curves for use of a virtual-reality ultrasound simulator in transvaginal ultrasound examination. Ultrasound Obstet Gynecol 2014;44:693-9.

Magee SR, Shields R, Nothnagle M. Low cost, high yield: simulation of obstetric emergencies for family medicine training. Teach Learn Med 2013;25:207–10.

Magill RA. Motor Learning and Control: Concepts and Applications. McGraw-Hill, New York USA. 9th edition. 2010.

McGaghie WC, Issenberg SB, Barsuk JH, Wayne DB. A critical review of simulation-based mastery learning with translational outcomes. Med Educ 2014;48:375-85.

McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. Acad Med 2011;86:706-11. **(McGaghie et al. 2011 A)**

McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Medical education featuring mastery learning with deliberate practice can lead to better health for individuals and populations. Acad Med 2011;86(11):e8-9. **(McGaghie et al. 2011 B)**

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.

Messick S: Validity; in Linn RL, (ed): Validity. New York: American Council on Education and Macmillan, Educational Measurement. 1989.

Miller GA. The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review 1956;63:81–97.

Moak JH, Larese SR, Riordan JP, Sudhir A, Yan G. Training in transvaginal sonography using pelvic ultrasound simulators versus live models: a randomized controlled trial. Acad Med 2014;89:1063-8.

Monteiro SM, Norman G. Diagnostic reasoning: where we've been, where we're going. Teach Learn Med 2013;25:S26-32.

Moore CL, Copel JA. Point-of-care ultrasonography. N Engl J Med 2011 24;364:749-57.

Mylopoulos M, Woods N. Preparing medical students for future learning using basic science instruction. Med Educ 2014;48:667-73.

NHS: Ten Years of Maternity Claims. 2012: http://www.nhsla.com/safety/Documents/Ten%20Years%20of%20Maternity%20Claims%20-20An%20Analysis%20of%20the%20NHS%20LA%20Data%20-%20October%202012.pdf. Accessed 24 April 2016.

Nodine CF, Kundel HL, Lauver SC, Toto LC. Nature of expertise in searching mammograms for breast masses. Acad Radiol 1996;3:1000-6.

Nodine CF, Kundel HL, Mello-Thoms C, Weinstein SP, Orel SG, Sullivan DC, Conant EF. How experience and training influence mammography expertise. Acad Radiol 1999;6:575-585.

Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. Ann Intern Med 2003;138:476-81.

Norcini JJ, Boulet JR, Opalek A, Dauphinee WD. Outcomes of cardiac surgery: associations with physician characteristics, institutional characteristics, and transfers of care. Med Care 2013;51:1034-9.

Norcini JJ, Boulet JR, Opalek A, Dauphinee WD. The relationship between licensing examination performance and the outcomes of care by international medical school graduates. Acad Med 2014;89:1157-62.

Norman G. Data dredging, salami-slicing, and other successful strategies to ensure rejection: twelve tips on how to not get your paper published. Adv Health Sci Educ Theory Pract 2014;19:1-5.

Norman G. RCT = results confounded and trivial: the perils of grand educational experiments. Med Educ 2003;37:582-4.

Norman GR, Coblentz CL, Brooks LR, Babcook CJ. Expertise in visual diagnosis: a review of the literature. Acad Med 1992;67:S78-83.

Norman GR, Trott A, Brooks L, Kinsey-Smith E. Cognitive differences in clinical reasoning related to postgraduate training. Teach Learn Med 1994;6:114–20.

O'Brien BJ, Drummond MF, Labelle RJ, Willan A. In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. Med Care 1994;32:150–63.

Page RL. Brief History of Flight Simulation. In SimTechT 2000 Proceedings. Sydney: The SimtechT 2000 Organizing and Technical Committee. 2000.

Plato. The Republic. Allen, RE (ed). New Haven: Yale University Press. Book VII. 2006.

Pocock S. Clinical Trials: A practical approach. John Wiley & Sons Ltd. 1st edition. 1983.

Räder SB, Henriksen AH, Butrymovich V, Sander M, Jørgensen E, Lönn L, Ringsted CV. A study of the effect of dyad practice versus that of individual practice on simulation-based complex skills learning and of students' perceptions of how and why dyad practice contributes to learning. Acad Med 2014;89:1287-94.

Regehr G. It's NOT rocket science: rethinking our metaphors for research in health professions education. Med Educ 2010;44:31-9.

Ringsted C, Pallisgaard J, Østergaard D, Scherpbier A. The effect of in-training assessment on clinical confidence in postgraduate education. Med Educ 2004;38:1261-9.

Ringsted C, Skaarup AM, Henriksen AH, Davis D. Person-task-context: a model for designing curriculum and in-training assessment in postgraduate education. Med Teach 2006;28:70-6.

Rizzolatti G, Craighero L. The mirror-neuron system. Annu Rev Neurosci 2004;27:169–92.

Royal College og Obstetricians & Gynaecologists (RCOG): Delivery of ultrasound training: https://www.rcog.org.uk/en/careers-training/resources-and-support-for-trainers/delivering-postgraduate-training-in-og/curriculum-resources-for-trainers/delivery-of-ultrasound-training-information-for-trainers/. Accessed April 25th 2016.

Salvesen KA, 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;36:525-9.

Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. Acad Med 1990;65:611-21.

Schwartz DL, Bransford J, Sears D. Efficiency And Innovation In Transfer. Transfer of Learning from a Modern Multidisciplinary Perspective. Information Age Publishing. 2005.

Sedlack RE, Kolars JC, Alexander JA. Computer simulation training enhances patient comfort during endoscopy. Clinical Gastroenterology and Hepatology 2004;2:348–52. **(Sedlack et al. 2004 A)**

Sedlack RE, Kolars JC. Computer Simulator Training Enhances the Competency of Gastroenterology Fellows at Colonoscopy: Results of a Pilot Study. Am J Gastroenterol 2004;99:33–7. **(Sedlack et al. 2004 B)**

Shanks D, Brydges R, den Brok W, Nair P, Hatala R. Are two heads better than one? Comparing dyad and self-regulated learning in simulation training. Med Educ 2013;47:1215-22.

Shea CH, Wulf G, Whitacre C. Enhancing training efficiency and effectiveness through the use of dyad training. J Motor Behav 1999;31:119–25.

Shilling V, Jenkins V, Fallowfield L. Factors affecting patient and clinician satisfaction with the clinical consultation: can communication skills training for clinicians improve satisfaction? Psychooncology 2003;12:599-611.

Simpson DE, Rich EC, Dalgaard KA, et al. The diagnostic process in primary care: A comparison of general internists and family physicians. Social Science & Medicine 1987;25:861–866.

Stefanidis D, Hope WW, Korndorffer JR, Markley S, Scott DJ. Initial laparoscopic basic skills training shortens the learning curve of laparoscopic suturing and is cost-effective. J Am Coll Surg 2010;210:436–40.

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.

Streiner DL, Norman G. Validity. In Health Measurement Scales: a Practical Guide to their Development and Use. Oxford Medical Publications: Oxford, UK, 2247 – 274. 2008.

Sweller J, Ayres P, Kalyuga S. Cognitive Load Theory. Springer New York Dordrecht Heidelberg London, UK. Volume 1. 2011.

Sweller J, Ayres PL, Kalyuga S, Chandler PA. The expertise reversal effect. Educational Psychologist 2003;38:23-31.

Sweller J. Cognitive Load During Problem Solving. Cognitive Science 1988;12:257–85.

Tegnander E, Eik-Nes SH. The examiner's ultrasound experience has a significant impact on the detection rate of congenital heart defects at the second-trimester fetal examination. Ultrasound Obstet Gynecol 2006;28:8-14.

Ten Cate O, Hart D, Ankel F, Busari J, Englander R, Glasgow N, Holmboe E, Iobst W, Lovell E, Snell LS, Touchie C, Van Melle E, Wycliffe-Jones K. Entrustment Decision Making in Clinical Training; International Competency-Based Medical Education Collaborators. Acad Med 2016;91:191-8.

Ten Cate OJ. Competency-based education, entrustable professional activities, and the power of language. Grad Med Educ 2013;5:6-7.

Teteris E, Fraser K, Wright B, McLaughlin K. Does training learners on simulators benefit real patients? Adv Health Sci Educ Theory Pract 2012;17:137-44.

Thorndike EL, Woodworth RS. The influence of improvement in one mental function upon the efficiency of other functions. Psychological Review 1901;8:247-61.

Tobler NS, Roona MR, Ochshorn P, Marshall DG, Streke AV, Stackpole KM. School-based adolescent drug prevention programs. J Primary Prev 2000;20:275–336.

Todsen T, Jensen ML, Tolsgaard MG, Olsen BH, Henriksen BM, Hillingsø JG, Konge L, Ringsted C. Transfer from point-of-care Ultrasonography training to diagnostic performance on patients-a randomized controlled trial. Am J Surg 2016;211:40-5.

Todsen T, Tolsgaard MG, Olsen BH, Henriksen BM, Hillingsø JG, Konge L, Jensen ML, Ringsted C. Reliable and valid assessment of point-of-care ultrasonography. Ann Surg 2015;261:309-15.

Tolsgaard MG, Arendrup H, Pedersen P, Ringsted C. Feasibility of self-directed learning in clerkships. Med Teach 2013;35:e1409-15. **(Tolsgaard et al. 2013 A)**

Tolsgaard MG, Bjørck S, Rasmussen MB, Gustafsson A, Ringsted C. Improving efficiency of clinical skills training: a randomized trial. J Gen Intern Med 2013;28:1072-7. **(Tolsgaard et al. 2013 B)**

Tolsgaard MG, Kulasegaram KM, Ringsted CV. Collaborative learning of clinical skills in health professions education: the why, how, when and for whom. Med Educ 2016;50:69-78 **(Tolsgaard et al. 2016 A)**

Tolsgaard MG, Madsen ME, Ringsted C, Oxlund BS, Oldenburg A, Sorensen JL, Ottesen B, Tabor A. The effect of dyad versus individual simulation-based ultrasound training on skills transfer. Med Educ 2015 Mar;49(3):286-95. **(Tolsgaard et al. 2015 A)**

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;43:444-51. **(Tolsgaard et al. 2014 A)**

Tolsgaard MG, Ringsted C, Dreisler E, Klemmensen A, Loft A, Sorensen JL, Ottesen B, Tabor A. Reliable and valid assessment of ultrasound operator competence in obstetrics and gynecology. Ultrasound Obstet Gynecol 2014;43:437-43. **(Tolsgaard et al. 2014 B)**

Tolsgaard MG, Ringsted C, Dreisler E, Nørgaard LN, Petersen JH, Madsen ME, Freiesleben NL, Sørensen JL, Tabor A. Sustained effect of simulation-based ultrasound training on clinical performance: a randomized trial. Ultrasound Obstet Gynecol 2015;46:312-8. **(Tolsgaard et al. 2015 B)**

Tolsgaard MG, Ringsted C, Rosthøj S, Nørgaard L, Møller L, Freiesleben NC, Dyre L, Tabor A. The Effects of Simulation-based Transvaginal Ultrasound Training on Quality and Efficiency of Care: A Multicenter Single-blind Randomized Trial. Ann Surg 2016 Jan 25. **(Tolsgaard et al. 2016 B)**

Tolsgaard MG, Tabor A, Madsen ME, Wulff CB, Dyre L, Ringsted C, Nørgaard LN. Linking quality of care and training costs: cost-effectiveness in health professions education. Med Educ 2015;49:1263-71. **(Tolsgaard et al. 2015 C)**

Tolsgaard MG, Todsen T, Sorensen JL, Ringsted C, Lorentzen T, Ottesen B, Tabor A. International multispecialty consensus on how to evaluate ultrasound competence: a Delphi consensus survey. PLoS One. 2013;8(2):e57687. **(Tolsgaard et al. 2013 C)**

Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA 2003;24;290:1624-32.

van der Gijp A, van der Schaaf MF, van der Schaaf IC, Huige JC, Ravesloot CJ, van Schaik JP, Ten Cate TJ. Interpretation of radiological images: towards a framework of knowledge and skills. Adv Health Sci Educ Theory Pract 2014;19:565-80.

van der Vleuten CPM, Driessen EW. What would happen to education if we take education evidence seriously? Perspect Med Educ 2014;3:222–32.

Van Hout BA, Al MJ, Gordon GS. Costs, effects and C/E-ratios alongside a clinical trial. Health Econ 1994; 3:309–19.

Walsh K, Levin H, Jaye P, Gazzard J. Cost analyses approaches in medical education: there are no simple solutions. Med Educ 2013;47:962–8.

Wayne DB, Butter J, Siddall VJ, Fudala MJ, Wade LD, Feinglass J, McGaghie WC.Mastery Learning of Advanced Cardiac Life Support Skills by Internal Medicine Residents Using Simulation Technology and Deliberate Practice. J Gen Intern Med 2006; 21: 251–256.

Wenger E. Communities of practice: learning, meaning, and identity. Cambridge, Cambridge University Press. 1998.

Woods NN, Neville AJ, Levinson AJ, Howey EH, Oczkowski WJ, Norman GR. The value of basic science in clinical diagnosis. Acad Med 2006;81:S124-7.

World Health Organization. Transforming and scaling up health professionals' education and training. Geneva, Switzerland: World Health Organization 2013:1–124. Retrieved from http://www.who.int Accessed April 26th 2016.

Wulf G. Attention and Motor Skill Learning. Human Kinetics. Champaign, IL. USA. 1st edition 2007;chapter 1.

Wynn BO, Smalley R, Cordasco KM. Does it cost more to train residents or to replace them? RAND Corporation 2013. www.rand.org. [Accessed 15 June 2015.]

Zendejas B, Brydges R, Wang AT, Cook DA. Patient Outcomes in Simulation-Based Medical Education: A Systematic Review. J Gen Intern Med 2013;28:1078–89. **(Zendejas et al. 2013 A)**

Zendejas B, Cook DA, Bingener J, Huebner M, Dunn WF, Sarr MG, Farley DR. Simulation-based mastery learning improves patient outcomes in laparoscopic inguinal hernia repair: a randomized controlled trial. Ann Surg 2011;254:502-9.

Zendejas B, Wang AT, Brydges R, Hamstra SJ, Cook DA. Cost: the missing outcome in simulation-based medical education research: a systematic review. Surgery 2013;153:160-76. **(Zendejas et al. 2013 B)**

Ziv A, Wolpe PR, Small SD, Glick S.Simulation-based medical education: an ethical imperative. Acad Med 2003;78:783-8.