Validation of Automated Metrics of the VIMEDIX Ultrasound Simulator

ESTABLISHING VALIDITY EVIDENCE FOR THE USE OF VIMEDIX-AR AUTOMATED METRICS IN ASSESSMENT OF FAST EXAM SKILLS

By MELLISSA A.R. WARD, BSc, MD

Thesis submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University Master of Science (2021) Hamilton, Ontario (Health Sciences Education)

Title: Establishing validity evidence for the use of VIMEDIX-AR automated metrics for assessment of FAST exam skills

Author: Mellissa A.R. Ward, BSc, MD

Supervisor: Dr. Paul Engels

Number of pages: 105

Lay abstract

Simulation has become ubiquitous in medical education, offering a safe environment to learn and practice new skills. With the increasing availability of point of care ultrasound and the need for significant training to generate and interpret images, simulation is becoming ever more important. We sought to assess an expert assessment tool for use with an ultrasound simulator and to validate automated metrics associated with the VIMEDIX-AR simulator. The expert assessment tool could reliably differentiate different expertise levels. Three of our automated metrics could discern different levels of expertise. Further work is needed to assess if a composite score of automated metrics could better differentiate skill.

Abstract

Introduction: Simulation has an increasing role in medical education. It offers the ability to learn and practice in a safe environment. Ultrasound is a key tool for many clinicians; however, it requires significant experience to gain expertise. The most common method to gain experience is by training courses with volunteers, where experts are present for one-on-one teaching. This is time and labour intensive. Commercial ultrasound simulators are increasingly available with software capable of generating automated metrics. We sought validity evidence to support the use of automated metrics as a tool for assessment of learners completing a Focused Assessment with Sonography in Trauma (FAST) exam.

Methods: Three groups with differing expertise were recruited to participate: novices with no ultrasound training, intermediates who had completed a formal course within six months, and experts with at least five years of clinical experience. All participants were recorded while completing a FAST exam. Automated metrics of time, path length, angular movement, and percent area viewed were obtained. This video was then scored using the Quality of Ultrasound Imaging and Competence (QUICk) by two expert assessors. Participants were also asked to complete ten find fluid exercises, where automated metrics were generated. Automated metrics from the recorded FAST and QUICk were compared using Kruskall-Wallis to assess for differences in expertise. Correlations between QUICk score and the automated

metrics were assessed using Pearson's correlation coefficient. Find fluid exercises were also assessed using repeated measures one-way ANOVA models.

Results: Time, angular movement, and percent area viewed left upper quadrant (LUQ) were significantly different with novices requiring more time and angular movement, and higher percent area viewed LUQ than experts. The QUICk scores were significantly higher for the experts and intermediates compared to the novices. The scores from the QUICk overall and checklist did not correlate with any automated metrics. Individual components of positioning and handling, probe handling, and image scrolling were negatively correlated with percent area viewed LUQ. Overall, the QUICk tool could differentiate novices from both intermediates and experts when using the VIMEDIX-AR simulator. Several automated metrics could differentiate expertise. Further work should develop a composite score of automated metrics to assess learners.

Acknowledgements

I would like to extend my sincere thanks to Dr. Paul Engels for providing valuable edits and administrative support throughout my thesis. Thank you to Dr. Lawrence Gillman for his methodological support, guidance in conducting the research and the helpful feedback during the writing process.

Thank you to my committee, Dr. Jim Lyons and Dr. Ilana Bayer for their insight and expertise to improve my research.

To the staff of the simulation center at the University of Manitoba, I would like to give my utmost gratitude for accommodating my last-minute requests. Thank you to Dr. Jung-Un Choi for her help with data collection and Justin Gawaziuk for his help with statical analysis.

Lastly, this thesis would not have been possible without the endless support of my husband.

Table of Contents

List of tables and figures

Introduction

Evolution of Medical Education Training

Apprenticeship model

Medical Education in North America began as an apprenticeship model with little standardization. The apprentice would begin providing service to a physician with menial tasks and as the term with the preceptor drew to a close, the apprentice would take part in the daily practice (1). With the return of Americans trained in Europe, the medical education system shifted focus to the empirical approach to disease (2). This new approach required medical education to be delivered within universities. To meet demand for this preclinical education, for-profit medical schools grew in number (3). The universities were variable in their curriculum both in terms of material covered and duration. The best medical schools required a three-year curriculum, with most others requiring only two years (3). Students were commonly choosing to attend the schools with the shortest time to obtain degrees.

In 1847, the American Medical Association (AMA) recommended standardization of the academic term to six months, completion of two courses of lectures, and that students needed to provide evidence of an apprenticeship with a qualified preceptor (2). At that time, the AMA had no way of enforcing this recommendation.

Several years later, in 1901, the AMA became the national representative body for physicians and sought to standardize the requirements for medical education. The first publication of what this education should comprise was brought forth in 1905 by the Council on Medical Education (CME) and included five years of medical work, with the last two years being

clinical with an apprenticeship model (4). This model was further legitimized in 1910 with the publication of Bulletin Number Four, also called the Flexner Report (1). The AMA approached the Carnegie Foundation for the Advancement of Teaching to carry out an independent assessment of all medical schools in the United States and Canada (2). Abraham Flexner was tasked with visiting all medical schools, totaling 155 (1). The recommendations from his report included a standardization of two years of basic medical and laboratory science followed by two years of clinical learning or apprenticeship. During this time of apprenticeship, the education each student received would vary significantly based on the patients the student was assigned. This assignment was determined by hospital needs and not student's educational needs.

Rotation model/time-based model

While the undergraduate medical education was being standardized, the Flexner report also had a significant impact on the standardization of internships (1). Internships were first created by Drs Osler, Halsted, and Kelly and consisted of twelve months spent working in the surgery, medicine, and gynecology departments (5). This was the first description of rotation-based education.

After their development at Johns Hopkins, internships became widespread but were variable in their design ranging from one to two years, rotating through different specialties or completing an entire internship in a single field (6). With the report on graduate medical education in 1940, the internship became standardized to one year with the objective of preparing individuals for general practice (6). This meant the intern would complete rotations in general medicine, general surgery, pediatrics, and uncomplicated obstetrics with a focus on each service as it relates to general practice.

In addition to the internship, Johns Hopkins Hospital was the first site of residency training. Residents were medical graduates who had already completed their first year of graduate training (i.e. internship) (7). Over the subsequent years, residents would complete rotations and learn the essential skills for practice in a specialty field (8). With the development of various specialty boards between 1913 and 1940, residency programs became further established and specific requirements for training were created (6). This rotation-based and time-based model remained in place with little change for most of the 20th century.

Competency based education

Since the publication of "To Err is human" in 1999 (9), the increased accountability to public safety has required a shift in training paradigm. Competency-based medical education could fill this role.

The first description of competency-based education long predates its implementation in medicine. Carroll, in 1963, noted that students with different aptitudes will require variable amounts of time to reach the same level of proficiency (10). To eliminate this variability in outcomes, the focus should be on attainment of the goal, providing the learner with the time necessary. Fifteen years later, McGaghie proposed a model for implementation of competency based education in medicine but there was only limited uptake (11). Only more recently has there been renewed interest in competency-based education. This interest stems from a need for greater public accountability, an emphasis on ensuring the curricula emphasize skills and abilities required for practice, the promotion of learner engagement, and a decreased emphasis on time in training (12).

What is competency-based education? There are a number of definitions of competency-based medical education in the literature (13). Frank and colleagues reviewed these many definitions and developed a definition encompassing the essential features (13). They suggest: *"[c]ompetency-based education (CBE) is an approach to preparing physicians for practice that is fundamentally oriented to graduated outcome abilities and organized around competencies derived from an analysis of societal and patient needs. It deemphasizes time-based training and promises greater accountability, flexibility, and learner-centredness*"(13).

Defining competence and competency

Similar to CBE, competence has many definitions (14). The varied definitions render a discussion of competence challenging. There are two conceptual models of competence: a task-based model and a general attributes model (15). The task-based model of competence defines competence by observable behaviors and completed tasks. This model is transparent and simple but has a number of weaknesses, including a reductionist approach, and ignoring the complexity of real world applications (16). The general attributes model focuses on essential qualities necessary for successful performance. These qualities are then applied to many situations. This model also has a number of weaknesses including the difficulty in the practical application of education around general attributes and the evidence suggesting expertise is non-transferable (17,18).

Within medical education, the application of a single definition of competence remains elusive, however most definitions focus on the general attributes model (19,20). These

definitions most often include knowledge and skills as key components of competence. Other components include abilities, attitudes, judgment, values, and character attributes (21–25). The inclusion of constructs such as attitudes, values, and character attributes in the definition of competence is problematic as the question arises: Can these constructs be taught?

In an attempt to create a definition all medical education scholars can agree with, Frank and colleagues suggest competence is: *"[t]he array of abilities across multiple domains or aspects of physician performance in a certain context. Statements about competence require descriptive qualifiers to define the relevant abilities, context, and stage of training. Competence is multi-dimensional and dynamic. It changes with time, experience, and setting"* (12). With competence being the successful application of abilities, competency is the demonstration or observation of the ability.

Evaluating competence

The dichotomy of general attribute and task-based competence becomes evident when attempting to evaluate a competency. Competencies tend to be broad and general when formulated but to evaluate they are "reduced to detailed skills" (22). To limit confusion, competencies are limited to general attributes, while activities are the specific skills that make up a competency. Within CBE, these activities are known as Entrustable Professional Activities (EPAs). EPAs are the essential elements that define a profession or specialty. They must be confined to qualified personnel, independently executable, measurable in both process and outcome and should reflect at least one competency (25). These discrete

tasks are able to be executed by a trainee unsupervised once competence is demonstrated (26).

Practically, assessing competence requires assessment of all levels of Miller's pyramid, with emphasis on the highest level, "doing" (27). Assessment of "doing" requires a shift in mindset from assessment *of* learning to assessment *for* learning (25). With this change in mindset, the focus then becomes working towards competence, not simply identifying incompetence. Assessment of competence requires that multiple assessments over multiple patient encounters occur, with a multitude of different assessors to combat the bias associated with the assessment. Assessors should also be trained as there can be significant variance between raters without training (28,29). A trained assessor has two important banks of knowledge to draw on: the knowledge of the competency being addressed and the knowledge of fundamental tasks associated with being an assessor (30). The trained assessor not only provides quantitative data but qualitative information about a learner's performance. The narrative data provides actionable feedback to the learner as well as information on non-medical expert CanMEDs roles.

We also need to consider the methods of assessment being used. The use of multiple assessment methods is essential in providing a comprehensive assessment and compensating for the limitations inherent in each method of assessment (31). While work-place based assessment is required to demonstrate competence of the "does" level of Miller's pyramid, the use of traditional standardized testing to assess "knows", "knows how" and "shows how" remains.

Assessment of clinical skills

The assessment of clinical skills requires direct observation of the learner. This observation allows the supervisor to provide formative feedback and can enhance skill acquisition for learners (32). There are many tools used in direct observation with differing validity evidence to support their use (33). Most tools use a checklist, a global rating scale (GRS), or both. With the ever-increasing use of simulators the role of automated metrics in assessment of competence is being evaluated. Historically, clinical skills were assessed by expert opinion with no guidance on what was a competent level of performance.

Expert opinion

With the apprenticeship model of training, assessment of competence was based on unstandardized tests and holistic judgments by the preceptors (31). This method of assessment is common place in surgical specialties but is limited by the lack of reliability when specific criteria are not defined (34). Using expert opinion without criteria creates a challenge when comparing scores of different trainees as the "norm is in the mind of the evaluator" (35). The other difficulty with expert opinion is that the subjective evaluations cluster around the mean. This phenomenon becomes more pronounced when the time between observation and evaluation increases (35). When acceptable and unacceptable performance criteria are defined, the validity and reliability of expert assessment improves (36). Despite this increase in validity and reliability, more objective assessments were

desired. With the creation of the Objective Structured Clinical Exam (OSCE), the use of checklists became more widespread.

Checklists

Checklists are procedure-specific tools that define the essential elements of a skill (34). Checklists provide objectivity by creating a list of observable elements that are either performed or not performed. A checklist can easily be used by individuals who are less familiar with the clinical task, decreasing the need for rater training. Additionally, checklists can be used to provide specific feedback as they outline observable behaviors essential for the skill being assessed (37). When there is a clearly defined best action, checklists are reliably able to discriminate between levels of performance (38). Checklists are often seen as rigid, require a significant time investment to develop and require a separate checklist for each skill being assessed (39). Through the creation of a checklist, each element of a skill needs to be converted to a binary (performed/not performed) or trinary outcome (not performed/performed poorly/performed). By creating a binary or trinary outcome, checklists reward thoroughness without consideration of the timeliness or efficiency of an action (40). Hodges *et al* (41) demonstrated that experts score significantly worse on checklists, while scoring significantly better on GRS. With expertise development, professionals rely on more focused information rather than the thoroughness of novices resulting in higher checklist scores for novices and higher GRS scores for experts (42).

Global Rating Scales

A GRS uses a scale to quantify learner behavior either by direct observation or by recalling performance by the trainee (43). Global ratings can consist of a single item assessing overall performance or more detailed global ratings of specific aspects of performance (44). The GRS is often criticized for its subjectivity as subjective judgments often demonstrate poor reliability (45). A review by van Der Vleuten *et al*. (44), demonstrated moderate correlation between GRSs and checklists, suggesting these subjective measurement tools are not inherently unreliable and may capture more nuanced details than those assessed with a checklist. Some of the underlying subjectivity associated with the GRS may be related to a lack of understanding of what is being measured (46). To improve reliability, rater-training has been employed using different strategies (29). The most efficacious strategy has been rater error training, which usually consists of a lecture followed by discussion of common rater errors (47). The goal of this training is to increase awareness of these errors, not to achieve a specific distribution of ratings (47). Some studies have failed to demonstrate an improvement with rater training (48,49).

Both technical and non-technical skills have validity evidence supporting the use of GRSs (39,50,51). When used to assess technical skills, GRS show greater expert-novice discrimination than checklists (52). GRS are often brief allowing for easy dissemination in the clinical environment. Finally, because of their general nature GRSs also offer the advantage of use across multiple tasks allowing for more robust validity evidence (46).

Combined tools

Assessments often use a combination of checklists and GRS. The use of combined checklists and GRS offer the advantages of each individual approach while minimizing the disadvantages. The use of checklist often results in a ceiling effect where superior performances are not captured by the checklist (53). By adding a GRS, this ceiling effect is minimized (53). Additionally, GRS offer the ability to assess skills that are not easily dichotomized. The limitation of subjectivity associated with a GRS is minimized by using the checklist of observable behaviors.

The approach of combined checklist and GRS has been used for both simulation and workplace-based assessment, technical, and non-technical skills (38,53–55). The Structured Technical Skills Assessment Form (STSAF) was developed for use in the operating room to assess specific procedural skills. This tool consists of a lengthy checklist that divides a procedure into its most fundamental components and is scored from 0-2 where 0 is the component was not completed, 1 where the component was completed poorly, or 2 if it was completed well. Part 2 is the GRS which consists of ten items summarizing the important aspects of surgical conduct. When assessing junior and senior residents both Part 1 and Part 2 could discriminate between junior and senior trainees (54).

In pediatric simulation, a combined tool is used to assess infant lumbar puncture skills (53). This tool uses a 4-point GRS and 15-item checklist. When used to assess beginners, intermediates, and experts, the GRS demonstrates concordance with level of expertise. With regards to the checklist, intermediates and experts had comparable scores both of

which were higher than beginners, demonstrating a ceiling effect that is common with checklists.

There is little literature around the use of combined tools compared to checklists or GRS alone. Most work to date has focused on checklist versus GRS and has shown little information is added about a trainee's competence with the addition of the checklist. Regehr *et al* (56) assessed the importance of the checklist compared to a GRS in an OSCE evaluation. Overall, GRS scored by experts showed higher inter-station reliability, construct validity and better concurrent validity than checklists and the addition of the checklist did not improve the reliability or validity of the GRS alone (56). Despite the evidence that checklists do not improve overall discrimination ability, combined tools continue to be used with regularity.

Simulation

Simulation, as defined by Gaba *"is a technique, not a technology, to replace or amplify real experiences with guided experiences, often immersive in nature, that evoke or replicate substantial aspects of the real world in a fully interactive fashion"* (57). Simulation in clinical education has its origin in the 1950s with the development of Resusci-Anne by toymaker Asmund Laerdal (58). A second movement in simulation occurred with the development of a more sophisticated simulator, SimOne, in the 1960s (59). SimOne showed effectiveness in training but was not widely accepted for training partly due to the cost (60). Twenty years later, "high-fidelity" simulation was revisited. The third movement in simulation has

been driven by the undergraduate medical education reform and the need to prepare students to be effective residents (60).

Simulation can take many forms including part-task trainers, which replicate a part of the real task allowing learners to focus on a key aspect of the task; computer-based systems, such as audio recordings for cardiac auscultation, interactive systems that can be manipulated providing feedback based on decisions made by the learner; virtual reality systems that create environments that the learner interacts with to complete the task; simulated patients, which are actors trained to present a history and mimic physical signs or patients who have received training to present their history in a reliable way; and integrated simulators, which use a mannequin and a computer to provide physical signs and physiologic variables (60,61).

The use of simulation has expanded over the last decade with increasing incorporation of simulation into the medical education curriculum. Simulation is used to teach specific technical skills such as suturing, laparoscopic surgery, and ultrasound. As well, it teaches non-technical skills, such as situational awareness and decision making (62–67). Simulation provides a safe environment for learners to practice and develop skills such that the mistakes made do not harm patients. Learning can take place at a rate set by the individual and at a range of difficulties allowing for deliberate practice and mastery learning (68). Simulation also offers the ability to develop scenarios on-demand and create scenarios for rare clinical events. Simulation not only takes place in the simulation facility but is now being brought into the clinical space in the form of in situ simulation (69,70). Simulation is used for both low-stakes and high-stakes assessment (52,71).

Despite the widespread use of simulation, the evidence supporting improved patient outcomes is limited. Teteris *et al* (72) reviewed the literature around simulation and patient outcomes based on three task categories, skills-based, rules-based, and knowledge-based. Overall, there is evidence supporting the use of simulation in skills-based and rules-based training, however the quality of this evidence is poor and often biased towards the simulation arm of training (72). Knowledge-based simulation has no strong evidence of transfer from simulation to patient outcomes. Despite the poor evidence base supporting its use, simulation continues to have an ever-increasing role in medical education. This increased use is driven by a number of different factors including mandates from educational accreditation bodies, the patient safety-movement, the need for standardized assessment, trainee access to individualized learning, and access to replicable rare or challenging cases (73–75).

Automated assessments

This increasing use of simulation has led to the increased access to simulators, including augmented reality and virtual reality simulators. Many of these simulators have built in metrics with the potential to be used to assess learner performance and provide feedback. Metrics can be divided into quality metrics, such as errors, outcome and task repetition, and efficiency metrics, such as path length, accelerations, and time to completion.

Efficiency metrics

Efficiency metrics are related to physical parameters, thus require tracking to obtain and are objective in their measurement (76). They include time, path length, economy of movement, economy of diathermy, speed, motion smoothness, instrument orientation, depth, angular path, angular area, volume, and force/torque (77). The most commonly used efficiency metrics are time, path length, and economy of movement. These metrics, however may not always show a significant difference in expert and novice performance (78). The difference in level of expertise may not be appreciated due to differences in surgical approaches among experts and what is defined as ideal for metrics.

Motion smoothness has shown differentiating ability for certain laparoscopic manipulation tasks such as transfer tasks, sharp dissection, and laparoscopic suturing (79–81). Force metrics show differentiating ability between novices and experts for suturing tasks and tissue dissection (82,83). Less commonly assessed metrics of depth, angular area, and volume can demonstrate a trainee's mastery of space but there is little evidence establishing their validity. The studies which have assessed these less often considered metrics show differences for tasks such as grasping, bimanual coordination, suturing, clipping, and cutting (84–86).

Efficiency metrics have most often been assessed for use with laparoscopic simulation as laparoscopic simulators are widely available, require video capture to complete the procedure, and more often are the subject of simulation. Efficiency metrics in non-laparoscopic procedures relies on the use of motion sensors or specifically modified equipment to allow measurement of these parameters. One of the earliest tools developed

for this purpose was the Imperial College Surgical Assessment Device (ICSAD) (87,88). The ICSAD uses an electromagnetic tracking system applied to the back of participants hands which then measures the position of the participants hands in three-dimensional space. Using custom software, the number of movements by each hand, path length efficiency, velocity, and total time can be calculated. When asked to perform a bench top simulation of small bowel anastomosis and vein patch insertion, both time and number of movements were able to discriminate between novices, intermediates, and experts (88). The ICSAD system has since been assessed in multiple contexts, including epidural and spinal anesthesia, central venous line insertion, and ureteroscopy (89–91). In each case ICSAD was able to discriminate between novices, intermediates, and experts.

The ICSAD has also been compared to non-automated methods of technical skills assessment. Specifically, ICSAD has been compared to OSATS in ophthalmology and vascular surgery simulation (92,93). In ophthalmology, residents were recruited after a microsurgical course and asked to complete a corneal suturing task. The OSATS GRS and ICSAD path length, economy of hand movement, and time were significantly correlated, demonstrating convergent validity of the OSATS tool and ICSAD parameters (92). Within vascular surgery simulation, the ICSAD was compared to OSATS and a task specific checklist. Participants were divided into four groups based on training milestones and asked to complete a vein patch to an artery. Number of movements, time taken, and the OSATS GRS were able to discriminate between the groups and were correlated (93). The checklist was not able to discriminate between level of experience.

Quality metrics

Competency assessment requires not only efficiency in performing a task but also completing the task completely and correctly. Quality metrics are a method of assessing the completeness and correctness of a task. They are defined by a task and its execution and are therefore procedure- or task-specific (76). These metrics include outcome, errors, idle states, task repetitions, and collisions/tissue damage (77). Errors and end-product analysis have been extensively evaluated across multiple platforms and tasks, ultimately showing discriminating ability in laparoscopic, hysteroscopic, endoscopic, and otologic procedures (94–100).

Using VR simulators, collisions and tissue damage has been assessed as a quality metric. Jensen *et al.* (101) used a VR simulator of a video-assisted thoracoscopic surgery (VATS) lobectomy. The following three groups were asked to complete a right upper lobe VATS lobectomy: medical students with no experience in VATS (novice), thoracic surgery trainees with varying levels of experience (intermediate), and experienced thoracic surgeons (experts). The simulator calculated 19 metrics, of which seven correlated with level of experience. Specific quality metrics showing significant difference between novices and experts were number of severed vessels, stretching of the bronchus, blood loss, and number of times a vessel or bronchus was stapled without removing the rubber band. When differentiating between intermediates and novices, stretch damage to the bronchus, number of severed vessels and blood loss were significant. When comparing intermediates and experts, the only quality metrics that were significant were number of severed blood vessels and blood loss. The difference in metrics that showed discriminating ability for the

three comparison groups demonstrate the importance of having metrics that are relevant to the trainee's level of learning. When beginning to learn a new procedure the time to complete the procedure is less important than the errors made.

A less often considered metric is idle time, defined as the time when instrument movements and interactions are minimal. Idle time was first described as a metric when Rosen *et al.* used Hidden Markov Models (HMM) from force/torque measurements to assess laparoscopic skill (102). Using specially developed laparoscopic instruments with sensors to measure force and torque, experts and three groups of residents completed a cholecystectomy on a pig. The different force/torque interactions were defined, creating 14 different states. These states were used to create a general architecture for the HMM which allowed comparison between the expert model and models created based on each participant's procedure. The statistical difference between the residents and the expert surgeons was then calculated. In addition to differences in the paths used to complete the procedure, the idle state (i.e. when the tool was being moved in space with no tissue contact) was used differently between the experts and the novices. Experts used the idle state as a transition between tool/tissue interactions, whereas novices spent more time in the idle state planning their next movement.

In open surgery simulation, idle time was assessed for a suturing task (103). Using three different materials to simulate different tissue types (tissue paper, balloon, dense foam), experts, residents, and medical students were asked to complete three interrupted sutures in each tissue. Hand motion was captured using a motion tracking system on the back of each participant's hands. When the motion capture data was assessed, participants of all groups had more idle time when working on the tissue paper. This did not vary with

experience level. When video analysis was performed and the task broken down into steps, experts had fewer idle periods when entering the tissue with the needle, driving the needle through the tissue, and pulling the suture through the tissue. Overall, residents and medical students spent more time planning the movement of their needle while the experts spent more time ensuring their knot was secure.

Gaze-tracking is another surrogate to assess attention allocation. Gaze-tracking has been assessed as both an assessment and training tool (104). In surgical simulation, gaze training has been used to teach laparoscopic skills. Gaze training involves demonstrating the strategies used by experts as well as providing feedback to learners explaining why these strategies are used. When expert-like gaze strategies are used by novices, training time decreases, completion time improves and fewer errors are made (105,106). Gaze training also improves performance during high-anxiety situations (107). This improvement has several mechanisms, including improved attention and focus during times of stress, control of emotional state, and limiting extraneous information. Training in expert-like gaze patterns can also improve lesion detection in mammography (108).

Differences in gaze-fixation between highly experienced and less experienced individuals also exist when interpreting images. Bell *et al.* assessed gaze fixation in fellowship sonographers (experts) and resident sonographers (intermediates) when interpreting a focused assessment with sonography in trauma (FAST) exam (109). Nine regions of interest were defined. The resident group had 24 participants, of which only 14 viewed all regions of interest. All participants in the fellowship group viewed all areas of interest. The regions of interest not visualized by the resident group included some of the most sensitive areas in the FAST exam. There was no difference in total time, or the time

fixated on the right upper quadrant, pericardium, and pelvis. Gaze tracking in ultrasound simulation shows promise but further studies should use images with varied pathology. Additionally, generating the image is a key skill necessary when performing the FAST exam which needs to be assessed to determine proficiency.

Combining metrics

Technical performance cannot be measured with a single metric. Technical competence is multi-factorial with components of efficiency and quality; therefore, any assessment tool should use measures of both efficiency and quality. Using a combination of metrics, an assessment tool was created for hysteroscopy (96). Diagnostic hysteroscopy was broken down into four modules (visualization, ergonomics, safety, fluid handling) representing key elements of the procedure. Visualization, ergonomics, and fluid handling modules each had at least one quality and one efficiency metric in their scoring. Safety only had one metric associated (time colliding) which is both an efficiency and quality metric. A group of novices and a group of experts then completed five trials on the simulator. In a diagnostic exercise, there was no difference in visualization or overall score between the groups. The experts scored higher in fluid handling and ergonomics, while novices scored higher in safety. This unexpected inverse difference in safety may relate to loss of haptic feedback or the module only having a single metric associated. When the safety module is removed from the scoring, the overall score then becomes significantly different, with experts scoring higher. The metrics chosen for the hysteroscopy multi-metric scoring system, weighting and scoring, and grading system were decided upon by two experts. There is no single standardized approach to performing hysteroscopy. The differences in technique

used by the experts deviated from the "standard" and resulted in lower scores contributing to the lack of difference between experts and novices. Overall, the multi-metric scoring system demonstrates that despite using both quality and efficiency metrics, differences in technique by experts needs to be accounted for through the development of these metrics.

To this effect, several tools have combined metrics to create programs that generate scores based on a few different metrics. In ophthalmology, the EyeSi™ simulator has been used to train learners to complete aspects of cataract surgery (110). Thirteen modules were assessed with between 21 and 33 metrics generated for each module. These metrics were broken down into four categories: target achievement, efficiency, instrument handling, and tissue handling. Three groups of individuals participated, novices (ophthalmological trainees), experienced cataract surgeons (expert), and vitreoretinal surgeons (intermediate groups). Of the thirteen modules, seven could differentiate between novices, intermediates, and experts. Based on the score for the expert group on the seven discriminating modules, a cut off score was determined, and a proficiency test was created to assess competence in cataract surgery and general microsurgical skills.

Simulators offer a vast number of unbiased metrics with each procedure performed. However, these metrics need to be assessed to identify the clinically relevant metrics and determine which metrics can discriminate between novices and experts. With this ever increasing access to simulators, limited duty-hours for residents and increasing demands of patient care, the use of automated metrics can have a number of potential advantages including providing formative feedback. This could avoid the need for constant supervision by a trained operator but still allow the trainee to continually improve on their skills; track

progress of trainee's performance; allow for evaluation of trainees in a standardized and unbiased environment; and has the potential to be used in defining competency.

Focused Assessment with Sonography in Trauma

Focused assessment with sonography in Trauma (FAST) is a point of care ultrasound exam used in unstable trauma patients to assess for the presence or absence of fluid in four specific regions: right upper quadrant, left upper quadrant, pelvis, and pericardium. The results of this exam are then used by the clinician to guide management decisions. A FAST exam is rapid, decreases time to surgical intervention, length of stay, and rates of computed tomography (111). When performed by experienced clinicians, the exam can be completed in under 5 minutes and has a sensitivity between 85-96% and specificity greater than 98% (112,113).

FAST is an important tool for general surgeons and emergency medicine practitioners as it is used frequently in day-to-day practice. For this tool to be useful a user must be able to generate an acceptable quality image. This can take as few as 10 exams (114). The unstable trauma patient however is not the ideal learning model. Several training courses have been developed to allow learners to familiarize themselves with the principles and techniques of the exam while being able to practice under the guidance of experts. Most of the available courses include some component of didactic teaching followed by simulation using volunteers (115). These courses are valuable tools for training but do have some limitations, specifically the volunteers are frequently individuals who are expected to have

a normal exam, thus limiting the exposure to abnormal findings. As well, these courses require significant resources in time and money.

Simulators may be an answer for some of the current limitations. Simulators that offer the ability to perform exams that are both normal and abnormal, reduce the need for volunteers and allow training in a safe environment (115). An additional potential benefit is the ability of simulators to generate automated metrics which can provide feedback to the learner about the completeness of their exam. There are limitations to simulators including the inability to simulate difficult exams, such as in obese patients or interference from rib shadows. Thus, simulators are not a replacement for clinical experience but may offer a safe environment to begin learning these technical skills.

Conclusion

As medical education evolves, and patient-safety comes to the forefront, the need for innovative methods of teaching and assessment is crucial. Training programs have a duty to the public to ensure that by the end of training, learners are competent and ready for practice. Competency is assessed in several ways and simulation is becoming a key tool for the assessment of learners. At present, the use of simulation is common in most training programs, especially where procedural skills are being taught. Simulation as an assessment tool is resource intensive, requiring direct observation of the learner outside the clinical environment. Educators, who are often clinicians, must balance clinical duties against participating in assessment. With the ever-increasing technology available automated metrics are an attractive method to assess learners.

The FAST exam is a key procedure for both emergency medical and general surgery trainees. This simple bedside procedure is a vital tool in unstable patients and provides crucial information for clinical decision making. The ability to generate accurate, clear images and interpret the exam requires practice with both normal and abnormal exams, which is not always possible in the clinical setting given the instability often present with abnormal findings. Simulation offers a possible solution to early learning of this important skill.

Here, our research aims to seek validity evidence for the use of automated metrics provided by the VIMEDIX-AR ultrasound simulator as tools for learner assessment. This will be accomplished by assessing the metrics in relation to the gold standard of expert assessment and the ability to differentiate between differing levels of expertise.

Methods

Objectives and Hypothesis

This study has four objectives. The first is to assess if automated metrics generated using an ultrasound simulator can differentiate between three different experience levels. Second, we will examine if expert assessment using the Quality of Ultrasound Imaging and Competence (QUICk) score for FAST exams is able to differentiate the three groups. Third, we will assess for differences in participants ability to discriminate abnormal exams based on their level of expertise. Finally, we will assess if there is correlation between expert assessment, automated metrics, and the ability of participants to identify abnormal exams.

We hypothesize the automated metrics will improve with increasing experience and that the QUICk expert assessment tool will differentiate the three groups. We expect the intermediate group to score higher in expert assessment as they have recently completed a training course. When asked to identify abnormal exams we hypothesize that experts will score higher and across the groups we expect as they proceed through ten exams, the scores will improve. Finally, we expect that both the expert assessment and automated metrics will correlate with ability to correctly identify the presence or absence of fluid.

Study design

Our protocol requires the recruitment of three cohorts of ultrasonographers: a novice group of medical students and residents at the University of Manitoba, Canada, who have no personal experience with FAST techniques; an intermediate group of family medicine,

surgery, or emergency medicine residents or attendings who have completed a training course for FAST examinations within the last six months; and an expert group of staff physicians with at least five years' experience using FAST. Research in skills training has demonstrated effect sizes between 1 and 1.2 standard deviations (SD) when comparing groups of differing levels of experience (116). To detect an effect size of 1.2 SD, using a 2-tailed alpha of 0.05 and a power of 0.8, we estimate that 12 subjects will be required in each arm.

Participants were recruited via e-mail communication to targeted departments. Novices were recruited via e-mail to the junior resident cohort of trainees in the Department of Surgery and Emergency Medicine. The intermediate group was recruited via e-mail through the University of Manitoba Canadian Point-of-Care Ultrasound course. Experts were recruited via e-mail to the Departments of Surgery and Emergency Medicine.

All participants signed consent forms indicating their agreement to participate in the study including video and ultrasound image recording in accordance with the University of Manitoba Health Research Ethics Board.

Trials were completed in the University of Manitoba simulation lab. The room was set up in advance with a consistent design allowing capture of the simulator, participant's hands, and the ultrasound image. The ultrasound simulator was connected to an Epiphan Lecture Recorder x2 (Epiphan Systems Incorporated, Canada) video capture terminal which was used to generate side-by-side images of the ultrasound generated images, and the video camera output.

The participants completed a brief questionnaire of their personal experience with ultrasound both educational and clinical use and previous video game experience

(Appendix A). All study participants were shown a five minute video illustrating the FAST procedure, including the specific views and the goals of imaging the four regions (117). Participants were then oriented to the VIMEDEX ultrasound simulator and the augmented reality goggles. Instructions were given on the hand gestures associated with the augmented reality goggles to start, stop, and select items on screen.

To conceal any identifying features that would allow expert reviewer to recognize the participants all participants donned gowns prior to the start of the exam. Participants were asked to complete a FAST exam using the simulator. Time began when the participant selected start and finished when they selected done. All participants were asked to announce when they finished as the observer is unable to see the augmented reality images. After each examination automated summary metrics were recorded for the examination including time, path length, probe angular movement, percentage area visualized for each anatomic area, and total percentage area visualized.

Upon completion of the single FAST exam, participants were then asked to participate in an automated free fluid identification exercise on the VIMEDEX-AR system. The participants were presented with 10 simulated trauma patients and asked to identify any FAST regions containing free fluid in each of those patients. The VIMEDIX-AR system would randomly generate an exam that was normal or abnormal with potential for multiple regions containing fluid. Metrics obtained from each exercise included probe angular movement, path length, time, and whether the individual was able to identify the presence or absence of fluid correctly.

Upon completion of all testing, participants were asked to complete a post-test questionnaire regarding their experience with the AR goggles (Appendix A).

Data scoring and video review

The thirty-five participant videos were randomly ordered and provided to two independent evaluators. The evaluators were blinded to the participants group and completed their assessment independently without consultation between reviewers or other researchers. The evaluators used the quality of ultrasound imaging and competence (QUICk) score for FAST (Appendix B). This tool was previously validated for use on live models for FAST (118). This tool has a checklist component and a global rating scale (GRS). The checklist is broken down into the four regions of the FAST exam with specific tasks the ultrasonographer must perform or images they must generate. The GRS has seven components with a 5-point behaviorally anchored scale.

Data Analysis

Statistical analysis was conducted using SPSS version 25 (IBM, Chicago IL) and MedCalc Statistical Software version 20.014 (MedCalc Software Ltd, Ostend, Belgium; https://www.medcalc.org; 2021). Significance was accepted as an alpha of 0.05. Inter-rater agreement results with scores of 0.6, were accepted as significant.

Kruskal-Wallis non-parametric test was used to assess difference across expertise for both the automated metrics and expert assessment scores as the data violated assumptions of normality. Dunn post hoc pairwise comparisons were used to identify significant differences between groups.

Internal consistency of the QUICk tool was assessed using Cronbach alpha. A value of 0.7-0.9 was deemed acceptable (119). Inter-rater reliability was assessed using weighted

Cohen's kappa. For summed scores of checklist total and GRS total, interclass correlation coefficient was used.

To assess for differences in automated metrics of the find fluid exercise, a general linear model was constructed for each of test 1-10. The between-subjects factor was expertise (Novice, Intermediate, Expert) and the within-subject factor was exercise with 10 consecutive measures for each study participant. The mean effect of exercise as well as the interaction between exercise × expertise were examined. For all 3 models, sphericity was assessed using Mauchly's Test of Sphericity. To test the hypothesis that there is a difference in automated metrics between different levels of expertise, pairwise comparisons for the estimated marginal means of expertise were made.

Correlations between expert assessment with QUICk, automated metrics and score of the find fluid exercise were assessed using Pearson's Correlation coefficient.

Results

Demographics

In total, 13 novice, 10 intermediate, and 12 expert ultrasonographers participated in the study. Novices included undergraduate medical students and residents in general surgery, urology, internal medicine, and emergency medicine. The intermediate group included both residents and attending physicians in emergency medicine and general surgery. The expert group consisted of attending physicians in emergency medicine and general surgery (Table 1).

Table 1: Participant* (13 novice, 10 intermediate, 12 expert) *demographic data.

	Novice	Intermediate	Expert
Medical student, % (n)	23.1 (3)	0	0
Resident, % (n)	76.9 (10)	40 (4)	0
PGY1	46.1 (6)	0	--
PGY2	23 (3)	40 (4)	--
> PGY2	7.7 (1)	0	--
% Attending	0	60	100
Years in practice, % (n)			
<2	--	30 (3)	25 (3)
2 - 5	--	20 (2)	33.3 (4)
6 - 10	--	10 (1)	25 (3)
> 10	--	0	16.7 (2)
Training with FAST, % (n)	0	100 (10)	75 (9)
Used FAST clinically, % (n)	46.1 (6)	40 (4)	100 (12)
Used US in training, % (n)	92.3 (12)	100 (10)	100 (12)
Used US clinically, % (n)	53.8 (7)	70 (7)	100 (12)
US certification, % (n)	0	100 (10)	91.7 (11)
Left handed, % (n)	7.7 (1)	30 (3)	16.7 (2)

Among the three cohorts, 76.9% of novices reported clinical ultrasound use once per month or less, while 100% experts used ultrasound at least once per week (Figure 1). The intermediate group reported ultrasound use varying from never to daily.

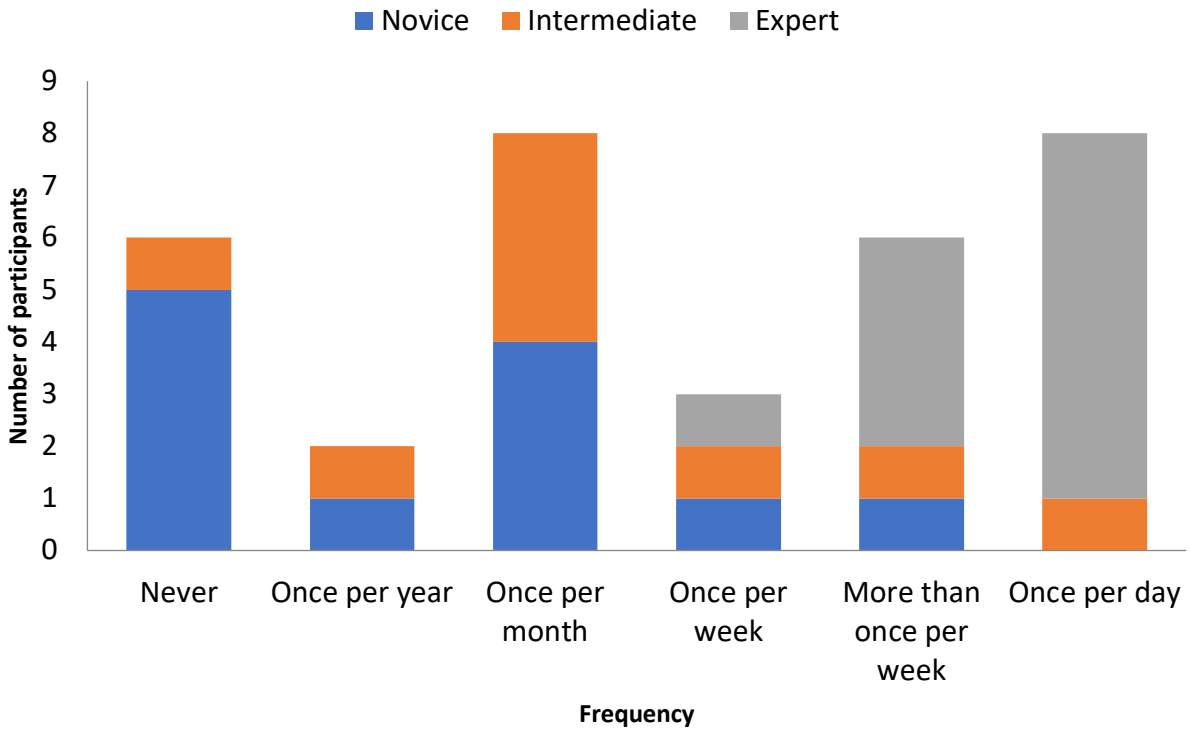

Figure 1: Clinical use of ultrasound reported by participants.

Regarding FAST use, 76.9% of novices used FAST once per month or less, compared to 91.7% of experts who used FAST more than once per week, and intermediates which 80% reported FAST use between once per month and more than once per week (Figure 2).

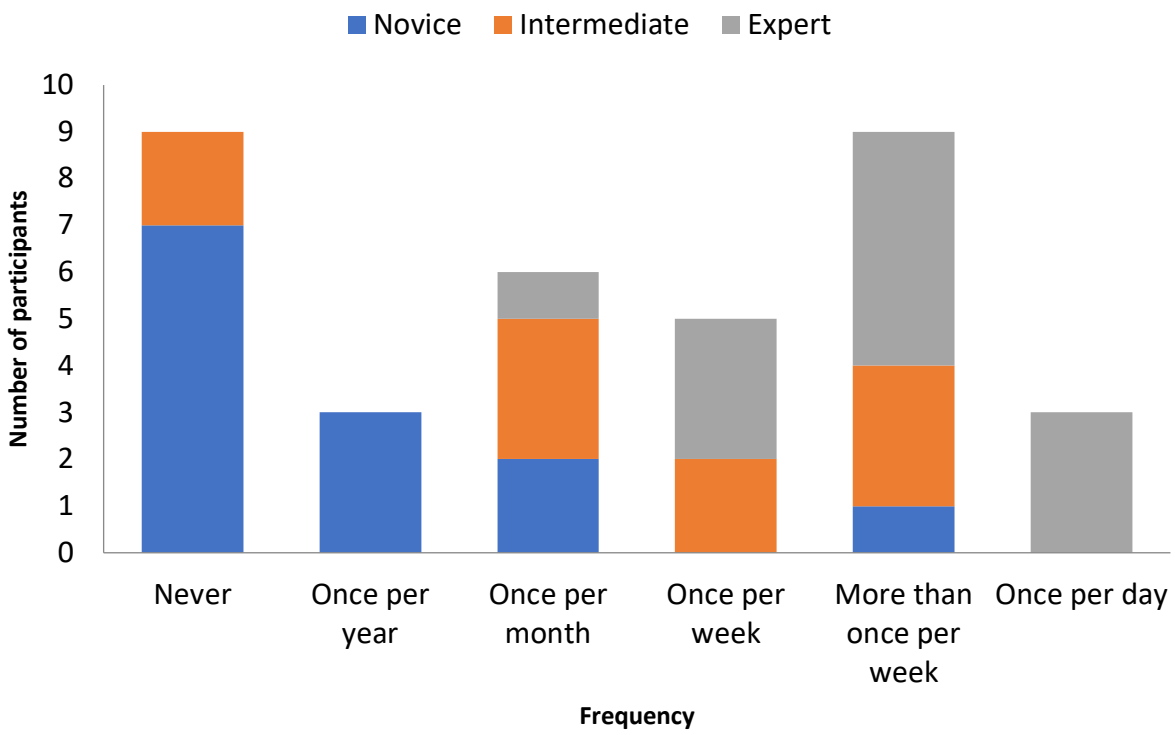

Figure 2: Clinical use of FAST reported by participants.

Regarding video game experience, 58.3%, 50%, and 46.2% of experts, intermediates, and novices, respectively use video games once per year or less (Figure 3). Zero participants in any group reported daily use of video games. No participants used augmented reality (AR) more than once per year (Figure 4).

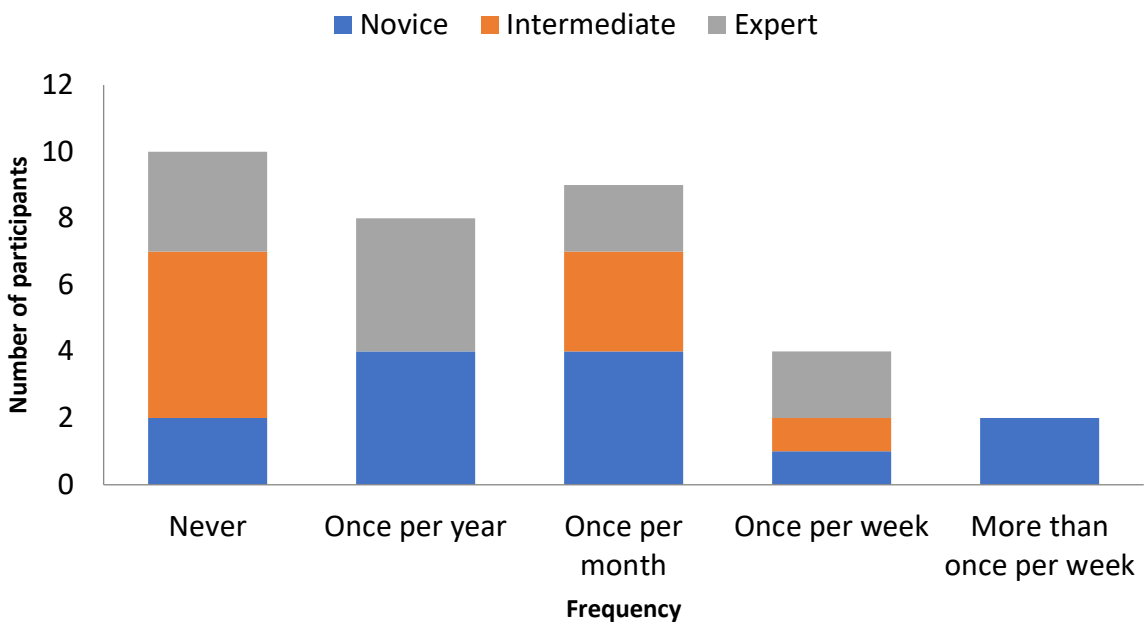

Figure 3: Participant reported use of video games.

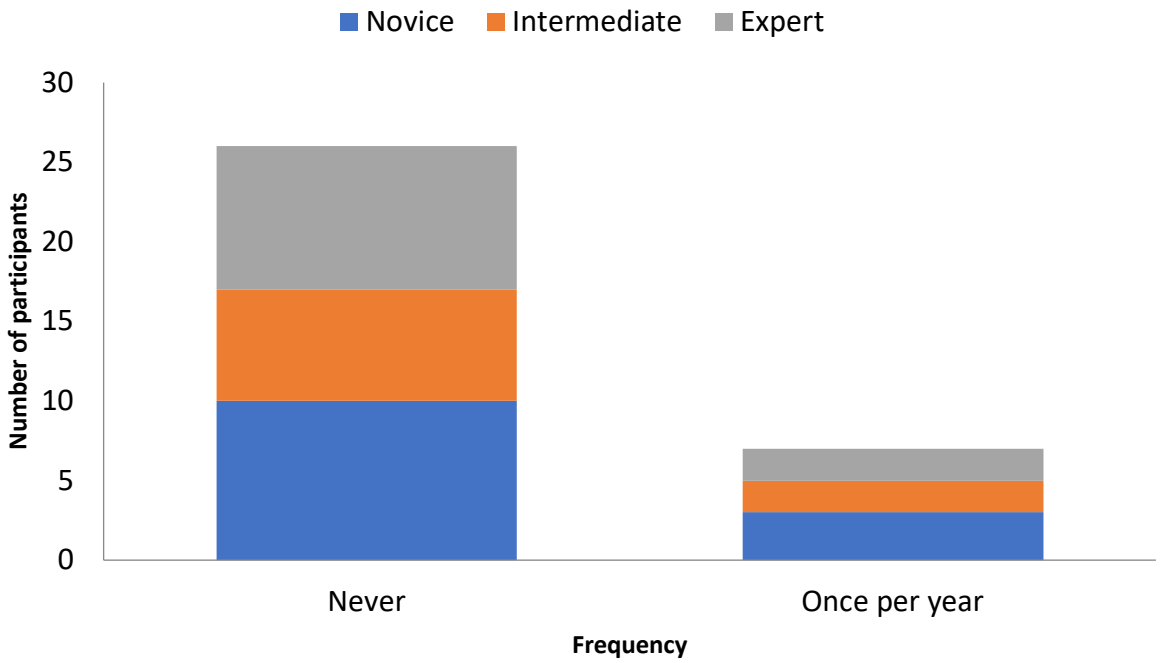

Figure 4. Participant reported use of augmented reality.

FAST automated metrics

The FAST automated metrics results are summarized in Table 2. Mean time ($p < 0.01$), mean angular movement ($p = 0.01$), and splenorenal percent area viewed were significantly different ($p = 0.02$). Post-hoc pairwise comparisons of mean time showed a difference between expert and novice and expert and intermediate. Regarding angular movement and percent area viewed for the LUQ, there was a significant difference between novice and expert.

***Table 2: Mean automated metric values for each participant group* (13 novice, 10 intermediate, 12 expert).**

Metrics	Novice (±SD)	Intermediate (±SD)	Expert (±SD)	*p-value*
Mean time (s)	304.08 (130.02)	339.70 (138.96)	176.50 (156.30)	< 0.01
Mean path length (cm)	639.07 (318.27)	611.80 (206.32)	561.92 (358.37)	0.28
Mean angular movement (degrees)	4741.00 (1370.80)	3573.70 (1139.89)	3635.42 (2478.40)	0.01
Percent area viewed - Total	87.45 (10.51)	84.62 (12.64)	86.57 (5.53)	0.45
Percent area viewed - Hepatorenal	90.77 (10.14)	93.90 (9.92)	92.75 (7.43)	0.73
Percent area viewed - Splenorenal	86.15 (10.15)	71.90 (26.44)	72.42 (12.29)	0.02
Percent area viewed - Pelvis	88.77 (21.24)	89.80 (12.60)	95.8 (8.2)	0.32
Percent area viewed - Pericardium	84.15 (17.76)	84.70 (15.66)	85.08 (9.35)	0.78

FAST expert assessment

Mean expert assessment scores are detailed in Table 3. Within the checklist domains, pelvis ($p = 0.01$) and pericardium ($p < 0.01$) were significant, in addition to total score ($p < 0.01$). Post hoc pairwise comparisons are shown in Table 4. Both the pericardial domain and total checklist score had pairwise differences between novice and intermediates as well as novice and experts.

Within the global rating scale, all domains were significantly ($p \leq 0.04$) different except time and flow. For the post hoc pairwise comparisons there were no differences between intermediates and experts (Table 4).

Table 3. Mean expert assessment score by participant group (13 novice, 10 intermediate, 12 expert) demographic data.

	Novice (±SD)	Intermediate (±SD)	Expert (±SD)	*p-value*
Checklist				
Hepatorenal	3.31 (0.63)	3.60 (0.97)	3.58 (0.67)	0.21
Splenorenal	2.62 (1.04)	2.50 (1.27)	2.83 (0.83)	0.79
Pelvis	1.69 (0.25)	3.10 (0.99)	2.83 (1.11)	0.01
Pericardial	1.69 (0.95)	2.90 (0.32)	2.92 (0.29)	< 0.01
Total	9.31 (2.60)	12.10 (1.73)	12.17 (2.08)	0.01
Global rating Scale				
Skin	3.54 (0.52)	4.0 (0)	3.92 (0.29)	0.01
Probe placement	2.54 (1.05)	3.20 (0.63)	3.58 (0.90)	0.04
Image scrolling	2.85 (1.07)	3.70 (0.48)	3.92 (0.90)	0.02
Positioning and Handling	3.23 (0.73)	3.80 (0.42)	3.92 (0.29)	0.01
Time	1.69 (1.11)	2.10 (0.99)	2.83 (1.59)	0.13
Flow	3.77 (0.44)	3.70 (0.67)	4.00 (0)	0.23
Overall	1.92 (1.12)	2.80 (1.03)	3.17 (1.27)	0.03
Total	19.54 (4.39)	23.3 (3.16)	25.34 (4.10)	P <0.01

Table 4. Post hoc pairwise comparisons for expert assessment.

	Novice-Intermediate	Novice-Expert	Intermediate-Expert
Checklist			
Hepatorenal	--	--	--
Splenorenal	--	--	--
Pelvis	0.02	0.07	1
Pericardial	< 0.01	< 0.01	1
Total	0.03	0.02	1
Global rating Scale			
Skin	0.02	0.06	1
Probe placement	0.41	0.04	1
Image scrolling	0.17	0.02	1
Positioning and Handling	0.08	0.01	1
Time	--	--	--
Flow	--	--	--
Overall	0.03	0.02	1
Total	<0.01	< 0.01	0.49

Internal consistency and inter-rater agreement

Cronbach alpha was calculated for both the checklist component and GRS. The checklist had an alpha of 0.75 while the GRS had an alpha of 0.84. Inter-rater agreement was assessed using weighted kappa (Table 5).

Table 5. Inter-rater agreement on QUICk score.

Domain	Kappa
Checklist	
Hepatorenal	0.38
Splenorenal	0.39
Pelvis	0.66
Pericardial	0.66
Total	0.83
Global rating Scale	
Skin	0.17
Probe placement	0.3
Image scrolling	0.33
Positioning and Handling	0.22
Time	0.37
Flow	0.48
Overall	0.39
Total	0.81

Automated metric – expert assessment correlation

Pearson's correlation was used for correlations between the automated metrics and the expert assessment (Tables 6 and 7). The percent area viewed RUQ and the expert score hepatorenal were correlated (r = 0.33, p = 0.05). The percent area viewed LUQ, pelvis, and pericardial did not correlate with their respective checklist scores. The total percent area viewed correlated with the hepatorenal expert score (r = 0.53, p < 0.05) but not the other checklist scores for the three other regions. The checklist total correlated with all components of the GRS except with GRS Flow (r = 0.32, p > 0.05). There was a trend toward significance in the correlation between checklist total and percent area viewed RUQ (r = 0.33, p = 0.06) as well as checklist total and percent area viewed suprapubic (r = 0.32, p =

0.06). The percent area viewed LUQ was negatively correlated with GRS components of probe handling (r = -0.37, $p = 0.03$), image scrolling (r =-0.40, $p < 0.01$), and positioning and handling (r = -0.37, $p = 0.03$).

When looking at the other automated metrics of path length, angular movement, and time, time was positively correlated with path length (r = 0.67, $p < 0.01$) and angular movement (r = 0.59, $p < 0.01$). The GRS probe handling component was negatively correlated with time (r = -0.42, $p = 0.01$), path length (r = -0.49, $p <0.01$), and angular movement (r = -0.61, $p < 0.01$). Neither the GRS overall or checklist total correlated with time, path length, or angular movement.

Table 6. Pearson correlation coefficients of percent area viewed component and each component of QUICk assessment.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. Percent Area View Total	1.00																
2. Percent Area View RUQ	0.40*	1.00															
3. Percent Area View LUQ	0.76**	0.15	1.00														
4. Percent Area View Pericardial	0.75**	0.12	0.38*	1.00													
5.Percent Area View Suprapubic	0.72**	0.12	0.30	0.44**	1.00												
6. Score Heptorenal	0.53**	0.33*	0.34*	0.52**	0.28	1.00											
7. Score Splenorenal	0.18	0.10	0.24	-0.09	0.25	0.19	1.00										
8. Score Pelvis	0.06	0.29	-0.26	0.08	0.26	0.30	0.09	1.00									
9. Score Pericardial	-0.10	0.13	-0.39*	0.14	0.02	0.34*	-0.02	0.50**	1.00								
10. Score Total	0.22	0.33	-0.07	0.20	0.32	0.63**	0.50**	0.79**	0.67**	1.00							
11. GRS Skin	0.02	0.04	-0.27	0.18	0.21	0.43*	0.25	0.48**	0.36*	0.59**	1.00						

12. GRS Probe Placement	-0.10	-0.04	-0.37*	0.17	0.09	0.30	0.26	0.32	0.59**	0.55**	0.41*	1.00				
13. GRS Image Scrolling	-0.11	0.14	-0.40*	0.09	0.05	0.45**	0.25	0.48**	0.59**	0.67**	0.53**	0.75**	1.00			
14. GRS positioning and Handling	-0.10	-0.02	-0.37*	0.07	0.15	0.22	0.31	0.40*	0.46**	0.54**	0.41*	0.71**	0.70**	1.00		
15. GRS Time	0.17	0.26	-0.06	0.08	0.28	0.32	0.31	0.38*	0.33	0.52**	0.35*	0.53**	0.54**	0.28	1.00	
16. GRS Flow	0.33	-0.08	0.18	0.14	0.57**	0.25	0.38*	0.05	0.21	0.32	0.13	0.37*	0.18	0.30	0.26	1.00
17. GRS Overall	0.19	0.20	-0.02	0.07	0.37*	0.41*	0.56**	0.65**	0.46**	0.82**	0.48**	0.49**	0.59**	0.43*	0.78**	0.29	1.00

* $p \leq 0.05$, ** $p \leq 0.01$

1 = Percent Area View Total, 2 = Percent Area View RUQ, 3 = Percent Area View LUQ, 4 = Percent Area View Pericardial, 5 = Percent Area View Suprapubic, 6 = Score Hepatorenal, 7 = Score Splenorenal, 8 = Score Pelvis, 9 = Score Pericardial, 10 = Score Total, 11 = GRS Skin, 12 = GRS Probe Placement, 13 = GRS Image Scrolling, 14 = GRS positioning and Handling, 15 = GRS Time, 16 = GRS Flow, 17 = GRS Overall

Table 7. Pearson correlation coefficients of automated metrics and each component of the QUiCK assessment.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Time	1.00														
Path Length	0.67**	1.00													
Probe Angular Movement	0.59**	0.73**	1.00												
Score Hepatorenal	-0.06	0.17	0.04	1.00											
Score Splenorenal	0.00	-0.25	-0.05	0.19	1.00										
Score Pelvis	0.04	-0.08	-0.14	0.30	0.09	1.00									
Score Pericardial	-0.10	-0.15	-0.43**	0.34*	-0.02	0.50**	1.00								
Score Total	-0.03	-0.14	-0.22	0.63**	0.50**	0.79**	0.67**	1.00							
GRS Skin	-0.14	-0.21	-0.26	0.43*	0.25	0.48**	0.36*	0.59**	1.00						
GRS Probe Placement	-0.42*	-0.49**	-0.61**	0.30	0.26	0.32	0.59**	0.55**	0.41*	1.00					
GRS Image Scrolling	-0.20	-0.17	-0.32	0.45**	0.25	0.48**	0.59**	0.67**	0.53**	0.75**	1.00				
GRS Position and Handling	-0.22	-0.22	-0.34*	0.22	0.31	0.40*	0.46**	0.54**	0.41*	0.71**	0.70**	1.00			
GRS Time	-0.31	-0.25	-0.31	0.32	0.31	0.38*	0.33	0.52**	0.35*	0.53**	0.54**	0.28	1.00		
GRS Flow	-0.35*	-0.25	-0.15	0.25	0.38*	0.05	0.21	0.32	0.13	0.37*	0.18	0.30	0.26	1.00	
GRS Overall	-0.08	-0.17	-0.18	0.41*	0.56**	0.65**	0.46**	0.82**	0.48**	0.49**	0.59**	0.43*	0.79**	0.29	1.00

* $p \leq 0.05$, **$p \leq 0.01$

1 = Time, 2 = Path Length, 3 = Probe Angular Movement, 4 = Score Hepatorenal, 5 = Score Splenorenal, 6 = Score Pelvis, 7 = Score Pericardial, 8 = Score Total, 9 = GRS Skin, 10 = GRS Probe Placement, 11 = GRS Image Scrolling, 12 = GRS Position and Handling, 13 = GRS Time, 14 = GRS Flow, 15 = GRS Overall.

Find fluid exercise

A repeated measures one-way ANOVA was completed for each of the automated metrics. Mauchly's test indicated the assumption of sphericity was violated and the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity. Results for time show a main effect of exercise number (df = 4.77, $p < 0.01$, partial eta squared = 0.30) with no significant interaction between exercise number and expertise (df = 9.54, $p = 0.08$, partial eta squared = 0.10). For post hoc testing, there was a significant difference between novice and expert and a trend toward significance between intermediate and experts.

With respect to path length and angular movement, there was a main effect of exercise number (df = 4.49, $p < 0.01$, partial eta squared 0.20; df = 4.86, $p < 0.01$, partial eta squared 0.20, respectively) with no interaction between exercise number and expertise for these two metrics (df = 8.97, p =0.68, partial eta squared 0.04; df = 8.97, $p = 0.68$, partial eta squared 0.04, respectively). Post hoc testing demonstrated no difference between expertise levels.

The find fluid exercise scores were then assessed for correlation with both the automated metrics and the expert assessment results (Table 8). The LUQ find fluid score correlated with the checklist score (r = 0.38, $p < 0.05$). There was no correlation between the scores on find fluid exercise and the automated metrics or the GRS overall.

Table 8. Correlation of find fluid exercise scores with automated metrics and expert assessment.

	1	2	3	4	5	6	7	8	9	10	11	12	13
LUQ Total	1.00												
RUQ Total	0.32	1.00											
Pelvic Total	0.54**	0.21	1.00										
Pericardial Total	-0.03	0.05	-0.21	1.00									
Time	-0.09	-0.18	-0.06	-0.03	1.00								
Path Length	0.06	-0.07	-0.03	-0.05	0.67**	1.00							
Probe Angular Movement	-0.08	-0.19	-0.11	-0.03	0.59**	0.73**	1.00						
Percent Area View RUQ	0.09	-0.08	0.14	-0.30	0.25	0.17	0.22	1.00					
Percent Area View LUQ	0.09	-0.21	-0.19	-0.09	0.23	0.35*	0.60**	0.16	1.00				
Percent Area View Pericardial	0.20	0.07	0.24	-0.01	-0.13	0.04	0.11	0.12	0.38*	1.00			
Percent Area View Suprapubic	0.09	0.17	0.03	-0.21	-0.24	-0.16	0.10	0.13	0.30	0.44**	1.00		
Score Total	0.38*	0.28	0.32	-0.33	-0.03	-0.14	-0.22	0.33	-0.07	0.20	0.32	1.00	
GRS Overall	0.26	0.25	0.23	-0.27	-0.08	-0.17	-0.17	0.20	-0.02	0.07	0.37*	0.82**	1.00

* $p \leq 0.05$, **$p \leq 0.01$

1 = LUQ Total, 2 = RUQ Total, 3 = Pelvic Total, 4 = Pericardial Total, 5 = Time, 6 = Path Length, 7 = Probe Angular Movement, 8 = Percent Area View RUQ, 9 = Percent Area View LUQ, 10 = Percent Area View Pericardial, 11 = Percent Area View Suprapubic, 12 = Score Total, 13 = GRS Overall

Participant experience

Following completion of the exam, participants were asked if they experienced any adverse effects from the AR, such as headache or nausea. Most participants did not experience any headaches during or after use of the AR headset (Figure 5). Two participants reported nausea during and after use (Figure 5).

Participants rated the value AR added value (Figure 6) and if they felt AR improved the experience (Figure 7). Experts disagreed more frequently with AR adding value and improving experience while most novices agreed or strongly agreed with both statements.

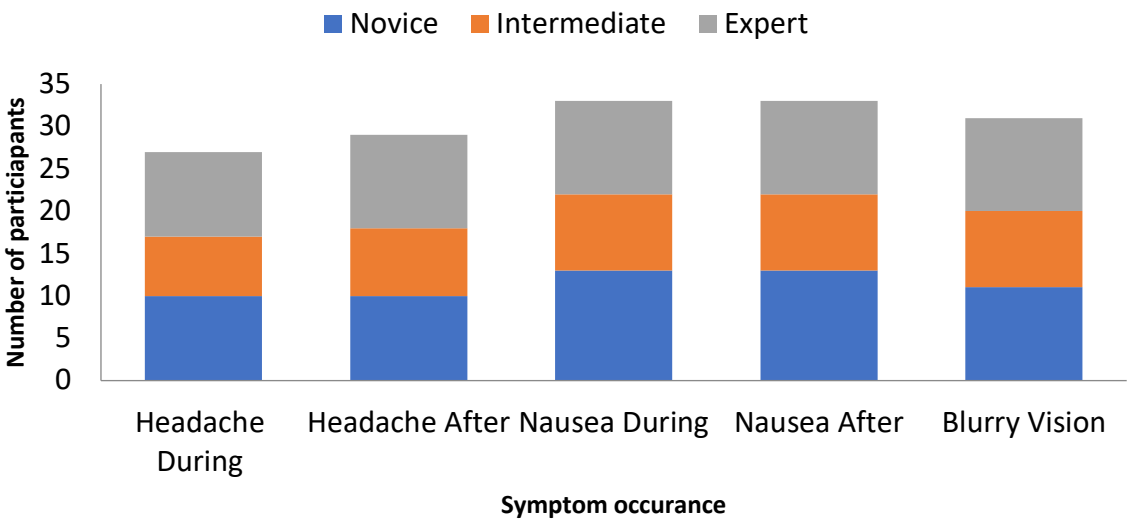

Figure 5. Adverse effects reported by participants during and after use of the VIMEDIX-AR simulator.

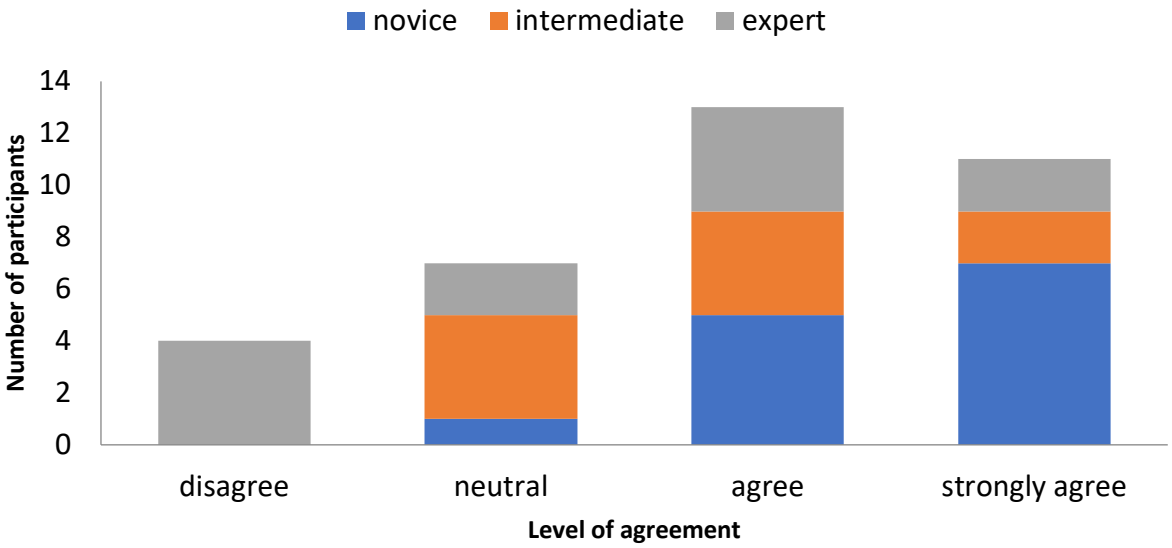

Figure 6. Participants response to the statement "The augmented reality goggles glasses added value to my experience with the ultrasound simulator."

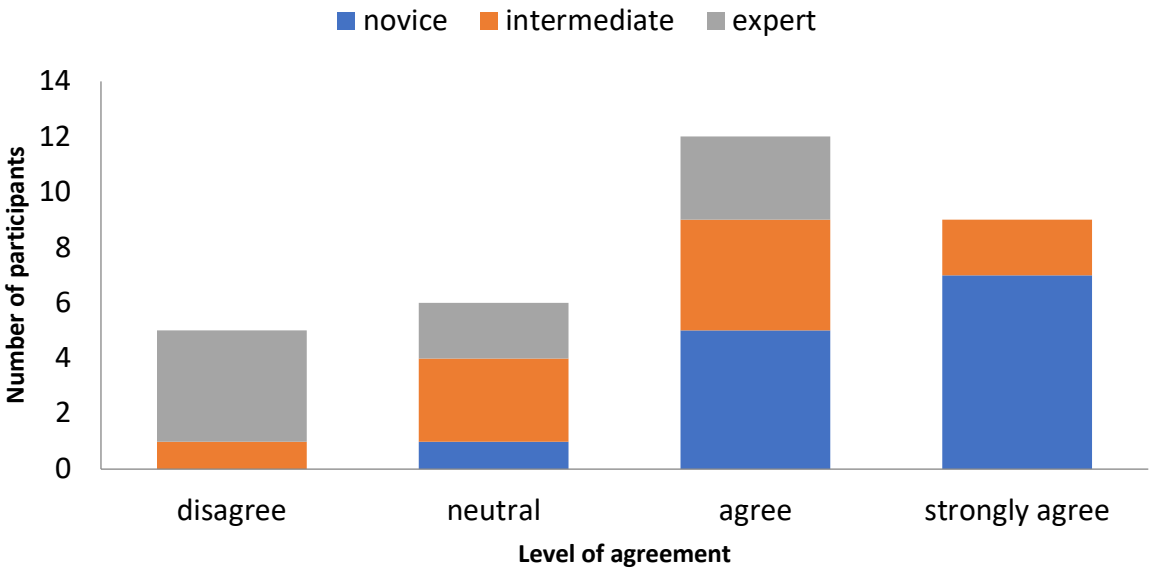

Figure 7. Participants response to the statement "The augmented reality goggles improved my experience with the ultrasound simulator."

Discussion

We sought validity evidence for the use of automated metrics in the assessment of FAST ultrasound skills. We first aimed to identify if the automated metrics were able to differentiate varying expertise. We then ascertained that our expert assessment tool, the QUICk, was able to identify the differing level of expertise on the ultrasound simulator. We used the simulator's ability to generate abnormal exams to create a 10-exercise find the fluid test. Finally, we assessed for correlations between the automated metrics, QUICk score, and score on the find fluid exercise. This thesis explores the results and relevant literature as well as a discussion of the implications of the results. Finally, the strengths, limitations, and future directions for research is outlined.

Our pre-participation assessment showed that novices used ultrasound less frequently than both intermediates and experts. As well, novices used FAST less often than the other two groups. The intermediate group used ultrasound more frequently than novices but less frequently than experts. These support our labels of "novice", "intermediate", and "expert" for the purpose of this study.

Automated metrics

Simulation is ubiquitous in medical education. Point of care ultrasound is an area where simulation with volunteers is being used to facilitate skill acquisition. The current approach to ultrasound skill acquisition is the use of volunteers and faculty trainers. Simulators offer the potential for independent self-study therefore identifying metrics that

can differentiate expertise are needed in developing validity evidence for these tools. Our study assessed four automated metrics for the VIMEDIX-AR simulator: time, angular movement, path length, and percent area viewed. We found time could differentiate between novice and experts as well as experts and intermediates. Novices and intermediates require the same amount of time to complete the exam, but experts are faster. The ultrasound course provides the foundation for FAST exams but more experience is required to create expertise. Time as a metric has been used as a method of assessment in a number of skills including, open surgery, laparoscopic surgery, arthroscopy, central line insertion, and peripheral nerve block (120–124). When looking specifically at ultrasound simulation, time has shown validity evidence for discrimination of expertise in obstetric ultrasound, transesophageal echocardiography, and FAST exams**.** (125–127). Time is an efficiency metric which does not provide information about the quality of an exam and thus is insufficient as a sole metric. Despite this, the role of time is clinically relevant in the setting of FAST exams as this is the first test used in unstable trauma patients to determine need for clinical intervention. Therefore, an expedient and complete exam is essential for the care of these patients.

Angular movement, another efficiency metric, is the tilting and rotating motion of the ultrasound probe. The operator begins by centering the organ of interest then tilts the probe to view the entire organ. We found a difference between novices and experts in angular movement, with novices having higher number of degrees of angular movement. This suggests novices are identifying their organ of interest but, to be confident they have viewed the entire organ, are tilting and rotating the probe more, either sweeping through

multiple times or angling further than necessary. Like angular movement with ultrasound, is tip angulation in colonoscopy. Obstein *et al.* (128) assessed kinematic data of a colonoscopy simulator and found that tip angulation was significantly different between a cohort of expert gastroenterologists and novice fellows when completing these exams. These data show that with expertise, individuals develop economy of motion. When performing FAST exams, this economy of motion translates to less probe angulation.

Path length is a commonly used automated metric. With the VIMEDIX-AR simulator, path length is the total distance the probe moves. We found no difference based on expertise. This is similar to reported data in obstetric ultrasound when assessing fetal heart anatomy (126). However, this is contrary to most data on path length as an automated metric in open, robotic and laparoscopic surgery, anesthesia, and central line insertion (88–91,129,130). When looking at data on FAST exams, previous reports using healthy volunteers and hand path length have found a difference between cohorts of novices and experts (125,131). All the above studies use hand path length. This is different than path length with our simulator. Hand path length includes all movements including distance the probe or instrument moves as well as the finer adjustments, such as tilting or angling. With the VIMEDIX-AR simulator the gross movements of the probe would be captured as path length whereas finer tilting movements, if the probe is stationary on the mannequin, would represent only angular movement. Path length also incorporates not only the distance the probe moves while in contact with the simulator but also the distance used to move between regions. These movements between regions are not required of all participants and are not necessarily reflective of expertise.

With increasing ultrasound expertise, motions are smoother and finer with less erratic movements (132). Our findings suggest that angular movement, which is fine movement of the probe, better captures expertise than path length alone. In another study looking at FAST exams, Zago and colleagues (125) used hand motion analysis in FAST exams and deconstructed the exam into regions focusing on each region individually and excluding the time between regions as they are not inherently related to the FAST exam. They found that path length was significantly different, however this again included angular movement as they used hand motion analysis.

Because a FAST exam does not have a concrete endpoint, rather the examiner stops when they believe they have completely visualized the area of interest, percent area viewed on a simulator represents a possible concrete endpoint and is thus a possible quality metric. Percent area viewed was composed of five metrics, a percent area viewed for each of the four regions of the FAST exam and a total percent area viewed. Percent area viewed calculations by the simulator have a time component to determine this metric. To ensure imaging an area was intentional a certain amount of time was set as the threshold to consider an area imaged, this was to ensure credit wasn't given for random movements of the probe that happened to image an area when it was unintentional.

Only percent area viewed for the left upper quadrant (LUQ) was significantly different with the difference between novice and experts, with novices having the higher mean percent area viewed. The LUQ is often a more technically challenging view to obtain, with the views often being inadequate (133). Despite this, our novices were able to visualize 86% of the area, compared to 71% by the intermediate group, and 72% by the

expert group. One possible explanation is that novices may have spent more time attempting to visualize this area and used higher angular movement to obtain those views. Our data was not collected in a way where the time to complete each individual region was collected and thus this hypothesis is not able to be tested in our study.

No other studies have looked at percent area viewed as a metric. The most similar accuracy metric used in assessment was on healthy volunteers, where Bell *et al*. developed points of interest, key anatomic areas that represent essential images for a complete FAST exam (134). These points of interest were developed using images from an expert sonographer and the points chosen. They were then recreated in 3-dimensional space on a healthy volunteer. The number of points of interest was not described however there were multiple points within each region. The percent of points visualized was significantly different between novice and intermediate groups across all four regions of the exam.

Expert Assessment

The QUICk was developed as an objective skill assessment tool for FAST (118). This tool has validity evidence for use on volunteers. Our study assessed if the QUICk assessment tool was able to differentiate expertise when used with ultrasound simulators. Because it was developed for use on volunteers, there were items within both the checklist and GRS which were not applicable to our study. Within the checklist, each region was composed of six items. With the VIMEDIX-AR simulator, participants did not

need to adjust depth or gain, thus these items were removed from all checklists. The pericardial region also had a checklist item for appropriate use of adjuncts, which was not necessary with the simulator. With these adjustments, the total possible score was 15 instead of the original 24. The GRS had nine domains in the original description. We excluded two: image adjustment and autonomy. Image adjustment assessed the learners appropriate use of gain and depth, while autonomy assessed the ability to complete the exam without direction or guidance. Despite removing these items, our Cronbach alpha was 0.84 for the GRS and 0.75 for the checklist, demonstrating good reliability without redundancy.

Two raters completed the QUICk tool on each participant. To assess inter-rater agreement we used a weighted kappa, which assesses for agreement that would occur by chance. The kappa score ranges from -1 to 1, with negative values representing disagreement, positive values being agreement, and zero representing chance alone (135). The strength of agreement is summarized in Table 9.

Table 9. Kappa correlation and interpretation from Viera et al. (135).

Kappa	Interpretation
<0	Less than chance agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.00	Near perfect agreement

Our findings show fair to moderate agreement for most of the domains except for GRS skin, which showed only slight agreement. The original description of QUICk

showed similar results with slight agreement for GRS skin and fair to moderate agreement in all other regions. Our findings are that the checklist total and GRS scores are tools that can be used in the assessment of novices using the ultrasound simulator.

Checklist

When looking at the checklist, two of the four region checklists were able to differentiate novices from either intermediate or experts, while the checklist total differentiates novices from intermediates and experts. The checklist for the hepatorenal and splenorenal spaces were not significantly different. Ziesmann *et al.* (118) reported significant differences in all checklist scores between novices and experts. Here, we used the same definition of novice and expert as those authors. The difference we found may be the result of either the simulator, the changes we made to the checklist, or the inclusion of third group resulting in insufficient power for subgroup analysis (118). Although the experts were familiar with the simulator, the use of the AR headset represents a new challenge, nonetheless one would expect the same principles of exam be applied regardless of the use of the simulator or a live volunteer.

The pelvis checklists showed there was a difference between novice and intermediates. The difference in the pelvis checklist is primarily the result of novices only imaging this area in one plane where the intermediate group more consistently obtained images in two planes. The experts were inconsistent in obtaining two views resulting in no difference between experts and either novices or intermediates. The instructional video demonstrated only a single view, and if able to completely image the bladder, a single

view is enough. The second plane is taught as there can be challenges in obtaining images in one plane or the other. Two views also improves the sensitivity of the exam (136). Intermediates, having recently completed a credentialling program would be familiar with the recommendation that two views of the bladder should be obtained. Experts, having significant experience performing these exams may prefer one view over the other but use the second view when the first is incomplete. Novices, who have none or very limited experience with FAST and no formal training, would only know the view taught in the video and thus would not readily examine the bladder in two planes.

The novice group had significantly lower total checklist scores than both the intermediate and expert groups. Having no training, other than our instructional video, it is expected they would perform worse. The video shows how to complete an exam, but five minutes is not enough time to teach all the principles of the FAST exam and how to obtain complete and correct images. We expected intermediates to score higher in expert assessment as they would have recently learned the technically correct and complete FAST exam. Experts, with time and experience, develop shortcuts that allow for an accurate exam but don't necessarily represent the most technically correct exam. This however was not the case. We found no difference between our intermediate and expert groups. When looking at the percent scores for the checklist, most of the scores from our study are higher than those reported by Ziesmann *et al.* (118, see Table 10). These differences show the experts and intermediates identify the anatomic landmarks and area of interest, orient the image correctly and, based on expert assessment, visualize the entire

region. The simulator may be too simple to identify the subtle differences in technique and image optimization is required to differentiate intermediates and experts.

Table 10: Comparison of Task checklist findings between Ziesmann et al. (118) and our study results.

	Our results			Ziesmann *et al.*	
	Percent score Novice	Percent score Intermediate	Percent score Expert	Percent score Novice	Percent score Expert
Heptaorenal	82.75	90	89.5	55.5	77.83
Splenorenal	65.5	62.5	70.75	57	77.17
Pelvis	42.25	77.5	70.75	41	60.5
Pericardial	56.3	96.67	97.33	31.33	72.17
Total	62.07	80.67	81.13	46.21	71.7

In a similar study, Burckett-St Laurent *et al*. (137) adapted a tool developed for clinical assessment for use with a simulator. Their tool used a combined checklist and global rating scale to assess ultrasound-guided regional anesthesia. A group of novices and a group of experts were asked to complete an ultrasound guided nerve block on high-fidelity simulator as well as on a volunteer. The two groups were then rated by an onsite clinician as well as by a blinded clinician off site via video recording. They found significantly higher checklist scores for the experts on a live patient but not the simulator. The GRS could differentiate the groups on the simulator as well as the live patient when assessed both on- and off-site. In their results, Burckett-St Laurent *et al*. (137) provided overall checklist score and not the scores on the three components of the checklist. The authors suggest familiarity with the simulator and cues from the simulated environment resulted in increased scores for the novices and thus the checklist was unreliable for use

with the simulator. In our study, some components of the checklist could differentiate the groups and the total score could discern novices from both intermediates and experts, thus the checklist as a whole is a reliable tool with the simulator but the individual components are not.

When comparing our results to similar checklists developed for assessment of ultrasound skills, our ratios of expert to novice scores are lower than other reported studies where the checklist was a reliable tool for assessment. These comparisons are summarized in Table 11. Burkett-St Laurent had the lowest expert-to-novice ratios and found the checklist to be unreliable for use with the simulator. Our lower ratios are unsurprising given the narrow range of scores on the checklist and the higher scores for our novice group. These higher novice scores are driven by fewer checklist items and higher scores in the hepatorenal and splenorenal region checklists.

Table 11: Comparison of checklist scores among similar ultrasound skills assessment studies

Assessment Tool	Score As a Percent of Total	Ratio of Expert to Novice Score	Inter Rater Agreement of Findings
Checklist, Experts	80.7%	1.31	
Checklist, Novices	62.1%		
Ziesmann *et al.* (118) Task Checklist, Experts	71.7%	1.55	0.7951
Ziesmann *et al.* (118) Task Checklist, Novices	46.2%		
Zago *et al.* (125) checklist, Experts female model	80.0%	1.15	NR
Zago *et al.* (125) checklist, Novice female model	69.6%		
Zago *et al.* (125) checklist, Expert male model	77.9%	1.21	NR
Zago *et al.* (125) checklist, Novice male model	64.2%		
Sultan *et al.* (138) Task Checklist, Experts	87.1%	1.98	0.842
Sultan *et al.* (138) Task Checklist, Intermediate	71.4%		
Sultan *et al.* (138) Task Checklist, Novices	43.9%		
Chin *et al.* (122) Task Checklist, Experts	93.1%	1.39	0.97
Chin *et al.* (138) Task Checklist, "Late" Fellows	91.0%		

Chin *et al.* (138) Task Checklist, "Early" Fellows	71.5%		
Chin *et al.* (138) Task Checklist, Novices	66.8%		
Burckett-St. Laurent el al. (137) Checklist, Experts off-site	78.2%	1.08	0.61
Burckett-St. Laurent el al. (137) Checklist, Novices off-site	72.5%		

*NR not reported

Two other studies have used the QUICk tool for assessment of FAST, both only using the checklist scores. The first, used QUICk as a method to assess the quality of images generated while assessing hand motion analysis between beginners and experts (125). They found a better overall score for experts, but their beginners had higher scores than the original report by Ziesmann *et al*. (118). To compare our checklist scores to those reported by Ziesmann *et al*. (118) and Zago *et al*. (125), we must look at percent scores rather than the total score as we modified the checklist to suit a simulator (Table 11). Experts in the three studies were defined similarly as having five years of experience with ultrasound. Despite the similar definition, our percent scores for the expert group are most similar to Zago *et al*. (125), which is higher than those reported by Ziesmann *et al*. (118). Zago *et al*. attributed this difference to the use of two models rather than one. Our increased scores are possibly the result of the modifications to the checklist. The simulator may have also played a role in the higher scores as it was not designed to create a difficult exam such as with overlying bowel gas, rib shadows or an obese patient. When looking at beginner scores, our beginner group is most similar in experience to Ziesmann

et al.'s (118) beginner group however our scores are higher than Ziesmann *et al*.'s (118) beginner group. Zago et al's (125) beginner group was recruited from a group of individuals who completed a formal ultrasound training course the same day as study participation but had limited ultrasound experience prior to the training. This is most similar to our intermediate group. The scores for our intermediate group however are higher than those of Zago *et al*. (125). This difference in score may be the result of our intermediate group having experience with ultrasound outside the course as well as the checklist changes we made. Overall, our increased scores are likely due to a combination of changes in the checklist as well as an "easy" exam on the simulator.

The second study used the QUICk tool checklist as part of an OSCE exam for paramedicine students. (139) The study aimed to assess the diagnostic accuracy as measured by sensitivity, specificity, positive and negative predictive value for FAST performed by paramedics. The OSCE exam used an ultrasound simulator and number of students who passed was reported, however the scores were not reported, therefore we cannot compare our experience with the checklist to those obtained in Buaprassert *et al*.'s study.

Global Rating Scale

All components of the GRS except flow and time were significantly different, with the difference being between novices and intermediates or novices and experts. None of the components showed a difference between intermediates and experts, suggesting that

training courses provide the foundation of ultrasound use necessary for a technically good exam.

The skin component assesses the probe contact with the skin and, when performing an exam on a volunteer, the appropriate use of ultrasound gel. Novices were different from intermediates with a trend toward significance between novices and experts. With the simulator, lower scores on this component mean the participant did not have consistent contact between the probe and the mannequin. We would expect some challenges with novices, however having some experience with ultrasound, the novice group had mean score of 3.54 meaning adequate skin contact most of the time. The experts scored higher but was still lower than expected. This finding is likely due to the simulator being binary in image generation meaning, if the probe is close enough, the simulator will generate an image but it cannot adjust the image if there is less or more contact.

Image scrolling assesses how smoothly the ultrasonographer fans through the area to obtain images. This would be similar to angular movement, with more angular movement required with more fanning or more staccato fanning. Similar to the automated metric, imaging scrolling had a difference between novice and experts, with novices having lower scores and thus more staccato movement.

Probe placement assesses the ultrasonographer's need to readjust the probe, with perfect score denoting correct placement of the probe with adequate views on the first attempt. Here, novices had significantly lower scores than experts. With increasing experience, one would expect a larger portion of experts to be able to correctly place the probe on first attempt. Our novice and intermediate group were not different, meaning the

ultrasound training course did not increase the ability of ultrasonographers to correctly place the probe on first attempt.

Positioning and handling assesses the ultrasonographer's ergonomics while performing the exam. Here again there was a difference between novice and experts, with more awkward body positioning and inappropriate handling of the ultrasound probe amongst the novices. There was trend toward significance between novice and intermediate groups, suggesting that the ultrasound course provides foundational knowledge on how to correctly position oneself as well as how to handle the ultrasound probe.

Flow assesses if the ultrasonographer frequently jumps between regions or if they move smoothly through the exam in a logical manner. There was no difference between the groups. The specific order of exam did not matter, provided there was an organized sequence to the exam. The video shown before the assessment described one approach to the order of the FAST exam, therefore, we would expect novices to follow the same order as provided by the video, which is one example of a logical approach. Intermediates and experts, with previous experience would also have an approach to order of the exam, thus no difference was found in this component.

For the time component the behavioral anchors are greater than ten minutes, between two and five minutes and less than two minutes, with higher scores awarded for shorter time. Additionally, if a participant scored two or less on any component, they would receive a score of one. No participants required greater than ten minutes and only experts had times less than one minute. Time is clinically relevant when performing a FAST

exam thus the importance placed on time as an individual GRS component. However, there was no differences in the scores across groups. This may be related to the second part of the behavioral anchor, where if a participant scored two or less on any other component their score was automatically dropped to one. The QUICk time component is different from other GRS tools, which often place time and motion as a single component. One of the most widely known GRS tools is the Objective Structured Assessment of Technical Skill (OSATS), which encompasses seven domains related to operative skill (52). The OSATS tool is often used as a non-specific GRS for technical skills. Within the OSATS and frequently when a generic GRS is used, time and motion are a single item that focus on economy of movement and efficiency rather than time alone. In most technical skills, completing the skill fast is less important than being efficient with the movement. Being more efficient however should decrease the time to complete a task. With the QUICk, the components of time and motion are broken into separate items to allow for specific assessment of these two skills independently, despite being inter-related.

Finally, overall performance is rated from unacceptable to exceptional, where the expected performance for safe practice is a four (118). Despite individual items not differentiating between all groups, the overall assessment, which does not define what is an unacceptable exam, shows difference between novice and both intermediate and expert. The mean score for our experts however did not meet the minimum expected standard of four. The mean score of 3.17 falls in the range of unacceptable with minor inadequacies, while intermediate score of 2.8 would be unacceptable with major

inadequacies. These scores are similar to those reported for the expert group in the initial validation of the QUICk tool (118). This lower-than-expected score may be due to experts, who over time, develop strategies that deviate from the algorithms, thus resulting in lower scores but clinically appropriate quality exams. It may also reflect experts not being as skilled as expected, having trained in a time without valid objective assessment tools for FAST or degradation of skill over time.

Similar to our checklist score ratio, the ratio of expert to novice scores for the GRS were lower than those previously reported by Ziesmann *et al*. (118) as well those reported for similar ultrasound skills assessment (Table 12). These differences in our ratio for the GRS are driven by lower scores for experts compared to the percent scores reported by Ziesmann *et al*. (118). These lower scores may be the result of shortcut or strategies used by experts. We also removed two components of the QUICk GRS as they were not relevant to the simulator. These two items, autonomy and image adjustment, may be important factors when assessing expertise. For our study, we did not provide any guidance when performing the exam thus, there was no way to assess autonomy, although, with increasing expertise one would expect higher scores for autonomy. Image adjustment also represent gain and depth adjustment, which were removed from the checklist. Our ratios however are similar to those by Burckett-St Laurent *et al*. (137), who adapted a regional anesthesia assessment tool for use with a simulator.

Table 12: Comparison of Global Rating Scales Among Similar Ultrasound Skills Assessment Studies

Assessment Tool	Score As a Percent of Total	Ratio of Expert to Novice Score	Inter Rater Agreement of Findings
Total GRS, Experts	72.4%	1.29	
Total GRS, Novices	55.8%		
Ziesmann *et al*. (118) GRS, Experts	88.2%	1.54	0.860
Ziesmann *et al*. (118) GRS, Novices	57.2%		
Sultan *et al.* (138) GRS, Experts	78.6%	1.97	0.795
Sultan *et al.* (138) GRS, Intermediate	52.6%		
Sultan *et al.* (138) GRS, Novices	39.8%		
Chin *et al*. (122) GRS, Experts	94.9%	1.67	0.98
Chin *et al.* (122) GRS, "Late" Fellows	92.1%		
Chin *et al*. (122) GRS, "Early Fellows	65.7%		
Chin *et al*. (122) GRS, Novices	56.9%		
Burckett-St Laurent *et al*. (137) GRS, Expert off-site assessment	88.0%	1.20	0.61
Burckett-St Laurent *et al*. (137) GRS, Novice off-site assessment	73.6%		

Find fluid exercise

The first part of our study assessed participant's ability to generate images, a key component of ultrasound skills. However, to be proficient in the use of ultrasound, the clinician must also be able to interpret the images they generate. With the find the fluid exercise we aimed to combine the image generating and interpretation skills. As part of the find fluid exercise, automated metrics of path length, angular movement and time were generated for each exercise. The first exercise in the series was the slowest with participants becoming faster as they grew accustomed to the exercise and the simulator. There was also a difference between novice and experts with experts being faster to complete the exercises. This is as we expected given experts were faster in the initial test of a normal exam. We also expected experts to be faster as they have experience identifying abnormal exams. There was a trend toward significance between intermediates and experts. This is different from our first exercise of a normal exam, where there was a difference between intermediates and experts. All participants for the first exercise did not have to interpret any images, simply demonstrate the ability to generate images. In our find fluid exercise, participants now had to both generate and interpret images. The fluid findings the simulator generates are not subtle, rather they represent somewhat extreme findings. A possible reason there was no difference in time between intermediate and expert groups is when these two groups encountered an area with fluid, they moved on to the next region, not fully scanning the positive region. This is a reasonable approach as the goal is to identify fluid, if fluid is present and if found early, completely examining that area doesn't change the answers. The novices may have

been less confident in their findings as few had previous experience with FAST and thus may not immediately recognize a positive finding of fluid.

Path length and angular movement also showed an effect of exercise number with the path length and angular movement being significantly higher for the first exercise compared to all subsequent exercises. Again, this likely represents familiarity with the simulator and experience with the exercise.

The importance of practicing with positive exams was demonstrated by Garcias *et al*. They found by simulating positive exams with peritoneal dialysis patients, participants who completed their modified program were able to detect smaller volumes of fluid than those who completed the standard course, where no abnormal exams were presented during training. Several other studies have assessed learners' ability to generate and interpret FAST exams, however, these studies have separated the image acquisition and interpretation components of the exam. For example, Damewood *et al*. (140) recruited medical students to participate in a study where half the group practiced on a healthy volunteer, while the other half used an ultrasound simulator. As their post course assessment, participants were asked to interpret pre-recorded FAST exams and then to complete a FAST exam on a healthy volunteer. They found no difference in participant's ability to interpret images or to generate images. Previous studies have used both simulators and peritoneal dialysis patients to demonstrate positive findings in the FAST exam. Salen *et al*. (141) used an ultrasound simulator and peritoneal dialysis patient and compared learners ability to identify positive findings on photographic images after practicing with one of the two models. After completing the test, the participants could

then practice on the model they hadn't yet used. Overall, there was no difference in participants ability to identify fluid on photographic images. Chung *et al*. (142) combined image acquisition and interpretation skills assessment. In their study, participants trained on either a healthy volunteer or an ultrasound simulator. Participants who trained using a healthy volunteer had shorter scan times when assessed on healthy volunteers and more participants had high quality windows in the RUQ. There was no difference in window quality for other regions or image interpretation between the two groups. The benefit to using a simulator over volunteers is the ability to practice on multiple normal and abnormal exams. This increased experience of abnormal exams is important as increased confidence is associated with improved accuracy (143).

The goal of the find fluid exercise was to combine image generation and interpretation to assess participants. Overall, we found participants became faster, used shorter path lengths and fewer angular movements after the first exercise. This exercise could serve as another tool in training learners in FAST, offering experience with positive exams which may improve confidence.

Correlation analysis

Our correlation analysis aimed to identify if the automated metrics were related to the QUICk assessment and if performance on the find fluid exercise was related to QUICk assessment. Within the automated metrics, percent area viewed total was correlated with all percent area viewed items. The total percent area viewed is the summation of each region, thus higher percent area viewed in each region intuitively should correlate with

higher total percent area viewed. The percent area viewed total was also correlated with the checklist scores for hepatorenal and splenorenal area. All region checklists included a point for completely viewing each region, thus each region should correlate with their checklist score, however only the percent area viewed RUQ correlated with its respective checklist score. Interestingly, the percent area viewed LUQ and percent area viewed pericardial also correlated with the hepatorenal score and the strongest correlation was between the percent area viewed pericardial and hepatorenal score. The reason for this correlation is not clear, as the checklist for each region is independent of the other regions and the four regions are separate enough that the scanned area should not overlap between the regions.

When looking at the correlations with the GRS, the percent area viewed LUQ was negatively correlated with three GRS component: probe placement, image scrolling and positioning and handling. Having limited experience, novices struggle with the ergonomics of reaching across the patient to place the probe as far posterior on the left flank to obtain LUQ views. These struggles resulted in low scores in the three above mentioned GRS components but had the highest percent area viewed for this region. Experts, on the other hand, had lower percent area viewed LUQ but scored higher in probe placement, image scrolling and positioning and handling. Together these differences contributed to this negative correlation between percent area viewed LUQ and the expert scores on probe placement, image scrolling and positioning and handling.

The other automated metrics, time, path length and angular movement were correlated. Intuitively this makes sense, the longer the path length the more time you

would take, and the more time spent moving the probe could result in more sweeping or tilting the probe. Time was negatively correlated with probe placement and flow. If the probe is placed incorrectly, it will take time to correct the placement and will interrupt the smooth flow of the exam. Path length was also negatively correlated with probe placement. Like time, incorrect placements require correction with more movement. With path length being the same across our three expertise levels, one would not expect to find correlations with the checklist total or GRS overall, which did show differences between expertise levels.

Angular movement was negatively correlated with the checklist pericardial score, suggesting tilting the probe more without being thoughtful and economical with the movement does not improve the quality of images. Probe placement and positioning and handling were also negatively correlated with angular movement. Meaning, correct probe placement will require less tilting to generate optimal images and correct handling and good ergonomics can lead to more economical movements.

The individual components of the QUICk, both checklist and GRS, correlated with each other with the exception of flow. The GRS overall and the checklist total had the strongest correlation of 0.82, meaning the checklist and the GRS are measuring similar things. None of the automated metrics correlated with the GRS overall or the checklist totals, suggesting automated metrics alone are insufficient to fully characterize the quality of a FAST exam with our simulator. In other hand motion analysis studies, path length and time correlate with expert assessment. In a study of ophthalmology residents and microsurgical technique, Ezra *et al*. found strong negative correlations between time, path

length and hand movements and a general procedural GRS. (92) Other studies using hand motion analysis and ultrasound skills assessment have been conducted in the clinical setting and have demonstrated strong correlations between GRS, checklist and path length as well as time (122,138).

Post-assessment feedback

The post-assessment feedback showed a large portion of participants experienced some adverse effect. Although less commonly reported with AR, headache and other adverse effects are possible (144)**.** The most common effect in our study was headache both during and after participation. The AR headset is heavy and must be tightened and adjusted to the user's eye level prior to beginning the exercises. With prolonged wear, as required to complete all parts of our study, the tightness of the headband may have contributed to headache, in addition to some effect of cybersickness. Cybersickness is a constellation of adverse effects that result from sensory mismatch (145). Symptoms can include nausea, discomfort, dizziness, and disorientation. Headache alone is rarely a symptom of cybersickness (145). Two important contributors to headache in AR are accommodation and convergence. Accommodation is what allows our eyes to shift between focus on near and far objects. This is accomplished by the ciliary muscles contracting and relaxing. When focusing on objects at a comfortable distance the ciliary muscles are relaxed. As we focus on objects that are closer, the ciliary muscles contract. Additionally, when focusing on objects that are closer to us, our eyeballs must rotate towards each other, or converge (146). These two movements, accommodation and

convergence, are important when using AR goggles as the images generated are close to the eye and we must alternate focus between the near images on the goggles and the surrounding environment which is further away. With the sustained muscle use over a period of time, individuals can develop eye strain or headaches.

The other symptom we asked about was nausea. Few participants experienced nausea as an adverse effect. Augmented reality goggles use the natural environment and overlays a simulation on to the environment. This allows the user to still fully interact with the natural environment. Although nausea has been reported as a cybersickness effect, it is commonly associated with other symptoms such as discomfort, dizziness, and disorientation. With AR using the natural environment, mismatch between sense is less common, resulting in less cybersickness (147).

All participants were asked to rate two statements about the AR goggles: the AR goggles added value and the AR goggles improved my experience with the simulator. More novices agreed or strongly agreed with both statements, while experts mostly disagreed or were neutral toward the statements. The AR goggles added a layer of complexity to completing the exercise. Experts, having significant experience with ultrasound, may find that the additional complexity from the goggles distracts from the exercise and adds to the cognitive load. Novices on the other hand, having limited experience with ultrasound and the simulator may find integrating the goggles with the simulator easier and thus find value in the experience.

Strengths and limitations

This study was the first to assess the QUICk tool for use with an ultrasound simulator. The tool could differentiate novices from experts most consistently but was inconsistent in differentiating intermediates and experts. Our intermediate group was required to complete an ultrasound training course in the 6 months prior to participation. The correct technique on completing a FAST exam would be fresh information and thus the intermediate group would have a technically correct exam similar to the experts. The inability to detect a difference between the intermediate and expert group may also stem from the smaller size of the intermediate group. We planned to recruit 12 participants per group however with the limitation placed on intermediates, we were unable to identify 12 participants. The incomplete intermediate group may also contribute to the non-significant differences in the automated metric assessment. Our sample size was calculated for an effect size of 1 to 1.2 standard deviations as suggested in the literature (116,148). This resulted in group sizes of 12. Despite the power calculation, this sample size may be too small for subgroup analysis such as item by item assessment of the checklist and GRS.

When defining our three groups we attempted to separate experience as much as possible however, when asked about experience there was overlap between the three groups. Although our novice group had limited ultrasound experience and no formal ultrasound training, the novice group did have participants that had some experience with FAST clinically and one participant reported use of FAST more than once per week. Our intermediate group completed an ultrasound training program within 6 months. Part of

this group participated immediately following their course while the remainder had time between their course and participation, resulting in more experience. In retrospect, the ideal intermediate group would participate immediately following their training course. Identifying a truly novice group is increasingly challenging due to the incorporation of ultrasound experience into medical student education. This is done with simulation as well as clinical exposure, thus our medical students may have more exposure to the simulator and thus perform better than our experts who are unfamiliar with the simulator.

Another possible contributor to the similar performance across expertise is the simulator is too simple. Simulators do not present the same challenges of rib shadowing or bowel gas and the intra-abdominal organs are clear with obvious interfaces. By having a simplified model, it may allow novices to score higher than if some or all of these challenges were presented. Despite the limitations of simulators, several authors have demonstrated simulators as a useful tool in training ultrasound skills (62,140,149,150).

Automated metrics have limitations. We chose path length as this is metric has been shown across skills, including in FAST exams, to differentiate expertise. The path length generated by the simulator includes transitions between regions as well as movements of the probe on the mannequin but does not include the fine movements of tilting the probe if the probe is stationary on the mannequin. Most previous studies have used hand tracking as a method of measuring path length, which would include all movements of the probe. The lack of difference in path length may be related to how path length was calculated by the simulator, thus this metric may not be an appropriate metric. Further

studies could use hand tracking with the simulator to assess for a relationship between hand movement and path length or angular movement.

The automated metrics of percent area viewed were developed with a time component to ensure the area being imaged was deliberately imaged and not simply the result of random probe movement. This sensitivity requires a balance to ensure the time is not so short that any quick image of the area gets credit but not so long that an expert who is efficient with their imaging does not receive credit. Given how high novices scored in percent area viewed this sensitivity likely needs further adjustment to be a useful metric.

A key limitation in our correlation analysis was the multiple comparisons we completed. We did not correct for these multiple comparisons which may result in us identifying a correlation when one is not truly present, thus we must be cautious in our interpretation of these correlations.

Despite these limitations, our study demonstrated the QUICk is a reliable assessment tool for use with the VIMEDIX-AR simulator. The automated metrics of time, path length and percent area viewed LUQ generated by the simulator can differentiate novices from experts but these metrics alone are insufficient to assess expertise.

Future directions

Future research should include analysis of a larger cohort of all groups. This could allow for narrower confidence intervals and potentially identify significance in metrics that were non-significant in our study. With more data, a reassessment of the correlation between the GRS, checklist and automated metrics may identify new or stronger

correlations. With the present study, the QUICk can be used to assess novices on the ultrasound simulator, creating an evidence-based assessment of competence.

Future work could also involve a cumulative sum analysis to identify the amount of repetition required with the simulator to reach a defined failure rate. This failure rate could be defined based on the automated metrics or on expert assessment. Cumulative sum analysis has been used with a number of skills both clinically and with simulation (151–154). This would allow for an adaptive curriculum for the learner. Next steps should also include assessment of trainees who complete a learning program with simulator to assess for skills transfer, which is the goal of simulation training.

Conclusion

We set out to assess the automated metrics associated with the VIMEDIX-AR simulator and the QUICk assessment tool for their ability to differentiate skill at the FAST exam. The automated metrics of time, angular movement and percent area viewed LUQ showed differences between novices and experts. The QUICk assessment tool checklist and GRS overall could differentiate the novices from both the intermediates and experts. Finally, although there were correlations between some parts of the QUICk tool and some of the automated metrics, none of the automated metrics correlated with the GRS overall or checklist total suggesting that, in their current format, the automated metrics may not be appropriate measures of skill for the FAST exam. Overall, the time and angular movement may serve a complimentary role to the QUICk assessment. The QUICk should be integrated into a learning curriculum for novices being introduced to

the FAST exam. Further work is needed to see if the automated metrics may have a role in assessing learners when it comes to FAST exam skills.

References

1. Flexner A. Medical Educaton in the United States and Canada. New York; 1910.
2. Schrewe B. From history to myth: Productive engagement with the Flexnerian metanarrative in medical education. Advances in Health Sciences Education. 2013;18(5):1121–38.
3. Custers EJFM, Cate O ten. The History of Medical Education in Europe and the United States, With Respect to Time and Proficiency. Academic Medicine. 2018;93(3):S49–54.
4. Bevan AD. Co-operation in Medical Education. The Journal of the American Medical Association. 1928;90(15):1173–7.
5. Wentz DK, Ford C V. A Brief History of the Internship. JAMA: The Journal of the American Medical Association. 1984;252(24):3390–4.
6. Commission on Graduate Medical Education. Graduate Medical Education: Report of the Commission on Graduate Medical Education. Chicago: University of Chicago Press; 1940.
7. Ludmerer KM. Let Me Heal : The Opportunity to Preserve Excellence in American Medicine. New York: Oxford University Press; 2015. 20 p.
8. Council on Medical Education and Hospitals. Medical Education in the United States. 1928;91(7):473–99.
9. Institute of Medicine. To err is human: Building a Safer health system. Institute of Medicine [Internet]. 1999 Mar 1 [cited 2021 Sep 30];(November):1–8. Available from: https://iom.nationalacademies.org/Reports/1999/To-Err-is-Human-Building-A-Safer-Health-System.aspx
10. Carroll JB. A model of school learning. Teachers College Record. 1963;64(8):723–33.
11. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003-2009. Vol. 44, Medical Education. 2010. p. 50–63.
12. Frank JR, Snell LS, Cate O Ten, Holmboe ES, Carraccio C, Swing SR, et al. Competency-based medical education: Theory to practice. Medical Teacher.

2010;32(8):638–45.

13. Frank JR, Mungroo R, Ahmad Y, Wang M, De Rossi S, Horsley T. Toward a definition of competency-based education in medicine: A systematic review of published definitions. Medical Teacher. 2010;32(8):631–7.

14. Fernandez N, Dory V, Ste-Marie LG, Chaput M, Charlin B, Boucher A. Varying conceptions of competence: An analysis of how health sciences educators define competence. Medical Education. 2012;46(4):357–65.

15. Gonczi AG. Competency based assessment in the professions in australia. Assessment in Education: Principles, Policy & Practice. 1994;1(1):27–44.

16. Ashworth PD, Saxton J. On 'Competence.' Journal of Further and Higher Education. 1990;14(2):3–25.

17. Ennis RH. Critical Thinking and Subject Specificity: Clarification and Needed Research. Educational Researcher. 1989;18(3):4–10.

18. Greeno JG. A Perspective on Thinking. American Psychologist. 1989;44(2):134–41.

19. Leung W-C. Competency based medical training: review. BMJ. 2002 Sep 28;325(7366):693–6.

20. Clements R, Mackenzie R. Competence in prehospital care: Evolving concepts. Emergency Medicine Journal. 2005;22(7):516–9.

21. Epstein RM, Hundert EM. Defining and Assessing Professional Competence. Journal of the American Medical Association. 2002;287(2):226–35.

22. Bhatti NI, Cummings CW. Viewpoint: Competency in surgical residency training: Defining and raising the bar. Academic Medicine. 2007;82(6):569–73.

23. Wood FP. The complex world of competence. AJR. 2010;194:1176.

24. Epstein RM, Dannefer EF, Nofziger AC, Hansen JT, Schultz SH, Jospe N, et al. Comprehensive assessment of professional competence: The Rochester experiment. Teaching and Learning in Medicine. 2004;16(2):186–96.

25. Ten Cate O, Scheele F. Viewpoint: Competency-Based Postgraduate Training: Can We Bridge the Gap between Theory and Clinical Practice? 2007;82(6):542–7.

26. Ten Cate O, Chen HC, Hoff RG, Peters H, Bok H, Van Der Schaaf M. Curriculum development for the workplace using Entrustable Professional Activities (EPAs):

AMEE Guide No. 99. Medical Teacher. 2015;37(11):983–1002.

27. Miller GE. Assessment of clinical skills/competence/performance. Academic Medicine. 1990;65(9):s63–7.

28. Gingerich A, Regehr G, Eva KW. Rater-Based Assessments as Social Judgments: Rethinking the Etiology of Rater Errors. Academic Medicine. 2011;86(10):S1–7.

29. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D. Rater training to support high-stakes simulation-based assessments. Journal of Continuing Education in the Health Professions. 2012;32(4):279–86.

30. Kogan JR, Holmboe E. Realizing the Promise and Importance of Performance-Based Assessment. Teaching and Learning in Medicine. 2013;25(SUPPL.1):68–74.

31. Van Der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Advances in Health Sciences Education. 1996;1:41–67.

32. Ericsson KA. Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains. Academic Medicine. 2004;79(Supplement):S70–81.

33. Holmboe ES, Hauer KE. Tools for Direct Observation and Assessment A Systematic Review. Jama. 2009;302(12):1316–26.

34. Reznick RK. Teaching and Testing Technical Skills. The American Journal of Surgery. 1992;165:358–61.

35. Cogbill TH, Malangoni MA, Potts JR, Valentine RJ. The general surgery milestones project. Journal of the American College of Surgeons. 2014;218(5):1056–62.

36. Neufeld V, Norman G. Assessing Clinical Competence. Springer Pub. Co;

37. Archer JC. State of the science in health professional education: Effective feedback. Medical Education. 2010;44(1):101–8.

38. Kim J, Neilipovitz D, Cardinal P, Chiu M. A comparison of global rating scale and checklist scores in the validation of an evaluation tool to assess performance in the resuscitation of critically ill patients during simulated emergencies (abbreviated as "CRM simulator study IB"). Simulation in Healthcare. 2009;4(1):6–16.

39. Szasz P, Louridas M, Harris KA, Aggarwal R, Grantcharov TP. Assessing technical competence in surgical trainees: A systematic review. Annals of Surgery. 2015;261(6):1046–55.

40. Yudkowsky R, Park YS, Riddle J, Palladino C, Bordage G. Clinically discriminating checklists versus thoroughness checklists: Improving the validity of performance test scores. Academic Medicine. 2014;89(7):1057–62.

41. Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE Checklists Do Not Capture Increasing Levels Of Expertise. Academic Medicine. 1999;74(10):1129–34.

42. Dreyfus SE. The five-stage model of adult skill acquisition. Bulletin of Science, Technology and Society. 2004;24(3):177–81.

43. Tabish SA. Assessment Methods in Medical Education. International Journal of Health Sciences. 2008 Jul;2(2):3–7.

44. Van Der Vleuten CPM, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. Medical Education. 1991;25(2):110–8.

45. Muzzin L. Oral Examinations In: Assessing clinical competence. Neufeld V, Norman G, editors. New York: Springer Publisher; 1984.

46. Boulet JR, Murray D. Review article: Assessment in anesthesiology education. Canadian Journal of Anesthesia. 2012;59(2):182–92.

47. Woehr DJ, Huffcutt AI. Rater training for performance appraisal: A quantitative review. Journal of Occupational and Organizational Psychology. 1994;67(3):189–205.

48. McLaughlin K, Ainslie M, Coderre S, Wright B, Violato C. The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. Medical Education. 2009;43(10):989–92.

49. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: A randomized, controlled trial. Journal of General Internal Medicine. 2009 Jan 11;24(1):74–9.

50. Steinemann S, Berg B, Ditullio A, Skinner A, Terada K, Anzelon K, et al. Assessing teamwork in the trauma bay: Introduction of a modified "NOTECHS" scale for trauma. Vol. 203, American Journal of Surgery. 2012. p. 69–75.

51. Kim J, Neilipovitz D, Cardinal P, Chiu M, Clinch J. A pilot study using high-

fidelity simulation to formally evaluate performance in the resuscitation of critically ill patients: The University of Ottawa Critical Care Medicine, High-Fidelity Simulation, and Crisis Resource Management I Study. Critical Care Medicine. 2006;34(8):2167–74.

52. Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. British Journal of Surgery. 1997;84(2):273–8.

53. Gerard JM, Kessler DO, Braun C, Mehta R, Scalzo AJ, Auerbach M. Validation of global rating scale and checklist instruments for the infant lumbar puncture procedure. Simulation in Healthcare. 2013;8(3):148–54.

54. Winckel CP, Reznick RK, Cohen R, Taylor B. Reliability and construct validity of a Structured Technical Skills Assessment Form. The American Journal of Surgery. 1994;167(4):423–7.

55. LeBlanc VR, Tabak D, Kneebone R, Nestel D, MacRae H, Moulton CA. Psychometric properties of an integrated assessment of technical and communication skills. American Journal of Surgery. 2009;197(1):96–101.

56. Regehr G, MacRae HM, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Academic Medicine. 1998;73(9):993–7.

57. Gaba DM. The future vision of simulation in healthcare. Simulation in healthcare : journal of the Society for Simulation in Healthcare. 2007;2(2):126–35.

58. Tjomsland N, Baskett P. The Resuscitation Greats Asmund S. Laerdal. Resuscitation. 2002;53:115–9.

59. Abrahamson S, Denson J, Wolf RM. Effectiveness of a simulator in training anaesthesiology residents. Journal of Medical Education. 1969;44:515–9.

60. Bradley P. The history of simulation in medical education and possible future directions. Medical Education. 2006;40(3):254–62.

61. P. Collins, R. M. Harden J. AMEE Medical Education Guide No. 13: real patients, simulated patients and simulators in clinical examinations. Medical Teacher. 1998;20(6):508–21.

62. Bentley S, Mudan G, Strother C, Wong N. Are Live Ultrasound Models Replaceable? Traditional vs. Simulated Education Module for FAST Exam.

Western Journal of Emergency Medicine. 2015 Nov 1;16(6):818–22.

63. Fried GM, Feldman LS, Vassiliou MC, Fraser SA, Stanbridge D, Ghitulescu G, et al. Proving the value of simulation in laparoscopic surgery. In: Annals of Surgery. 2004. p. 518–28.

64. Brunt LM, Halpin VJ, Klingensmith ME, Tiemann D, Matthews BD, Spitler JA, et al. Accelerated Skills Preparation and Assessment for Senior Medical Students Entering Surgical Internship. Journal of the American College of Surgeons. 2008;206(5):897–904.

65. Yule S, Sacks GD, Maggard-Gibbons M. Innovative approaches for modifying surgical culture. Vol. 151, JAMA Surgery. American Medical Association; 2016. p. 791–2.

66. Rosqvist E, Lauritsalo S, Paloneva J. Short 2-H in Situ Trauma Team Simulation Training Effectively Improves Non-Technical Skills of Hospital Trauma Teams. Scandinavian Journal of Surgery. 2019;108(2):117–23.

67. Lewis R, Strachan A, Smith MM. Is High Fidelity Simulation the Most Effective Method for the Development of Non-Technical Skills in Nursing? A Review of the Current Evidence. The Open Nursing Journal. 2012;6(12):82–9.

68. Issenberg SB, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. Medical Teacher. 2005;27(1):10–28.

69. Kalaniti K. In situ simulation: Let's work, practice and learn together. Acta Paediatrica, International Journal of Paediatrics. 2014;103(12):1219–20.

70. Spurr J, Gatward J, Joshi N, Carley SD. Top 10 (+1) tips to get started with in situ simulation in emergency and critical care departments. Vol. 33, Emergency Medicine Journal. BMJ Publishing Group Ltd and the British Association for Accident & Emergency Medicine; 2016. p. 514–6.

71. O'Leary F. Simulation as a high stakes assessment tool in emergency medicine. EMA - Emergency Medicine Australasia. 2015 Apr;27(2):173–5.

72. Teteris E, Fraser K, Wright B, McLaughlin K. Does training learners on simulators benefit real patients? Advances in Health Sciences Education. 2012;17(1):137–44.

73. Ziv A, Wolpe PR, Small SD, Glick S. Simulation-Based Medical Education: An Ethical Imperative. Simulation in Healthcare: The Journal of the Society for

Simulation in Healthcare. 2006;1(4):252–6.

74. Kahn K, Pattinson T, Sherwood M. Simulation in medical education. Medical Teacher. 2011;33:1–3.

75. Willis RE, Van Sickle KR. Current Status of Simulation-Based Training in Graduate Medical Education. Surgical Clinics of North America. 2015;95(4):767–79.

76. Fried GM, Feldman LS. Objective assessment of technical performance. World Journal of Surgery. 2008;32(2):156–60.

77. Oropesa I, Sánchez-González P, Lamata P, Chmarra MK, Pagador JB, Sánchez-Margallo JA, et al. Methods and tools for objective assessment of psychomotor skills in laparoscopic surgery. Journal of Surgical Research. 2011;171(1):e81–95.

78. Thijssen AS, Schijven MP. Contemporary virtual reality laparoscopy simulators: quicksand or solid grounds for assessing surgical trainees? American Journal of Surgery. 2010;199(4):529–41.

79. Van Sickle KR, McClusky DA, Gallagher AG, Smith CD. Construct validation of the ProMIS simulator using a novel laparoscopic suturing task. Surgical Endoscopy and Other Interventional Techniques. 2005;19(9):1227–31.

80. Chmarra MK, Klein S, De Winter JCF, Jansen F-WW, Dankelman J. Objective classification of residents based on their psychomotor laparoscopic skills. Surgical Endoscopy. 2010;24(5):1031–9.

81. Pellen MGC, Horgan LF, Barton JR, Attwood SE. Construct validity of the ProMIS laparoscopic simulator. Surgical Endoscopy and Other Interventional Techniques. 2009;23(1):130–9.

82. Horeman T, Rodrigues SP, Jansen FW, Dankelman J, Van Den Dobbelsteen JJ. Force measurement platform for training and assessment of laparoscopic skills. Surgical Endoscopy. 2010;24(12):3102–8.

83. Singapogu RB, Smith DE, Long LO, Burg TC, Pagano CC, Burg KJLL. Objective differentiation of force-based laparoscopic skills using a novel haptic simulator. Journal of Surgical Education. 2012;69(6):766–73.

84. Kowalewski KF, Garrow CR, Schmidt MW, Benner L, Müller-Stich BP, Nickel F. Sensor-based machine learning for workflow detection and as key to detect expert level in laparoscopic suturing and knot-tying. Surgical Endoscopy.

2019;33(11):3732–40.

85. Oropesa I, Escamirosa FP, Sánchez-Margallo JA, Enciso S, Rodríguez-Vila B, Martínez AM, et al. Interpretation of motion analysis of laparoscopic instruments based on principal component analysis in box trainer settings. Surgical Endoscopy. 2018;32(7):3096–107.

86. Maithel S, Sierra R, Korndorffer J, Neumann P, Dawson S, Callery M, et al. Construct and face validity of MIST-VR, Endotower, and CELTS: Are we ready for skills assessment using simulators? Surgical Endoscopy and Other Interventional Techniques. 2006;20(1):104–12.

87. Datta V, Mackay S, Darzi A, Gillies D. Motion analysis in the assessment of surgical skill. Computer Methods in Biomechanics and Biomedical Engineering. 2001;4(6):515–23.

88. Datta V, Mackay S, Mandalia M, Darzi A. The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. Journal of the American College of Surgeons. 2001;193(5):479–85.

89. Hayter MA, Friedman Z, Bould MD, Hanlon JG, Katznelson R, Borges B, et al. Validation of the Imperial College Surgical Assessment Device (ICSAD) for labour epidural placement. Canadian Journal of Anesthesia. 2009 Jun;56(6):419–26.

90. Varas J, Achurra P, León F, Castillo R, Fuente N, Aggarwal R, et al. Assessment of central venous catheterization in a simulated model using a motion-tracking device: An experimental validation study. Annals of Surgical Innovation and Research. 2016;10(1).

91. Salvadó JA, Oyanedel F, Sepúlveda S, Toledo H, Saavedra Á, Astroza G, et al. Validation of a high-fidelity model in ureteroscopy incorporating hand motion analysis. Int Urol Nephrol. 2015;3:1265–9.

92. Ezra DG, Aggarwal R, Michaelides M, Okhravi N, Verma S, Benjamin L, et al. Skills Acquisition and Assessment after a Microsurgical Skills Course for Ophthalmology Residents. Ophthalmology. 2009;116(2):257–62.

93. Datta V, Chang A, Mackay S, Darzi A. The relationship between motion analysis and surgical technical assessments. American Journal of Surgery. 2002 Jul;184(1):70–3.

94. Duffy AJ, Hogle NJ, McCarthy H, Lew JI, Egan A, Christos P, et al. Construct validity for the LAPSIM laparoscopic surgical simulator. Surgical Endoscopy and Other Interventional Techniques. 2005;19(3):401–5.

95. Woodrum DT, Andreatta PB, Yellamanchilli RK, Feryus L, Gauger PG, Minter RM. Construct validity of the LapSim laparoscopic surgical simulator. American Journal of Surgery. 2006;191(1):28–32.

96. Bajka M, Tuchschmid S, Fink D, Székely G, Harders M, Stefan AE, et al. Establishing construct validity of a virtual-reality training simulator for hysteroscopy via a multimetric scoring system. Surgical Endoscopy. 2010;24(1):79–88.

97. Varoquier M, Hoffmann CP, Perrenot C, Tran N, Parietti-Winkler C. Construct, Face, and Content Validation on Voxel-Man® Simulator for Otologic Surgical Training. International Journal of Otolaryngology. 2017;2017:1–8.

98. Verdaasdonk EGG, Stassen LPS, Schijven MP, Dankelman J. Construct validity and assessment of the learning curve for the SIMENDO endoscopic simulator. Surgical Endoscopy and Other Interventional Techniques. 2007;21(8):1406–12.

99. Zhang A, Hünerbein M, Dai Y, Schlag PM, Beller S. Construct validity testing of a laparoscopic surgery simulator (Lap Mentor®): Evaluation of surgical skill with a virtual laparoscopic training simulator. Surgical Endoscopy and Other Interventional Techniques. 2008;22(6):1440–4.

100. Suzuki T, Egi H, Hattori M, Tokunaga M, Sawada H, Ohdan H. An evaluation of the endoscopic surgical skills assessment using a video analysis software program. Surgical Endoscopy. 2015;29(7):1804–8.

101. Jensen K, Bjerrum F, Hansen HJ, Petersen RH, Pedersen JH, Konge L. Using virtual reality simulation to assess competence in video-assisted thoracoscopic surgery (VATS) lobectomy. Surgical Endoscopy. 2017;31(6):2520–8.

102. Rosen J, Solazzo M, Hannaford B, Sinanan M. Task Decomposition of Laparoscopic Surgery for Objective Evaluation of Surgical Residents' Learning Curve Using Hidden Markov Model. Computer Aided Surgery. 2002;7(1):49–61.

103. D'angelo A-LD, Rutherford DN, Ray RD, Laufer S, Kwan C, Cohen ER, et al. Idle Time: An Underdeveloped Performance Metric for Assessing Surgical Skill. American Journal of Surgery. 2015;209(4):645–51.

104. Ashraf H, Sodergren MH, Merali N, Mylonas G, Singh H, Darzi A. Eye-tracking technology in medical education: A systematic review. Medical Teacher. 2018;40(1):62–9.

105. Vine SJ, Masters RSW, McGrath JS, Bright E, Wilson MR. Cheating experience: Guiding novices to adopt the gaze strategies of experts expedites the learning of technical laparoscopic skills. Surgery. 2012;152(1):32–40.

106. Wilson MR, Vine SJ, Bright E, Masters RSW, Defriend D, McGrath JS. Gaze training enhances laparoscopic technical skill acquisition and multi-tasking performance: A randomized, controlled study. Surgical Endoscopy. 2011;25(12):3731–9.

107. Causer J, Vickers JN, Snelgrove R, Arsenault G, Harvey A. Performing under pressure: Quiet eye training improves surgical knot-tying performance. Surgery. 2014;156(5):1089–96.

108. Soh BLP, Reed WM, Poulos A, Brennan PC. E-tutorial improves students' ability to detect lesions. Radiologic Technology. 2013 Sep;85(1):17–26.

109. Bell CR, Szulewski A, Walker M, McKaigney C, Ross G, Rang L, et al. Differences in gaze fixation location and duration between resident and fellowship sonographers interpreting a FAST. AEM Education and Training. 2020;5(1):28–36.

110. Thomsen ASS, Kiilgaard JF, Kjærbo H, La Cour M, Konge L. Simulation-based certification for cataract surgery. Acta Ophthalmologica. 2015;93(5):416–21.

111. Melniker LA, Leibner E, McKenney MG, Lopez P, Briggs WM, Mancuso CA. Randomized Controlled Clinical Trial of Point-of-Care, Limited Ultrasonography for Trauma in the Emergency Department: The First Sonography Outcomes Assessment Program Trial. Annals of Emergency Medicine. 2006;48(3):227–35.

112. Boulanger BR, McLellan BA, Brenneman FD, Wherrett L, Rizoli SB, Culhane J, et al. Emergent abdominal sonography as a screening test in a new diagnostic algorithm for blunt trauma. Journal of Trauma. 1996;40(6):867–74.

113. Pearl WS, Todd KH. Ultrasonography for the initial evaluation of blunt abdominal trauma: A review of prospective trials. Annals of Emergency Medicine. 1996;27(3):353–61.

114. Shackford SR, Rogers FB, Osler TM, Trabulsy ME, Clauss DW, Vane DW, et al.

Focused abdominal sonogram for trauma: The learning curve of nonradiologist clinicians in detecting hemoperitoneum. Journal of Trauma - Injury, Infection and Critical Care. 1999;46(4):553–64.

115. Mohammad A, Hefny AF, Abu-Zidan FM. Focused assessment sonography for trauma (FAST) training: A systematic review. World Journal of Surgery. 2014;38(5):1009–18.

116. Matsumoto ED, Hamstra SJ, Radomski SB, Cusimano MD. The effect of bench model fidelity on endourological skills: A randomized controlled study. The Journal of Urology. 2002;167:1243–7.

117. Center HCM. The FAST Exam [Internet]. Online. [cited 2019 Mar 13]. Available from: https://vimeo.com/1044031

118. Ziesmann MT, Park J, Unger BJ, Kirkpatrick AW, Vergis A, Logsetty S, et al. Validation of the quality of ultrasound imaging and competence (QUICk) score as an objective assessment tool for the FAST examination. Journal of Trauma and Acute Care Surgery. 2015;78(5):1008–13.

119. Tavakol M, Dennick R. Making sense of Cronbach's alpha. International Journal of Medical Education. 2011;2:53–5.

120. Clinkard D, Holden M, Ungi T, Messenger D, Davison C, Fichtinger G, et al. The development and validation of hand motion analysis to evaluate competency in central line catheterization. Academic Emergency Medicine. 2015;22(2):212–8.

121. Dosis A, Bello F, Moorthy K, Münz Y, Gillies D, Darzi A. Real-time synchronization of kinematic and video data for the comprehensive assessment of surgical skills. Studies in Health Technology and Informatics. 2004;98:82–8.

122. Chin KJ, Tse C, Chan V, Tan JS, Lupu CM, Hayter M. Hand motion analysis using the Imperial College surgical assessment device: Validation of a novel and objective performance measure in ultrasound-guided peripheral nerve blockade. Regional Anesthesia and Pain Medicine. 2011;36(3):213–9.

123. Aggarwal R, Grantcharov TP, Eriksen JR, Blirup D, Kristiansen VB, Funch-Jensen P, et al. An evidence-based virtual reality training program for novice laparoscopic surgeons. Annals of Surgery. 2006;244(2):310–4.

124. Chang J, Banaszek DC, Gambrel J, Bardana D. Global rating scales and motion analysis are valid proficiency metrics in virtual and benchtop knee arthroscopy

simulators. Clinical Orthopaedics and Related Research. 2016;474(4):956–64.

125. Zago MM, Sforza C, Mariani D, Marconi M, Biloslavo A, Greca A La, et al. Educational impact of hand motion analysis in the evaluation of FAST examination skills. European Journal of Trauma and Emergency Surgery. 2020;46(6):1421–8.

126. Le Lous M, Despinoy F, Klein M, Fustec E, Lavoué V, Jannin P, et al. Impact of physician expertise on probe trajectory during obstetric ultrasound: A quantitative approach for skill assessment. Simulation in Healthcare. 2021;16(1):67–72.

127. Matyal R, Mitchell JD, Hess PE, Chaudary B, Bose R, Jainandunsing JS, et al. Simulator-based Transesophageal Echocardiographic Training with Motion AnalysisA Curriculum-based Approach. Anesthesiology. 2014;121(2):389–99.

128. Obstein KL, Patil VD, Jayender J, Estpar RSJ, Spofford IS, Lengyel BI, et al. Evaluation of colonoscopy technical skill levels by use of an objective kinematic-based system. Gastrointestinal Endoscopy. 2011;73(2):315-321.e1.

129. Hovgaard LH, Andersen SAW, Konge L, Dalsgaard T, Larsen CR. Validity evidence for procedural competency in virtual reality robotic simulation, establishing a credible pass/fail standard for the vaginal cuff closure procedure. Surgical Endoscopy. 2018;32(10):4200–8.

130. Moorthy K, Munz Y, Dosis A, Bello F, Darzi A. Motion analysis in the training and assessment of minimally invasive surgery. Minimally Invasive Therapy and Allied Technologies. 2003;12(3–4):137–42.

131. Ziesmann MT, Park J, Unger B, Kirkpatrick AW, Vergis A, Pham C, et al. Validation of hand motion analysis as an objective assessment tool for the Focused Assessment with Sonography for Trauma examination. Journal of Trauma and Acute Care Surgery. 2015;79(4):631–7.

132. Holden MS, Portillo A, Salame G. Skills Classification in Cardiac Ultrasound with Temporal Convolution and Domain Knowledge Using a Low-Cost Probe Tracker. Ultrasound in Medicine and Biology. 2021;47(10):3002–13.

133. O'Brien KM, Stolz LA, Amini R, Gross A, Stolz U, Adhikari S. Focused assessment with sonography for trauma examination: Reexamining the importance of the left upper quadrant view. Journal of Ultrasound in Medicine. 2015;34(8):1429–34.

134. Bell CR, McKaigney CJ, Holden M, Fichtinger G, Rang L. Sonographic Accuracy as a Novel Tool for Point-of-care Ultrasound Competency Assessment. AEM Education and Training [Internet]. 2017 [cited 2020 Jan 6];1(4):316–24. Available from: www.aem-e-t.com

135. Viera AJ, Garrett JM. The kappa statistic. Family Medicine. 2005;37(5):360–3.

136. Bloom BA, Gibbons RC. Focused assessment with sonography for trauma [Internet]. Stat Pearls. StatPearls Publishing; 2021 [cited 2021 Oct 15]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470479/

137. Burckett–St.Laurent DA, Niazi A, Cunningham M, Jaeger M, Abbas S, Mcvicar J, et al. A valid and reliable assessment tool for remote simulation-based ultrasound-guided regional anesthesia. Regional Anesthesia and Pain Medicine. 2014;39(6):496–501.

138. Buaprasert P, Sri-On J, Sukhuntee J, Asawajaroenkul R, Buanhong O, Khiaodee T, et al. Diagnostic accuracy of extended focused assessment with sonography for trauma performed by paramedic students: A simulation-based pilot study. Open Access Emergency Medicine. 2021;13:249–56.

139. Damewood SC, Lewiss RE, Huang J V. Ultrasound simulation utilization among point of care ultrasound users: Results of a survey. Journal of Clinical Ultrasound. 2018 Nov 1;46(9):571–4.

140. Salen P, O'Connor R, Passarello B, Pancu D, Melanson S, Arcona S, et al. Fast education: A comparison of teaching models for trauma sonography. Journal of Emergency Medicine. 2001;20(4):421–5.

141. Jang T, Naunheim R, Sineff S, Aubin C. Operator Confidence Correlates with More Accurate Abdominal Ultrasounds by Emergency Medicine Residents. Journal of Emergency Medicine. 2007;33(2):175–9.

142. Sultan SF, Iohom G, Saunders J, Shorten G. A clinical assessment tool for ultrasound-guided axillary brachial plexus block. Acta Anaesthesiologica Scandinavica. 2012;56(5):616–23.

143. Vovk A, Wild F, Guest W, Kuula T. Simulator sickness in Augmented Reality training using the Microsoft HoloLens. In: Conference on Human Factors in Computing Systems - Proceedings. 2018.

144. Howarth PA, Costello PJ. The occurrence of virtual simulation sickness symptoms

when an HMD was used as a personal viewing system. Displays. 1997;18(2):107–16.

145. Kore. The human eye's understanding of space for Augmented Reality | by Kore | UX Collective [Internet]. 2018 [cited 2021 Oct 26]. Available from: https://uxdesign.cc/human-eyes-understanding-of-space-for-augmented-reality-d5ce4d9fa37b

146. Moro C, Štromberga Z, Raikos A, Stirling A. The effectiveness of virtual and augmented reality in health sciences and medical anatomy. Anatomical Sciences Education. 2017;10(6):549–59.

147. Grober ED, Hamstra SJ, Wanzel KR, Reznick RK, Matsumoto ED, Sidhu RS, et al. The educational impact of bench model fidelity on the acquisition of technical skill: The use of clinically relevant outcome measures. Annals of Surgery. 2004;240(2):374–81.

148. Le CK, Lewis J, Steinmetz P, Dyachenko A, Oleskevich S. The use of ultrasound simulators to strengthen scanning skills in medical students: A randomized controlled trial. Journal of Ultrasound in Medicine. 2019;38(5):1249–57.

149. Paddock MT, Bailitz J, Horowitz R, Khishfe B, Cosby K, Sergel MJ. Disaster response team FAST skills training with a portable ultrasound simulator compared to traditional training: Pilot study. Western Journal of Emergency Medicine. 2015;16(2):325–30.

150. Hu Y, Goodrich RN, Le IA, Brooks KD, Sawyer RG, Smith PW, et al. Vessel ligation training via an adaptive simulation curriculum. Journal of Surgical Research. 2015;196(1):17–22.

151. Sivaprakasam J, Purva M. CUSUM analysis to assess competence: What failure rate is acceptable? Clinical Teacher. 2010 Dec 1;7(4):257–61.

152. Hu Y, Jolissaint JS, Ramirez A, Gordon R, Yang Z, Sawyer RG. Cumulative sum: A proficiency metric for basic endoscopic training. Journal of Surgical Research. 2014;192(1):62–7.

153. Renaud M, Reibel N, Zarnegar R, Germain A, Quilliot D, Ayav A, et al. Multifactorial analysis of the learning curve for totally robotic roux-En-Y gastric bypass for morbid obesity. Obesity Surgery. 2013;23(11):1753–60.

Appendix A

FAST Study Questionnaire – Pre-assessment

Novice Test Subject_____

Intermediate Test Subject _____

Expert Test Subject_____

If you are a resident, what is your PGY year?

Circle one

1 2 3 4 5 6 7 8

If you are an attending, how many years have you been in practise?

Circle one

<2 2-5 6-10 11-15 16-20 21-25 26-30 30+

Are you right-hand or left-hand dominant?

Circle one

Right / Left

Have you ever performed a FAST exam in a training, simulation, or practise session?

Circle one

Y / N

Have you ever performed a FAST exam in a clinical setting?

Circle one

Y / N

Are you Certified to perform FAST? If "Yes" please describe certification body and date of certification.

Circle one

Y / N Body_____

Date_____

Have you used the Ultrasound device in a training, simulation, or practise session?

Circle one

Y / N

Have you used the Ultrasound device in a clinical setting?

Circle one

Y / N

Have you taken a formal Ultrasound training course? If "Yes" please describe course and date of training

Circle one

Y / N Course_____

Date_____

How frequently do you estimate that you use the Ultrasound device in clinical settings?

Circle one

Once/day More than Once/week Once/week Once/month Once/year

How frequently do you estimate that you perform the FAST exam in clinical settings?

Circle one

Once/day More than Once/week Once/week Once/month Once/year

Have you used augmented reality or virtual reality simulators?

Circle one

Y / N

How often do you use augmented reality or virtual reality simulators?

Circle one

Once/day More than Once/week Once/week Once/month Once/year Never

How frequently do you play video games?

Circle one

Once/day More than Once/week Once/week Once/month Once/year

FAST Study Questionnaire – Post-assessment

Novice Test Subject_____

Intermediate Test Subject _____

Expert Test Subject_____

Did you experience headaches while using the augmented reality goggles?

Circle one

Y / N

Did you experience headaches after using the augmented reality goggles?

Circle one

Y / N

Did you feel nauseous while using the augmented reality goggles?

Circle one

Y / N

Did you feel nauseous after using the augmented reality goggles?

Circle one

Y / N

Did you have any blurry vision while using the augmented reality goggles?

Circle one

Y / N

The augmented reality goggles glasses added value to my experience with the ultrasound simulator.

Strongly disagree	disagree	neutral	agree	strongly agree
1	2	3	4	5

The augmented reality goggles improved my experience with the ultrasound simulator.

Strongly disagree	disagree	neutral	agree	strongly agree
1	2	3	4	5

Appendix B

Structured Assessment of Image Acquisition Tool
Part A - FAST Image Acquisition Evaluation Checklist

Study ID number: _____
Examiner: _____

HEPATORENAL SPACE

- ❑ Orients image with the liver to the left and kidney to the right
- ❑ Visualizes the interface between the liver and kidney clearly
- ❑ Visualizes the interface between the liver and kidney in entirety by sweeping through the entire kidney
- ❑ Visualizes the caudal tip of the liver clearly

Score ____ /4

SPLENORENAL SPACE

- ❑ Orients image with the spleen to the left and kidney to the right
- ❑ Visualizes the interface between the spleen and kidney clearly
- ❑ Visualizes the interface between the spleen and kidney in entirety by sweeping through the entire kidney
- ❑ Clearly visualizes between the diaphragm and spleen

Score ____ /4

PELVIS

- ❑ Visualizes the bladder in longitudinal section
- ❑ Visualizes the bladder in entirety in longitudinal section by sweeping through the entire bladder
- ❑ Visualizes the bladder in transverse section
- ❑ Visualizes the bladder in entirety in transverse section by sweeping through the entire bladder

Score ____ /4

PERICARDIUM

- ❑ Orients image such that the apex of the ventricles point towards the right of the image
- ❑ Visualizes both the inferior and superior pericardium
- ❑ Visualizes the pericardium in its entirety by sweeping through the entire heart

Score ____ /3

Total Score ____ /15

Structured Assessment of Image Acquisition Tool
Part B - Global Rating Scale Of FAST Image Acquisition

Study ID number: _____

Evaluator: _____

Note: a score of 4 is considered an "acceptable performance" for an adequate exam.

1. Skin contact

1	2	3	4	5
Consistently uses *insufficient* amounts of gel or achieves *inadequate* skin contact		Uses *appropriate* amounts of gel and achieves *adequate* skin contact *most* of the time		*Consistently* uses *appropriate* amounts of gel and achieves *adequate* skin contact

2. Probe placement

1	2	3	4	5
Frequently readjusts probe position on the skin or obtains *inadequate* views		*Correctly* places the probe to obtain *adequate* views but *occasionally* requires readjustment		*Correctly* places the probe to obtain *adequate* views on *1st attempt* with *minimal* readjustment

3. Image scrolling

1	2	3	4	5
After establishing probe position *continues* to reposition in a *staccato* manner		After establishing probe position has *mostly smooth* image sweeping but makes occasional *staccato* movements		After establishing probe position makes *subtle* probe movements with *smooth* sweeping

4. Sonographer positioning and probe handling

1	2	3	4	5
Repeatedly assumes an *awkward* body position or holds the probe in an *awkward or inappropriate* manner		*Occasionally* assumes an *awkward* position or holds the probe in an *inappropriate* manner		Assumes a *comfortable* body position and holds the probe in a *appropriate* manner

5. Time of exam

1	2	3	4	5
Takes an *excessively* long time to complete the exam (generally greater than 10 minutes)		Completes the exam in an *average* amount of time (generally between 2-5 minutes)		*Rapidly* completes the exam *with* acceptable performance (generally less than 2 minutes)

6. Flow of procedure

1	2	3	4	5
Is *consistently unorganized* with *frequent* jumps between anatomic regions		Mostly organized but jumps *occasionally* between anatomic regions		Completes the procedure by moving *smoothly* from region to region

7. Overall performance

1	2	3	4	5
Unacceptable performance; multiple major inadequacies	Unacceptable performance; some major inadequacies	Unacceptable performance; minor inadequacies only	Acceptable performance	Exceptional performance; expert FAST performer